ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA DOPRAVNÍ

Ústav letecké dopravy

Zvyšování úspěšnosti textové analýzy nestrukturovaných dat v letecké údržbě

Diplomová práce

Bc. Markéta Adamcová

Vedoucí práce: doc. Ing. Andrej Lališ, Ph.D.; Mgr. Miroslav Blaško, Ph.D

Praha 2022

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
Fakulta dopravní
děkan
Konviktská 20, 110 00 Praha 1

K621.....Ústav letecké dopravy

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení studenta (včetně titulů):

Bc. Markéta Adamcová

Studijní program (obor/specializace) studenta:

navazující magisterský – PL – Provoz a řízení letecké dopravy

Název tématu (česky): **Zvyšování úspěšnosti textové analýzy nestrukturovaných dat v letecké údržbě**

Název tématu (anglicky): Increasing Text Analysis Relevancy of Aircraft Maintenance Non-Structured Data

Zásady pro vypracování

Při zpracování diplomové práce se řiďte následujícími pokyny:

- Cílem práce je navrhnout řešení pro zvyšování úspěšnosti textové analýzy vyhodnocení nestrukturovaných dat v letecké údržbě za účelem identifikace neočekávaných nálezů v údržbě dopravních letadel.
- Analyzujte postupy pro záznam nálezů v procesu údržby dopravních letadel ve vybrané letecké organizaci.
- Analyzujte dostupné slovníky pro klasifikaci záznamů dat.
- Proveďte analýzu a vyhodnocení vzorků dat ve vybrané letecké organizaci a identifikujte nedostatky současného postupu textové analýzy.
- Navrhněte řešení pro zvyšování úspěšnosti textové analýzy vyhodnocení nestrukturovaných dat v letecké údržbě.
- Navržené řešení ověřte a vyhodnoťte.

ČVUT FD

Rozsah grafických prací: dle pokynů vedoucího diplomové práce

Rozsah průvodní zprávy: minimálně 55 stran textu (včetně obrázků, grafů a tabulek, které jsou součástí průvodní zprávy)

Seznam odborné literatury: R. Doc Palmer. Maintenance Planning and Scheduling Handbook, McGraw Hill Professional, 1999.
T. Vojtěch. Textová analýza nestrukturovaných závadových dat v letecké údržbě. Diplomová práca, ČVUT v Praze, 2020.

Vedoucí diplomové práce: **doc. Ing. Andrej Lališ, Ph.D.**
Mgr. Miroslav Blaško, Ph.D.

Datum zadání diplomové práce: **16. července 2021**
(datum prvního zadání této práce, které musí být nejpozději 10 měsíců před datem prvního předpokládaného odevzdání této práce vyplývajícího ze standardní doby studia)

Datum odevzdání diplomové práce: **30. listopadu 2022**
a) datum prvního předpokládaného odevzdání práce vyplývající ze standardní doby studia a z doporučeného časového plánu studia
b) v případě odkladu odevzdání práce následující datum odevzdání práce vyplývající z doporučeného časového plánu studia

L. S.

doc. Ing. Jakub Kraus, Ph.D.
vedoucí Ústavu letecké dopravy

prof. Ing. Ondřej Přibyl, Ph.D.
děkan fakulty

Potvrzuji převzetí zadání diplomové práce.

Bc. Markéta Adamcová
jméno a podpis studenta

V Praze dne 17. května 2022

Poděkování

Ráda bych poděkovala vedoucím své práce doc. Ing. Andreji Lališovi, Ph. D a Mgr. Miroslavu Blaškovi, Ph. D. za jejich pomoc, podporu a cenné rady poskytované během psaní této práce. Také bych ráda poděkovala všem zaměstnancům Ústavu letecké dopravy na Fakultě dopravní ČVUT za sdílení jejich odborných znalostí během celého mého studia. Nakonec bych chtěla poděkovat své rodině a přátelům za jejich trvalou podporu během studia a neustálou víru ve mne.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracovala samostatně a že jsem uvedla veškeré informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných pracích.

Nemám závažný důvod proti užití tohoto školního díla ve smyslu zákona § 60 Zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon).

V Praze dne 30.11.2022

Bc. Markéta Adamcová

Abstrakt

Na letadle je prováděno mnoho úkonů a o všech se vedou záznamy, které se uchovávají. V případě, že je během provádění prací na letadle nalezena neočekávána závada, je třeba o tom vytvořit záznam. Tento záznam má podobu nestrukturovaného textu, kde kam jsou zaneseny údaje, které pro údržbovou organizaci mohou mít velkou hodnotu s ohledem na plánování údržby a větší předvídatelnosti neočekávaných závad. Jedním ze způsobů, jak tyto informace z velkého objemu nestrukturovaného textu získat, je textová analýza. Již bylo navrženo řešení, které aplikuje automatickou textovou analýzu na záznamech letecké údržby. Toto řešení vykázalo výsledky, které by bylo třeba pro praktické využití zvýšit. Z toho důvodu tato práce na navržené řešení navazuje a jejím cílem je nelézt způsob, jak lze úspěšnost automatické textové analýzy zvýšit. Dále je cílem navrhnout obecná doporučení pro údržbové organizace, které by chtěly využít textové analýzy na svých datech.

Klíčová slova: komponenta, nestrukturovaná data, TermIt, textová analýza, údržba letadel, závada

Abstract

Many operations are carried out on an aircraft and records are kept of all of them. In the event that an unexpected failure is found during work on the aircraft, a record of this failure must be made. This record has form of unstructured text, but these records can contain information of greate value to maintanance organizations, mostly with impact on maintanance planning and better predictability of unexpected failures. One of the options to get this information from big volume of unstructured text, is text analysis. A solution that applies automatic text analysis to aircraft maintanence records has already been proposed. This solution showed results that would need to be scaled up for practical use. For that reason, this theses follows up this proposed solution and its goal is to find solution that increases the success rate of automatic text analysis. Furthermore, the goal is to propose general recommendations for maintanance organizations that would like to use text analytics on their data.

Keywords: aircraft maintanance, component, failure, text analysis, TermIt, unstructured data

Obsah

Seznam obrázků 5

Seznam tabulek 6

Seznam použitých zkratek 7

Úvod 9

1 Údržba letadlové techniky 10

1.1 Nařízení Komise EU č. 1321/2014 12

2 CSAT 16

3 Proces údržby 19

3.1 Plán údržby 20

3.2 Technická dokumentace 22

3.3 Plánovaná údržba 24

3.4 Neplánovaná údržba 24

3.4.1 Záznam závady 25

3.5 Softwary pro údržbové organizace 26

4 Textová analýza 28

4.1 Textová analýza v kontextu technických dat 29

4.2 Dostupná literatura 29

4.2.1 Limitace současného stavu 31

5 Programy 33

6 TermIt 35

6.1 Dostupné slovníky 44

7 Dostupná data 45

7.1 Tvorba slovníku 46

8 Zhodnocení úspěšnosti analýzy 52

8.1 Výpočet precision a recall 52

8.2 Výstupy automatické analýzy 54

8.3 Výsledky automatické textové analýzy 55

8.3.1 Výpočet precision a recall v kontextu řádku ... 55
8.3.2 Výpočet precision a recall v kontextu pojmu ... 57
8.3.3 Úspěšnost výběru pojmů ... 58
9 Možná řešení pro zvýšení úspěšnosti ... 63
9.1 Doporučení na tvorbu a úschovu dat ... 63
9.2 Doporučení na tvorbu a údržbu slovníku ... 67
9.3 Doporučení pro software Termit ... 67
9.3.1 Aplikované patterny ... 69
9.4 Doporučení pro zvýšení úspěšnosti precision a recall v kontextu řádku ... 72
10 Diskuse ... 73
Závěr ... 78
Zdroje ... 80

Seznam obrázků

Obrázek 1: schéma procesu a komunikace organizací zahrnutých do údržby letadel ... 20
Obrázek 2: Schéma vzniku OAMP ... 21
Obrázek 3: Organizace dle normy ATA100 (kapitola 56 - křídla) ... 23
Obrázek 4: Prostředí softwaru AMOS [16] ... 27
Obrázek 5: Prostředí softwaru „Voyant“ [10] ... 33
Obrázek 6: Prostředí softwaru "Word and phrase" [11] ... 34
Obrázek 7: Práce s nástrojem TermIt ... 36
Obrázek 8: TermIt – správa slovníků ... 37
Obrázek 9: TermIt – podrobnosti o slovníku ... 38
Obrázek 10: TermIt – nadřazené pojmy ... 40
Obrázek 11: TermIt – nový zdroj ... 41
Obrázek 12: TermIt – zobrazení obsahu dokumentu a legendy výskytu pojmů ... 42
Obrázek 13: TermIt – pop-up menu navrhovaného výskytu existujícího pojmu ... 42
Obrázek 14: TermIt – pop-up menu po označení textu v dokumentu ... 43
Obrázek 15: Možnosti po označení zájmového textu ... 47
Obrázek 16: Zobrazení přiřazeného pojmu ... 47
Obrázek 17: Návrhy komponent a závad ... 48
Obrázek 18: Možnosti výstupu textové analýzy [17] ... 53
Obrázek 19: Výpočet precision [17] ... 53
Obrázek 20: Výpočet recall [17] ... 53
Obrázek 21: Určení pozitivity/negativity v kontextu řádku ... 56
Obrázek 22: Porovnání potvrzených anotací s automatickou analýzou ... 58
Obrázek 23: Porovnání výsledků této práce a T. Vojtěcha ... 60
Obrázek 24: Pojem 'missing' ... 61
Obrázek 25: Částečná shoda pojmů ... 62
Obrázek 26: Příklad patternu 'removal/inspection' ... 68
Obrázek 27: Pozice komponentu za spojkou nebo předložkou ... 68
Obrázek 28: Umístění pojmů před a za znakem ‚:‘ ... 68
Obrázek 29: Pojem v závorkách ... 68
Obrázek 30: Výskyt pojmu 'wet' ... 69

Seznam tabulek

Tabulka 1: Příklad záznamů v poli pro volný text 29
Tabulka 2: Výsledky návrhu výskytu pojmů po spuštění automatické textové analýzy aktuálním slovníkem u všech datasetů 50
Tabulka 3: Výsledek výpočtu precision a recall v kontextu řádku 56
Tabulka 4: Výsledky precision a recall v kontextu pojmu 57
Tabulka 5: Porovnání výsledků práce 59
Tabulka 6: Porovnání výsledků práce prováděné stejným vyhodnocením 61
Tabulka 7: Příklady nerelevantních záznamů 63
Tabulka 8: Příklady záznamů s pokyny a záznamy o sejmutí dílu 64
Tabulka 9: Klíčová slova a příklady záznamů 65
Tabulka 10: Porovnání výsledků po aplikaci patternů 71
Tabulka 11: Porovnání výsledků precision a recall po aplikaci patternů 72

Seznam použitých zkratek

AD	Airworthiness Directive
AFM	Aircraft Flight Manual
AMM	Aircraft Maintenance Manual
AMO	Approved Maintenance Organization
	Organizace provádějící údržbu letadel
APU	Auxiliary Power Unit
ATA	Air Transport Association
AWL	Airwothiness limitations
CAMO	Continuing Airwothiness Management Organization
	Organizace pro řízení zachování letové způsobilosti
CDL	Configuration Deviation List
CMM	Component Maintenance Manual
CMR	Certification Maintenance Requirements
CSAT	Czech Airlines Technics
EASA	Europen union Aviation Safety Agency
	Agentura Evropské unie pro bezpečnost v letectví
EK	Evropská komise
FAA	Federal Aviation Administartion
	Federální letecká správa
FIM	Fault Isolation Manual
ICA	Instruction for Continued Airworthiness
	Instrukce pro zachování letové způsobilsoti
ICAO	International Civil Aviation Organization
	Mazinárodní organizace pro civilní letectví
IPC	Illustrated Parts Catague
ITEM	Illustrated Tool and Equipment Manul
LRU	Line Replaceable Unit
MEL	Minimul Equipment List
	Seznam minimálního vybavení
MM	Maintenance Manual

MMEL	Master Minimul Equipment List
	Základní seznam minimálního vybavení
MPD	Maintenance Planning Document
MRBR	Maintenance review board report
MTO	Maintenance Training Organization
NDTM	Non-Destructive Testing Manual
NPL	Natural Language Processing
OAMP	Operator Approved Maintenance Program
OHM	Overhaul Manual
PZZ	Příkaz k Zachování letové Způsobilosti
SB	Service Bulletine
SL	Service Letter
SRM	Structural Repair Manual
SSM	System Schematic Manual
ÚCL	Úřad pro Civilní Letectví
WBM	Weight and Balance Manual
WDM	Wiring Diagram Manual
WO	Workorder

Úvod

Tato práce je zaměřena na odvětví letecké údržby. Letectví je obecně odvětví, které je silně legislativně regulované a letecká údržba není výjimkou. Údržbu na letadle civilního letectví mohou provádět pouze úřadem schválené organizace pro údržbu. Jednou za takových organizací je i Czech Airlines Technics (CSAT).

Během údržby letadel vzniká velké množství digitálních a papírových záznamů, které se dle legislativy musí po určitou dobu uchovávat. V těchto záznamech jsou obsažené pro údržbovou organizaci cenné informace. Záznamy obsahují informace o selhaných letadlových celcích a komponentech, typu letadla, datumu. Analýzou historických záznamů je možné určit trendy a předpovídat přesněji možné neočekávané závady. Tyto informace mohou být dále efektivně využity při plánování údržby. Záznamy o závadách, které byly objeveny během plánované revize, jsou zaznamenány ve formě volného textu. Z toho důvodu je nelze automaticky filtrovat. Analyzovat velké množství textu a dokumentů manuálně není efektivní. Řešení nabízí automatická textová analýza, která dokáže podle předem definovaných parametrů z nestrukturovaného textu získat požadované informace.

Řešení, které aplikovalo textovou analýzu na datech z letecké údržby skrz vhodný nástroj, již bylo navrženo. Výsledky, které toto řešení vykázalo, poukázaly na potenciál využití textové analýzy v tomto odvětí, avšak pro praktické využití je třeba úspěšnost zvýšit.

Tato práce na zmíněné řešení navazuje a jejím cílem je navrhnout postup, který by úspěšnost analýzy zvýšil. Jako dalším dílčím cílem je obecně definovat doporučení pro údržbové organizace, které by měly zájem do svého prostředí na svých datech textovou analýzu aplikovat.

1 Údržba letadlové techniky

Pojem „údržba" má několik definic. Jedna z nich zní: *Údržba je činnost udržování vozidla, budovy nebo jiného objektu v dobrém stavu jejich pravidelnou kontrolou a v případě potřeby opravou. [1]* Tím je myšlena údržba převážně technických zařízení a přístrojů. Cílem je udržení přístroje nebo zařízení ve stavu, který umožní zachování požadované funkce.

Mezi oblasti, kde je údržba důležitým článkem celého procesu, patří i letectví. Letadla jsou komplexní stroje se složitými interakcemi. Jejich cena je v porovnání s jinými dopravními prostředky několikanásobně vyšší, navíc při případné nehodě je ohroženo zdraví velkého počtu lidí. Je tedy v zájmu leteckých společností udržovat letadlo ve spolehlivém stavu. Případná nemožnost provést let stojí velké množství prostředků. Celý proces údržby letadla je prováděn s cílem udržet letadlo ve stavu, který umožní vydání a zachování osvědčení letové způsobilosti. Jedná se o směs preventivních opatření k zjištění, zda nedošlo k nezjištěným náhodným poruchám, a nápravných prací (k odstranění již vzniklých poruch).

Údržba letadlové techniky se týká komplexního objektu-letadla a všech jeho komponent, a popisuje tedy komplexní proces, který zahrnuje práci od vytvoření údržbového plánu, přes samotnou fyzickou údržbu, až po uchování záznamů o provedené údržbě. Údržba, generální údržba, veškeré prohlídky, opravy, testy jsou prováděny na letadle nebo jeho části s cílem zachovat letovou způsobilost konkrétního letadla.

Základní informace k procesu údržby a intervalů, které je potřeba dodržet jsou uvedeny v dokumentu MRBR (maintanance review board report), který je schválen výrobcem letadla a dodán uživateli spolu s ním. Na základě tohoto dokumentu údržbová organizace vytvoří MPD (maintanance planning document). Ten slouží jako základ k vytvoření již konkrétních údržbových plánů ke všem jednotlivým letadlům v letadlovém parku.

Dalším důležitým dokumentem pro údržbu je AMM (aircraft maintanance manual). Jedná se o základní a nejrozsáhlejší příručku pro údržbu. Všechny úkony na letadle musí být provedeny v souladu s informacemi uvedenými v AMM. Jsou zde uvedeny všechny postupy ke všem schváleným metodám oprav. Je zde popsáno vše od podrobného popisu letadlového celku přes potřebné nářadí, celkovou časovou náročnost úkonu až po požadavky na prostory a skladování. U složitých motorových letadel (např. B737) by

bylo nepraktické všechny informace uložit pouze do jedné příručky. Proto jsou k základní AMM připojeny další příručky logicky spojující úkony na stejné části letadla. AMM tedy popíše jen základní informace a dále se odkazuje na jinou konkrétní příručku. Často se jedná např. o OHM (overhaul manual) nebo SRM (structural repair manual).

Údržba letadel se dělí do dvou kategorií: [2]

- Traťová údržba
- Údržba na základně

Traťová údržba se někdy taktéž označuje jako lehká údržba. Part 145, který je součástí nařízení EU č. 1321/2014, který je podrobně popsán níže, definuje traťovou údržbu jako úkoly údržby, které lze provést venku pod širým nebem mimo hangár. Letadlo tedy během této údržby zůstává v provozním prostředí, proto jsou tyto úkony většinou prováděny na letišti na stojánce, tak aby nedošlo k narušení plánovaného provozu. Velkou část traťové údržby tvoří rutinní provozní inspekce a každodenní kontrolní akce jako např. před a po letová prohlídka. Tyto kontroly jsou prováděny pomocí check-listů. Pokud je během těchto prohlídek nalezena závada (která ovlivňuje letovou způsobilost), mohou nastat dvě situace. Za prvé, oprava je možná v rámci traťové údržby. Tedy že oprava je natolik jednoduchá, že je možné ji provést venku a s dostupným personálem pro traťovou údržbu. Tento úkon se pak dále označuje LRU (line replaceable unit). V provozu se často jedná např. o výměnu poškozené pneumatiky. Za druhé, oprava není možná v rámci traťové údržby. V tom případě není letadlo způsobilé k letu, a musí být přetaženo do hangáru na údržbu na základně.

Údržba na základně je označovaná taktéž jako těžká údržba. Sestává se z úkolů, které jsou obecně podrobnější, důkladnější, složitější a časově náročnější než v případě lehké údržby. Z letadla jsou zpravidla vymontovávány celé celky. Může dojít k modifikacím, implementaci nového zařízení nebo k nedestruktivnímu zkoušení. Na základě toho musí být hangár dostatečně vybaven speciálním nářadím a pomůckami, stejně tak úkoly může provádět pouze personál s potřebnou kvalifikací.

Základním parametrem pro plánování údržby jsou požadované intervaly. Používané jednotky jsou letové hodiny, letové cykly, kalendářní čas, provozní hodiny. Základní minimální požadavky jsou uvedeny v MRBR. Provozovatel ale může tyto intervaly upravit na základě vlastní zkušenosti pro potřebu plánování a sdružení úkolů s podobným intervalem kontrol. Tyto prohlídky dělíme tzv A, B, C, D checks. Nejjednodušší A zahrnuje

preventivní kontrolu základních systémů, nejsložitější D naopak představuje v podstatě kompletní rozebrání celého letadla.

Ne všichni provozovatelé využívají všechny checky. Cílem je co nejvíce minimalizovat čas letadla strávený v hangáru, a tím i minimalizovat finanční ztráty a optimálně plánovat i technické a personální vybavení. Vše ale za předpokladu dodržení základních požadovaných intervalů. Pokud by provozovatel tyto intervaly nedodržel, letadlo by přišlo o letovou způsobilost.

Údržbu ale neovlivňují jen pravidelné intervaly plánované údržby. V praxi je běžné, že za nějakou dobu v provozu je odhalena skrytá závada nebo vada. V mnoha takových případech není možné čekat až na další termín pravidelné údržby, ale je třeba pokyn vydat okamžitě. Takovéto dokumenty, které ovlivňují letovou způsobilost, se označují ICA (instructions for continued airworthiness) a obecně jsou to instrukce a pokyny pro zachování letové způsobilosti. Mezi ICA patří:

- AD (airworthiness directive)

  Česky příkaz k zachování letové způsobilosti (PZZ) je dokument, který nařizuje provozovatelům letadel (letadlových celků/komponentů daného sériového čísla) odstranění nežádoucích vlastností nebo nařizuje uzemnění určitého typu letadel. Provedení AD je závazné ve stanovené lhůtě, v případě neprovedení AD dojde ke ztrátě letové způsobilosti. AD pro letadla registrována v EU jsou vydávána úřadem EASA (jinde AD vydává příslušný úřad) na základě rozhodnutí příslušného národního regulačního úřadu členského státu, návrhu držitele schválených konstrukčních údajů nebo AD vydaného státem projekce.
- SB (service bulletine), SL (service letter)

  Servisní bulletin a servisní dopis jsou dodatečné ICA vydávané držitelem schválených konstrukčních údajů. Obsahem servisního bulletinu jsou většinou informace o změnách v obsahu technické dokumentace, informace o doporučených případně nařízených kontrolách nebo úpravách, informace pro realizaci modifikací. Provedení SB není pro provozovatele povinné s výjimkou kdy, se na daný SB odkazuje AD, což je běžné. Servisní dokument je dokument, který obvykle obsahuje méně důležité informace než servisní bulletin.

1.1 Nařízení Komise EU č. 1321/2014

Údržba letadlové techniky je v procesu letového provozu kritickým článkem. Je potřeba zajistit, aby za provozu nedošlo k nenadálým událostem, jako je například porucha

letadlové části. Protože letový provoz je globální fenomén, vzniká zde silná potřeba harmonizace a sjednocení úrovně prováděné údržby letadlové techniky napříč státy. Jedná se tedy o prostředí silně regulované.

Základním regulátorem je Mezinárodní organizace pro civilní letectví (International civil aviation organisation-ICAO). Tato organizace vydává mezinárodní standardy, které členské státy implementují do svých mezinárodních předpisů s cílem regulovat všechny subjekty a procesy, které jsou v letectví nějakým způsobem zainteresované. Mezi to patří úkoly údržby, organizace, které údržbu provádějí, včetně jejího personálu, a auditní systém.

V Evropě je hlavní regulátor Evropská unie (EU). Ta na základě mezinárodních standardů ICAO dále reguluje údržbu v rámci konkrétních nařízení. Tato nařízení jsou automaticky platná ve všech členských státech EU, tedy i v České republice. [3]

Nařízení komise č.1321/2014 ze dne 26. listopadu 2014 o zachování letové způsobilosti letadel a leteckých výrobků, letadlových částí a zařízení a schvalování organizací a personálu zapojených do těchto úkolů (dále jen nařízení), zrušilo platnost starého nařízení Komise č. 2042/2003. Veškeré činnosti spojené s údržbou letadel se řídí tímto nařízením. [3]

Nařízení v sobě nese veškeré informace potřebné během procesu získání a zachování letové způsobilosti letadel. Jsou zde obsaženy definice jednotlivých kategorií letadel, požadavky na zachování letové způsobilosti, proces schvalování organizací zapojených do zachování letové způsobilosti atd.

Nařízení obsahuje čtyři přílohy, které budou popsány níže: [4]

- Part CAMO

  Tato příloha se zabývá požadavky na organizace řídící zachování letové způsobilosti. Tato organizace má zkratku CAMO (continuing airwothiness management organization). Jejím úkolem je řídit zachování letové způsobilosti všech letadel, ke kterým je smluvně vázaná. Všechny letecké společnosti podnikající v obchodní letecké dopravě mají povinnost zavést CAMO, a to buď jako vlastní oddělení, nebo externím dodavatelem.

  CAMO vede a uchovává veškeré záznamy o letadlech, vytváří plány údržby, řídí příkazy k zachování letové způsobilosti, které do plánů údržby implementuje, hlídá termíny a požadavky. Kromě toho že letadlo je uvolněno do provozu organizací oprávněnou k údržbě letadel, musí být osvědčení o uvolnění do provozu vydáno

i nezávislým osvědčujícím personálem CAMO. CAMO úzce spolupracuje s národním regulačním úřadem (v České republice Úřad pro civilní letectví). Úřad schvaluje všechny plány údržby, změny ve výkladu organizace, určuje požadavky na personál atd. Úřad provádí plánované i neplánované audity.
Je nutné podotknout, že nařízení bylo v roce 2019 změněno prováděcími nařízeními Komise č. 1383/2019 a 1384/2019. Tyto změny se týkají především přílohy M, kde definují 3 části. Část CAMO byla popsána výše a týká se obchodní letecké dopravy. Další dvě části jsou CAO a ML a týkají se všeobecného letectví. CAO umožňuje jedné organizaci řídit jak zachování letové způsobilosti, tak provádět samotnou údržbu. Část ML se týká letadel jiných než složitých motorových.

- Part 145

Část 145 se zabývá požadavky na společnosti AMO (approved maintenance organization), organizace provádějící údržbu letadel. Organizace provádí fyzickou údržbu na letadlech, pro které má oprávnění. To definuje i rozsah prací, který musí být podrobně popsán ve výkladu organizace spolu s dalšími organizačními a provozními informacemi. Výklad organizace a jeho změny schvaluje příslušný úřad. Nařízení definuje na AMO požadavky na prostory, vybavení, skladování, personál. AMO má podobně jako CAMO povinnost uchovávat veškeré záznamy o údržbě a hlásit příslušnému úřadu provozní události. AMO dostává informace a pokyny k údržbě od CAMO ve formě plánu údržby. Mezi nejdůležitější práva organizace patří právo provádět údržbu na typech letadel, pro které má organizace oprávnění schválné úřadem, v rámci stanoveného rozsahu prací, a následné uvolnění letadla do provozu. Všechny společnosti podnikající v obchodní letecké dopravě musí mít zřízenou společnost AMO, nebo AMO nasmlouvat externě. AMO může také nasmlouvat externí práce u jiné AMO na práce, pro které nemá oprávnění.

- Part 66

Část 66 se týká požadavků na osvědčující personál v údržbě letadel. Osvědčující personál musí být držitelem průkazu způsobilosti. Tyto průkazy mají několik kategorií.

Pro každou jednotlivou kategorii průkazu nařízení stanovuje podmínky pro získání, jako jsou požadavky na základní znalosti a požadavky na praxi. K získání práv údržby na určitý typ letadla musí být k průkazu způsobilosti zapsána ještě příslušná kvalifikace na letadlo. Po splnění všech požadavků průkaz držiteli přináší právo

uvolňovat letadlo do provozu v rámci jeho oprávnění. Základní právo držitele průkazu způsobilosti je provádět fyzickou údržbu na letadle dle jeho oprávnění a vydávat osvědčení o uvolnění do provozu.

Part 66 dále stanovuje požadavky na výcvik personálu, který je prováděn výcvikovou organizací. Jsou zde podrobně popsány osnovy teoretické a praktické výuky dle jednotlivých kategorií průkazů, dále požadavky na minimální počet hodin teoretické výuky, zkoušky, zápočty atd. Výuka pro kategorie A, B1, B2, B2L, B3 a C je rozdělena do 17 modulů.

Kategorie a podkategorie průkazů způsobilosti jsou:

A – Osvědčující mechanik traťové údržby

B1 – Osvědčující technik traťové údržby – drak/motor/systémy

B2 – Osvědčující technik traťové údržby – avionika

B2L – Osvědčující technik traťové údržby – avionika – kromě letadel skupiny 1[1]

B3 – Osvědčující technik traťové údržby – letouny s pístovými motory bez přetlakové kabiny s maximální vzletovou hmotností 2000 kg a nižší

L – Osvědčující technik traťové údržby pro letadla dle podkategorií definovaných předpisem

C – Osvědčující technik údržby na základně

- Part 147

Tato část nařízení se týká organizace, která provádí výcvik personálu údržby, který je držitelem průkazu způsobilosti, dle výcvikových požadavků definovaných v části 66. Organizace provádějící odborný výcvik je MTO (maintanance training organization). Part 147 podrobně definuje požadavky na prostory a personál, uchování záznamů, vybavení, zkoušky, výcvikové postupy. Organizace musí mít výklad, který schvaluje příslušný úřad (ÚCL). Ve výkladu jsou uvedeny všechny informace o organizaci, jako organizační schéma, odpovědný vedoucí, vedoucí výcviku, popis prostorů, podrobný

[1] *Skupina 1: Nařízení komise č.1321/2014: složitá motorová letadla, vícemotorové vrtulníky, letouny s maximální letovou výškou větší než FL290, letadla vybavená elektroimpulzivními systémy řízení, plynové vzducholodě kromě ELA2 a jiná letadla vyžadující typovou kvalifikaci na letadlo, pokud to agentura stanoví.*

popis učebních osnov atd. Mezi organizace MTO schválené pro poskytování výcviku a provádění zkoušek patří i ČVUT, Fakulta dopravní.

2 CSAT

CSAT je zkratka společnosti Czech Airlines Technics [2], byla založena v roce 2010 jako dceřiná společnost Českých aerolinií. Celkově je ale zkušenost v oboru mnohem delší, přes 90 let, dříve bylo dnešní CSAT součástí technického úseku Českých aerolinií. V současné době je jediným akcionářem společnosti společnost Letiště Praha a.s. Celkově zaměstnává přes 800 zaměstnanců včetně osvědčujícího leteckého personálu certifikovaného podle Nařízení komise č. 1321/2014 part 66. [12]

Dle nařízení part 145 je společnost CSAT certifikovanou organizací AMO, organizací provádějící údržbu letadel. Má osvědčení provádět traťovou i těžkou údržbu, a to u následujících typů letadel.

Těžká údržba: [13]

- Airbus A318/A319/A320/A321 (CFM56)
- Airbus A319/A320/A321 (IAE V2500/ CFM LEAP 1-A)
- Boeing B737-300/400/500 (CFM-56)
- Boeing B737-600/700/800/900 (CFM-56)
- Boeing B737-8/9 MAX (CFM LEAP 1-B)
- ATR 42/72 (PWC PW120)

Traťová údržba: [13]

- Airbus A318/A319/A320/A321 (CFM56)
- Airbus A319/A320/A321 (IAE V2500/ CFM LEAP 1-A, IAE PW1100G)
- Airbus A330 (GE CF6/ PW 4000/ RR Trent 700)
- Boeing B737-300/400/500 (CFM-56)
- Boeing B737-600/700/800/900 (CFM-56)
- Boeing B737-8/9 MAX (CFM LEAP 1-B)
- Boeing B757-200/300 (PW 2000/ RR RB211)
- Boeing B767-200/300/400 (PW 4000/ PX JT9D/ GE CF6)
- Boeing B777-200/300 (GE 90)
- Boeing B787-8/9/10 (GEnx/ RR Trent 1000)
- ATR 42/72 (PWC PW120)

[2] https://www.csatechnics.com/cs/about-us

- Embraer ERJ-170/190 (GE CF34)

Dále je zde oprávnění k provádění údržby komponentů na letadlových celcích jiných než kompletní motory a APU: [13]

- Klimatizace a přetlakování
- Automatické řízení letu
- Spojení a navigace
- Dveře – nouzové východy
- Elektrické zdroje
- Vybavení
- Motor – APU
- Řízení letadla
- Palivo – drak
- Hydraulika
- Přístroje
- Přistávací zařízení
- Kyslík
- Vrtule
- Pneumatické systémy
- Ochrana proti námraze/dešti/požáru
- Konstrukce draku

A oprávnění specializovaných služeb – nedestruktivní testování: [13]

- Metoda prozařování
- Kapilární metoda
- Elektromagnetická metoda
- Termografická metoda
- Ultrazvuková metoda
- Elektroindukční metoda

K těmto pracím má společnost prostory na území Letiště Václava Havla a to hangár S, kde se provádí především traťová údržba, a hangár F určený především pro těžkou údržbu, kam se vedle sebe vejde až 6 letadel. Mezi doplňkové služby patří údržba podvozků, drakové opravy, prodej a zapůjčení materiálu a nářadí, podpora provozovatele letadla, kam lze zařadit služby správy údržbového plánu, program spolehlivosti atd. [12]

Jednou z priorit společnosti je zachování úrovně kvality a postupů, které mohou mít vliv na bezpečnost letového provozu, životní prostředí nebo ochranu před protiprávními činy. Tohoto je docíleno pomocí Safety management systému a Quality management systému. [12]

3 Proces údržby

V této kapitole bude popsán celý proces údržby letadla od předání dokumentace výrobce až po samotnou fyzickou údržbu. Tento proces není platný jen pro společnost CSAT, ale v obecném měřítku je tento proces stejný pro všechny společnosti oprávněné k údržbě letadel.

V centru údržby letadla stojí provozovatel. Ten je povinný udržovat letadlo v provozuschopném stavu a dodržovat intervaly údržby, v opačném případě je letadlu odejmuta letová způsobilost a provozovatel musí letadlo uzemnit.

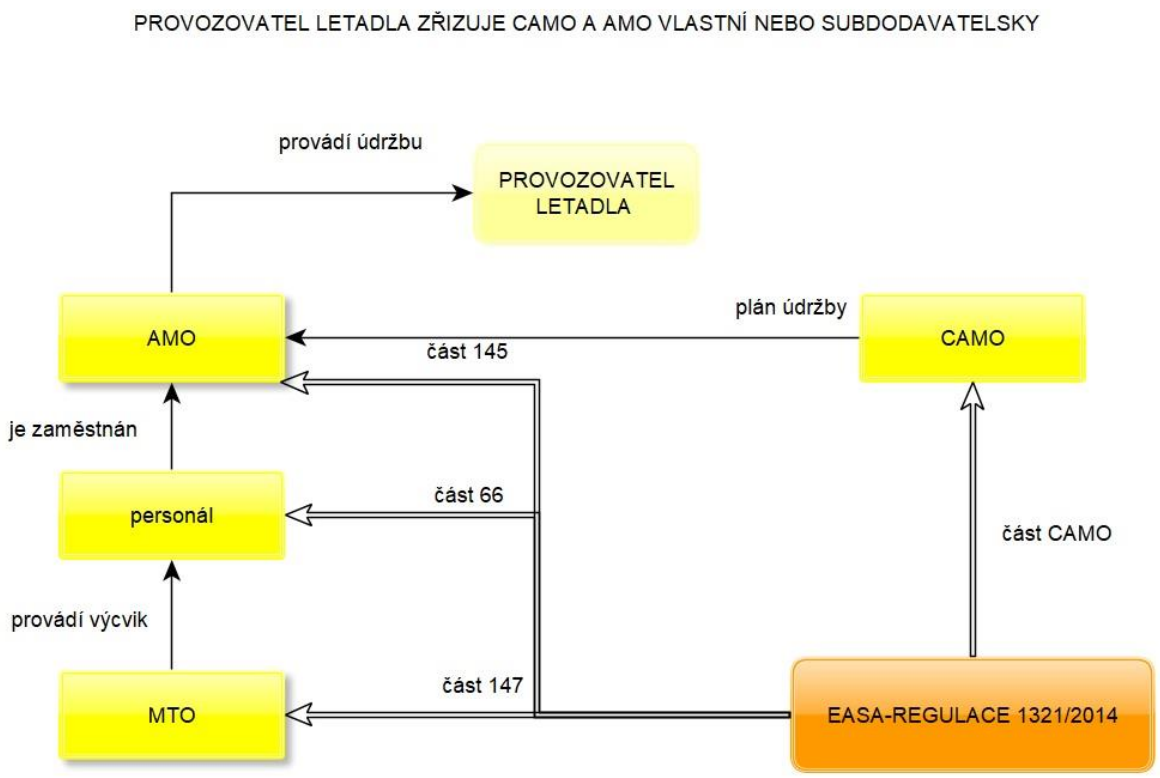
Legislativní požadavky na údržbu jsou v odvětví civilního letectví velmi přísné, proto do toho procesu zasahuje několik organizací jako provozovatel letadla, CAMO, AMO, MTO. Nad těmito organizacemi v konkrétním státě dohlíží místní Úřad pro civilní letectví, v Evropě pak nad vším dohlíží EASA.

V civilním letectví je provozovatel letadla povinen zřídit organizaci CAMO (nebo subdodavatelsky). Legislativní požadavky na tuto organizaci vycházejí z Nařízení komise (EK) č. 1321/2014 část CAMO, byla popsána výše. CAMO je zodpovědné za vytvoření a udržování údržbového plánu, řízení příkazů zachování letové způsobilosti, hlášení závad, řízení modifikací, zpracování programu spolehlivosti, plánování údržby, řízení MELu (minimum equipment list), přejímku a předání letadel, a uchovává veškerou dokumentaci, a to pro každé letadlo.

Provozovatel letadla je rovněž povinen zřídit organizaci AMO (nebo subdodavatelsky). Legislativní požadavky na tuto organizaci vycházejí z Nařízení komise (EK) č. 1321/2014 část 145, byla popsána výše. Hlavní činností této organizace je provádění fyzické údržby na letadle a letadlových celcích, a to na základě pokynů organizace CAMO.

Organizace AMO je povinná zaměstnávat techniky provádějící údržbu na letadle s certifikací dle Nařízení komise (EK) š. 1321/2014 část 66. Tento personál je školen organizací MTO dle Nařízení komise (EK) š. 1321/2014 část 147.

Schéma tohoto procesu a komunikace mezi organizacemi je zobrazeno na obrázku 1.



Obrázek 1: schéma procesu a komunikace organizací zahrnutých do údržby letadel

3.1 Plán údržby

Důležitým dokumentem pro organizaci AMO je již zmíněný plán údržby. Ten sestavuje organizace CAMO. Jeho sestavení je ale složitý proces a je potřeba do něj zahrnout informace z mnoha zdrojů.

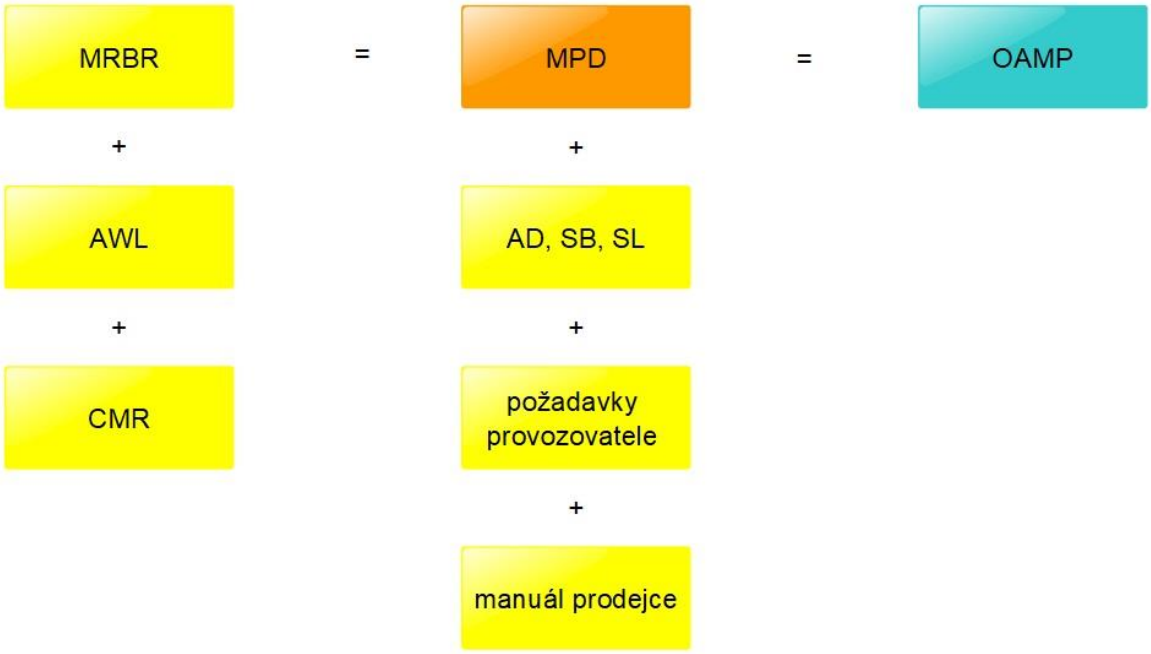
Základním zdrojem informací o požadovaných úkonech a jejich frekvencích je dokument Maintanance review broad report (MRBR). Tento dokument zahrnuje požadavky EASA a FAA zahrnuté v předpisech letové způsobilosti. Před uvedením nového letadla na trh je výrobce povinen navrhnout a předložit počáteční minimální požadavky na údržbu. Ty jsou schváleny a následně do MRBR zahrnuty. MRBR schvaluje výbor pro přezkoumání systému údržby a jsou v něm zastoupeni zástupci výrobce, provozovatelů, certifikačního úřadu a dodavatelů. MRBR schválený místními regulačními úřady se používá jako rámec, na jehož základě provozovatele sestavují vlastní programy údržby. Mohou se ve výsledku lišit, počáteční požadavky jsou ale vždy stejné.

Další dokumenty, které je třeba zahrnout jsou další dokumenty schvalované certifikačním úřadem, v Evropě EASA, a doporučení výrobce. Jedná se o CMR (certification maintanance requirements) a AWL (airwothiness limitations). Jedná se o doplňkové

prohlídky nad rámec základního programu. Spojením těchto dokumentů vznikne nový dokument MPD (maintanance planning document).

OAMP (operator approved maintanance program) je program údržby leteckého provozovatele. Je sestavený konkrétnímu provozovateli na míru a zahrnuje další informace, které nejsou v MPD zahrnuty. Jedná se o požadavky provozovatele, dodatečné ICA (AD, SB, SL), informace obsažené v manuálech motoru, APU a manuálu prodejce. Z OAMP vycházejí jednotlivé plány údržby pro jednotlivá letadla, které musí schválit místní regulační úřad.

Schéma návaznosti dokumentů a vznik OAMP je zobrazen na obrázku 2.



Obrázek 2: Schéma vzniku OAMP

3.2 Technická dokumentace

Prvotním zdrojem informací po zakoupení letadla je dokumentace, která je dodána výrobcem spolu s letadlem, tzv. průvodní technická dokumentace. Mezi tyto dokumenty patří:

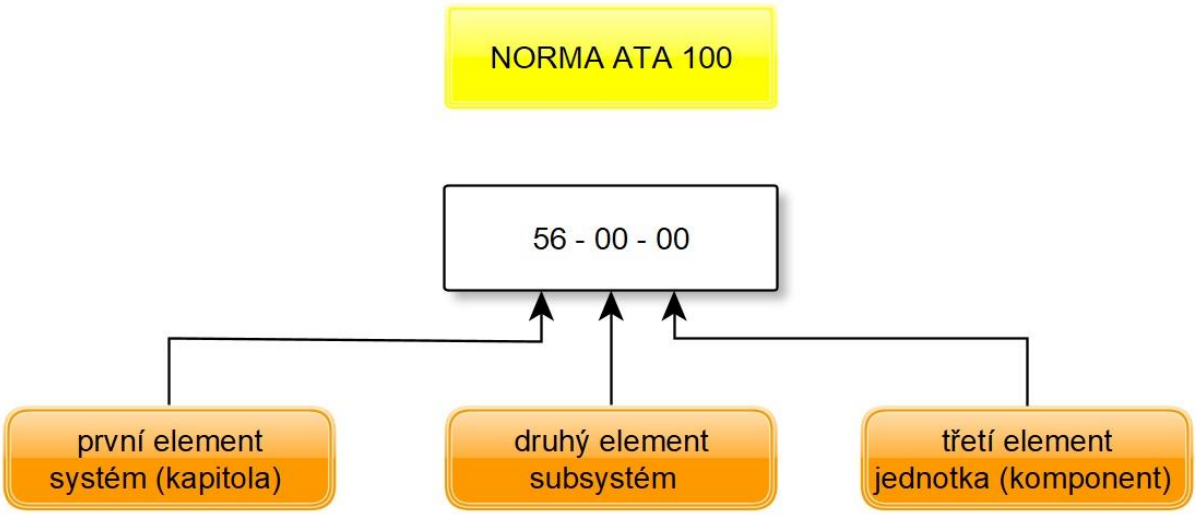
- Letová příručka (AFM – aircraft flight manual)
- Instrukce pro zachování letové způsobilosti (ICA – instruction for continued airworthiness)
- Ilustrovaný kusovník (IPC – ilustrated parts catalogue)
- Základní seznam minimálního vybavení (MMEL – master minimum equipment list)
- Seznam povolených odchylek na draku (CDL – configuration deviation list)
- Příručka nakládání a vyvažovaní (WBM – weight and balance manual)
- Příručka údržby (MM – maintenance manual)

Další dokumentace, kterou vydává výrobce letadla jsou dodatečné ICA. Podrobně byly popsány výše.

Veškerá technická dokumentace v letecké dopravě je členěna dle americké normy ATA 100. Smyslem této normy je organizace a orientace dat v příručkách a jejich vzájemná kompatibilita. Jedná se společný standard, který udává systém číslování a odkazovaní. Je stejná pro všechny typy letadel.

Systém číslování kapitol a organizace je zobrazen na obrázku 3. První element udává systém (kapitolu), jedná se o větší funkční celek na letadle. Kapitola 0-20 se věnuje letadlu obecně, 21–49 je věnována letadlovým systémům, kam patří např. hydraulický systém, palivový systém nebo protipožární systémy. Kapitoly 50-57 se věnují konstrukci (křídla, okna, stabilizátory) a kapitoly 71-85 pohonným jednotkám. [14]

Druhý element udává subsystém, tedy menší část definovaného systému, který sám o sobě může fungovat jako systém. U posledního elementu se jedná o konkrétní letadlový komponent.



Obrázek 3: Organizace dle normy ATA100 (kapitola 56 - křídla)

Nejdůležitějším dokumentem z hlediska údržby je příručka údržby (MM). Jedná se o nejrozsáhlejší příručku s údaji pro údržbu. Jejím obsahem je popis jednotlivých systémů, umístění a popis komponentů a jejich funkce. Dále podrobně popsané postupy provádění údržby, kontrol, montáže, demontáže a oprav. U velkých složitých letadel, která jsou jako celek velice komplexní, by byla tato příručka příliš obsáhla, což by mělo za následek špatnou orientaci. Proto tato základní příručka neobsahuje veškeré informace, ale odkazuje na další podrobnější manuály:

- OHM (overhaul maintenance manual)
  Zde je možné nalézt informace o opravách na komponentech, které lze sejmout z letadla
- SRM (structural repair manual)
  Obsahuje informace pro provádění oprav na konstrukci letadla.
- WDM (wiring diagram manual)
  Tento manuál slouží pro znázornění elektrických zařízení a okruhů na letadle.
- FIM (fault isolation manual)
  Tento manuál slouží k identifikaci poruchy. Po nalezení příčiny uvádí postup odstranění závady nebo se dále odkazuje na ostatní manuály.
- SSM (system schematic manual)
  Graficky a schematicky popisuje funkce jednotlivých systémů.
- NDTM (non-destructive testing manual)
  Obsahuje obecné i podrobné informace o metodách nedestruktivního testování. Popis postupů jedlových metod při použití na konkrétní části letadla.
- ITEM (illustrated tool and equipment manual)

Ilustrovaný seznam veškerého nářadí a vybavení včetně testovacího vybavení.

- CMM (component maintenance manual)
  Jedná se o instrukce pro údržbu komponentů sejmutých z letadla. Tento manuál vytváří výrobce komponentů dodávaných externě.

Všechny výše zmíněné manuály jsou osvědčujícímu personálu během práce k dispozici a ti jsou povinni je používat a přesně se jimi řídit.

3.3 Plánovaná údržba

Plánování údržby provádí CAMO. Úkolem je plánovat předepsanou údržbu (a co nejefektivněji zahrnovat i neplánovanou údržbu). Prakticky po objednání údržby přijde společnosti AMO od CAMO seznam požadovaných prací. Ty musí být převedeny na konkrétní postup v souladu s dokumentem AMM (aircraft maintanance manual).

Jednotlivé úkoly jsou rozděleny na tzv. technologické karty nebo task (job) cards. Jedná se o přizpůsobený popis úkolu údržby připravený z původní technické dokumentace (AMM) odpovědným oddělením AMO. Jsou zpracovány tak, aby se usnadnilo správné dokončení tohoto úkolu osobami pověřenými jeho provedením. Zároveň se dají používat opakovaně, neboť práce, ke kterým se task karty vztahují, se opakují.

Task cards jsou jednoduchý způsob, jak provést údržbu (v souladu s předpisy), zároveň je to způsob přehledný. Konkrétnímu mechanikovi dle jeho kvalifikace je přidělena task karta, na té je přehledně zobrazený seznam vyžadovaných prací, odkaz do manuálu údržby, požadované vybavení, časová náročnost. Task karty poskytují tak přehledně vše, co je potřeba o úkolu vědět. Po provedení prací jsou archivovány, takže je jednoduché zjistit, kdo konkrétně práci vykonal.

3.4 Neplánovaná údržba

Kdykoli během práce na letadle nebo jeho užívání může dojít k nálezu závady. Dle závažnosti a času nálezu se závada může odložit (dle podmínek MEL), pokud je to možno opravit závadu v rámci traťové údržby nebo letadlo uzemnit a závadu odstranit v rámci údržby na základně.

V případě AMO dojde k nálezu závady nejčastěji při jiných pracích na letadle (provádění revizí, kontrol, testů). Jedná se jakoukoli závadu, nesrovnalost nebo nevyhovující stav. V takovém případě musí dojít k zaevidování závady a kontaktování zákazníka. Je třeba zákazníka informovat o nálezu a zároveň domluvit i případnou požadovanou formu odstranění.

3.4.1 Záznam závady

Každou nalezenou závadu je třeba řádně zaznamenat. Závada může být detekovaná kdykoli při preventivních kontrolách, při provádění plánovaných i neplánovaných oprav a údržeb nebo i při pochůzce posádky. V případě nálezu je osvědčující personál povinen vytvořit záznam o nálezu. Z provozních důvodů a časové náročnosti práce k záznamu dochází nejčastěji do papírového formuláře. Cílem je co nejpřesněji zaznamenat a popsat nález. Jako zdroj informací technického charakteru a přesné názvy letadlových komponentů slouží průvodní technická dokumentace, především pak příručky údržby.

Na základě těchto příruček dojde k identifikaci závady a poškozeného komponentu. Formulář má kromě pole volného textu pro popsání nálezu, který je v rámci této práce nejdůležitější, i řadu strukturovaných polí. Tato pole slouží k formálnímu záznamu nálezu a později snadné orientaci. Jedná se o datum nálezu, typ záznamu, registrační číslo letadla, zařazení nálezu dle ATA100, vygenerované číslo záznamu apod.

Tento záznam je nutné převést do elektronické podoby, nejčastěji přepisem papírového formuláře do elektronického , který je propojen s interní databází. Na základě tohoto elektronického záznamu dojde později ke zhodnocení závažnosti a vygenerování pracovního příkazu.

Vzhledem k tomu, že nález je neplánovaná práce, neexistuje k němu předpřipravená task karta, která by se dala použít, je tady třeba využít něčeho jiného.

V CSAT se pro záznam nálezu využívá formulář „workorder" dále jen WO. Je to formulář, který má strukturovanou hlavičku, kde se zaznamená vše potřebné. Dále je zde pole volného textu, kde je žádoucí co nejpřesněji popsat daný problém. Záznam je odeslán do systému, kde je nález vyhodnocen, a je mu přiřazená odpovídající akce (například popis opravy, výměny, reference na určitou kapitolu manuálu).

V případě nálezu je vygenerovaný nový WO a je mu přiřazeno unikátní číslo, je vyplněna hlavička se všemi požadovanými údaji. Dále zápisem závady vzniká ve WO tzv. work step. Ten představuje výchozí záznam pro další akce, které z nálezu vyplynou. Jako reakce na nález jsou poté k work stepu přiřazeny akce, kde je popsáno, co se udělalo, co se má udělat, popis, instrukce, odkaz do technické dokumentace atd. V případě odlišného nálezu, který je objeven např. při výkonu předepsané opravy, je možné do WO přidat další work step, na který budou navazovat jiné akce. WO lze uzavřít osvědčujícím pracovníkem až ve chvíli, kdy jsou vyřešeny a uzavřeny všechny work stepy.

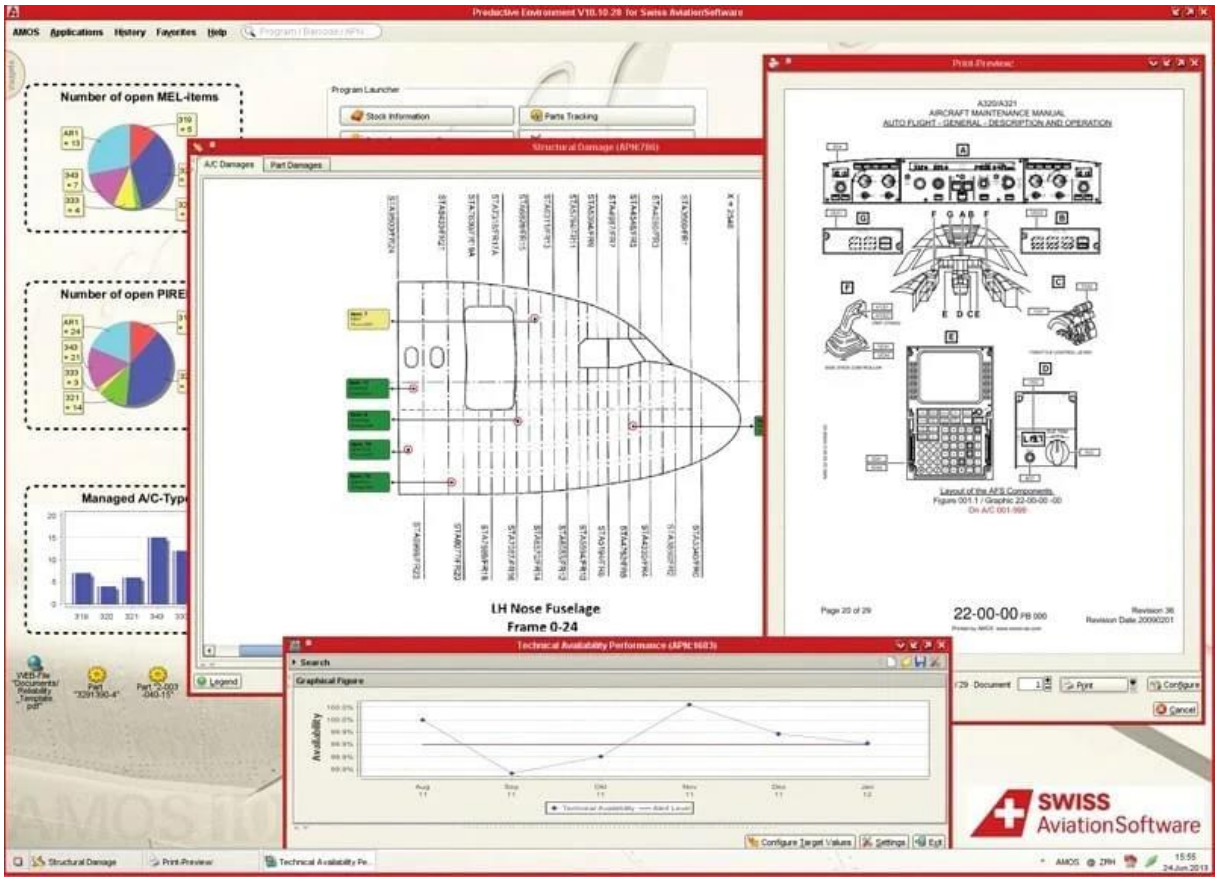
3.5 Softwary pro údržbové organizace

K uchovávání velkého množství údajů o provedených pracích je žádoucí využití vhodného softwaru hlavě ze strany větších AMO. Během údržby vzniká velké množství záznamů a dokumentů, které je třeba po určitou dobu uchovávat. Uchovávání velkého množství papírových záznamů je nepraktické. Navíc vhodný software umožňuje mnohem více než jenom ukládání. Umožňuje plánování revizí, vytváření task karet, workorderů, distribuci, systém elektronických podpisů a potvrzování práce atd. Zpětně lze v případě potřeby jednoduše vyhledat starší záznamy dle požadovaných parametrů.

Na trhu je dostupných několik softwarů určených pro AMO organizace. CSAT využívá software AMOS od společnosti Swiss avition industries [15]. Prostředí toho softwaru je na obrázku 4.

AMOS má pro toto odvětví široké využití. V CSAT je využíván mimo samotnou fyzickou údržbu napříč všemi organizačními jednotkami (např. příjem a výdej, produkční plánování, finanční kontrola, marketing). Umožňuje jednoduchou komunikaci a sdílení informací mezi těmito útvary. V AMOS je možné plánování revizí, kapacit, lidských zdrojů, práce s legislativními požadavky a technickou dokumentací. Jsou zde předdefinované formuláře pro vytvoření potřebných dokumentů jako task karty nebo workordery. Ty je pak možné distribuovat jednotlivým pracovníkům. Zároveň mají všichni uživatelé unikátní login, který zajistí jednoznačný záznam práce. Je možné nahrávat i externí soubory jako fotografie nebo objednávky a ty ukládat k dotyčným formulářům. Jednotlivé dokumenty je možné navzájem propojit nebo vytvářet reference. To je důležité především u workorderu, který byl popsán výše.

Při vytvoření záznamu o nálezu dojde k vygenerování nového WO s unikátním číslem. Po popisu závady je třeba přidat referenci na konkrétní task kartu nebo jiný WO, při které k nálezu došlo. Jedním kliknutím se tak uživatel dostane na danou referenci. Poté je na nález vytvořena akce, například výměna. Opět je zde možné vytvořit referenci, např. na konkrétní část manuálu, kam se uživatel jednoduše dostane. Po provedení práce je toto potvrzeno a WO je buď uzavřen, nebo dle potřeby dále rozšířen. Po dokončení je uzavřen osvědčujícím pracovníkem a až poté je třeba tento soubor vyexportovat, fyzicky podepsat a předat zákazníkovi spolu s další dokumentací. Během práce, která může trvat i několik týdnů tedy není nutné vytvářet nové formuláře pro jednotlivé akce a reakce, veškerá komunikace se odehrává v AMOS.



Obrázek 4: Prostředí softwaru AMOS [16]

4 Textová analýza

Textová analýza má za cíl extrahování strojově čitelných dat a faktů z nestrukturovaného volného textu. Účelem je tedy vytvoření strukturovaných dat z nestrukturovaného textu. Jako synonymum k procesu textové analýzy se používá i výraz „dolování textu“ (z anglického „text mining“). Tento proces využívá přístupu zpracování přirozeného jazyka (natural language processing – NLP), kde cílem není úplné porozumění textu, ale způsob jak získat data co nejpraktičtějším způsobem. To znamená rovnováha mezi složitostí vytvoření analytického postupu a jeho náklady jako je např. výkon nebo paměť a samotnou přesností. Příkladem jednoduchého NLP může být počítání slov nebo frekvencí, zatímco vysoce pokročilý NLP může být odpovědí na otázky v lidském jazyce. Příklady aplikací NLP jsou webové vyhledávače, strojový překlad. [5]

Textová analýza jde provést dvěma způsoby: [6]

- Manuálně: tento přístup spočívá v ručním procházení záznamů člověkem. Často se jedná o záznamy ve formě databází nebo tabulek. V případě velkého množství dat je tento způsob neefektivní, časově náročný a nepřesný. S manuální textovou analýzou se ale setkáváme v každodenním životě. Např. při čtení zpráv. Při procházení článku automaticky informace zpracováváme a filtrujeme tak, abychom si zapamatovali to důležité.
- Automaticky: tento přístup využívá počítače a softwarů určených k textové analýze. Jejich základem je algoritmus, který využívá strojového učení a přístupu zpracování přirozeného jazyka, což je přístup, který strojům umožňuje číst, porozumět a replikovat řeč člověka.

Textová analýza (často automatická) je využívána společnostmi, které pracují s velkými objemy dat, jako jsou marketingové a obchodní společnosti. Tato potřeba se dále zvyšuje s rostoucím zájmem obchodování a obecně fungování v online prostoru, který generuje obrovské množství dat. To má za následek informační přetížení, tedy stav, kdy máme tolik informací, že nejsme schopni najít ty podstatné a definovat trendy. Zde se otevírá prostor pro využití textové analýzy. [6]

Proces extrahuje strojově čitelná fakta z velkého množství textu a umožňuje tato fakta dále zadávat do databází, tabulkových programů nebo jiných programů. Toto je dále využito pro analýzu trendů v datech.

Automatická textová analýza se může zaměřit na konkrétní data, která nás v danou chvíli zajímají např. jména, čísla apod. Časová výhoda automatické analýzy je, že výsledek je

k dispozici v podstatě okamžitě. To umožňuje efektivně alokovat ostatní zdroje, jako jsou lidské zdroje nebo peníze.

4.1 Textová analýza v kontextu technických dat

V kontextu této práce budou automatické textové analýze podrobena data z letecké údržby. Jedná se o data, která jsou částečně strukturovaná. Většinou se jedná o formu tabulky. Strukturovaná data v záznamech jsou informace o registraci letadla, číslo záznamu, datum apod. V záznamech se ale vyskytuje i pole pro volný text. Jedná se o pole, kam technik popíše co nejpřesněji objevenou závadu. V poli by se měla vyskytovat informace o tom, co je poškozené (komponenta) a jak je to poškozené (failure). Příklad těchto záznamů je v tabulce 1.

Jedná se tedy o text, který obsahuje technické pojmy, který zaznamenává člověk, což s sebou nese fakt, že je do textu vložena určitá neurčitost. Tedy překlepy, nepřesné zadání technického pojmu, skloňování apod. Proto je na tato data vhodné aplikovat nástroj na textovou analýzu, a ne jen jednoduchý nástroj pro počítání četnosti výskytu pojmů.

Tabulka 1: Příklad záznamů v poli pro volný text

WO text
peeled-off paint found on rh inb main and aft flap lower and upper surface(marked with tape)
ext - worn-out seal found on rh inbd main flap inbd side (marked with tape)
ldg: the placard with content:flow regulator ldg gear sys of nww ceiling was found missing.
ext - worn-out seal found on lh inbd main flap inbd side (marked with tape)
ldg: block clamp above lh mlg fwd trunnion housing was found with tron block.
aft outb block clamp near lh mlg aft bearing was found with unserviceable block
ext.tail: the fairing strip was found cracked at several location on the te of the access panel 343cb.

V případě, že by se na tento nestrukturovaný text aplikovala automatická textová analýza, podařilo by se vysledovat trendy závad. Tedy jaké komponenty nejčastěji selhávají, jakým způsobem, jaký je poměr k ostatním závadám apod. Toto umožní letecké společnosti a údržbové organizaci efektivně alokovat zdroje, v tomto případě převážně finance a lidské zdroje. Vysledovaný vzorec nejčastějších selhání povede k efektivní úpravě preventivních úkonů a ve výsledku k úspoře financí.

4.2 Dostupná literatura

Při procházení dostupné vědecké literatury nebyl nalezen žádný zdroj, který by se věnoval využití dat z letecké údržby pro automatickou textovou analýzu. Existují ale

články zabývající se zpracováním dat technického charakteru z údržby z odvětví jiného než letectví.

První článek s názvem „*Natural Language Processing of Maintenance Records Data*“ se zabývá obecně záznamem, uchováním a strukturou dat z údržby. Struktura záznamu obsahuje kromě zaškrtávacích seznamů a polí také pole s volným textem, která mohou obsahovat libovolné množství znaků. Právě na toto pole je vhodná aplikace automatické textové analýzy, neboť manuální by byla příliš zdlouhavá a náročná. [7]

Kvalita analýzy závisí na kvalitě zdrojových dat. Tento fakt může hrát klíčovou roli v procesu rozhodování v údržbové organizaci. Rozhodující je distribuce správných dat správnému uživateli ve správném čase a ve správné kvalitě.

Základem analýzy je seznam tzv. tokenů (unikátních pojmů), výstupem je statistika výskytu těchto pojmů získaných přístupem NLP. Výsledkem této analýzy byl fakt, že analýza textu za pomocí NLP odhalí až o 40 % více pojmů než jednoduché vyhledávání četnosti výskytu konkrétního pojmu jednoduchým např. tabulkovým softwarem (pozn. Originální text byl analyzován ve švédštině). NLP dokázalo eliminovat přehlédnutí výrazů např. z důvodu překlepu, množného čísla, skloňování. [7]

Druhý článek s názvem „*Natural Language Processing and Classification Methods for the Maintenance and Optimization of US Weapon Systems*“ využívá data z oblasti údržby amerických zbraňových systémů. Hlavním problémem při zpracování těchto údajů je určení způsobu, jak extrahovat užitečné informace z neuspořádané krátké formy textu za účelem optimalizace jejich systému údržby. [8]

V základním slovníku se vyskytují dva druhy pojmů, a to objekt a akce. Z těch se vytváří páry. Výstupem je poté statistika četnosti a cílem nalezení základních vzorů v kombinacích objektů a akcí.

Tento přístup vytvoření páru dvou druhů pojmů ve slovníku je vhodný i pro řešení této práce. Umožňuje vytvořit pojmy typu komponent a závada a jejich následné nalezení a spárování v jednotlivých záznamech. To umožní hledání souvislostí a vzorců ve výskytu pojmů, které by za použití ruční analýzy nebyly zaznamenány.

Výsledkem analýzy vědeckých článků z oblasti automatické textové analýzy nestrukturovaných dat pro tuto diplomovou práci je, že automatická textová analýza je pro tento typ dat z letecké údržby vhodná a je pravděpodobné, že bude vykazovat vyšší přesnost než manuální zpracování. Základem pro analýzu je referenční slovník pojmů, na

základě kterého program analýzu provede. Ve slovníku se rovněž budou vyskytovat dva druhy pojmů. Vytvoření tohoto referenčního slovníku je jedním z cílů této práce.

Dále je zde diplomová práce s názvem „*Textová analýza nestrukturovaných závadových dat v letecké údržbě*". Autor má k dispozici data ve formě, která byla popsána výše. V programu TermIt, který bude podrobně popsán níže, vytváří referenční slovník pojmů, obsahující anotace typu komponent a závada. Na základě toho slovníku program provede automatickou textovou analýzu. Po provedení analýzy jsou data zanalyzována ručně a porovnána s výsledky automatické analýzy. Automatická analýza označila správně v 69 % řádků komponentu i závadu v případě závislé sady dat, ze které byl slovník vytvářen, respektive správně označil 46 % řádků komponentu i závadu v případě nezávislé sady dat. [9]

Jak již bylo popsáno výše, v datech tohoto typu mohou být ukryty informace, které by za okolností jen obyčejného uskladnění dat zůstaly neodhaleny. Tyto informace mohou být pro údržbovou společnost velice cenné. Mohou pomoci v procesu plánování údržby, zefektivnění celého procesu a ve výsledku šetří peníze. Aby ale byl celý proces v praxi efektivní a především použitelný, je potřeba, aby analýza vykazovala spolehlivé výsledky. Je tedy třeba získat vyšší procento úspěšnosti označení pojmů.

Tato diplomová práce navazuje na práci „*Textová analýza nestrukturovaných závadových dat v letecké údržbě*". Zde bylo cílem vytvořit referenční slovník pojmů a na sadách dat poskytnutých společností CSAT provést analýzu a uvést výsledky. V závislosti na výsledcích této práce je cílem této práce navrhnout řešení, které bude mít vyšší úspěšnost značení pojmů než v uvedené diplomové práci. Dalším dílčím cílem je sepsat doporučení generalizované pro všechny provozovatele, jak zvýšit efektivitu textové analýzy.

4.2.1 Limitace současného stavu

Jedna z limitací současného stavu se skrývá v úspěšnosti textové analýzy ve zmíněné diplomové práci. Na tuto limitaci je práce zaměřena. Aktuálních 69 % u závislé a 46 % u nezávislé sady dat není špatný výsledek. Pro praxi je ale žádoucí, aby byla úspěšnost co nejvyšší, aby mohlo být co nejvíce dat zpracováno. Cílem práce je tuto limitaci zmírnit, aby se úspěšnost co nejvíce zvýšila a bylo možné rozhodnout o aplikovatelnosti automatické textové analýzy na tento typ dat.

Jak bylo popsáno výše v kapitole o procesu záznamu závady, záznam vznikne na základě preventivní prohlídky letadla nebo plánované či neplánované údržbě. Případná závada

nebo poškození musí být řádně zaznamenána a popsána. Problém je, že příslušná dokumentace pro záznam není mechanikovi vždy při ruce. K záznamu tedy z pravidla dochází až po nějakém čase např. po provedení celé kontroly nebo opravy. Tato skutečnost zvyšuje pravděpodobnost, že poté dojde k vynechání, překlepům nebo pozměnění důležitých informací. Celou skutečnost ještě umocňuje situace, kdy dojde odhalení více závad najednou.

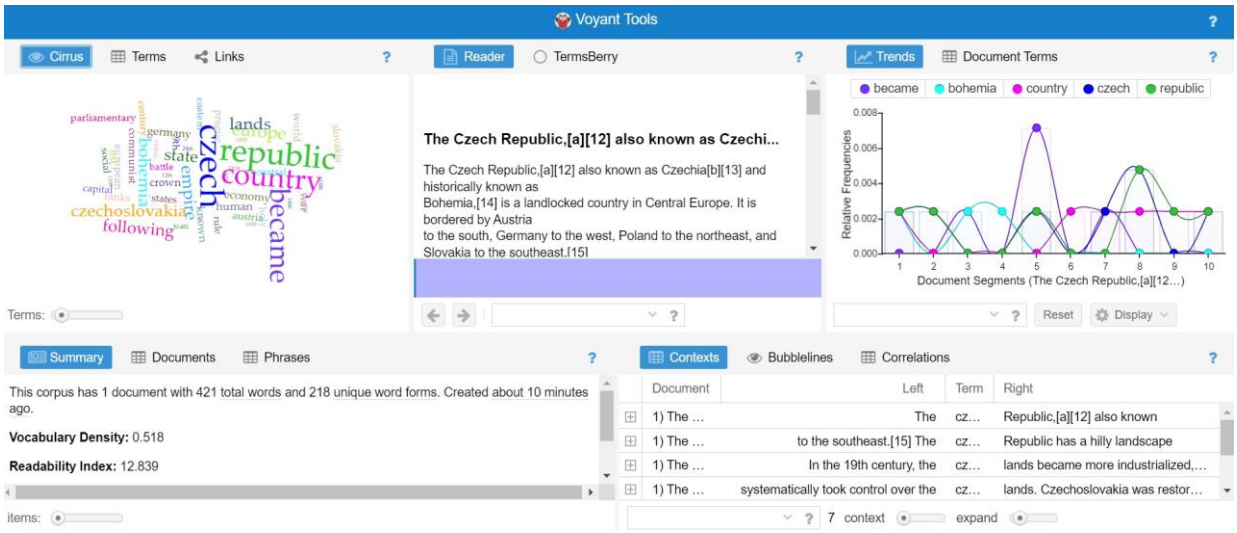
Tento problém by částečně vyřešilo přenosné zařízení s přímým přístupem do systému. Toto zařízení např. tablet by bylo k dispozici kdykoli během prohlídky, nedocházelo by tedy k prodlevě mezi nálezem závady a záznamem.

Další limitací je fakt, že záznamy o závadách nevytváří v průběhu času jeden mechanik, který bude stejnou skutečnost popisovat neustále stejně. Záznam o závadě může vytvořit jakýkoli mechanik. Tato skutečnost přináší fakt, že každý člověk je individuální ve vnímání a jazykovém vyjadřování. Může se tedy stát, že jedna stejná závada bude popsána více různými způsoby za použití různých slov a synonym. V tomto odvětví neexistuje žádné ustálené názvosloví ani přesná pravidla pro záznam závady.

Velký objem dat z tohoto prostředí prakticky znemožňuje využití manuální analýzy. Je tedy vhodné využít automatické analýzy. K tomu abychom získali relevantní informace, je třeba dobře znát výše uvedené limitace. K tomu, aby bylo možné automatické analýzy využít, je třeba mít k dispozici definiční slovník pojmů, které nás zajímají. V tomto případě to je slovník komponent a závad. Ke správnému provedení analýzy a k získání co největšího počtu relevantních informací je třeba mít k dispozici nástroj, který umožní definovat a anotovat pojmy tak, aby se co nejvíce eliminovaly výše zmíněné limitace. Správné nastavení atributů umožní programu odhalit i překlepy nebo synonyma, které by za jiných okolností zůstali nepovšimnuté.

5 Programy

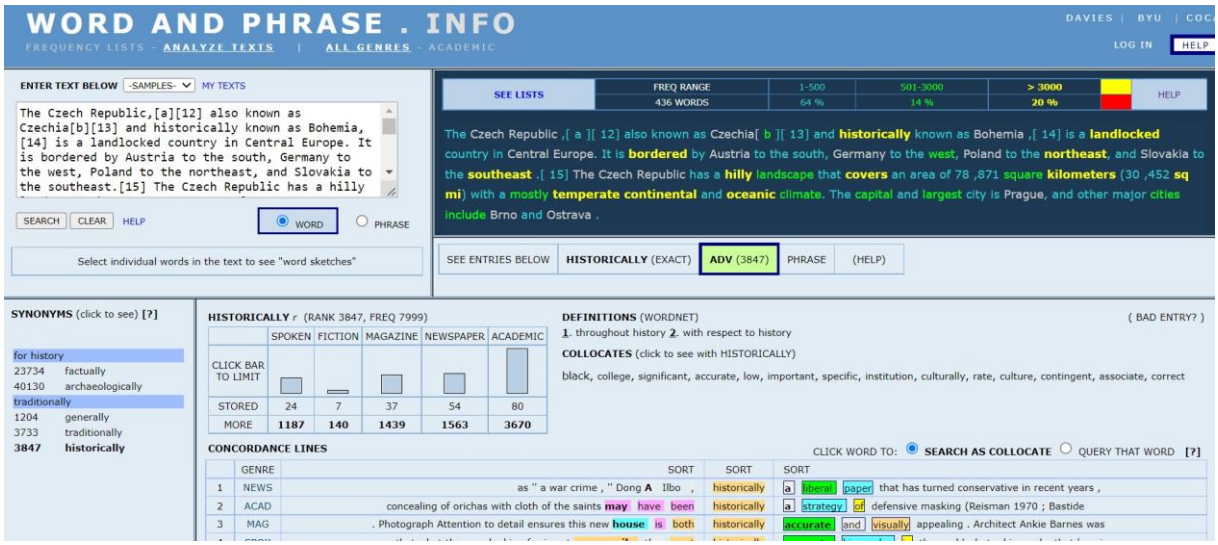
Existuje široká škála programů, které nabízejí nějakou formu automatické analýzy volného textu. Jedná se desktopové i webové aplikace . Většina těchto softwarů pracuje na základě NLP. Jednodušší webové aplikace poskytují funkci vložení volného textu a následnou automatickou analýzu, která většinou zahrnuje hloubkovou analýzu četnosti výskytu pojmu a vývoj trendu. V rámci výsledků lze vyhledat jednotlivé pojmy, zjistit jejich čestnost, umístění v textu apod. Pro ilustraci budou uvedeny jako příklad dva různé nástroje a následně podrobně popsán nástroj, který byl zvolen pro tuto práci. Mezi takovéto nástroje patří online volně dostupný software Voyant[3]. Výstup analýzy z toho programu je na obrázku 5. [10]



Obrázek 5: Prostředí softwaru „Voyant" [10]

Další webovou aplikací je „Word and phrase[4]". Podobně jako výše zmíněná aplikace provede automatickou textovou analýzu volného textu. Ve výstupu ale nabízí interaktivní prostředí. V konzoli je možno vybrat jednotlivá slova a zobrazovat statistiky. Analýza určí druh a online vyhledá i definici výrazu. Je možné slova označit štítky a roztřídit je tak do podkategorií. Prostředí tohoto softwaru je na obrázku 6. [11]

[3] https://voyant-tools.org/
[4] https://www.english-corpora.org/coca/



Obrázek 6: Prostředí softwaru "Word and phrase" [11]

6 TermIt

K tomu, abychom mohli efektivně provádět analýzu na datech záznamů závad, je třeba provést analýzu účelně. Tedy neanalyzovat celý text, ale jen tu část nebo ty fráze, které jsou pro nás klíčové. V tomto případě komponenty a závady. Je potřeba referenční slovník pojmů. Je nezbytné najít nástroj, který umožňuje vytvoření tohoto referenčního slovníku a umožní vytvářet a upravovat atributy a podmínky vyhledávání. Nástroj, který toto umožňuje je TermIt [18].

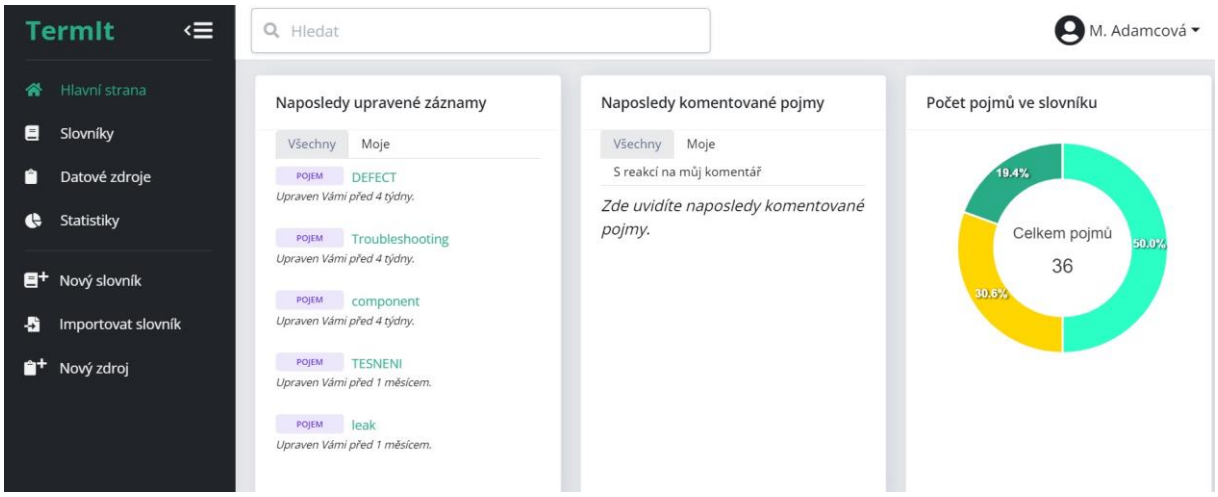
Tato práce navazuje na práci „*Textová analýza nestrukturovaných závadových dat v letecké údržbě*“. Zde byl TermIt představen a byly uvedeny důvody, proč je tento nástroj v tomto případě vhodný. Jedná se především o to že, ostatní aplikace umožňují širokou škálu funkcí, a není jednoduché se s nimi naučit pracovat, oproti tomu Termit je jednoduchý, uživatelsky příjemný a poskytuje veškeré funkce, které jsou pro tento účel potřeba.

Další z důvodu pro výběr TermItu je i validace výsledků. K tomu, aby bylo možné porovnat výsledky z této a práce „*Textová analýza nestrukturovaných závadových dat v letecké údržbě*“ je nejlepší vybrat stejný nástroj. V opačném případě by bylo třeba popsat softwarové odlišnosti a algoritmy analýzy. Výsledky a navržená řešení této práce tak budou platné pro nástroj TermIt, do jisté míry bude možné je přeupoužít pro ostatní aplikace.

TermIt je webová aplikace, která byla vytvořena v rámci projektu na Fakultě elektrotechnické ČVUT. Umožňuje vytvořit několik zdrojových textů a slovníků, čímž umožňuje práci na několika projektech najednou. Je možné aplikovat jeden slovník na několik zdrojů nebo naopak na jeden zdroj aplikovat několik slovníků. Uživatel je přihlášen pod přihlašovacími údaji a jeho práce je ukládána, takže je možné se k ní vracet.

Slovník je možné průběžně upravovat podle požadavků a udržovat ho aktuální včetně úpravy parametrů. Po provedení analýzy nástroj zvýrazní všechny nalezené pojmy, po rozkliknutí je možné zobrazit podrobnosti. Nástroj vyhledá jen ty pojmy, které jsou předdefinované ve slovníku, avšak často se vyskytující pojmy navrhne jako potenciální pojem. Podrobně bude práce s nástrojem popsána níže.

TermIt je pod neustálým vývojem. Během psaní této práce probíhala oboustranná komunikace s vývojáři a byly implementovány návrhy na zlepšení a opravy nalezených bugů a chyb. Prostředí nástroje je na obrázku 7.



Obrázek 7: Práce s nástrojem TermIt

Termit má několik funkcí, pro snadnou orientaci a přepínání mezi funkcemi slouží hlavní panel, je vidět v levé části obrázku 7.

Hlavní strana

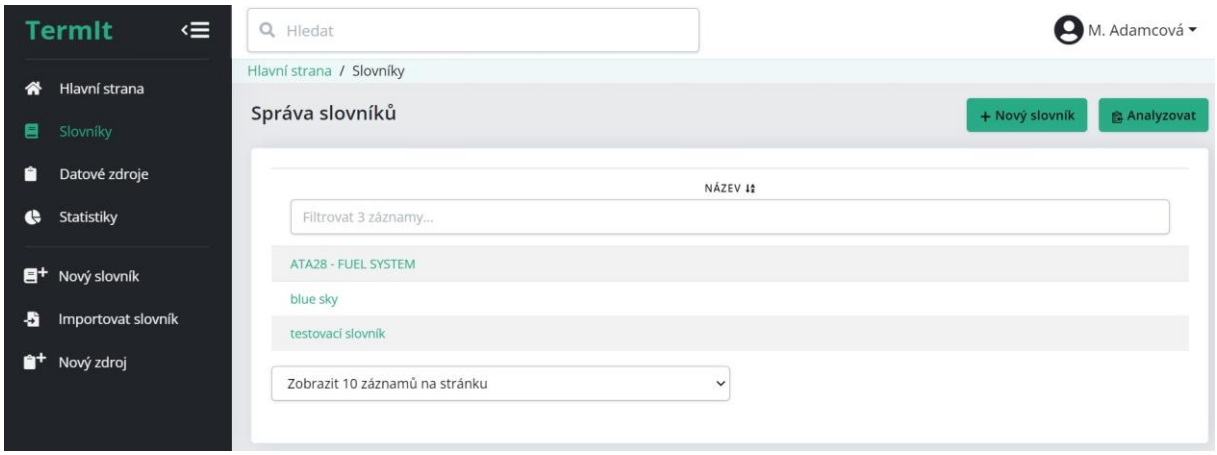
První z nabídky je panel hlavní strany. Jeho rozložení je rovněž vidět na obrázku 7. V části „Naposledy upravené záznamy" je přehled všech úprav, které byly jako poslední upraveny, zároveň je možné i zobrazit, který uživatel změnu provedl, pokud má přístup více uživatelů.

Další část je nazvaná „Naposledy komentované pojmy". Podobně jako v sekci o úpravách záznamů zde se zobrazují všechny komentáře, které uživatelé vytvořili.

V poslední části je zobrazen graf a celkový počet pojmů uložených na profilu. Tento počet zahrnuje všechny pojmy ve všech slovnících. Zobrazený graf znázorňuje rozložení těchto pojmů v jednotlivých slovnících. V tomto případě jsou vytvořené celkem tři slovníky. Pro zobrazení podrobností je třeba najet myší na graf a zobrazí se název slovníku a počet jeho pojmů.

Slovníky

V této sekci najdeme vše, co je potřeba k vytvoření a úpravě slovníků. Nabídka zobrazená po vybrání sekce slovníků je na obrázku 8.

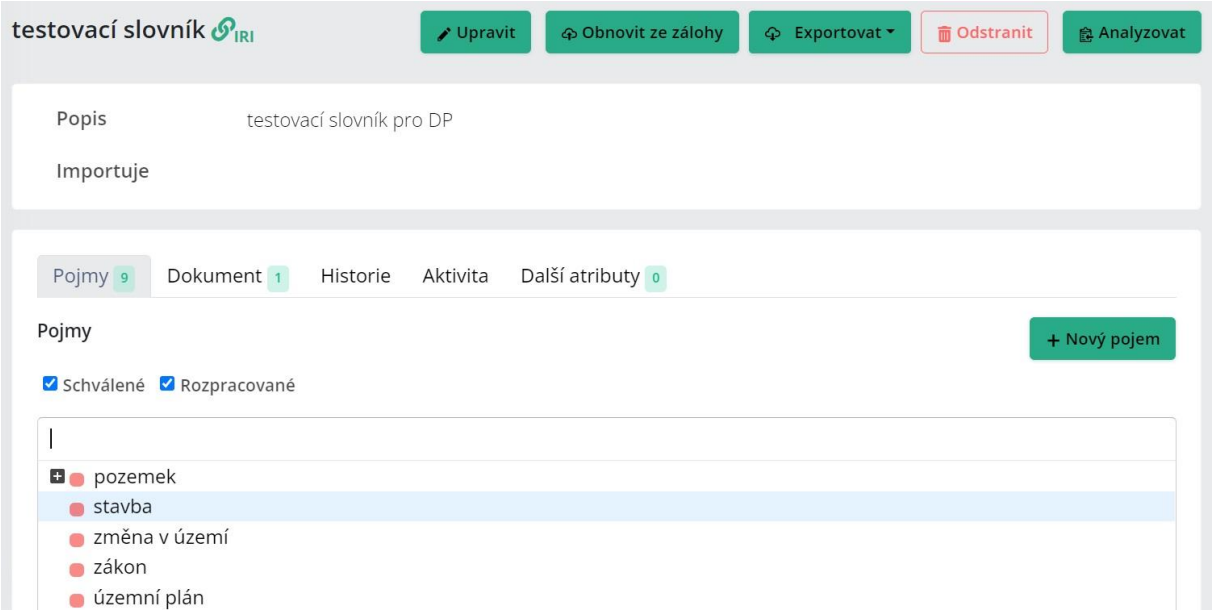


Obrázek 8: TermIt – správa slovníků

Zobrazí se seznam všech vytvořených slovníků a jejich názvy, v tomto případě jsou vytvořeny tři slovníky. V pravém horním rohu se zobrazí dvě tlačítka. „Nový slovník“ otevře panel pro vytvoření nového slovníku. Je třeba zvolit název, rovněž je možné připojit popis. Slovník je možné načíst z externího souboru. Soubor je možné připojit z počítače, aby byl TermIt schopen soubor přečíst, je nutné mít soubor ve formátu UTF-8 do velikosti 10 MB. Stejnou nabídku k vytvoření nového slovníku dostaneme, klikneme-li myší na nabídku „Nový slovník“ v hlavním ovládacím panelu. Druhé tlačítko „analyzovat“ spustí nově analýzy definic všech pojmů ve všech slovnících.

Pro zobrazení podrobností a editaci slovníku klikneme na název a zobrazí se nabídka zobrazená obrázku 9. V horní části obrázku je vidět 5 rozklikávacích záložek. Záložka „upravit“ otevře podobné menu jako při vytváření slovníku. Je možné upravit název, popis, importovat obsah dalších vytvořených slovníků a vytvářet atributy. Záložka „obnovit ze zálohy“ umožňuje nahrát externí soubor z počítače ve formátu UTF-8.

Funkce „exportovat“ vyexportuje celý slovník se všemi pojmy a informacemi o nich. Je možné vybrat z několika formátů, a to: CSV, excel, SKOS. Červeně zvýrazněná nabídka „smazat“ smaže kompletně celý slovník. Tlačítko analyzovat spustí textovou analýzu definic všech pojmů ve vybraném slovníku.



Obrázek 9: TermIt – podrobnosti o slovníku

Na obrázku 9 je vidět, že aktivní je záložka „pojmy". Tato záložka zobrazuje seznam všech definovaných pojmů ve slovníku. V tomto seznamu je možné vyhledávat. Záložka „dokument" zobrazí informace o dokumentu, ke kterému je slovník přiřazen (může jít o více dokumentů). Jedná se o název a popis. Pokud se jedná o dokument nahraný externě z počítače, je zde množnost tento dokument od slovníku odpojit nebo zobrazit jeho obsah. Záložka aktivita ukazuje poslední úpravy provedené na vybraném slovníku a uživatele, který změnu provedl. „Aktivita" ukazuje na časové ose vytvořené a aktualizované pojmy.

Po kliknutí na vybraný konkrétní pojem se zobrazí podrobnosti o něm. Je to definice, pokud byla definice vyznačená ve zdrojovém dokumentu, lze zobrazit i konkrétní definici v textu. Další podrobnosti jsou: typ pojmu, pojmy se stejným významem, nadřazené pojmy, související pojmy, doplňující poznámka a slovník. Všechny tyto atributy pojmu lze zadat během vytváření pojmu nebo je kdykoli upravit a odstranit.

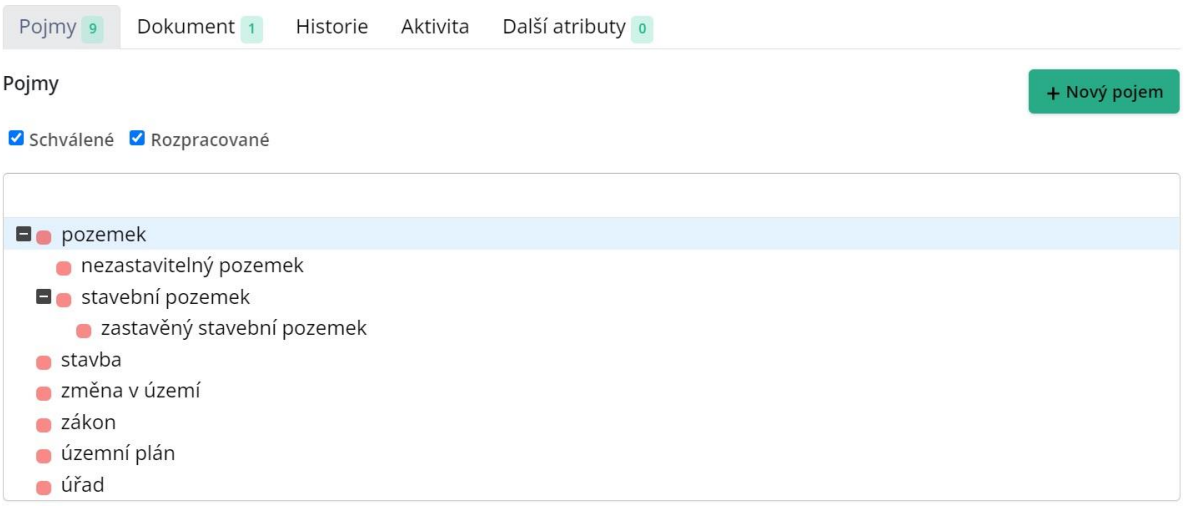
Tlačítko „nový pojem" je nejčastěji používanou funkcí v TermItu, protože umožňuje vytvoření nového pojmu do vybraného slovníku. Při vytváření je možné změnit jazyk z výchozího anglického jazyka na další jazyky z nabídky včetně češtiny. Dále je na řadě zadat název, toto pole je povinné. Jedná se o označení, které daný pojem jednoznačně v rámci slovníku identifikuje. Název pojmu není vhodné později měnit, protože by to mohlo ovlivnit význam dat, která jsou tímto pojmem popsána.

Název je jediný povinný atribut, pokud ostatní nepovinné atributy zůstanou nevyplněné, slouží také jako jediný identifikátor pro textovou analýzu, je tedy vhodné ostatní atributy vyplnit, pokud je to možné.

Atribut „synonyma“ slouží pro definici synonym k názvu atributu. Jedná se o slova nebo fráze, která mají v daném jazyce stejný nebo téměř stejný význam, např. slova ‚obvyklý‘ a ‚běžný‘. Vývojáři bylo doporučeno zahrnout do atributu synonyma také zkratky a zkrácené verze slov, které by samy o sobě neměly být názvem pojmu, např. ‚OSN‘ – ‚Organizace spojených národů‘, nebo ‚např.‘, ‚atd.‘. Dále slova, která se píší odlišně, ale znamenají totéž. A dále nepravidelné tvary množných čísel. Toto se týká hlavně angličtiny, ve které jsou ale data z údržby letadel zaznamenána např. ‚person‘, ‚people‘.

Dále je zde sekce pro vytvoření definice pojmu. Definice se může vyplnit ručně, ale TermIt nabízí i funkci vyznačení definice přímo ve zdrojovém textu. Je možné také připojit zdroj definice. Odkazuje na konkrétní místo v textu dokumentu např. na kapitolu knihy či konkrétní odstavec v zákoně. V případě označení definice přímo v textu dokumentu je zdroj vyplňován automaticky.

Dalšími atributy upřesňujícími pojem jsou „doplňující poznámka“ a „nadřazené pojmy“. V případě doplňující poznámky jde o volny nedefiniční text, který má upřesnit význam pojmu. Díky nadřazeným pojmům můžeme upřesnit vztah mezi jednotlivými definovanými pojmy nebo vytvořit určitou hierarchii. Nadřazený pojem je pojem s širším významem. Slouží k zachycení vazby na obecnější pojem např. ‚kostel‘ -> ‚budova‘, k přiřazení pojmu jeho typu např. ‚kostel sv. Mikuláše‘ -> ‚kostel‘ nebo k vyjádření části celku např. ‚klika‘ -> ‚dveře‘. Všechny tyto vytvořené vztahy a hierarchie jsou vidět v záložce slovníky po rozkliknutí konkrétního slovníku v seznamu pojmů. Nadřazený pojem bude mít vedle názvu symbol ‚+‘. Po jeho rozkliknutí zobrazíme všechny podřazené pojmy. Zobrazeno na obrázku 10.



Obrázek 10: TermIt – nadřazené pojmy

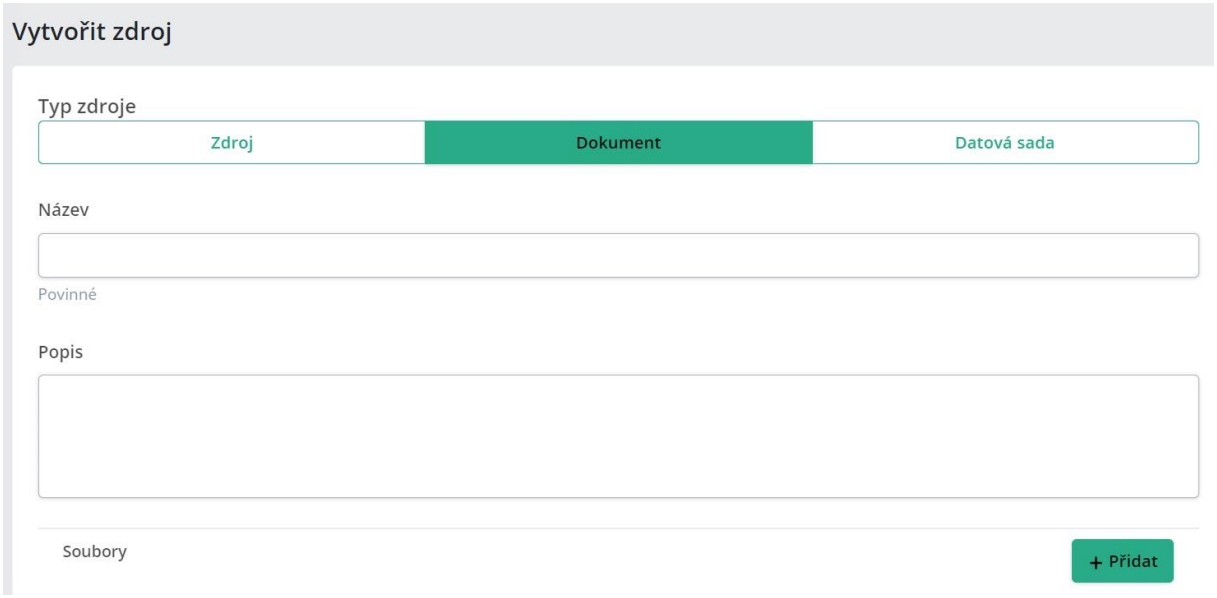
Velmi důležitým atributem je atribut „vyhledávácí texty". Jde o text, který není určen pro vizuální prezentaci pojmu a slouží výhradně pro vyhledávání. Dle vývojářů by měly být do této kategorie zahrnuty překlepy, zastaralé termíny a archaismy a vše ostatní, co se nehodí do jiných atributů.

Dalším atributem, který nabízí úpravu vztahů atributů, je „typ pojmu". Jedná se v podstatě o charakter pojmu. TermIt rozlišuje typy a individuály. Typ je v tomto kontextu vždy obecnější. Typem může být např. ‚*actuator*' individuálem je pak konkrétní ‚*slat actuator*' nebo ‚*trim actuator*'. Individuály jsou dále rozděleny na objekt, vlastnost, vztah a událost. Analogicky lze dále rozlišit typ objektu, typ vlastnosti atd. Objekty jsou nezávislé v čase se měnící prvky např. lidé, auta, dokumenty. Oproti tomu vlastnosti jsou závislé na objektech např. barva auta. Podobně jsou na tom vztahy, ty propojují více objektů, na kterých jsou závislé např. manželství. Události jsou plně neměnné prvky, které proběhly v minulosti např. Olympijské hry 2020.

Datové zdroje

Po rozkliknutí záložky „datové zdroje" v hlavním naváděcím panelu se zobrazí seznam všech vytvořených datových zdrojů, které lze filtrovat a vyhledávat. Rovněž se zobrazí tlačítko „nový zdroj", po jeho rozkliknutí se zobrazí prostředí pro vytvoření a úpravu nových zdrojů, viz obrázek 11. Nejprve je potřeba v horní liště zvolit, jestli se jedná o „zdroj", „dokument" nebo „datovou sadu". V případě této práce byla použita možnost dokument, protože umožňuje nahrání externího dokumentu. Prvním atributem je název, jedná se o povinné pole volného textu. Dalším je popis, ten podobně jako popis u pojmů slouží k podrobnějšímu popisu a specifikaci zdroje. Záložka „soubor" umožňuje

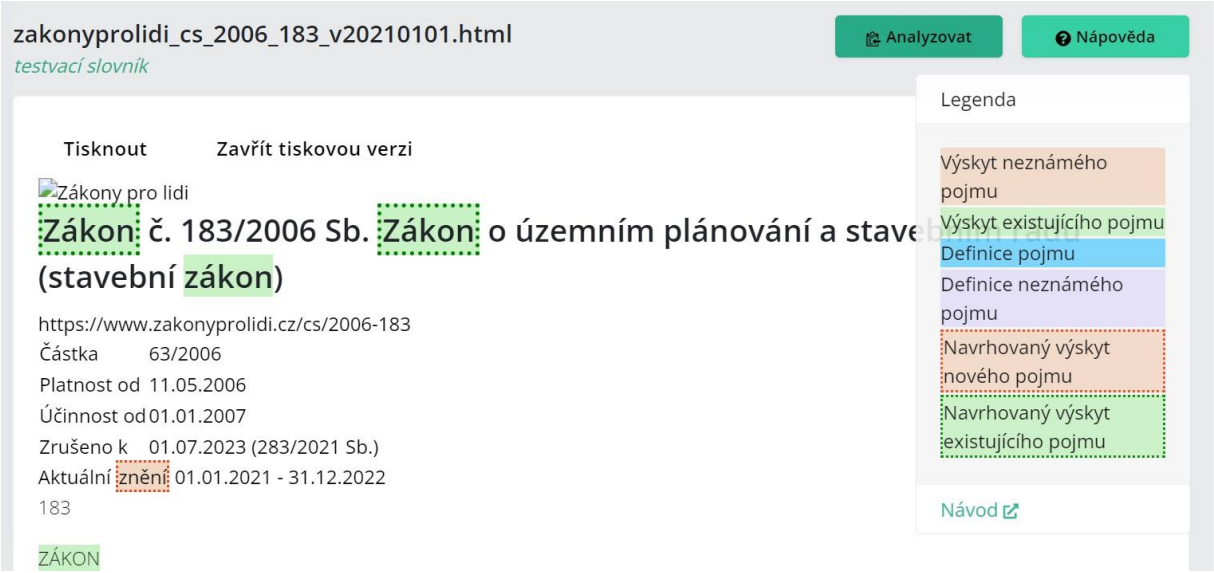
(v případě dokumentu) nahrání externího dokumentu z počítače do velikosti 10 MB ve formátu UTF-8.



Obrázek 11: TermIt – nový zdroj

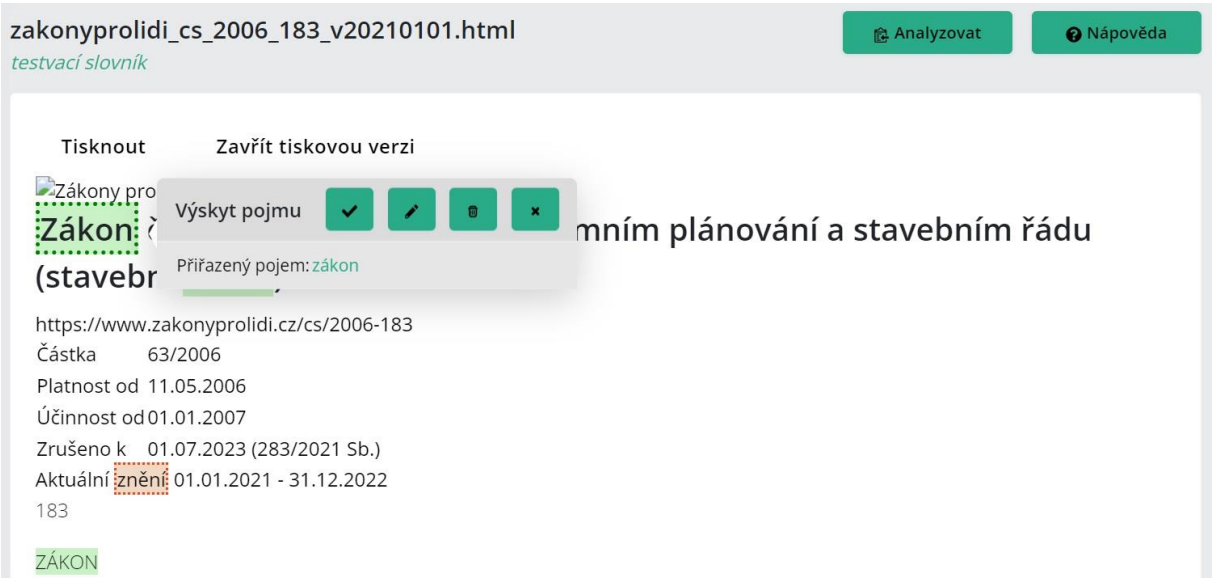
Po vybrání již vytvořeného zdroje se zobrazí panel se základními informacemi, jako je název datového zdroje, popis, pokud byl zadán, dokumentový slovník a již přiřazené pojmy. Tyto atributy lze upravit klinutím na tlačítko „upravit" a celý zdroj smazat klinutím na možnost „smazat". Dále je možné zobrazit již připojené soubory. Ty je možné přidávat nebo odstraňovat a zobrazovat obsah.

Po klinutí na možnost „zobrazit obsah" se zobrazí plný text nahraného souboru. V případě, že již na dokumentu proběhla analýza, se zobrazí i nalezené pojmy. Viz obrázek 12. Pojmy jsou barevně rozlišeny podle výskytu, který TermIt zaznamenal, viz legenda na obrázku 12.



Obrázek 12: TermIt – zobrazení obsahu dokumentu a legendy výskytu pojmů

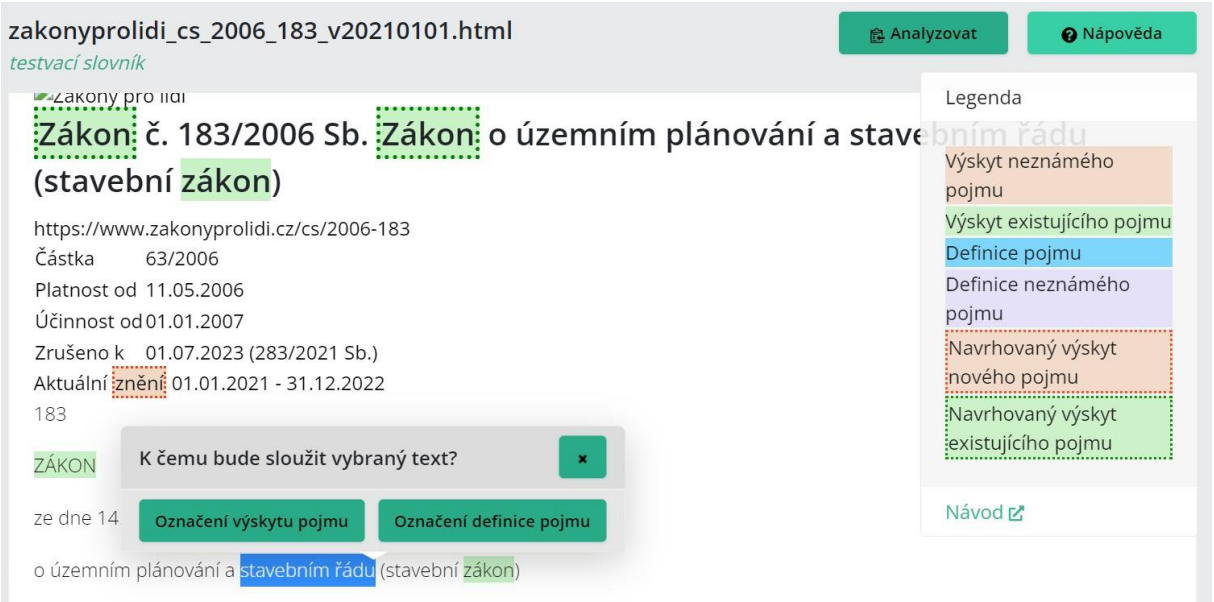
V dokumentu jsou zeleně označené výskyty pojmů a modře jejich definice. Pokud klikneme na takto označený pojem nebo definici, otevře se pop-up menu, kde můžeme pojem nebo definici upravit či odstranit. Pojmy, které jsou označeny zeleně s čárkovaným ohraničením jsou navrhované výskyty existujících pojmů. Po kliknutí na takový pojem se otevře podobné pop-up menu jako u vyskytujícího se pojmu. Zobrazeno na obrázku 13. Zde je vidět, který pojem byl automaticky TermItem přidělen. Kliknutím na symbol ‚fajfky' se výskyt pojmu odsouhlasí, v takovém případě zmizí čárkované ohraničení pojmu. Dále je možné editovat výskyt pojmu. Je možné změnit pojem na jiný existující nebo vytvořit úplně nový. Klinutím na ikonu popelnice se výskyt pojmu vymaže.



Obrázek 13: TermIt – pop-up menu navrhovaného výskytu existujícího pojmu

Analogicky pracujeme s pojmy označenými oranžově s čárkovaným ohraničením. Zde se jedná o návrh výskytu neexistujícího pojmu. Jedná se o část textu, která se vyskytuje tak často, že byla automaticky vyhodnocena jako vhodná k vytvoření pojmu. Po klinutí se opět otevře menu. Tentokrát nebude automaticky vyplněn existující pojem. Můžeme vybrat a přidělit již existující nebo vytvořit úplně nový, nebo výskyt pojmu úplně odmítnout. V případě že označíme nový výskyt pojmu, barva se změní z oranžové na zelenou, při odmítnutí výskytu pojmu barevné podbarvení zmizí úplně.

Zobrazení obsahu zdroje má jednu velkou výhodu. Umožňuje označování a vytváření pojmu přímo v textu. Pokud se při procházení narazí na pojem, který je vhodné vytvořit a zařadit do slovníku, není potřeba soubor zavírat, otevírat záložku se slovníky a vytvořit nový pojem, a následně se vracet zpět k dokumentu. Pojem lze z textu vytvořit tak, že je požadovaný text označen tak, jako bychom ho chtěli kopírovat, otevře se pop-up, který se zeptá, zda chceme označený text využít k označení výskytu pojmu nebo definice, viz obrázek 14.



Obrázek 14: TermIt – pop-up menu po označení textu v dokumentu

V případě, že chceme označit výskyt pojmu, je dále možnost využít již vytvořené pojmy, nebo vytvořit nový. V případě, že chceme označit definici, je nutné vybrat již existující pojem. Tato možnost označování výskytu pojmů přímo v dokumentu je velice vhodná pro práci se záznamy z údržby letadel a byla využita během vytváření slovníku pro tuto práci.

Klinutím na tlačítko analyzovat v pravé horní části obrazovky se obrazí menu, kde je nutné k analýze vybrat vytvořený slovník, poté analýza proběhne. Analýza trvá

v závislosti na velikosti souboru, velikosti slovníku a rychlosti internetového připojení. Výsledkem analýzy je označení nalezených pojmů a návrhy výskytu.

Statistiky

Záložka „statistiky" na hlavním naváděcím panelu, slouží k zobrazení hlavních statistik na uživatelském profilu. Je zde zobrazen na jednom místě celkový počet slovníků, celkový počet pojmů a počet uživatelů. Pojmy jsou dále graficky rozděleny do jednotlivých slovníků a barevně rozlišeny podle typu (objekt, vlastnost, vztah, událost).

6.1 Dostupné slovníky

Z kapitoly výše o dostupné vědecké literatuře na téma automatické textové analýzy technických dat vyplynulo, že pro úspěšnou textovou analýzu je potřeba mít kvalitní slovník se specifickými pojmy, které nás zajímají. Z toho vyplívá, že tento slovník je velmi konkrétní pro danou tématiku. Dosud byla nalezena pouze jedna práce, která se věnuje tématu zpracování a analýze dat z letecké údržby, a to diplomová práce „Textová analýza nestrukturovaných závadových dat v letecké údržbě" Tomáše Vojtěcha.

Pro potřeby této práce vznikl slovník nový. Každý slovník vytvářel jiný uživatel a je pravděpodobné, že k vytváření a údržbě slovníku zvolil trochu odlišný přístup.

7 Dostupná data

Tato kapitola se zabývá popisem, úpravou a prací s dostupnými daty.

Data pro tuto práci byla poskytnuta společností CSAT. Data obsahují záznamy z let 2017, 2018, 2019 a 2020 a byla ze systému AMOS generována do tabulkového formátu. Jejich zpracování bylo tedy možné v tabulkovém editoru Microsoft Excel ve formátu .xlsx.

Data byla poskytnuta pod podmínkou mlčenlivosti. To znamená, že data použitá pro tuto práci nefigurují jako příloha této práce. Zároveň bylo třeba zaručit, že data budou anonymizovaná a nebude možné je spojit s žádným konkrétním letadlem. Tato podmínka byla zajištěna prvotní filtrací pro tuto práci nadbytečných dat. Z hlavičky souboru byly odstraněny sloupce obsahující registraci letadla, záznamy o datech a další informace, které by mohly vést ke spojení záznamu s konkrétním letadlem. Naopak některé sloupce bylo třeba ponechat, aby mohlo být v případě potřeby možné zpětně výstupy této práce spárovat s původním záznamem. Zdrojové záznamy již byla filtrovány a upraveny do požadované podoby během vzniku práce T. Vojtěcha. Pro tuto práci byly vybrány takové vzorky dat, které T. Vojtěch nepoužil. Je to pro to, aby nebyly stejné datasety analyzovány a vyhodnocovány dvakrát.

Ze zdrojových dat byly vytvořeny celkem 4 datasety, jejichž vznik bude podrobně popsán níže v podkapitole „Vznik slovníku“. U všech vytvořených datasetů bylo dbáno na to, aby obsahovaly alespoň 1000 řádků. To zajistí dostatečnou jistotu výsledků analýzy a minimalizuje vliv statistické chyby, která by se zde mohla vyskytnout. Jde například o případ, kdy se stejný nebo velmi podobný řádek vyskytne vícekrát, jeho zhodnocení by tak mohlo ovlivnit celkový výsledek.

Z dostupných dat byl vybrán první vzorek dat čítající 1018 řádků nazvaný dataset 1. Tento dataset sloužil jako výchozí k tvorbě refenčního slovníku, jeho vznik je popsaný v podkapitole níže. Celkově byly z dostupných dat vytvořeny 4 datasety nazvané dataset1, dataset2, dataset3 a dataset4 čítající 1018, 1080, 1041, respektive 1089 řádků, celkem tedy bylo hodnoceno 4227 řádků. Soubor, bylo třeba rozdělit do několika menších proto, aby práce s nimi byla jednodušší. Při vyhodnocování výsledků byla snadnější orientace, TermIt při zobrazování a analyzování nemusel pracovat najednou s tak velkým objemem a celkově tento proces nebyl tak zdlouhavý. Dále bylo třeba datasety exportovat z formátu .xlsx do formátu .html. Nástroj Termit umožňuje nahrání

souboru pouze ve formátu .html. Tento export proběhl pomocí online exportéru[5] vyvinutém na Fakultě elektrotechnické ČVUT přímo pro tento účel.

7.1 Tvorba slovníku

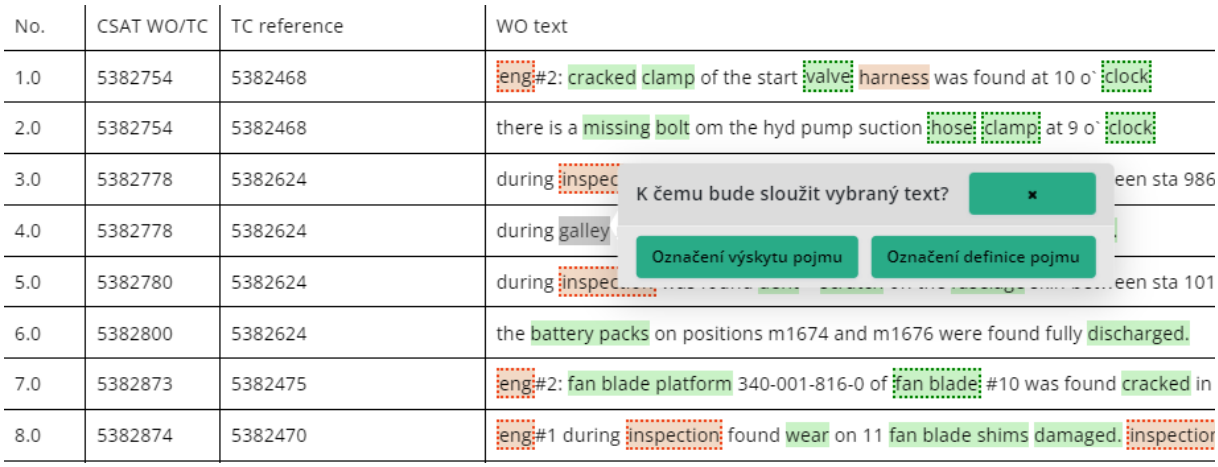
V nástroji Termit byla dle postupu popsaném výše vytvořena nová záložka slovníku. Je to pro to, aby se slovník jasně odlišil od ostatních aktivních slovníků a nedošlo omylem k úpravě cizího slovníku, neboť v TermItu pracuje zároveň více uživatelů . Ke slovníku byly postupně nahrávány datasety ve formátu .html.

Ve slovníku byly vytvořeny dva základní pojmy, které budou dále figurovat jako nadřazené všem ostatním, a to ‚komponenta' a ‚závada'. Je to z toho důvodu, že se předpokládá, že v každém řádku představující jeden záznam se nachází jedna komponenta a jedna závada. Dalším důvodem je snadnější orientace v seznamu pojmů, protože slovník obsahuje několik set pojmů. Dále toto rozdělení zjednodušuje analýzu výsledků, kde je možné uvést výsledky a statistiky zvlášť pro komponenty a zvlášť pro závady.

Slovník vznikl ručním procházením jednotlivých řádků souboru dataset1. Nástroj TermIt umožňuje vytvoření pojmu do slovníku ze souboru označením požadované části textu. Není tedy nutné zobrazený soubor zavírat nebo přeskakovat mezi záložkami a vytvářet pojem přes možnost tlačítka ‚nový pojem'. Tento způsob je pro tento případ pohodlnější a umožňuje plynulé procházení souboru. Zároveň je možné si kontrolovat, které řádky již byly anotované, a nedojde tak přehlédnutí nebo chybě.

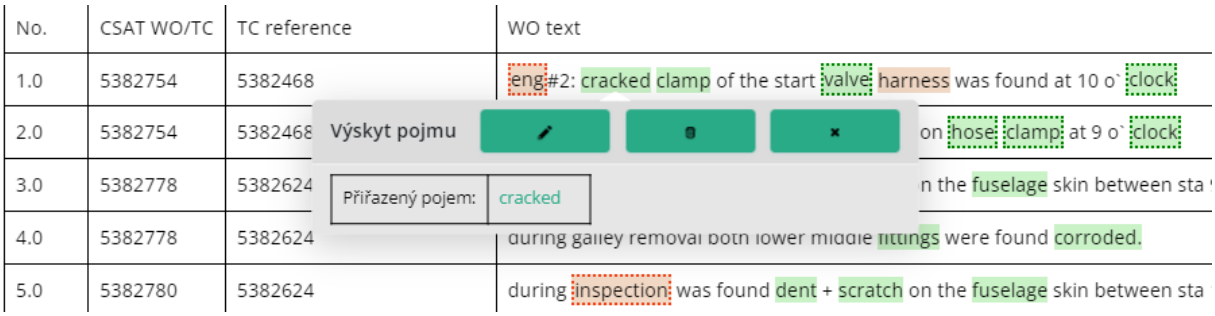
V případě že je v textu nalezen pojem, který nás zajímá, je třeba ho myší označit. Následně se zobrazí možnost, zda chceme označený text použít k označení výskytu pojmu nebo definice pojmu. Zobrazeno na obrázku číslo 15.

[5] https://kbss.felk.cvut.cz/html-exporter/



Obrázek 15: Možnosti po označení zájmového textu

V tomto případě vybíráme možnost označit výskyt pojmů. Následně máme možnost vybrat označení již existujícího pojmu nebo vytvoření zcela nového. Přiřazením existujícího pojmu nebo vytvořením nového pojmu se text zvýrazní zelenou barvou a anotace se potvrdí. Kliknutím na zvýrazněný pojem máme možnost si přiřazený pojem zobrazit, viz obrázek 16.



Obrázek 16: Zobrazení přiřazeného pojmu

Vybráním možnosti ‚nový pojem' se otevře nové dialogové okno pro vytvoření nového pojmu a vytvoření všech potřebných atributů, jako vyhledávací texty nebo synonyma. Po zavření tohoto okna nedojde k zavření souboru a je možné pokračovat v práci v místě, kde byla přerušena.

Zanalyzováním všech řádků v datasetu 1 a zaznamenáním všech zájmových pojmů vznikl první slovník čítající 545 pojmů. Následně bylo třeba ověřit úspěšnost návrhu existujících pojmů ze slovníku na nahraný dokument, tedy to, zda termit po zanalyzovaní dokumentu vybraným slovníkem nalezne a navrhne výskyt správného pojmu. K tomuto účelu byla vytvořena kopie datasetu 1, ale bez potvrzených anotací, které v původním dokumentu vznikly během vytváření slovníku. Tato kopie je závislým dokumentem, neboť slovník vznikal právě z něho. Výsledkem bylo, že automatická analýza nalezla

a správně navrhla pojem u 96 % řádků v případě komponenty, v 98 % řádků v případě závady. Celkově byla správně nalezena komponenta i závada u 95 % řádků. U 4 % řádků byla komponenta nebo závada špatně přiřazena nebo nenalezena. Celkové zhodnocení úspěšnosti u všech datasetů je zobrazeno v tabulce 2.

Dále bylo zjištěno, že TermIt navrhnul výskyt pojmu ve 46 % řádků v případě komponenty a 10 % řádků v případě závady u pojmu, který v daném případě nebyl ten, který bylo třeba najít. Může se jednat například o lokaci komponentu nebo popis okolností nalezení. Zde se nejedná o chybu, TermIt v textu pouze označí návrhy výskytu všech pojmů, které zná. Příklad je uveden na obrázku 17. Potvrzení anotace je pak na uživateli. Možnosti jak zamezit nebo omezit výskyt takovýchto anotací, budou popsány níže v kapitole ‚Možná řešení pro zvýšení úspěšnosti'.

167.0	5347736	53-866-00	gromets blind from lower latch of main mirror were found missing .
168.0	5347736	53-866-00	lens from lavatory a lights was found cracked and missing their part.

Obrázek 17: Návrhy komponent a závad

Na obrázku 7 na řádku 167 je vidět, že TermIt navrhuje 4 pojmy. V tomto případě je závada ‚missing' a komponenta ‚gromets blind'. Ostatní navrhované pojmy pouze upřesňují umístění daného komponentu. Podobný případ je na řádku 168, zde je hledanou závadou ‚cracked' a komponentou ‚lens'.

Pro potvrzení úspěšnosti návrhu výskytu pojmů na závislém dokumentu bylo třeba vytvořit další dokument, který bude na slovníku nezávislý. Proto vznikl soubor dataset 2. Na něm byla spuštěna automatická textová analýza původním slovníkem a úspěšnost návrhu výskytu pojmů zhodnocena stejně, jako v prvním případě. Výsledkem bylo, že v 67 % řádků byla správně označena komponenta, v 92 % řádků byla správně označena závada, celkově byla správně označena komponenta i závada v 62 % řádků. Dále bylo zjištěno, že u 49 % řádků byla označena komponenta, která nebyla požadována, ve stejném případě to bylo u 8 % řádků u závady. Celkově u 36 % řádků nebyla správně nalezena komponenta nebo závada nebo byl přiřazen nesprávný pojem. Největší část těchto řádků tvořily případy, kdy pojmy nemohly být označeny jednoduše proto, že ve slovníku nebyly. Průnik datasetu 1 a 2 nebyl dost velký, aby tyto pojmy obsáhnul. Tyto chybějící pojmy byly doplněny a vznikl aktualizovaný slovník čítající 681 pojmů. Z datasetu 2 se stal na slovníku závislý dokument.

Účinnost tohoto kroku byla ověřena stejně jako v případě datasetu 1. Byla vytvořena kopie datasetu 2, ale bez potvrzených anotací, na které byla puštěna automatická textová analýza aktualizovaným slovníkem. Výsledky jsou uvedeny v tabulce 2.

U všech sledovaných parametrů se podařilo dosáhnout zlepšení, kromě parametru navrhovaný výskyt pojmů komponenty. Toto jsou případy, kdy navržený výskyt pojmu komponenta není požadovaná komponenta, ale v daném případě pouze popis nebo upřesnění lokace nebo popis okolností, během kterých k objevení závady došlo. Toto zvýšení pouze o 6 % (ze 49 na 55) ale není nelogické. Tím, že se zvýšil počet pojmů ve slovníku, se i zvýšila šance, že v nějakém řádku bude daný pojem figurovat jako popis nebo lokace, a ne jako komponent a bude tak navržený k potvrzení.

Dále celý proces popsaný výše proběhl ještě jednou na dalším datasetu 3, nejprve jako nezávislém. Poté byl opět slovník doplněn o chybějící pojmy a spuštěn na datasetu 3 – závislém. Slovník byl rozšířen na konečných 947 pojmů a následně puštěn na posledním datasetu 4 – nezávislém, pro konečné ověření úspěšnosti označení pojmů na nezávislém dokumentu. Výsledky sledovaných parametrů jsou v tabulce 2.

Tabulka 2: Výsledky návrhu výskytu pojmů po spuštění automatické textové analýzy aktuálním slovníkem u všech datasetů

DATASET	SPRÁVNĚ NALEZENA KOMPONENTA	SPRÁVNĚ NALEZENA ZÁVADA	SPRÁVNĚ OZNAČNO KOMPONENTA I ZÁVADA	NAVRHOVANÝ VÝSKYT KOMPONENTA	NAVRHOVANÝ VÝSKYT ZÁVADA	NENALEZENO NEBO NESPRÁVNĚ OZNAČENO	POJMŮ VE SLOVNÍKU
dataset 1 - závislý	96%	98%	95%	46%	10%	4%	545
dataset 2 - nezávislý	67%	92%	62%	49%	8%	36%	545
dataset 2 -závislý	94%	97%	91%	55%	8%	9%	681
dataset 3 - nezávislý	82%	91%	76%	57%	5%	24%	681
dataset 3 - závislý	98%	99%	97%	66%	8%	4%	947
dataset 4 - nezávislý	89%	94%	83%	66%	12%	18%	947

Celkově lze konstatovat, že u parametru procenta řádků se správně označenou komponentou se hodnota u nezávislých dokumentů podařila zvýšit, u závislých dokumentů se zdržela na podobné úrovni kolem 95 %. U parametru správného označení závady se u závislých dokumentů hodnota držela kolem 94 % a nezávislých 92 %. Procento řádků, kde byla správně označena komponenta i závada se podařilo zvýšit u nezávislých datasetů, u závislých se hodnota pohybuje kolem 95 %.

V případě návrhu pojmu, který v daném případě není komponentem nebo závadou, se u komponent hodnota zvyšovala u každého nového datasetu jak závislého, tak nezávislého. Tento nárůst je způsoben rozšířením slovníku o nové pojmy. Slovník byl doplňován především o pojmy typu komponenta, těch je celkově jich ve slovníku mnohem více (z 947 pojmů ve slovníku je 829 typu komponenta). To zvyšuje pravděpodobnost překryvu pojmů, které jsou v jednom případě komponenta, ale v jiném řádku pouze popis umístění. Vychází to i ze samotného charakteru záznamu. V záznamu by mělo být co nejpodrobněji uvedeno, o co se jedná – co je to za komponentu, umístění okolnosti a další relevantní informace. Naopak u závad se hodnota parametru drží u všech datasetů kolem 5–10 %. Je to dáno tím, že závad je ve slovníku řádově méně než komponent (z 947 pojmů ve slovníku je 118 typu závada) a obecně termíny pro označení závady lze využít při popisu jen omezeně.

Důležité je, že se podařilo u nezávislých datasetů rozšířením slovníku snížit procento řádků, kde nebyla nalezena komponenta nebo závada, nebo byl přiřazen nesprávný pojem. Jedná se totiž o řádky, kde není možné při následném zpracování využít informace v něm obsažené. Při automatickém zpracování statistiky by byly tyto řádky přeskočeny a označeny za nerelevantní, a mohla by se tak ztratit cenná informace.

8 Zhodnocení úspěšnosti analýzy

Tato kapitola se zabývá uvedením a zhodnocením výsledků automatické textové analýzy, která proběhla na datových sadách, které byly vybrány z dostupných dat popsaných výše. Hodnocení proběhla v několika parametrech.

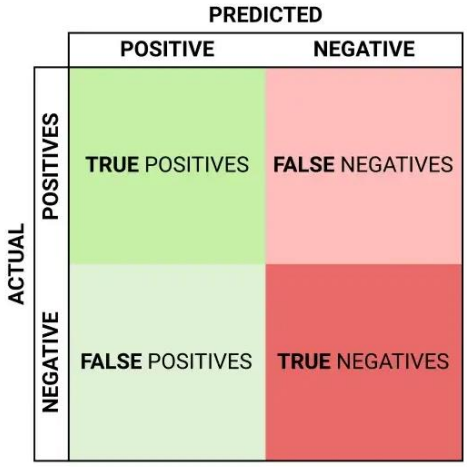
Parametry, které byly použity v práci T. Vojtěcha a které budou pro porovnání použity i zde jsou:

- Procento řádků, kde je správně vybrána komponenta
  Zde se jedná o procento řádků, kde byla správně automaticky přiřazena komponenta bez ohledu na to, zda byla nebo nebyla správně přiřazena závada. Procento je vypočítáno ze všech řádků, které jsou relevantní a obsahují informaci o závadě i komponentě.
- Procento řádků, kde je správně vybrána závada
  Zde se jedná o procento řádků, kde byla správně automaticky přiřazena závada bez ohledu na to, zda byla nebo nebyla správně přiřazena komponenta. Procento je vypočítáno ze všech řádků, které jsou relevantní a obsahují informaci o závadě i komponentě.
- Procento řádků, kde je správně vybrána komponenta i závada
  Zde se jedná o procento řádků, kde byla správně přiřazena komponenta zároveň i závada. Procento je vypočítáno ze všech řádků, které jsou relevantní a obsahují informaci o závadě i komponentě.

8.1 Výpočet precision a recall

Dále budou pro vyhodnocení použity i další parametry, které v práci T. Vojtěcha nebyly, ale které je vhodné vyhodnotit, neboť jsou to statistické parametry, které se standardně používají pro hodnocení textové analýzy jako celku. Díky tomu se dají jednotlivé textové analýzy porovnat mezi sebou. Tyto parametry jsou ‚precision' a ‚recall'.

Pro použití těchto parametrů je třeba nejprve definovat možné výstupy textové analýzy. Pro tyto potřeby jsou údaje definované jako pozitivní (positive) a negativní (negative). Pro další hodnocení je třeba určit, zda předpoklad pozitivity a negativity byl správný (true) anebo špatný (false). Tedy toto rozdělení určuje, zda byl údaj správně nebo špatně označený jako pozitivní nebo negativní. Vznikají 4 možnosti určení a to: TP (true positives), FP (false positives), TN (true negatives) a FN (false negatives). Tyto možnosti jsou rovněž znázorněny na obrázku 18.

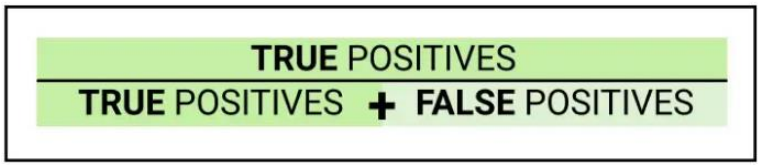


Obrázek 18: Možnosti výstupu textové analýzy [17]

Parametry precision a recall jsou vypočítány právě z toho rozdělení výstupů.

- Precision

  Precision je poměr mezi skutečně pozitivními údaji a všemi údaji, které byly předpovězené jako pozitivní. Vzorec je na obrázku 19. Výstupem je údaj, který říká, jak platné jsou výsledky, tedy jak moc jsou relevantní. Vysoká přesnost precision znamená, že pokud je řádek označený jako kladný, je s velkou pravděpodobností kladný i ve skutečnosti. [17]



Obrázek 19: Výpočet precision [17]

- Recall

  Recall popisuje zlomek skutečně pozitivních a všech pozitivních údajů, vzorec je obrázku 20. Tento parametr říká, jak kompletní výsledky jsou. Pokud má recall vysokou hodnotu, pak to vypovídá o tom, že se podařilo zachytit velkou část skutečně pozitivních údajů. [17]

Obrázek 20: Výpočet recall [17]

Pro potřeby dalšího postupu a aplikace analýzy je třeba nejprve definovat, které údaje jsou brány jako pozitivní a které jako negativní. Precision a recall bude níže vyhodnocen v kontextu řádků a v kontextu pojmů.

V případě řádků analýza určuje, zda je řádek jako celek pro další zpracování relevantní či nikoli. Tento postup lze dělat ručně, ale v tomto případě to není postup efektivní, neboť se zde jedná o datovou sadu čítající desetitisíce záznamů. Je tedy praktičtější definovat, v jakém případě lze řádek automaticky vyřadit nebo ponechat. Definujeme předpoklad, že na relevantním řádku lze určit právě jednu komponentu a k tomu jednu závadu, poměr je tedy 1:1. Takové řádky jsou označeny jako relevantní. Dalšími relevantními řádky jsou ty, kde jsou komponenty a závady v poměru 1:n nebo n:1. Nerelevantní jsou řádky, kde nelze určit nic nebo lze určit pouze jeden pojem, nebo řádky, kde lze určit více komponent i více závad (n:n). V takovém případě analýza nemůže jednoznačně rozhodnout, k jakému komponentu patří jaká závada.

V kontextu řádků jsou true positives všechny relevantní řádky, které byly analýzou označeny jako positive, false positives jsou nerelevantní řádky, které byly nesprávně označeny jako positive. False negatives jsou relevantní řádky, které byly označeny jako negative a true negatives jsou nerelevantní řádky, které byly označeny jako negative.

V případě kontextu pojmů je třeba definovat pozitivitu jinak. Zde jsou jako positive počítány všechny řádky, kde byl přiřazen pár, true positive je takový řádek, kde je pojem (v případě, že počítáme parametry jen pro komponentu nebo závadu) přiřazen správně, nebo pár je přiřazen správně (v případě že počítáme parametry pro pár). False positive jsou řádky, kde byl pojem přiřazen nesprávně (v případě párů alespoň jeden pojem nesprávně). Precision je určený poměrem všech true positive řádků ku všem řádkům, kde byl přiřazen celý pár. Recall je určený poměrem všech true positive řádků ku všem řádkům, které obsahují informaci o páru (jsou relevantní).

8.2 Výstupy automatické analýzy

Po provedení analýzy jsou výsledky reprezentovány ve výstupní statistice. Tato statistika je generována do formátu .xlsx, tudíž je možné další zpracování v programu Microsoft Excel. Tato statistika obsahuje na každém řádku jeden záznam z analyzovaného souboru. V tomto případě byla analýza spuštěna na Datasetech 1, 2 a 3, tedy závislých souborech na slovníku, kde byly manuálně potvrzené anotace. To představuje statistiku tréninkového datasetu. Dále byla analýza spuštěna na celém zdrojovém souboru

obsahujícím 74 000 řádků bez potvrzeních anotací (součástí tohoto souboru je i ta část, ze které byly vytvořeny datasety 1, 2 a 3). To představuje statistiku celého datasetu.

Statistika je kromě řádků členěna i do sloupců. Zde jsou uvedené informace, které již obsahoval původní soubor. Jsou to informace o čísle WO, TC reference a WO text. Další sloupce reprezentují již výsledky automatické textové analýzy. Mezi ně patří ‚ComponentLabel' a ‚FailureLabel'. Zde je uvedený pojem, který na daném řádku textová analýza přiřadila. Pokud na daném řádku vybírala z více pojmů, ty jsou vedeny pod ‚MultipleComponent' a ‚MultipleFailure'. Každému výběru pojmu je přiřazena pravděpodobnost, s jakou si je analýza daným výběrem jistá. To je reprezentováno sloupci ‚ComponentScore' a ‚FailureScore'. Tento parametr nabývá hodnot od 0 do 1, kde 1 znamená jistotu výskytu pojmu. Celkové zhodnocení řádku je ‚AgregatedScore', kde jsou předchozí dva parametry mezi sebou vynásobeny. Nakonec je u každého řádku uvedena informace, jestli je anotace manuálně potvrzena, ve sloupci ‚IsConfirmed'. Potvrzené anotace se vyskytují pouze ve statistice s trénovacím datasetem.

Tyto statistiky reprezentují chování automatické textové analýzy provedené vybraným slovníkem na vybraném souboru. V případě statistiky tréninkového datasetu se vyskytují potvrzené i nepotvrzené anotace (anotace byly manuálně potvrzené u relevantních řádků). Pokud je anotace manuálně potvrzena, je v ‚ComponentLabel' a ‚FailureLabel' zobrazen tento potvrzený pojem. V ‚MultipleComponent' a ‚MultipleFailure' jsou uvedeny všechny pojmy daného typu, které analýza na řádku nalezla. V případě, že je anotace nepotvrzená, je do ‚ComponentLabel' a ‚FailureLabel' vybrán pojem automatickou analýzou a v ‚MultipleComponent' a ‚MultipleFailure' jsou uvedeny všechny pojmy, které byla na řádku nalezeny. Díky tomu lze mezi sebou porovnat manuálně potvrzené anotace ze statistiky tréninkového datasetu s chováním analýzy na stejném datasetu s nepotvrzenými anotacemi ze statistiky celého datasetu, což bude popsáno níže v další kapitole.

8.3 Výsledky automatické textové analýzy

Pro následující výpočty byl vybrán nezávislý dataset čítající 2000 řádků ze statistiky celého datasetu tak, aby se nepřekrýval s datasety 1, 2 a 3.

8.3.1 Výpočet precision a recall v kontextu řádku

Na této sadě byla manuálně zhodnocena positivita/negativita řádků dle pravidel popsaných výše. Následně bylo rozhodnuto, zda automatická analýza určila tento parametr správně. O tomto bylo rozhodnuto na základě obsahu sloupce

‚ComponentLabel', ‚MultipleComponent', ‚FailureLabel' a ‚MultipleFailure'. Pokud je na daném řádku pole pro ‚ComponentLabel', nebo ‚FailureLabel' prázdné, analýza předpovídá negativitu. Stejně tak předpovídá negativitu u řádků, kde je více pojmů u ‚MultipleComponent' a zároveň u ‚MultipleFailure'. Ostatní případy jsou předpovídány jako pozitivní. Příklad je uveden na obrázku 21.

	ComponentLabel	MultipleComponents	FailureLabel	MultipleFailures	POSITIVE	PREDICT POSITIVE	PREDICT SPRÁVNĚ
2	test button	test button; battery pack; emergency cylinder; check valve	not illuminating	not illuminating; low capacity	0	0	1
3	oxy bottle	oxy bottle; emergency cylinder	low pressure		1	1	1
4	label				0	0	1
5	check valve				0	0	1
6	refuel door				0	0	1
7	check valve		wet		0	1	0

Obrázek 21: Určení pozitivity/negativity v kontextu řádku

Řádek 2 byl manuálně vyhodnocen jako negativní, automatická analýza ho taktéž vyhodnotila jako negativní, protože ve sloupci ‚MultipleComponent' i ‚MultipleFailure' je přiřazeno více pojmů. Řádek 3 byl manuálně vyhodnocen jako pozitivní, taktéž byl vyhodnocen jako pozitivní i automatickou analýzou a to pro to, že v poli pro ‚ComponentLabel' i ‚FailureLabel' je určený pojem a zároveň v poli pro ‚MultipleFailure' není více pojmů, jedná se tedy o případ n:1. Řádky 4, 5 a 6 jsou manuálně vyhodnoceny jako negativní, stejně tak je i analýza označila jako negativní a to proto, že pole pro ‚FailureLabel' je zde prázdné. Rozdíl mezi manuálním vyhodnocením a automatickým vyhodnocením nastává u řádku 7, který je negativní, ale analýza ho označila jako pozitivní.

Takto byla vyhodnocena celá nezávislá sada a vypočítány parametry precision a recall, výsledky jsou uvedeny v tabulce 3.

Tabulka 3: Výsledek výpočtu precision a recall v kontextu řádku

relevantní řádky	1264		
správný predict POSITIVE/NEGATIVE	1418		
správný predict procentuálně	0,709		
true positive	921	PRECISION	0,794650561
false positive	238	RECALL	0,728639241
false negative	343		
true ngative	497		

Celkově bylo z 1264 relevantních řádků určeno 921 správně jako positive a 343 nesprávně jako negative. Ze 736 nerelevantních řádků bylo 497 správně určeno jako negative a 238 nesprávně určeno jako positive. Celkově bylo z 2000 řádků správně určena pozitivita/negativita u 1418 řádků, což se rovná procentuální úspěšnosti 71 %.

Precision a recall byl vypočítán dle vzorců uvedených výše. Výsledek precision 79 % říká, že v se 79 % skutečně pozitivních řádků podařilo správně určit pozitivitu. Výsledek recall 73 % říká, že se podařilo zachytit 73 % všech skutečně pozitivních řádků.

8.3.2 Výpočet precision a recall v kontextu pojmu

Na stejné nezávislé datové sadě jako v předchozím případě byl vypočítán precision a recall v kontextu pojmu. Precision je určený poměrem všech true positive řádků ku všem řádkům, kde byl přiřazen celý pár. Recall je určený poměrem všech true positive řádků ku všem řádkům, které obsahují informaci o páru (jsou relevantní).

Výsledky jsou zobrazeny v tabulce 4.

Tabulka 4: Výsledky precision a recall v kontextu pojmu

správně určeno komponent			758
správně určeno závad			1103
správně určeno párů			679
přiřezno párů	1495		
relevantních řádků	1264		
	pár	komponena	závada
precision	0,454181	0,507023411	0,737793
recall	0,537184	0,599683544	0,872627

Výsledky z tabulky 4 se dají interpretovat takto: precision 50,7 % u komponenty a 73,8 % u závady znamená, že v 50,7 % respektive 73,8 % byl správně přiřazen pojem ze všech řádků, kde byl přiřazen pár. To znamená že, tento parametr je vypočítán i z řádků, které jsou nerelevantní a nenesou informaci o páru. Může totiž nastat situace, že i v případě, že je řádek ve skutečnosti nerelevantní a neobsahuje informaci o komponentě a závadě, analýza nalezne shodu a pojem přiřadí.

Recall 60 % u komponenty a 87,3 % u závady znamená, že v 60 % respektive 87,3 % byl správně určen pojem ze všech případů, kdy mohl být správně určen (všech relevantních řádků). To znamená že do recall jsou počítány jen relevantní řádky, ale jsou mezi nimi takové řádky, kde pojem nebyl z nějakého důvodu určen a ‚ComponentLabel', ‚FailureLabel' zůstal prázdný.

Nejdůležitější informace je pro uživatele v případě určení parametrů pro páry (určení komponenty i závada), protože to je informace, pro kterou je analýza určena. U páru byl precision vypočítán 45,4 %, což znamená že, v tolika případech byl pár určen správně (ze všech případů, kdy byl pár určen). Recall vykázal hodnotu 53,7 %, což znamená že, v tolika

případech byl pár určen správně (ze všech případů, kdy mohl být určen správně, tedy relevantních řádků).

8.3.3 Úspěšnost výběru pojmů

Zde budou uvedeny a zhodnoceny výsledky, které automatická textová analýza vykázala na různých datasetech, a porovnání s výsledky z práce T. Vojtěcha. Aby byly tyto výsledky porovnatelné, je třeba je vyhodnotit ve stejných parametrech a stejným způsobem.

Závislá sada

Výsledek analýzy na závislé sadě dat byl v této práci prováděn automaticky porovnáním statistiky trénovacího datasetu s potvrzenými anotacemi se statistikou celého datasetu. V programu Microsoft Excel byly do jednoho souboru vloženy sloupce ‚ComponentLabel' a ‚FailureLabel' z obou statistik ze stejné části souboru. Poté se na každém řádku provedlo porovnání obou pojmů a bylo rozhodnuto, zda automatická textová analýza nalezla správný pojem. V případě že, byly pojmy shodné, bylo rozhodnuto že ano. Příklad je uveden na obrázku 22.

F	G	K	L	Q	T	U
ComponentLabel -	ComponentLabel-z analýz	FailureLabel - potvrzer	FailureLabel-z analýzy	OriginalText	komponent	závada
clamp	harness	cracked	cracked	eng#2: cracked clamp of	NEPRAVDA	PRAVDA
bolt	bolt	missing	missing	there is a missing bolt on	PRAVDA	PRAVDA
fuselage skin	fuselage skin	dent	scratch	during inspection was fou	PRAVDA	NEPRAVDA
fitting	fitting	corroded	corroded	during galley removal bol	PRAVDA	PRAVDA
fuselage skin	fuselage skin	dent	scratch	during inspection was fou	PRAVDA	NEPRAVDA
battery pack	battery pack	discharged	discharged	the battery packs on pos	PRAVDA	PRAVDA
fan blade platform	lug	cracked	cracked	eng#2: fan blade platform	NEPRAVDA	PRAVDA
fan blade shim	shim	worn	worn	eng#1 during inspection	NEPRAVDA	PRAVDA
sidewall panel	sidewall panel	worn	out of the limit	these sidewall panels wer	PRAVDA	NEPRAVDA
sidewall panel	sidewall panel	loosened	loosened	during sidewall panel rep	PRAVDA	PRAVDA
paint	paint	corroded	corroded	were found worn and cor	PRAVDA	PRAVDA
sealing tape	sealing	poorly installed	installed	in aft c / c has been foun	NEPRAVDA	NEPRAVDA
sidewall panel	sidewall panel	worn	out of the limit	these sidewall panels wer	PRAVDA	NEPRAVDA
ceiling shrouds	shroud	broken	broken	these ceiling shrouds wer	NEPRAVDA	PRAVDA
sidewall panel	sidewall panel	loosened	loosened	during sidewall panel rep	PRAVDA	PRAVDA
paint	paint	corroded	corroded	were found worn and cor	PRAVDA	PRAVDA
sealing tape	sealing	poorly installed	poorly installed	in fwd c / c has been four	NEPRAVDA	PRAVDA
fuselage skin	fuselage skin	dent	dent	during inspection was fou	PRAVDA	PRAVDA
access door	leading edge	delamination	delamination	ext: delamination was fou	NEPRAVDA	PRAVDA
wheel well panel	panel	erosion	erosion	ext: edge erosion was fou	NEPRAVDA	PRAVDA
wheel well panel	panel	delamination	delamination	ext: delamination was fou	NEPRAVDA	PRAVDA
probe	probe	separation	out of the limit	ext: the wear,cracking an	PRAVDA	NEPRAVDA

Obrázek 22: Porovnání potvrzených anotací s automatickou analýzou

První sloupec u komponentu a závady reprezentuje potvrzený pojem, druhý sloupec pojem vybraný analýzou. V případě, že jsou pojmy shodné, je shoda označena PRAVDA v opačném případě jako NEPRAVDA.

Celkový výsledek včetně porovnání s výsledky T. Vojtěcha jsou uvedeny v tabulce 5.

Tabulka 5: Porovnání výsledků práce

	TATO PRÁCE			T. VOJTĚCH		
	SPRÁVNĚ OZNAČENO KOMPONENT	SPRÁVNĚ OZNAČENO ZÁVAD	SPRÁVNĚ OZNAČENA KOMPONENTA I ZÁVADA	SPRÁVNĚ OZNAČENO KOMPONENT	SPRÁVNĚ OZNAČENO ZÁVAD	SPRÁVNĚ OZNAČENA KOMPONENTA I ZÁVADA
ZÁVISLÁ SADA	51,50%	91,90%	47,90%	74%	79,40%	59,50%
NEZÁVISLÁ SADA	60%	86,30%	53,70%	60,10%	77,30%	45,70%

Z tabulky 5 lze vyčíst, že výsledky obou analýz nevykázaly na závislé sadě diametrálně odlišné výsledky, i když rozdíly zde jsou. Automatická analýza u slovníku T. Vojtěcha na závislé sadě vykázala vyšší procentuální výsledek u správného označení komponent a u celkového výsledku. Naopak vyšší úspěšnost správného označení závady vykázala analýza na slovníku této práce.

Rozdíl ve výsledcích je pravděpodobně způsoben rozdílným přístupem k vyhodnocení. V této práci byla závislá sada hodnocena zcela automatickým porovnáváním. Z tohoto důvodu nebyly některé řádky posouzeny dostatečně individuálně a s kontextem. Proto v některých případech mohlo dojít k nesprávnému negativnímu ohodnocení pojmu. Např. v některých řádcích, kde se komponentů nebo závad vyskytovalo více (případy 1:n a n:1). Ve sloupci ‚ComponentLabel' a ‚FailureLabel' je možné ale uvést pouze jeden pojem, proto se může stát, že v takovém případě je automaticky pojem vyhodnocen jako špatně vybraný. Zároveň v některých řádcích lze více pojmů označit za správný. Např. pokud je manuální anotací přiřazen pojem ‚access panel' a automatickou analýzou ‚panel', automatické hodnocení vyhodnotí tento pár jako neshodu. Pokud by se tento řádek vyhodnocoval manuálně, v určitém případě by se i toto dalo označit na shodu, protože ‚access panel' je pouze upřesnění původního pojmu ‚panel'. Nižší procento shody komponentů snižuje i procento shody u páru. Proto je výsledek v tomto případě nižší než výsledek T. Vojtěcha.

Z tabulky 5 vyplývá, že metoda vyhodnocení ovlivňuje výsledek znatelně, neboť výsledek automatické analýzy na závislé sadě vykazuje nižší hodnoty než automatická analýza na nezávislé sadě. A to i přes to, že v závislé sadě jsou eliminovány řádky, kde analýza nenalezla správný pojem, protože ve slovníku nebyl obsažený.

Nezávislá sada

Výsledek nezávislé sady byl hodnocen na vzorku 2000 řádků vybraných ze statistiky celého datasetu. Protože zde nebyla možnost porovnat statistiky mezi sebou, muselo být hodnocení provedeno manuálně řádek po řádku. Výsledky jsou v tabulce 5. V případě shody u komponentů jsou výsledky téměř shodné s T. Vojtěchem, závady jsou obdobně jako u závislé sady vyšší v této práci. To zapříčiňuje i lehce vyšší výsledek u vyhodnocení párů v této práci.

K tomu, aby mohly být tyto dvě práce efektivně porovnány mezi sebou, bylo třeba vyhodnotit je stejnou metodou na stejné sadě dat. Proto byla vybrána sada 1000 řádků ze statistiky celého datatsetu této práce a T. Vojtěcha (která byla k dispozici), tak aby se nepřekrývala se závislými sadami, na kterých vznikaly slovníky. Podobně jako při vyhodnocení závislé sady se do jednoho souboru vložily sloupce s ‚ComponentLabel' a ‚FailureLabel' a navzájem se porovnávali. Porovnání probíhalo manuálně, protože každý slovník vznikal s trochu jiným přístupem. Stejné pojmy nebyly nazvány vždy stejně, stejně tak vyhledávací texty byly použity jinak atd. Proto bylo třeba ke každému řádku přistupovat individuálně a porovnávat i se slovníkem nahraným v TermItu. Příklad je na obrázku 23.

Component	Component	Failure	Failure	OriginalText	SPRÁVN	SPRÁVNĚ	SPRÁVNĚ FA	SPRÁVNĚ F
Fairing	fairing	Worn	corroded	hinges from bin doors were found mal	1	1	1	1
Upper skin of spo	upper skin	Dent	dent	several shallow hail strike dents have	1	1	1	1
Plug	window	Dent	dent	two dents were found on skin plug win	1	0	1	1
Door sill	floor	Missing		int . paint on cocpit floor sill was found	0	0	0	0
Cargo door skin	doorstop	Missing	failured adj	finding (nrc) taskcard 53-030-00-01 (	0	0	0	1
Rivet	rivet	Corrosion	corroded	were found the corrosions signs of riv	1	1	1	1
Door frame	door frame	Missing	torn	lavatory d (door frame): the seal was	0	0	0	1
		Missing		fwd c/c-the capstrips were found shab	0	0	0	0
Wedge surface o	slat	Corrosion	corroded	ext>slat#1 has been found spread cor	0	0	1	1
Plate	plate	Worn	worn	was found worn plate assy-wiggle on	1	1	1	1
Bolt	bolt	Worn	worn	ext /kle:the upper tab, bolts and serra	1	1	1	1
Aft rod	panel	Dent	dent	eng2 dent has been observed on inne	0	1	1	1
Cover	cover	Broken	broken	during removal lav a were found broke	1	1	1	1

Obrázek 23: Porovnání výsledků této práce a T. Vojtěcha

První sloupec v pořadí patří pojmům přiřazeným analýzou T. Vojtěcha, druhý sloupec představuje pojmy z analýzy této práce. U obou případů bylo rozhodnuto, zda se shoduje komponenta a závada, pokud je pojem přiřazen správně, je do příslušného sloupce zapsána 1, v opačném případě 0. To umožňuje správně přiřazené pojmy na konci sečíst a vypočítat úspěšnost.

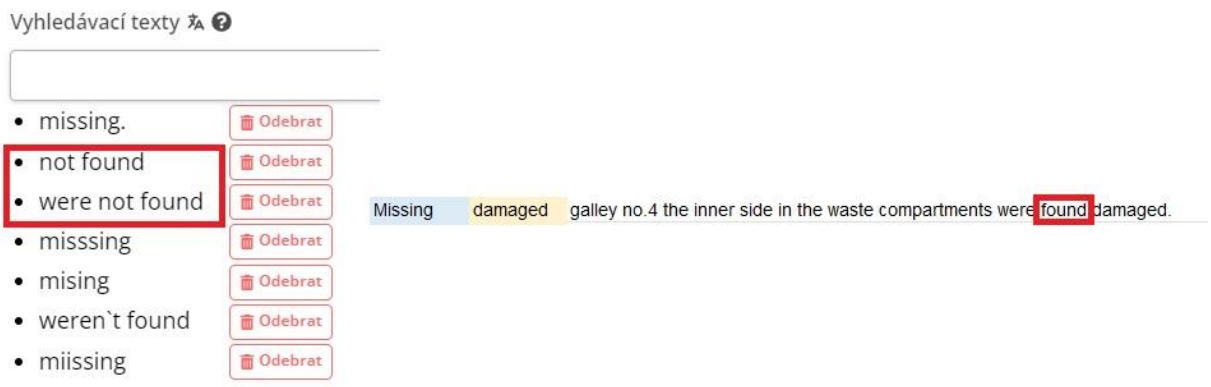
Výsledky, vzniklé tímto porovnáním stejnou metodou, jsou uvedené v tabulce 6.

Tabulka 6: Porovnání výsledků práce prováděné stejným vyhodnocením

	TATO PRÁCE			T. VOJTĚCH		
	SPRÁVNĚ OZNAČENO KOMPONENT	SPRÁVNĚ OZNAČENO ZÁVAD	SPRÁVNĚ OZNAČENA KOMPONENTA I ZÁVADA	SPRÁVNĚ OZNAČENO KOMPONENT	SPRÁVNĚ OZNAČENO ZÁVAD	SPRÁVNĚ OZNAČENA KOMPONENTA I ZÁVADA
NEZÁVISLÁ SADA	60%	90,60%	54,70%	54,70%	77,60%	43,90%

Z tabulky 6 vyplývá, že výsledky nezávislé sady předchozí (tabulka 5) a této jsou u této práce téměř shodné. U T. Vojtěcha jsou výsledky nezávislých sad taktéž téměř totožné, lehký pokles je vidět pouze u správně určení komponenty.

Během vyhodnocování bylo zjištěno, že rozdíl ve výsledcích u správného určení závady, byl způsobený pojmem ‚missing'. Je to jeden z nejčastěji se vyskytujících pojmů a v práci T. Vojtěcha tento pojem často figuruje ve ‚FailureLabel' jako špatně určený. To je způsobeno tím, že v jeho slovníku je v atributu vyhledávací text uveden i výraz ‚not found'. Tato formulace je v tomto případě nevhodná, protože slovo ‚found' se objevuje ve většině záznamů. TermIt tedy nalezne část vyhledávacího textu, která se shoduje, a označí ho jako správný. Příklad je uvedený na obrázku 24.



Obrázek 24: Pojem 'missing'

Ve slovníku z této práce takto upravený vyhledávací text není, tudíž se nestalo, že by analýza pojem ‚missing' často nesprávně přiřazovala. Proto je celkové procento správně určených komponent vyšší. Lehce vyšší úspěšnost slovníku této práce u označování komponent je způsobeno tím, že ve slovníku jsou více zahrnuty pojmy komponentů, které jsou z pohledu kontextu záznamu obecnější. Ve slovníku T. Vojtěcha je více konkrétních pojmů. Ty mají často i několik slov. Pokud analýza nalezne na řádku částečnou shodu s takto definovaným pojmem, často ho přiřadí. V některých případech absence obecnějšího pojmu nutí analýzu hledat pojem s ne stoprocentní shodou. Příklady jsou uvedeny na obrázku 25.

Flap track fairing	bearing	during removal aft outb. flap have been found unsrvicable bearings on flap track .
Ceiling light	panel	fwd cargo the cieling panels p/n : 453a1220-39, 453a1220-229, 453a1220-156 where found in bad condition (punctured)

Obrázek 25: Částečná shoda pojmů

Na obrázku 25 je v prvním případě vidět, že analýza u T. Vojtěcha se rozhodla přiřadit pojem ‚Flap track fairing', protože nalezla shodu s částí pojmu. V druhém případě byl přiřazen pojem ‚ceiling light' se shodou s částí pojmu, protože slovník neobsahuje pojem ‚panel'.

9 Možná řešení pro zvýšení úspěšnosti

Tato kapitola se bude věnovat všem možným řešením, která vedou nebo by mohla vést ke zvýšení úspěšnosti automatické textové analýzy. Budou zde definována doporučení pro organizace, které záznamy vedou a uchovávají. Dále budou navrhnuta řešení a doporučení pro tvorbu a úpravu referenčního slovníku, a nakonec řešení a doporučení pro samotnou textovou analýzu.

9.1 Doporučení na tvorbu a úschovu dat

Jedním z omezení této práce je, že i přes prvotní filtraci dat na pouze závadové WO, kde je předpoklad obsahu komponenty a závady se i tak dostává do následného zpracování velké množství nerelevantních záznamů, které následně zvyšují procento špatně určených řádků. Příklady takovýchto záznamů jsou v tabulce 7.

Tabulka 7: Příklady nerelevantních záznamů

NDT: 1. PERFORM AN HFEC COUNTERSINK INSPECTION AT ALL REMOVED FASTENER LOCATIONS PER 737 NDT MANUAL PART 6, 53-30-00, PROCEDURE 1, 3 OR 4 TO CONFIRM
SHM: PERFORM THE REPAIR OF THE FUSELAGE SKIN BETWEEN STA 986-1006 PER BOEING REPAIR INSTRUCTIONS MESSAGE NUMBER CZT-MA6-20-0001-10B AND BOEING REPAIR
STRINGER S14L STA 992.80-1016 WAS REMOVED DUE TO ACCESS.
PERFORM DOUBLE INSPECTION OF THE REPAIR
REF AIRBUS REFERENCE : 80792274/008

Jedním z doporučení jak toto omezit alespoň částečně, by mohl být způsob, jak přistupovat k záznamu závady a jiných zápisů do dokumentů. Většina z těchto pro tuto práci nerelevantních záznamů nějakým způsobem souvisí s nalezenou závadou. Vyžaduje další práce nebo popisuje provedenou opravu nebo jinou akci. Standardně probíhá zápis dle popisu uvedeného v kapitole 3.4.1 Záznam závady. Ne vždy platí, že jako výchozí zápis work stepu je závada. V praxi se do WO zaznamenávají i jiné informace jako požadavky zákazníka, požadavky na další práce, instrukce, poznámky, reference. Pokud se tyto informace zapíší jako první záznam ve work step, budou při exportu dat figurovat v kolonce WO text, což je část, která se jako volný text používá pro automatickou textovou analýzu, a tento záznam pro analýzu vytvoří nerelevantní záznam.

Doporučení směřuje k procesu zápisu do WO. V případě, že by se podařilo nalézt způsob, jak zaznamenávat ostatní provozní informace kromě samotných nálezů jinam než do

work step, zamezilo by to průniku těchto dat do dalšího zpracování. Po konzultaci s pracovníky z provozu bylo ale toto řešení zamítnuto, protože systém ani software AMOS není na takovou změnu připravený. Zároveň by to znamenalo zvýšené nároky na personál, který záznamy vytváří. Další pravidla a složitost v postupech by mohla způsobit chybu při vytváření záznamu. Navíc současný systém je nastavený tak, že všechny potřebné informace jsou dostupné na jednom místě a všem zainteresovaným pohodlně přístupné. V případě, že by se tento návrh aplikoval, znamenalo by to vytvoření nového záznamového dokumentu nebo WO jiného typu než závadový. Pro propojení potřebných informací by bylo nutné tyto WO navzájem propojit, což by vedlo ke zvýšené pracovní zátěže a prodlevě v provozu. Celkově návrh na změnu postupu zápisu do WO byl vyhodnocen jako pro aplikaci do provozu nevhodný.

Další možností, jak se zbavit nepotřebných dat před samotnou analýzou, je filtrace po vygenerování dat ze systému. Tato data se obvykle uchovávají ve formě tabulky ve formátu .xlsx, tudíž je lze zpracovat v programu Microsoft Excel. To umožňuje v datech hledat nebo je filtrovat.

Pro tento účel je možné použít ruční procházení záznamů a manuální vymazání nepotřebných řádků. Tento postup je ale neefektivní, protože při procházení tisíců řádků záznamů je velice časově náročný. Je vhodnější nalézt způsob, jak část odfiltrovat automaticky. Při procházení záznamů bylo zjištěno, že většina řádků nepotřebných pro další zpracování se odkazovala na provedení práce nebo sejmutí určitého dílu, což bylo v záznamu explicitně uvedeno. Na začátku záznamu je uvedena konkrétní instrukce, někdy i požadavek na kvalifikaci personálu. Příklady těchto řádků, které byly při analýze dat objeveny, jsou uvedeny v tabulce 8.

Tabulka 8: Příklady záznamů s pokyny a záznamy o sejmutí dílu

SHM: PERFORM REPAIR PER SRM 53-00-01-2R-3 REPAIR 3
B1: PERFORM DOUBLE INSPECTION OF THE REPAIR
SHM+NDT: PERFORM DAMAGE REMOVAL AND NDT INSPECTION OF THE DAMAGED FUSELAGE SKIN BETWEEN STA 986-1006 PER BOEING INSTRUCTIONS CZT-MA6-20-0001-08B AND B
SHM: PERFORM THE REPAIR OF THE FUSELAGE SKIN BETWEEN STA 986-1006 PER BOEING REPAIR INSTRUCTIONS MESSAGE NUMBER CZT-MA6-20-0001-10B AND BOEING REPAIR
FLAP SLAT COMMAND SENSOR UNIT (CSU) REMOVED DUE TO WORKSHOP INSPECTION. P/N 780A0000-02 S/N 00888
LAVATORY MODULE A HAS BEEN REMOVED DUE TO FLOOR STRUCTURE INSPECTION
STOWAGE S5 HAS BEEN REMOVED DUE TO ACCESS TO CENTER SECTION

Tyto záznamy se dají zobecnit na všechna data, neboť se v podobném tvaru objevují napříč celou referenční datovou sadou. Filtraci lze dále provádět hledáním klíčových slov textu, dále se nepotřebné řádky odstraní. Jako klíčová slova, která nejčastěji indikují řádek bez záznamu závady, jsou ta uvedená v tabulce 9. Dále jsou k nim uvedené příklady záznamů.

Tabulka 9: Klíčová slova a příklady záznamů

perform	PERFORM CABIN PRESSURE LEAK TEST ACC TO AMM 05-51-91-790-801.
do	ENG#1 DO AN IDLE LEAK CHECK PER AMM TASK 71-00-00-710-012,R.00 AND EXAMINE THE INSTALLED CRANK COVER FOR OIL LEAKS.
please (pls)	PLEASE SEND CARGO DOOR FOR REPAINT
restore	PLS RESTORE PAINT ON LH+RH WING SPOILERS LOWER AND UPPER SURFACES
remove	THE SIDEWALL PANELS HAS BEEN REMOVED IAW AMM 25-52-06 REV 15-FEB-2020.
replace	WING ACCESS PANEL 540JB HAS BEEN REPLACED IAW AMM TASK 57-27-11-400-001-A REVISION DATE : MAY 01, 2020 REVISION NUMBER : 74
instruction	SEE WORK STEP #2# FOR REPAIR INSTRUCTIONS
issued	WO ISSUED FOR LABEL BOOKING PURPOSE
reference	REFERENCES: /A/ DWG 182A2401-12_SH 11 /B/ PSDL 182A2401
follow	FOLLOW STEP #9 FOR NUTPLATE INSTL.
continue	PLS CONTINUE FROM POINT 4.H (FLAP PEENING) SRM 57-20-01-2R-1.

Filtraci ale nelze provést pouze vyhledáním klíčových slov a odstraněním všech řádků, které ho obsahují, neboť některé záznamy obsahují najednou jak informaci o komponentě a závadě tak i záznam podobný těm, které jsou uvedeny v tabulkách výše.

Z tohoto důvodu nelze ani tímto způsobem dosáhnout úplného vyčištění dat. Navíc, je velice pravděpodobné, že se během této práce nepodařilo odhalit všechna klíčová slova, protože byla analyzována jen část ze všech záznamů. V záznamech jsou i řádky, u kterých nelze klíčová slova identifikovat. Tento způsob by byl velice časově náročný, ale vedl by k částečnému vyčištění datasetu, což by pozitivně ovlivnilo textovou analýzu.

Možným řešením tohoto problému by mohla být kombinace prvního a druhého doporučení. V případě prvního doporučení by se jednalo a jednoznačné rozlišení řádku, který neobsahuje popis komponenty a závady. Mohlo by se jednat o označení slovním spojením např. ‚NOT FINDING' nebo speciálním znakem. Toto označení by mohlo být i přednastavené v AMOS popřípadě v jiném softwaru, který organizace využívá. Při zápisu

by bylo potřeba pouze vybrat, zda se jedná o FINDING nebo NOT FINDING a program by automaticky přiřadil identifikaci. Toto označení by bylo zároveň univerzální, protože by se dalo použít pro všechny zápisy, jak pro instrukce, odkazy tak pro chybně vytvořené záznamy. Při následném zpracování by se data označená jako NOT FINDING jednoduše odfiltrovala. Tím by se zajistila téměř stoprocentní úspěšnost filtrace nepotřebných dat.

Dalším problémem je, že popis závady je podrobný a obsahuje popis umístění a ostatní okolnosti při nálezu závady. To má za následek vysoký počet navrhovaných výskytů pojmů ve statistice pod ‚MultipleComponents' a MultipleFailure', které jsou TermItem nabízeny po automatické analýze. S rozšiřováním slovníku se dále zvyšuje počet řádků, kde je kromě skutečné komponenty a závady navrhován i jiný pojem. Dále více pojmů na řádku, které slovník zná, znamená i více možností pro automatickou textovou analýzu. Výsledné přiřazené pojmy mohou být pouze jedna komponenta a jedna závada na řádku, v případě že je navrhovaných pojmů více, musí se algoritmus rozhodnout, kterému dá přednost.

Nabízelo by se řešení, které by omezilo rozsah záznamu závady pouze na popis komponenty a závady. Např. „LEAK WAS FOUND ON VACUUM TUBE" nebo upřesnění pomocí označení dílu např. „FLOOR BEAM Y 503.2 FR 12-16 WAS FOUND CORRODED", v takovém případě by se na jednom řádku nevyskytovalo více pojmů. Jak ale bylo popsáno výše v kapitole o vzniku záznamu závady, v provozu je žádoucí, aby byl popis co nejpodrobnější. Závadu běžně opravuje jiný technik, než který závadu objevil a vytvořil záznam, tudíž je pro další postup důležité mít na jednom místě co nejpřesnější popis. Proto není možné omezit zápis tímto způsobem.

Další možností by bylo rozdělení zápisu na více řádků. Do jednoho work step popsat jednoduše závadu a komponentu a do dalšího přidat podrobnosti. Zde ale narážíme na podobný problém jako výše u filtrace dat. Provoz ani software není připravený na změnu postupu zápisu. Toto řešení není za současného provozu možné.

Další možností je určit pravidla postupu zápisu, tedy v jakém pořadí se budou informace zapisovat. V praxi by to mohlo vypadat tak, že by se v první větě uvedla komponenta a závada a dále podrobnosti. Při analýze dat by analytik, který bude potvrzovat navrhované pojmy věděl, že jsou uvedeny na prvním místě. Tato znalost by mohla být zakomponována i do algoritmu TermItu, který přiřazuje pojmy. V případě, že by na daném řádku bylo více navržených pojmů, upřednostnil by ten uvedený na prvním místě.

Opět ale má toto doporučení pouze omezený dosah, protože se jedná o zásah do zavedených postupů a praxe.

9.2 Doporučení na tvorbu a údržbu slovníku

Z výsledků popsaných v kapitole výše vyplynulo i několik drobných doporučení pro tvorbu slovníku. Z pohledu vyhodnocení je lepší, když slovník obsahuje i obecnější pojmy, protože v případě, kdy by slovník tento pojem neznal, dal by přednost jinému s částečnou shodou. Příklad lze uvést na pojmu ‚panel'. Druhů panelů je na letadle více a více se jich vyskytuje i záznamech. Během tvorby slovníku na referenční sadě není pravděpodobné, že by se podařilo zachytit všechny vyskytující se druhy panelů (‚floor panel', ‚lining panel', ‚access panel' atd.). Pokud by analýza probíhala na řádku, kde by se vyskytoval druh panelu, který není zahrnutý ve slovníku (např. ‚grill panel') analýza by buď neoznačila nic, nebo vybrala pojem s částečnou shodou. V každém případě by se jednalo a špatné určení. V případě, že by slovník obsahoval pojem ‚panel', mohl by tento pojem přiřadit. V daném případě by se jednalo a správně určený pojem, i když s menším detailem. Vznikla by zde ale možnost tento neznámý pojem do slovníku doplnit pro další analýzy. Je tedy výhodné mít ve slovníku zahrnuté jak pojmy s větším detailem, tak obecnější pojmy.

9.3 Doporučení pro software Termit

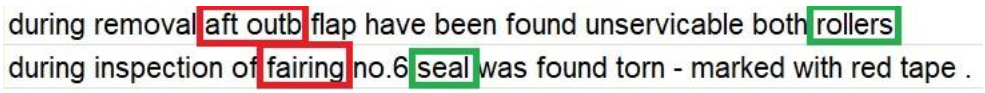
Jedním ze způsobů, jak zvýšit úspěšnost automatické textové analýzy při zachování současného stavu ve tvorbě a úschově dat, je definice pravidel výskytu pojmů v závislosti na různých okolnostech. Tato pravidla (patterny) mohou upravovat vztahy mezi pojmy nebo určovat priority. Tyto vztahy mohou být různého typu, v této práci se podařilo definovat patterny diakritické a sémantické. Diakritické patterny určují pravidla pro výběr daného pojmu na základě výskytu určitých znaků, pozici pojmu v záznamu nebo skladby věty. Sémantické využívají definici vztahů jednotlivých pojmů na základě technické znalosti nebo upravují priority v případě výskytu určitého pojmu.

- Diakritické

Diakritické patterny jsou patterny, které říkají, za jakých okolnosti vybrat jaký pojem na základě výskytu určitých znaků, slov, slovních spojení nebo pozici ve větě. Tato pravidla není možné aplikovat do nástroje TermIt, proto budou popsaná jen teoreticky.

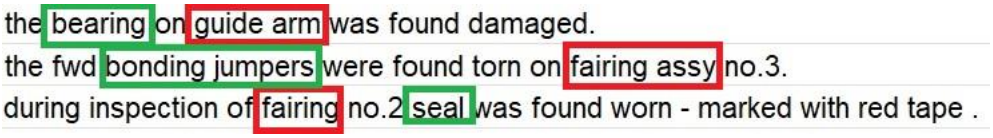
Na základě analýzy záznamů bylo identifikováno několik pravidel, která by v případě jejich aplikace mohla vést ke zvýšení úspěšnosti automatické textové analýzy.

Po výrazu ‚removal' a ‚inspection' navazuje komponent, který je jen popisem. Tento pattern říká, že pokud analýza nalezne více pojmů typu komponenta, první uvedený za výrazem ‚removal' a ‚inspection' nemá označovat. Příklad je uveden na obrázku 26. Červeně jsou vyznačeny popisné komponenty, zeleně požadované komponenty.



Obrázek 26: Příklad patternu 'removal/inspection'

Dalším nalezeným patternem je pravidlo polohy za spojkou a předložkou. V záznamech bylo nalezeno opakující se pravidlo pozice pojmu při použití slovního druhu spojky nebo předložky. V takovém případě je komponent za spojkou nebo předložkou pouze popisem nebo lokací (v případě že je nalezeno více komponent). Příklad je uveden na obrázku 27.



Obrázek 27: Pozice komponentu za spojkou nebo předložkou

V datech byly nejčastěji nalezeny výrazy ‚on' a ‚of', ale pravidlo je platné pro celé spektrum těchto slovních druhů (above, inside, under, below…).

Některé záznamy začínají úvodem, který popisuje umístění závady, které je uvedeno před znaky ‚:' a ‚;'. Příklad je na obrázku 28.

Obrázek 28: Umístění pojmů před a za znakem ‚:'

Tento pattern říká, že pokud je na řádku nalezeno více komponent, ten uvedený před znakem ‚:' nebo ‚;' je jen popisem nebo lokací a je možné ho vynechat.

Posledním diakritickým patternem, který se podařilo definovat, je význam výrazu uvedený v závorkách. Obecně, pokud je pojem nalezený v závorkách, jedná se pouze o doplnění nebo upřesnění a je tak možné tento pojem vynechat. Příklad je na obrázku 29.

Obrázek 29: Pojem v závorkách

Na obrázku 29 je také vidět, že diakritické patterny platí najednou a na jednom řádku je možné jich uplatnit více. Zde je možné použít pravidlo pozice pojmu za dvojtečkou (‚:'), pozice za spojkou (on mirror) i pozice v závorkách.

- Sémantické

Sémantické patterny upravují výběr pojmu na základě vztahů mezi definovanými pojmy. Na rozdíl od diakritických jsou tyto patterny více konkrétní a často svojí aplikací ošetřují pouze malou část záznamů. Proto zde nebudou uváděny všechny, ale budou popsány obecně petterny jednotlivých skupin a uvedeny příklady.

Part of: Patterny typu ‚part of' popisují vazbu mezi větším a menším celkem. Obecně se jedná o vztah, kdy jeden menší pojem je součástí většího. V případě nálezu obou definovaných pojmů na jednom řádku a aplikace patternu bude upřednostněn menší pojem, protože se jedná konkrétnější komponentu. Příkladem může být dvojice komponent ‚door' a ‚door spring'. Pattern by byl definován jako ‚door spring part of door'. V případě že by byly tyto dva pojmy nalezeny na jednom řádku, analýza by upřednostnila pojem ‚door spring'.

Jako další byly definovány patterny, které upravují prioritu určitého pojmu bez ohledu na výskyt ostatních. Jedná se o takové pojmy, které mají vždy bez ohledu na okolnosti přednost a budou vždy vybrány. Z pravidla se jedná o drobné dále nedělitelné komponenty jako (těsnění, kostření, šroub apod.). Naopak od pozitivní prioritizace je možné některé pojmy i upozadit. To je v případě, kdy daný pojem je za určitých okolností pouze popisem lokací nebo jinou akcí. Příkladem je závada ‚wet', která často popisuje stav izolace. Velice často je ale pojem ‚wet' použit jako popis prostoru. Příklad je na obrázku 30.

Obrázek 30: Výskyt pojmu 'wet'

Jako poslední sémantický pattern je pattern, který upravuje sílu vazby. Jedná se dvojice komponenta a závada, jejichž spojení je silnější než spojení ostatních dvojic. V případě že je na řádku identifikován jeden pojem z dvojice, je upřednostněn druhý před ostatními. Příkladem je silné spojení ventil a netěsnost (‚valve', ‚leak').

9.3.1 Aplikované patterny

Pro ověření vlivu nalezených patternů na výsledek automatické textové analýzy bylo třeba několik patternů vybrat a nasimulovat jejich vliv v případě implementace do

nástroje TermIt. Během analýzy dat bylo vybráno 6 patternů, u kterých byl předpoklad, že budou mít největší vliv na výsledek. Z toho důvodu nebyly aplikovány patterny typu part of. Petterny typu part of jako takové mají širokou aplikovatelnost, ale jeden konkrétní pattern toho typu (např. ‚door spring part of door) nebude ovlivňovat takový počet řádků, jako patterny, které byly aplikovány. K tomu, aby patterny typu part of měly větší vliv, bylo by třeba jich definovat a aplikovat větší množství. Aplikované patterny jsou popsány níže:

Pattern 1: priorita pojmu ‚seal', ‚sealant', a všech víceslovných pojmu obsahující seal před ostatními pojmy. Tento pojem byl identifikován jako jeden z nejčastějších, které byly chybně určeny. Vyskytuje se často v záznamech, kde je více komponent najednou, a je třeba mezi nimi vybírat. Během analyzování záznamů nebyl nalezen žádný záznam, kde by úplná priorita tohoto pojmu vedla k chybně označené komponentě.

Pattern 2: priorita pojmu ‚bushing' a všech víceslovných pojmů obsahujících bushing před ostatními pojmy. O aplikaci toho patteru bylo rozhodnuto ze stejných důvodů, jako u patternu 1.

Pattern 3: snížení priority pojmu ‚out of limit' na minimum. Tento pattern znamená, že v případě výskytu více závad na jednom řádku nebude označen pojem ‚out of limit' jako závada. Tento pojem může být označen jako závada, až když žádný jiný pojem není k dispozici.

Pattern 4: priorita pojmu ‚skin' a všech ostatních víceslovných pojmů obsahujících skin před ostatními komponentami. Pojem skin se vyskytuje v záznamech velice často, protože je využíván k popisu povrchových vad na letadla. Často se ale vyskytuje v záznamu s větším množstvím jiných komponent, které jsou díky shodě s jinými pojmy ve slovníku upřednostněny.

Pattern 5: priorita pojmu ‚paint' před ostatními komponentami. Zde byly důvody k výběru patteru stejné jako u předchozího patternu 4.

Pattern 6: priorita pojmu ‚hole' snížena na minimum. Použití toho patternu je stejné jako u patternu 3. Důvodem k výběru bylo, že pojem ‚hole' se často vykytuje i jako komponenta nebo popis komponenty.

Vyhodnocení probíhalo na sadě 1000 řádků. Jedná se o stejnou sadu, na které proběhlo srovnání stejnou metodou vyhodnocení u této práce a T. Vojtěcha. Nejprve byl každý pattern vyhodnocen zvlášť a následně všechny dohromady. Vyhodnocení probíhalo

individuálně na každém řádku, kde bylo rozhodnuto, zda aplikace příslušného patternu změní nebo nezmění výběr pojmu. V závislosti na tom byl i upraven záznam správnosti výběru. Výsledky jsou v tabulce 10.

Z tabulky lze vyčíst, že největší vliv měl pattern 1, jehož implementací se podařilo zvýšit úspěšnost u správně označených komponent o 5 %. Celkově se u všech patternů podařilo dosáhnout zlepšení nebo výsledek zůstal stejný, tudíž lze konstatovat, že aplikace těchto patternů nezhorší celkový výsledek analýzy. Aplikací všech šesti patternů se celkově podařilo zvýšit úspěšnost označení komponent o 8,1 %, u označení závad o 0,6 % a úspěšnost u správě označeného páru o 7,6 %.

Tabulka 10: Porovnání výsledků po aplikaci patternů

	SPRÁVNĚ OZNAČENO KOMPONENT	SPRÁVNĚ OZNAČENO ZÁVAD	SPRÁVNĚ OZNAČENA KOMPONENTA I ZÁVADA
PŮVODNÍ VÝSLEDEK	60,40%	90,60%	54,70%
PATTERN 1	65,80%	90,60%	59,70%
PATTERN 2	61,40%	90,60%	55,40%
PATTERN 3	60,40%	90,60%	54,80%
PATTERN 4	60,70%	90,60%	54,80%
PATTERN 5	62,50%	90,60%	56,40%
PATTERN 6	60,40%	90,90%	54,70%
VŠECHNY PATTERNY	68,50%	91,20%	62,30%

Po aplikaci těchto patternů byl zhodnocen jejich vliv i na precision a recall. Jejich původní hodnota na této datové sadě a jejich hodnota po aplikování patternů je v tabulce 11. U všech patternů se podařilo jejich aplikací zvýšit hodnoty precion a recall nebo je ponechat na původní hodnotě. Po aplikaci všech šesti patternů se podařilo u párů dosáhnout zlepšení u precision o 6 % a recall o 7,5 %. Největší vliv měl opět pattern 1, který zvýšil precision komponentu o 4,4 % a u páru o 4,1 %. Recall se zvýšil u komponentu o 5,4 % a u páru o 5,3 %.

Celkově lze konstatovat, že aplikace patternů má pozitivní vliv jak na určení precision a recall, tak pro samotnou procentuální úspěšnost přiřazování páru a vede tak ke zvýšení úspěšnosti automatické textové analýzy.

Tabulka 11: Porovnání výsledků precision a recall po aplikaci patternů

	PRECISION			RECALL		
	KOMPONENTA	ZÁVADA	PÁR	KOMPONENTA	ZÁVADA	PÁR
PŮVODNÍ	48,40%	72,60%	43,90%	60,40%	90,60%	54,70%
PATTERN 1	52,80%	72,60%	48%	65,80%	90,60%	60%
PATTERN 2	49,30%	72,60%	44,40%	61,40%	90,60%	55,40%
PATTERN 3	48,50%	73,90%	44%	60,40%	90,90%	54,80%
PATTERN 4	48,70%	72,60%	44%	60,70%	90,60%	54,90%
PATTERN 5	50,10%	72,60%	45,20%	62,50%	90,60%	56,40%
PATTERN 6	48,50%	72,90%	43,90%	60,40%	90,90%	54,70%
VŠECHNY PATTERNY	54,90%	73,10%	49,90%	68,50%	91,20%	62,20%

9.4 Doporučení pro zvýšení úspěšnosti precision a recall v kontextu řádku

Pro zvýšení úspěšnosti analýzy v aspektech precision a recall v kontextu řádků je třeba definovat trochu jiná doporučení než ta uvedená výše, protože zde nahlížíme na úspěšnost analýzy v kontextu celého řádku a ne pojmu. Jak je vidět v tabulce 3 nejvíce řádků bylo určeno falešně negativních. Zde se jedná především o řádky, kde nebyl přiřazen žádný pojem v dané kategorii, především proto, že nebyl ve slovníku označený. Tento problém řeší rozšiřování a úprava slovníku.

U řádků označených false positive je to složitější. Jako falešně pozitivní jsou označeny nejčastěji řádky, kde je více pojmů pod ‚MultipleComponent' a ‚MultipleFailure', je tedy nutné pojmy eliminovat a ne rozšiřovat. Snazší způsob je eliminovat pojmy pod ‚MultipleFailure', protože těchto řádků je méně než těch s ‚MultipleComponent' a obecně pojmů typu závada je ve slovníku méně. Zde by se daly částečně přepoužít některé z jichž definovaných patternů. Konkrétně patterny 3 a 6, které se týkají právě závady. V kontextu řádku by tento pattern mohl fungovat tak, že v případě výskytu pojmu ‚hole' a ‚out of limit' v ‚MultipleFailure' by se tento pojem vymazal. Tím by se dala eliminovat část řádků označená jako false positive.

10 Diskuse

Cílem této práce bylo navrhnout řešení a doporučení, které by vedlo ke zvýšení úspěšnosti automatické textové analýzy prováděné na nestrukturovaných datech z letecké údržby společnosti CSAT. Tato práce navázala na diplomovou práci T. Vojtěcha, jejímž cílem bylo seznámení se s možnostmi textové analýzy na těchto datech a zhodnocení jejího využití v praxi. Výsledkem této práce byla hladina úspěšnosti u závislé sady dat 74 % v případě určení komponenty, 79 % v případě určení závady a 59 % v případě určení páru. U nezávislé sady dat byla hladina úspěšnosti 60 % v případě určení komponenty, 77 % v případě určení závady a 46 % v případě určení páru.

Hodnocení úspěšnosti textové analýzy v práci T. Vojtěcha bylo omezeno pouze na relevantní řádky, tedy na ty, u kterých bylo manuálně rozhodnuto o tom, jestli obsahují informaci o komponentě a závadě. Z pohledu praxe ale toto není úplně vypovídající statistika, protože v praxi při použití analýzy nebude manuálně o relevantnosti řádků rozhodováno. Bylo by to možné, ale v případě velkého množství dat velice časově náročně. Pro praktické využití je více vypovídající vytvoření statistiky, která bude brát v potaz celý dataset a bude vypovídat o výsledcích textové analýzy objektivně. K tomuto účelu byly zvoleny parametry precision a recall, které počítají a berou v potaz i nerelevantní řádky a zobrazují úspěšnost analýzy na celém analyzovaném datasetu. Pro potřeby porovnání byl vyhodnocen i parametr úspěšnosti výběru pojmů u relevantních řádků.

V této práci proběhlo vyhodnocení na těchto parametrech: úspěšnost výběru komponenty (z relevantních řádků), úspěšnost výběru závady (z relevantních řádků), úspěšnost výběru páru (z relevantních řádků), precision a recall. Poslední dva parametry byly vyhodnoceny v kontextu řádků a v kontextu pojmů. V kontextu pojmů proběhlo vyhodnocení zvlášť pro komponenty, závady a páry.

Porovnání výsledků této práce bylo provedeno pomocí výše zmíněných parametrů. Byl použit i jiný slovník, který pro potřeby této práce vnikl, lze tedy porovnat, jak jiný slovník ovlivňuje úspěšnost a rozdíly v přístupu k jeho vytváření.

Slovník vznikl na závislé sadě dataset 1, celkem bylo vytvořeno 545 pojmů. Překryv toho slovníku byl následně 2x ověřen na dvou nezávislých datasetech a 2x rozšířen na celkových 947 pojmů. V tabulce 2 jsou vypsány výsledky, které slovník vykazoval na datasetech během jeho rozšiřování. Výsledky prokázaly, že rozšiřováním slovníku se zvyšuje úspěšnost označování pojmů v textu automatickou analýzou.

Jako první byla vyhodnocena závislá sada z trénovací statistiky. Porovnání probíhalo automaticky porovnáním textu ‚ComponentLabel' a ‚FailureLabel' z trénovací statistiky a statistiky celého datasetu. Výsledky i s porovnáním z výsledků textové analýzy T. Vojtěcha jsou zapsány v tabulce 5. Nižší procenta výsledků analýzy této práce jsou způsobeny automatickým porovnáním textu. Protože závislé sady byly vyhodnocovány odlišnou metodou, nelze jejich výsledky efektivně porovnat. Více vypovídající je až vyhodnocení na nezávislé sadě dat, výsledky jsou zapsány v tabulce 5.

U nezávislé sady bylo dosaženo téměř totožného výsledku jako u T. Vojtěcha v případě komponent. V případě závad tato analýza vykázala téměř o 10% lepší výsledek. Vyšší úspěšností u závad měla za následek i zvýšení úspěšnosti u párů o 8 %. K tomu, aby se eliminovaly veškeré rozdíly ve vyhodnocení dvěma různými uživateli, byla vyhodnocena ještě jedna nezávislá sada dat, kde bylo manuálně rozhodováno o správnosti přiřazeného pojmu jak statistiky této práce, tak statistiky práce T. Vojtěcha. Výsledky jsou uvedené v tabulce 6.

Z této tabulky vyplývá, že výsledky jsou téměř totožné s těmi z předchozí tabulky. V případě T. Vojtěcha jsou taktéž téměř totožné, menší pokles úspěšnosti je vidět jen u komponent. Celkově na nezávislé sadě dat tato analýza vykázala o 5,3 % lepší výsledek v případě komponent, o 13 % lepší výsledek v případě závad a 10, 8 % a v případě párů. V případě, že by byly za výchozí výsledky považovány ty uvedené v jeho práci, jednalo by se o totožný výsledek v případě komponent, zlepšení o 13,3 % v případě závad a zlepšení o 9 % v případě párů. Lepší výsledek u komponent je zapříčiněn jiným přístupem k tvorbě slovníku, kde jsou zahnuty ve větší míře i obecnější pojmy, které lze efektivně využít v případech, kdy ve slovníku není přesný termín. Lepší výsledek u závad je z největší části způsoben vyhledávacím textem ‚not found' pojmu ‚missing' ve slovníku T. Vojtěcha, který má za následek nesprávně přiřazení pojmu ‚missing' v řádcích, kde se vyskytuje slovo ‚found'. Celkově se podařilo jiným přístupem k tvorbě slovníku zlepšit úspěšnost analýzy v porovnání s T. Vojtěchem.

Dalšími parametry byly precision a recall. Ty byly vyhodnoceny v kontextu řádků. Zde by měla být textová analýza schopna na základě definovaných parametrů rozhodnout, zda je řádek relevantní nebo není. Výsledky jsou zapsány v tabulce 3. Celkově dokázala analýza v 71 % případů správně rozhodnout o relevantnosti řádku. Precision byl vypočítán s výsledkem 79,5 % a recall s výsledkem 72,9 %. Tyto výsledky jsou zajímavé pro uživatele v případě, že by chtěl automaticky roztřídit řádky na relevantní a

nerelevantní. V případě využití textové analýzy to je schopen udělat s relativně vysokou přesností. Takový postup by byl vhodný, pokud by chtěla společnost automaticky určit poměr mezi relevantními a nerelevantními řádky v datasetu nebo pokud by chtěla analyzovat dále pouze analýzou určené relevantní řádky.

Dále byl precision a recall vypočítán na nezávislé sadě v kontextu pojmu. Výsledky jsou v tabulce 4. Tyto parametry mají pro koncového uživatele větší vypovídající hodnotu, protože uvažují celý dataset. V případě T. Vojtěcha byla úspěšnost počítána pouze na relevantních řádcích, u kterých bylo ale o relevantnosti rozhodnuto manuálně. Z časového hlediska není praktické manuálně rozhodovat o relevantnosti řádků celého datasetu, který jich může obsahovat i desetitisíce. Pokud by ale toto bylo provedeno, výsledem by vykázaly lepší hodnoty.

Způsobů, jak zvýšit úspěšnost analýzy, je několik. První z nich je možnost změny postupů zápisu nálezů tak, aby se nerelevantní záznamy zapisovaly jinam, nebo aby se od relevantních jednoznačně odlišily. Další možností je omezení složitosti záznamů. Více pojmů na jednom řádků je aspekt, který nejvíce snižuje přesnost analýzy. Jednalo by se o opatření, které by vyžadovalo zkrácení zápisu. Zbytek požadovaných informací by bylo nutné zaznamenat jinak. Další možností je nerelevantní část záznamu jednoznačně odlišit. Ideálním řešením by byla kombinace obou výše zmíněných doporučení. V praxi mají doporučení, která mění postup záznamu, nebo doporučení, která definují požadavky na zavedený software, značně omezenou možnost aplikace.

V práci bylo rovněž dokázáno, že úspěšnost je možné zvýšit i přístupem k tvorbě slovníku. V případě, že jsou do slovníku zahrnuty pojmy, které mají obecnější charakter, a zároveň je zamezeno tomu, aby se název nebo část názvu pojmu shodovala s často se vyskytujícími slovy, má analýza s tímto slovníkem vyšší úspěšnost.

Možností, jak zvýšit úspěšnost za současného zachování postupu záznamu, je definovat patterny a ty poté aplikovat během přiřazování pojmů. Patterny popsané v této práci byly definované v průběhu analyzování a ručního procházení záznamů. Do práce byly vybrány ty, které se na daném vzorku vyskytly nejčastěji. Jejich efekt byl posouzen na výsledné procentuální úspěšnosti označení pojmů a na precision a recall. Výsledky jsou zapsány v tabulce 10 a 11. Každý pattern byl nejprve vyhodnocen zvlášť a následně všechny najednou. V případě aplikace všech šesti patternů se podařilo dosáhnout zlepšení u párů u procentuální úspěšnosti o 7,6 %, v případě precision o 6 % a recall o 7,5 %.

Nalezení patternů je časově náročná záležitost. Patterny vznikaly a byly posuzovány během celé doby, kdy tato práce vznikala. I přesto, že byly vybrány ty, které se vyskytovaly nejčastěji, nebyl jejich efekt takový, že by zásadně ovlivnil výsledek analýzy. Ostatní patterny, které nebyly aplikovány a které se nevyskytovaly tak často, by měly menší efekt než ty aplikované. Navíc aplikací většího množství patternů, by nastaly situace, kdy by definované patterny byly spolu v konfliktu. To by vedlo k potřebě definovat navíc i prioritu jednotlivých patternů.

Obecně během vzniku práce bylo nalezeno několik konfliktních požadavků na textovou analýzu.

Za prvé se jedná o velikost slovníku. Bylo dokázáno, že čím obsáhlejší slovník je, tím větší má na nezávislé sadě úspěšnost v nalezení správného pojmu (uvedeno v ‚MultipleComponent' a ‚MultipeFailure'), tím se eliminují případy, kdy není na řádku nalezen správný pojem. Zároveň ale obsáhlejší slovník a více pojmů v ‚MultipleComponet' a ‚MultipleFailure' znamená pro analýzu více shod na daném řádku, tudíž více nutností rozhodování a více nesprávně přiřazených pojmů.

Za druhé se jedná o samotná data. Z provozu je požadavek na co nejpřesnější zápis závady. Z pohledu analýzy jsou ale takovéto řádky ty nejvíce komplikované a nejvíce chybové. Dokud budou do analýzy vstupovat řádky obsahující velké množství popisu a nerelevantní řádky, bude úspěšnost analýzy omezena na hodnoty, které byly v této práci popsány. Jednoduchým řešením pro filtraci nerelevantních řádků by bylo je při zápisu jednoznačně odlišit. Například určitým znakem nebo slovním spojením. Takové opatření by nezasahovalo do softwarových možností programu a z pohledu provozu to není velký zásah do provozního postupu. Takovéto jednoduché opatření by umožnilo filtraci dat např. v Microsoft Excel před samotnou analýzou. Dalším problémem je nekonzistence relevantních řádků. Záznamy jsou vytvářeny různými mechaniky, to znamená různé přístupy k zápisu. Každá situace si žádá trochu jiný přístup, jinou míru detailu atd. To způsobuje, že záznamy jsou každý trochu jiný a nelze v nich identifikovat strukturu nebo pravidelnosti výskytu určitých pojmů. Některé řádky jsou složité, obsahují rozsáhlý popis lokace a okolností další instrukce a informace. V jiném případě se zase jedná o jednoduchý zápis o čtyřech slovech.

Za třetí je to samotný požadavek na zvyšování úspěšnosti. Obecně manuální práce s velkým množstvím dat je zdlouhavá. Úspěšnost analýzy se dá zvýšit manuálním roztřízením relevantních řádků. Pokud by se toto před samotnou analýzou provedlo,

vedlo by ke zlepšení výsledků. Je to tedy kolize časové úspory a lepších výsledků. Identifikace patternů je taktéž časově náročné. K tomu, aby byly vybrány ty, které budou mít největší vliv, je třeba analyzovat několik tisíc řádků manuálně. V případě, že je uživatel tento čas ochotný věnovat a identifikovat patterny, přinese to celkové zlepšení řádově několika procent.

Závěr

Tato práce byla zaměřena na zvýšení úspěšnosti automatické textové analýzy prováděné na nestrukturovaných záznamech dat letecké údržby. Tato práce navázala na výsledky práce T. Vojtěcha a jejím cílem bylo navrhnout řešení pro zvýšení úspěšnosti.

Z této práce byl využit návrh provádění textové analýzy a využit stejný software. Pro porovnání různých přístupů tvorby referenčního slovníku dvou uživatelů byl v této práci vytvořen nový slovník.

Výsledky obou prací byly porovnány na závislé a nezávislé sadě dat. V případě nezávislé sady dat analýza na slovníku této práce vykázala lepší výsledky v řádu několika procent. Různé přístupy k tvorbě slovníku tedy ovlivňují výsledek textové analýzy a je třeba dbát na efektivní tvorbu a údržbu slovníku.

V této práci byly určeny a vypočítány další parametry (precision a recall), které mají pro koncového uživatele větší vypovídající hodnotu než vyhodnocení použité v práci T. Vojtěcha. Ten ve své práci vyhodnocení prováděl pouze na relevantních řádcích, které ale v praxi určeny nebudou. Parametry precision a recall jsou vypočítány z celé datové sady, navíc se dají použít k porovnání s jinými textovými analýzami.

V práci bylo definováno několik doporučení pro zvýšení úspěšnosti, z nichž některé jsou navzájem konfliktní. Jedním z doporučení je změna postupu záznamu závady. V praxi jsou ale změny zavedených postupů komplikované. Některá z doporučení také naráží na technické možnosti využívaného údržbového softwaru.

Doporučení, které se podařilo aplikovat a jeho účinnost ověřit, bylo vytvoření patternů. V práci bylo použito celkem 6 patternů, u kterých bylo posouzeno, že budou mít na výsledek největší vliv. Celkově se po aplikaci těchto patternů podařilo dosáhnout zlepšení výsledků o několik procent. Z pohledu praxe to není zanedbatelné, protože i pár procent v celkovém objemu dat představuje několik tisíc záznamů.

Je důležité zmínit, že aplikace patternů byla pouze nasimulovaná. Jejich vliv byl posuzován individuálně na každém řádku datasetu. Patterny nebyly z časové náročnosti aplikovány do algoritmu nástroje TermIt. V případě, že by byly do TermIt implementovány předpokládám zlepšení, které by se pohybovalo na podobné úrovni, jako v této práci. Úroveň zlepšení by byla ovlivněna četností řádků s výskytem aplikovaných patternů na analyzovaném datasetu.

Pro budoucí využití této analýzy bych doporučila nejprve změnu postupů záznamu závady jedním ze způsobů navržených v této práci. V případě, že by bylo možné změnit postupy záznamu a omezit množství nerelevantních řádků vstupujících do analýzy a pokud by se podařilo nalézt způsob, jak jednoznačně označit nebo odstranit nepoužitelnou část záznamu, mělo by smysl ověřit znovu úspěšnost analýzy, případně vliv identifikovaných patternů. Změnu přístupu k záznamu závady považuji za klíčovou k dalšímu potenciálnímu zlepšení a praktickému využití.

Zdroje

[1] Dictionary cambridge english maintenance. *Dictioary cambridge* [online]. [cit. 2022-02-17]. Dostupné z: https://dictionary.cambridge.org/dictionary/english/maintenance

[2] Types of Aviation Maintenance Checks. *National aviation academy* [online]. 2020 [cit. 2022-02-17]. Dostupné z: https://www.naa.edu/types-of-aviation-maintenance-checks/

[3] Nařízení Komise (EU) č. 1321/2014. *CAA - Úřad pro civilní letectví* [online]. [cit. 2022-02-17]. Dostupné z: https://www.caa.cz/dokumenty/predpisy/zakladni-informace-k-narizenim-eu/zachovani-letove-zpusobilosti/narizeni-komise-eu-c-1321-2014/

[4] *Nařízení komise č.1321/2014 ze dne 26. listopadu 2014 o zachování letové způsobilosti letadel a leteckých výrobků, letadlových částí a zařízení a schvalování organizací a personálu zapojených do těchto úkolů*. In:. EU: Evropská komise, ročník 2014, číslo 1321. Dostupné také z: https://eur-lex.europa.eu/legal-content/CS/TXT/PDF/?uri=CELEX:02014R1321-20211202&qid=1642680112277&from=en

[5] What is Text Analysis?. *ONTOTEXT* [online]. [cit. 2022-02-24]. Dostupné z: https://www.ontotext.com/knowledgehub/fundamentals/text-analysis/

[6] CHEN, Michelle. A Guide: Text Analysis, Text Analytics & Text Mining. *Towards data science* [online]. 2020 [cit. 2022-02-24]. Dostupné z: https://towardsdatascience.com/a-guide-text-analysis-text-analytics-text-mining-f62df7b78747

[7] STENSTRÖM, Christer, Mustafa ALJUMAILI a Aditya PARIDA. Natural Language Processing of Maintenance Records Data. International journal of COMADEM: Division of Operation and Maintenance Engineering, Luleå University of Technology. Dostupné také z: https://www.diva-portal.org/smash/get/diva2:975548/FULLTEXT01.pdf

[8] BRUNO, Nicola, Tommy JUN a Henry TESSIER. *Natural Language Processing and Classification Methods for the Maintenance and Optimization of US Weapon Systems* [online]. University of Virginia, April 2020 [cit. 2022-03-08]. Dostupné z: doi:978-1-7281-0998-5

[9] VOJTĚCH, Tomáš. *Textová analýza nestrukturovaných závadových dat v letecké údržbě*. Praha, 2021. Diplomová práce. ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE, FAKULTA DOPRAVNÍ, Ústav letecké dopravy.

[10] *Voyant: see through your text* [online]. [cit. 2022-03-22]. Dostupné z: https://voyant-tools.org/

[11] *Word and phrase* [online]. [cit. 2022-03-22]. Dostupné z: https://www.wordandphrase.info/analyzeText.asp

[12] *Czech Airlines Technics* [online]. [cit. 2022-04-05]. Dostupné z: https://www.csatechnics.com/cs

[13] Osvědčení organizace údržby: CZ.145.0067. *Úřad pro civilní letectví* [online]. 25.2.2022 [cit. 2022-04-05]. Dostupné z: https://www.caa.cz/wp-content/uploads/2022/03/CZ.145.0067.pdf?cb=1a7ba36f1d17b7f40560748511e94a6a

[14] List Of ATA 100 Chapters. *Aerospace unlimited* [online]. [cit. 2022-04-13]. Dostupné z: https://www.aerospaceunlimited.com/ata-chapters/

[15] *Swiss aviation software* [online]. [cit. 2022-11-30]. Dostupné z: https://www.swiss-as.com/

[16] *Swiss as- AMOS* [online]. [cit. 2022-11-30]. Dostupné z: https://softwareconnect.com/aviation-mro/swissas-amos/

[17] SHAFI, Adam. How to Learn the Definitions of Precision and Recall. *Towards data science* [online]. [cit. 2022-11-25]. Dostupné z: https://towardsdatascience.com/precision-and-recall-88a3776c8007

[18] LEDVINKA, Martin, Petr KŘEMEN, Lama SAEEDA a Miroslav BLAŠKO. TermIt: : A Practical Semantic Vocabulary Manager. ICEIS 2020 - 22nd International Conference on Enterprise Information Systems: Department of Computer Science, Faculty of Electrical Engineering, Czech Technical University in Prague. Dostupné také z: https://www.scitepress.org/Papers/2020/95637/95637.pdf