FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání bakalářské práce

Název:	Vizualizace extrakce témat Korpusu českého verše
Student:	Jonáš Sirko
Vedoucí:	Ing. Magda Friedjungová, Ph.D.
Studijní program:	Informatika
Obor / specializace:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2023/2024

Pokyny pro vypracování

Cílem této práce je zmapovat různé možnosti vizualizace shlukovacích algoritmů. Vybrané metody budou demonstrovány na korpusu české poezie, nad kterým je řešena nesupervizovaná úloha shlukování za účelem extrakce témat a motivů (tzv. "topic modeling"). Výstupem bude ucelená vizuální prezentace řešení zmíněné úlohy.

1. Proveďte rešerši vizualizačních metod a algoritmů pro shlukování dat s důrazem na aplikaci v oblasti zpracování přirozeného jazyka (NLP), např. [1,2].
2. Seznamte se s Korpusem českého verše https://github.com/versotym/corpusCzechVerse/ a s metodami modelování poetických témat (tzv. "topic modeling"), např. [3].
3. Vybrané metody aplikujte na Korpus českého verše. Lze využít již hotové implementace.
4. Na základě rešerše z bodu 1 zvolte a implementujte vhodné vizualizační metody, kterými doplníte vybrané kroky (např. průzkum zdrojových dat a příslušných embeddingů, stanovení vhodného počtu shluků, vyhodnocení evaluační metriky apod.) v procesu modelování poetických témat včetně vizualizace výsledků. Pokud to metoda umožňuje, experimentujte s nastavením parametrů a své závěry komentujte.

Reference

[1] https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_617
[2] https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf
[3] https://arxiv.org/pdf/2006.15732.pdf

Elektronicky schválil/a Ing. Karel Klouda, Ph.D. dne 21. října 2022 v Praze.

Bakalářská práce

VIZUALIZACE EXTRAKCE TÉMAT KORPUSU ČESKÉHO VERŠE

Jonáš Sirko

Fakulta informačních technologií
Katedra teoretické informatiky
Vedoucí: Ing. Magda Friedjungová, Ph.D.
29. června 2023

Odkaz na tuto práci: Sirko Jonáš. *Vizualizace extrakce témat Korpusu českého verše*. Bakalářská práce. České vysoké učení technické v Praze, Fakulta informačních technologií, 2023.

Obsah

Seznam obrázků

Seznam tabulek

Chtěl bych poděkovat především Ing. Magdě Friedjungové, Ph.D. za vedení a pomoc při tvorbě této práce.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze dne 29. června 2023 .....

Abstrakt

Tato bakalářská práce se zabývá procesem vizualizace shlukovacích metod s důrazem na aplikaci v oblasti zpracování přirozeného jazyka, které jsou použity pro řešení nesupervizované úlohy shlukování za účelem extrakce témat a motivů z korpusu českého verše. V práci jsou představeny jak shlukovací metody samotné, tak možné způsoby jejich vizualizací. Tyto metody jsou dále aplikovány na vektorizovaný korpus českého verše a s pomocí vizualizací porovnány.

Klíčová slova shlukování, vizualizace, extrakce témat, zpracování přirozeného jazyka, korpus českého verše

Abstract

This bachelor thesis deals with the visualization process of clustering methods with emphasis on the application in natural language processing, which are used to solve the unsupervised clustering task in order to perform a topic modeling on a corpus of Czech verse. The paper presents both the clustering methods themselves and possible ways of visualizing them. These methods are further applied to embedded corpus of Czech verse and compared with the help of visualizations.

Keywords clustering, visualization, topic modeling, natural language processing, Czech corpus verse

Seznam zkratek

KČV	Korpus českého verše
PCA	Principal Component Analysis
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction
PaCMAP	Pairwise Controlled Manifold Approximation Projection
BERT	Bidirectional Encoder Representations from Transformers
LDA	Latent Dirichlet allocation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise

Úvod

Extrakce témat je užitečný nástroj pro práci s textovými dokumenty. Dává rychlou představu o jejich obsahu, usnadňuje jejich třídění a umožňuje efektivní vyhledávání ve velkých archivech textů. Tento proces však nemusí být sám o sobě transparentní a poskytovat informace o relevanci a kvalitě témat a motivů v dokumentech nalezených. Proto je důležité průběh tohoto procesu monitorovat v celém jeho průběhu a srozumitelně vyhodnocovat a vysvětlovat výsledky jeho jednotlivých kroků. Tato práce se především zaměří na vysvětlování jednoho takového kroku – shlukování – a bude tak činit hlavně pomocí vizualizací.

Cílem této práce je popsat soudobé metody extrakce témat (převážně ty s využitím nesupervizovaného shlukování), metody shlukování samotné a představit možnosti jejich vizualizací. V praktické části je pak cílem demonstrovat využití těchto metod na Korpusu české poezie, nad kterým je úloha nesupervizovaného shlukování řešena, včetně porovnání a vyhodnocení výsledků a prezentace uceleného vizuálního řešení zmíněné úlohy.

Kapitola 1
Teoretická část

V této kapitole budou představeny některé nesupervizované metody extrakce témat a shlukování a možností jejich vizualizace.

1.1 Extrakce témat

Obecně lze metody extrakce témat rozdělit na supervizované a nesupervizované. U obou je předpokládáno použití na korpusu, který je složen z velkého množství různých textů – dokumentů. Cílem nesupervizovaných metod je odhalit významná témata, jimiž se tyto dokumenty zabývají [1]. Supervizované metody jsou pak vhodné pro řešení klasifikační úlohy – dělení dokumentů na základě předem známých kategorií (témat)[2]. Tato práce se bude zabývat pouze nesupervizovanými metodami.

1.1.1 LDA

LDA [3] (Latent Dirichlet Allocation) je hierarchický, trojvrstvý bayesovský pravděpodobnostní model. LDA lze zjednodušeně vysvětlit fiktivním generativním procesem, který popisuje vznik jednotlivých dokumentů. Předpokladem LDA je, že v každém dokumentu je obsaženo více témat. Tématem se v LDA rozumí pravděpodobnostní rozdělení nad pevně danou množinou slov a předpokládá se, že tato témata jsou určena ještě před vznikem dokumentů samotných. Proces generování jednotlivých dokumentů je následující:

1. Nad tématy je náhodně zvoleno pravděpodobnostní rozdělení
2. Pro každé slovo v dokumentu:
   a. Je náhodně zvoleno téma na základě rozdělení vybraného v prvním kroku
   b. Z tohoto tématu [1] je náhodně zvoleno slovo.

Všechny dokumenty tedy sdílejí stejná témata, ale v každém z dokumentů se vyskytují s různou pravděpodobností, v závislosti na pravděpodobnostním rozdělení nad tématy.

LDA lze intuitivně chápat jako „obrácení“ tohoto fiktivního generativního procesu – pozorované jsou dokumenty – a témata, pravděpodobnostní rozdělení témat nad jednotlivými dokumenty a přiřazení témat k jednotlivým slovům v dokumentech zůstávají skryta.

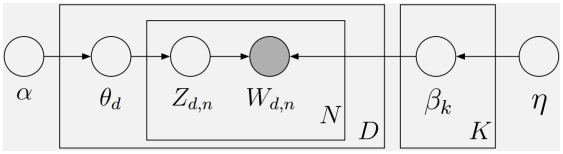
[1] Které, jak je popsáno výše, je samo pravděpodobnostním rozdělením nad konečnou množinou slov

Formálně můžeme generativní proces vyjádřit jako sjednocené pravděpodobnostní rozdělení skrytých[2] a pozorovatelných veličin

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:k}, z_{d,n}) \right),$$

kde každé β_k je téma, θ_d rozdělení témat pro d-tý dokument ($\theta_{d,k}$ je rozdělení tématu k v dokumentu d). Přiřazení konkrétního tématu v dokumentu d je pak značeno z_d (kde $z_{d,n}$ je téma n-tého slova v dokumentu d) a w_d jsou pak pozorovaná slova v dokumentu d ($w_{d,n}$ je pak opět n-té slovo v dokumentu).

V tomto rozdělení se vyskytuje několik podmíněných proměnných, konkrétně přidělení tématu $z_{d,n}$ (závislá na θ_d) a jednotlivá pozorovaná slova $w_{d,n}$, závislá na $z_{d,n}$ a všech tématech $\beta_{1:k}$. Závislosti jsou vystiženy v obrázku 1.1



Obrázek 1.1 Grafický model LDA. Vrcholy zobrazují jednotlivé náhodné veličiny. Jejich barva pak určuje, zda se jedná o přímo pozorované (šedá), nebo skryté veličiny (bílá).Obdélníky pak znázorňují opakovaný výskyt – N pro slova v dokumentu a D pro dokumenty v datasetu.

Obecně lze říci, že LDA se snaží najít pravděpodobnost podmíněného rozdělení témat na základě pozorovaných dokumentů:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}. \tag{1.1}$$

Výraz ve jmenovateli odpovídá pravděpodobnosti výskytu pozorovaného korpusu pod libovolným tématem. V praxi je jeho výpočet velmi náročný. Tento problém je společný pro mnoho pravděpodobnostních modelů extrakce témat. Řešení spočívá v aproximaci rovnice 1.1 zvolením alternativního pravděpodobnostního rozdělení nad skrytými náhodnými veličinami popisující témata. [1, 3]

1.1.2 BERTopic

BERTopic [4] je jedna z metod extrakce témat, která volí jiný přístup než LDA. Grootendorst ve studii [4], která BERTopic představuje, zmiňuje, že metody extrakce témat, která popisují dokument jako *bag of words* a modelují ho jako směs latentních témat trpí nedostatky. Jako jeden z nich uvádí, že při nahlížení na dokument jako bag of words nejsou uvažovány sémantické vazby mezi slovy a větný kontext, tudíž tento způsob může vyústit v nepřesnou interpretaci dokumentu. Jako řešení tohoto problému představuje způsob extrakce témat, který je založen na vektorizaci dokumentů [3]. Vektorizace dokumentů je způsob jejich reprezentace jako vektory reálných čísel tak, že sémanticky podobné dokumenty se ve vektorovém prostoru vyskytují blízko u sebe a autor dále pracuje s předpokladem, že dokumenty obsahující stejná témata jsou si sémanticky podobná. Celý proces extrakce témat probíhá následujícím způsobem:

1. Vektorizace jednotlivých dokumentů pomocí modelu Sentence-BERT (SBERT) [6].

[2] neboli latentních

[3] Tato myšlenka vychází ze studie představující model BERT[5]

2. Dimenzionální redukce. Dimenze vektorového prostoru, ve kterém nově vzniklé vektory leží je obvykle dostatečně vysoká na to, aby se projevily některé jevy spojené s tzv. *prokletím dimenzionality*, autor zmiňuje případ popsaný ve studiích [7, 8]: Se zvyšující se dimenzí vektorového prostoru se vzdálenost k nejbližšímu bodu blíží vzdálenosti k nejvzdálenějšímu bodu. Koncept vzdálenosti v závislosti na poloze bodů přestává dávat smysl a metriky vzdálenosti se liší pouze nepatrně. Vzhledem k následujícímu kroku je tedy vhodné provést dimenzionální redukci vzniklých vektorů. Jako výchozí je zvolena metoda UMAP [9].

3. Shlukování (viz sekce 1.2) redukovaných vektorů. Vzhledem k předpokladu, že dokumenty obsahující stejná témata jsou si sémanticky podobná a jejich vektorové reprezentace se ve vektorovém prostoru vyskytují u sebe lze intuitivně předpokládat, že dokumenty zabývající se podobnými tématy budou v tomto prostoru tvořit shluky. K jejich identifikaci je použit algoritmus HDBSCAN (viz podsekce 1.2.4.2).

4. Každému shluku je přiřazeno jedno téma tak, aby co nejlépe vystihovalo všechny obsažené dokumenty. K tomuto účelu je použit modifikovaný algoritmus TF-IDF [10], který měří důležitost slova W v dokumentu d

$$W_{t,d} = tf_{t,d} \cdot \log\left(\frac{N}{df_t}\right),$$

kde $tf_{t,d}$ je frekvence termu v dokumentu, N je celkový počet dokumentů v datasetu a df_t je počet dokumentů v datasetu obsahujících term. Význam slova roste, pokud se velmi často vyskytuje v dokumentu a klesá, pokud se příliš často vyskytuje v ostatních dokumentech. Autor představuje upravenou verzi pro shluky dokumentů c-TF-IDF (class-based TF-IDF). Všechny dokumenty ve shluku jsou spojeny dohromady a jsou nadále považovány za jeden dokument. TF-IDF je nadále upraven tak, aby místo slovy v dokumentech pracoval s dokumenty ve shlucích:

$$W_{t,c} = tf_{t,c} \cdot log\left(1 + \frac{A}{tf_t}\right),$$

kde $tf_{t,c}$ je četnost termu t ve shluku c[4], A je průměrný počet slov ve shluku a tf_f je frekvence termu ve všech shlucích. Tato úprava tedy umožňuje měření důležitosti slova ve shluku a tím i tvorbu témat, kdy jako témata jsou určena slova s největším významem. Určitého (menšího) počtu témat je možné dosáhnou postupným slučováním nejpodobnějších nejméně významných slov

Výhodou této metody je její modularita. Oddělení jednotlivých kroků umožňuje jak experimentování s různými metodami pro jednotlivé úlohy (různé modely pro vektorizaci, různé typy shlukovacích algoritmů, ...)[6], tak flexibilní práci s modelem obecně, například odlišnou úpravu dat pro vektorizaci a tvorbu témat. Nevýhodou pak je, že BERTopic nepředpokládá existenci více různých témat v rámci jednoho dokumentu a dále skutečnost, že BERTopic sice kontext původních dokumentů díky jejich vektorizaci uchovává, nicméně při tvorbě témat s ním nepracuje; témata jsou tvořena z bag of words reprezentace[7]. Slova použitá při tvorbě témat tak téma nemusí vystihnou přesně a i když pravděpodobně budou podobná, mohou se opakovat a být pro dané téma nadbytečná [4].

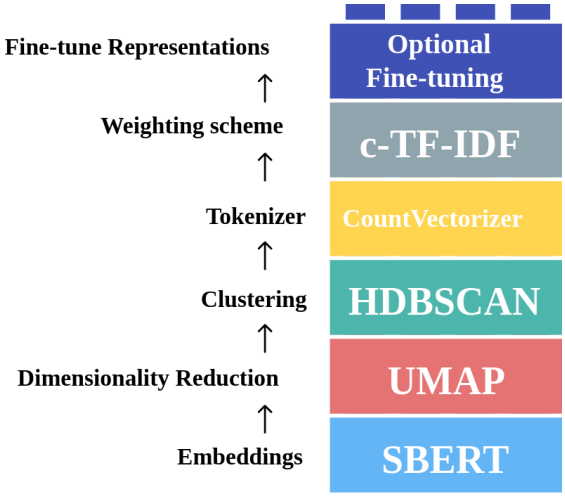
1.1.3 Top2Vec

Top2Vec [11] je založen na podobné myšlence jako BERTopic – pracuje s vektorizovanými dokumenty. Jednou z hlavních odlišností je, že top2vec využívá k vektorizaci model doc2vec [12],

[4] Dokumentu vzniklého spojením všech dokumentů ve shluku.

[6] Autor konkrétně zmiňuje, že model dosahuje dobrých výsledků pro různé modely použité pro vektorizaci a konstatuje, že s vývojem nových modelů se může BERTopic zlepšovat spolu s nimi.

[7] Vytvořené ze spojených dokumentů v rámci jednotlivých shluků.



Obrázek 1.2 Základní schéma BERTopicu s výchozími metodami.[5]

který vytváří reprezentace jak dokumentů, tak jednotlivých slov ve stejném vektorovém prostoru a to takovým způsobem, že vektory zastupující dokumenty leží blízko slovních vektorů, s nimiž jsou si sémanticky podobné. [11] Extrakce probíhá v následujících krocích:

1. Vektorizace dokumentů a slov z datasetu.
2. Dimenzionální redukce a shlukování. Podobně jako v BERTopicu je předpokládáno, že vektorové reprezentace dokumentů budou ve vektorovém prostoru přirozeně tvořit shluky na základě sémantické podobnosti. Z důvodu efektivity shlukování opět proběhne dimenzionální redukce (pomocí UMAP) a pro identifikaci shluků je použit HDBSCAN .
3. Volba vhodného tématu pro shluk. V tomto bodě se top2vec opět odchyluje od přístupu, který je použit v BERTopicu – díky tomu, že v tomto případě jsou k dispozici i vektorové reprezentace jednotlivých slov, je možné je k popisu témat využít a objevit slova, která jsou si s nimi sémanticky podobná. Vektory nejpodobnějších slov leží nejblíže ke středovému bodu jednotlivých shluků. [8]

Autor argumentuje, že díky práci s vektorovými reprezentacemi dokumentů a témat dosahuje top2vec při jejich extrakci lepších výsledků, než klasické generativní modely typu LDA, jelikož ty se při hledání témat snaží najít taková, která s největší pravděpodobností původní dokumenty vytváří a díky tomu volí slova, která se v dokumentech vyskytují nejvíce, ale téma nemusí vystihovat, často jsou příliš specifická, nebo naopak moc obecná. Jako další výhody pak uvádí automatické nalezení počtu témat a již zmíněný fakt, že z dokumentů není potřeba předem odstraňovat slova bez zvláštního významu.

[7] https://maartengr.github.io/BERTopic/algorithm/algorithm.html

[8] Autor dále zmiňuje, že není nutné odstraňovat slova bez zvláštního významu, jelikož ta se vyskytují téměř ve všech dokumentech a ve většině případů se vyskytují rovnoměrně vzdálena od všech vektorů dokumentů a nebudou tedy v názvu tématu zahrnuta.

1.2 Shlukování

Shlukování je typicky nesupervizovaná metoda, jejímž cílem je dané objekty rozdělit na skupiny tak, aby si objekty v každé skupině mezi sebou byly co nejvíce podobné a zároveň neobsahovaly objekty, jež si podobné nejsou. Některé shlukovací metody, pak do speciální skupiny řadí „odlehlé“ objekty, které se nehodí do žádné skupiny a mezi sebou si podobné nejsou[9]. Obecně mohou být objekty numerické, nebo kategorické, nicméně často (a také v tomto případě) se jedná o vektory příznaků v d-dimenzionálním vektorovém prostoru $D \subset S = \mathbb{R}^d$ [13, 14]. V takovém případě je možné k vyhodnocování (ne)podobnosti bodů použít eukleidovskou, nebo jinou vzdálenost. Na objekty (body) lze nahlížet tak, jako by pocházely z nějakého neznámého sdruženého pravděpodobnostního rozdělení $p(x)$, které je složeno z k rozdělení $p_i(x)$, kde každé odpovídá jednomu z k shluků v datech. [14]

Proces shlukování lze obecně popsat následovně:

1. Výběr příznaků, dle kterých bude shlukování probíhat.
2. Výběr shlukovacího algoritmu.
3. Interpretace výsledků. Vzhledem k tomu, že během shlukování vznikají skupiny, které nejsou předem známé, je zhodnocení jejich kvality klíčové.

Vzhledem k tomu, že se tato práce zabývá shlukováním vektorizovaných lemmat, jsou v tomto případě příznaky chápány jako souřadnice v d-dimenzionálním vektorovém prostoru. Cílem je tedy tvořit shluky tak, aby pohromadě byly vektory, které leží blízko u sebe a naopak ve stejných shlucích nebyly vektory, které jsou od sebe vzdálené. [13, 15]

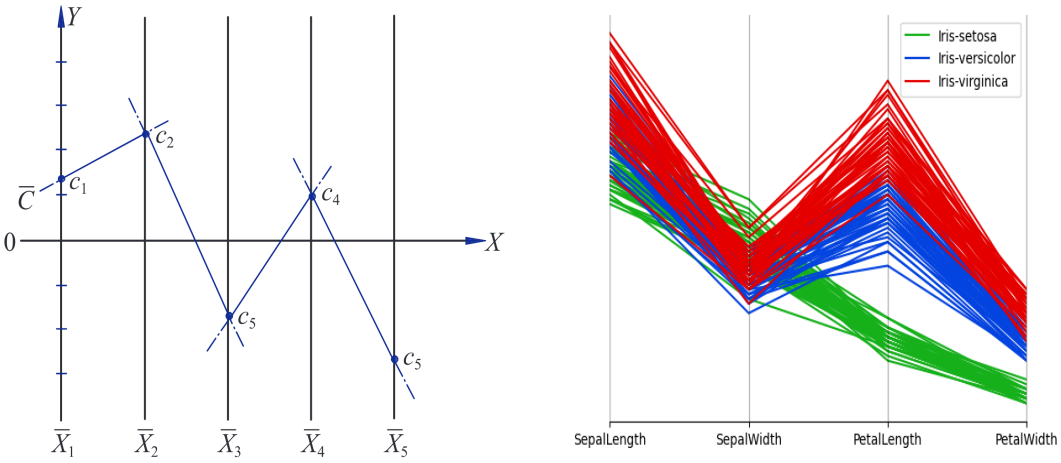
1.2.1 Vizualizace shlukování

Velmi častým a jednoduchým způsobem vizualizace shluků je bodový graf, který body na základě jejich souřadnic přímo zobrazí. Takové zobrazení je přehledné a intuitivní, nicméně jeho nevýhodou je, že bez další úpravy dat je vhodné pouze pro zobrazování bodů z $\mathbb{R}^2$ a $\mathbb{R}^3$. Jednou z možností zobrazení dat z prostoru vyšší dimenze d je nutné buď zobrazit projekce všech párových kombinací jednotlivých atributů. Těchto párů je však $\frac{d*(d-1)}{2}$, tedy počet grafů velmi rychle narůstá a přehlednost a intuitivnost klesá.[16] Další možností je před vizualizací redukovat dimenzi prostoru, ve kterém shlukování probíhá pomocí k tomu určených metod (např. PCA[17],t-SNE[18],UMAP[9], PACMaP[19]), s cílem co nejvíce zachovat strukturu dat. Výsledkem je jediná dvoudimenzionální projekce, tedy přehlednost a intuitivita zůstává zachována za cenu určitého možného zkreslení.

Jedním ze způsobů, díky kterému je možné vyhnout se jak redukci, tak zobrazování více grafů je zobrazení pomocí paralelních souřadnic (viz obrázek 1.3a). Tento způsob pracuje s alternativním zobrazováním, kdy pro n příznaků používá více os X značených jako $\overline{X}_1, \ldots, \overline{X}_n$, kdy tyto osy jsou rozmístěny s rovnoměrnými odstupy kolmo k původní ose X. Bod C se souřadnicemi $(c_1, \ldots, c_n)$ je pak zobrazen jako lomená čára s vrcholy ležícími na $(\overline{X}i, c_i)$ pro $i = 1, \ldots, n$.[20, 16]

Výše zmíněné metody pracovaly s přímým zobrazováním bodů shluků, nicméně k jejich vizualizaci lze využít i metriky jejich (ne)podobnosti, například vzdálenost. K tomuto účelu je využita matice vzájemných vzdáleností, z čehož plyne hlavní nevýhoda této metody – velikost této matice roste kvadraticky vzhledem k počtu bodů, nicméně tomu je možné se částečně vyhnout zobrazením vzdáleností pouze mezi jednotlivými shluky. Výhodou této metody pak je, že umožňuje vizualizaci různorodějších typů dat – data nemusí být vektory čísel, stačí aby existovala metrika určující jejich (ne)podobnost. [16]

[9] outliers

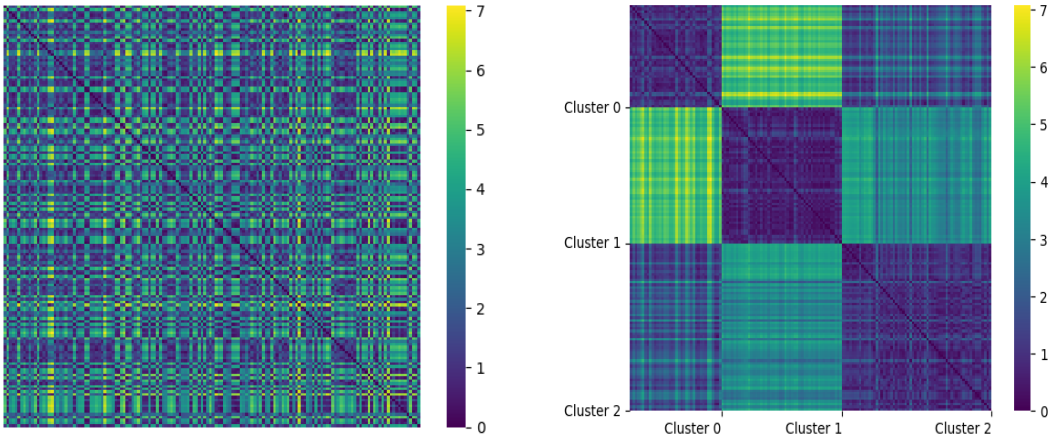


(a) Ukázka zobrazení pětidimenzionálního bodu C systémem paralelních souřadnic

(b) Paralelní zobrazení ukázkového datasetu iris se čtyřmi atributy

Obrázek 1.3 Zobrazení pomocí paralelních souřadnic

Vzdálenosti jsou pomocí barevné mapy převedeny na barvu (resp. odstín šedé), kdy změna barvy (resp. intenzity) odpovídá rostoucí (resp. klesající) vzdálenosti. Body jsou dále seřazeny podle vzniklých shluků. V případě, kdy přiřazení bodů do shluků dává smysl, je na první pohled patrný rozdíl mezi náhodnou permutací sloupců a řádků matice oproti částečnému seřazení dle shluků (viz obrázek 1.4). V takovém případě shluky tvoří čtvercové bloky podél diagonály matice. [16]



(a) Matice s náhodně prohozenými řádky (a odpovídajícími sloupci)

(b) Matice vzdáleností s řádky a sloupci seřazenými dle příslušnosti do shluku

Obrázek 1.4 Ukázka vizualizace matice vzdáleností na datasetu Iris, shlukování bylo provedeno pomocí K-means pro tři shluky.

1.2.2 Přímé dělení na části

Tento typ algoritmů se snaží data rozdělit do předem daného počtu shluků na základě funkce určující jejich kvalitu. [21, 13] Po zvolení takové funkce se úloha de facto mění na optimalizační

(např. minimalizace vzdálenosti, maximalizace korelace příznaků atd). [21] Tyto optimalizační úlohy jsou však NP-těžké [22], proto jsou pro implementaci algoritmů využívány hladové heuristiky ve formě iterativních optimalizací. [23]. Nejpoužívanějším v této kategorii je algoritmus K-means.

1.2.2.1 K-means

K-means dělí data na K shluků, které jsou určeny pomocí center. Centra shluků jsou vypočtena jako průměr všech bodů patřících do shluku

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_i,$$

kde N_k je počet bodů v k-tém shluku. Cílem je umístit středy shluků tak, aby součty vzdáleností mezi středem a body ve shluku ke středu náležícím byl co nejmenší, tedy minimalizovat reziduální součet čtverců[10]. Samotný algoritmus pak lze zapsat jako

1. Je zvoleno K shlukovacích center[11]
2. Dokud není splněna ukončující podmínka[12]:
   - Body se stejným nejbližším centrem jsou zařazeny do stejného shluku
   - Na základě nově vzniklých shluků je aktualizována poloha center

Na K-means je možné pohlížet jako na úlohu gradientního sestupu, kdy se s každou iterací klesá hodnota chybové funkce. [23]. Výhody K-means jsou převážně jeho lineární asymptotická složitost[13], jednoduchost, rychlost a možnost snadně ho implementovat. Hlavní nevýhody pak jsou citlivost na výchozí umístění center (hladový přístup může způsobit konvergenci do suboptimálního lokálního minima) a mimolehlé body – několik málo mimolehlých bodů může výrazně vychýlit centra shluků, možnost použití pouze na datech s numerickými příznaky a nutnost specifikovat dopředu počet shluků. [23, 24]

1.2.2.2 K-medoids

K-medoids [25] je algoritmus velmi podobný K-means, který také pracuje s minimalizací RSS, ovšem na rozdíl od K-means místo vzdálenosti ke geometrickému centru shluků minimalizuje vzdálenost k bodu zvanému medoid. Medoid je bod, jehož součet vzdáleností k ostatním bodům ve shluku je nejmenší z celého shluku Formálně je medoid shluku C definován jako

$$\arg\min_{x_m \in C} \sum_{x_i \in C} d(x_i, x_m),$$

kde d je nějaká míra odlišnosti, např. eukleidovská vzdálenost [25]. K-medoids je méně citlivý na odlehlé body, jelikož na rozdíl od průměru medoid tolik nevychýlí, nicméně výpočetní náročnost je vyšší. [23]

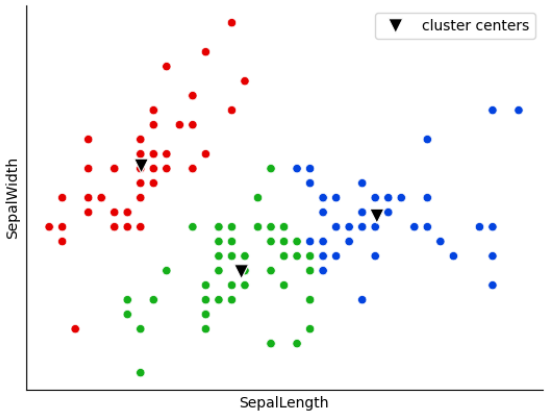
---

[10] $$\sum_{k=1}^{K} \sum_{\forall x_i \in C_k} \|x_i - \mu_k\|^2,$$
kde K je celkový počet shluků a C_k je k-tý shluk.

[11] Náhodně, či za pomoci nějaké heuristiky po předchozí analýze dat.

[12] Např. pokud se RSS dále nezmenšuje (v tom případě bylo dosaženo lokálního optima), nebo po určitém počtu iterací

[13] Pro T iterací pro N objektů s m příznaky rovna $O(T{\cdot}K{\cdot}m{\cdot}N)$



Obrázek 1.5 Shlukování na datasetu Iris pomocí algoritmu K-means pro tři shluky s vyznačenými středy.

1.2.3 Hierarchické shlukování

Hierarchické shlukování se dělí do dvou kategorií:

- Aglomerativní – na začátku je jako shluk uvažován každý jednotlivý bod. V každém dalším kroku jsou pak shluky po dvou slučovány do větších shluků.
- Rozdělující – probíhá přesně opačným způsobem než aglomerativní – všechny body začínají v jednom shluku a shluky jsou postupně děleny do menších a menších shluků.

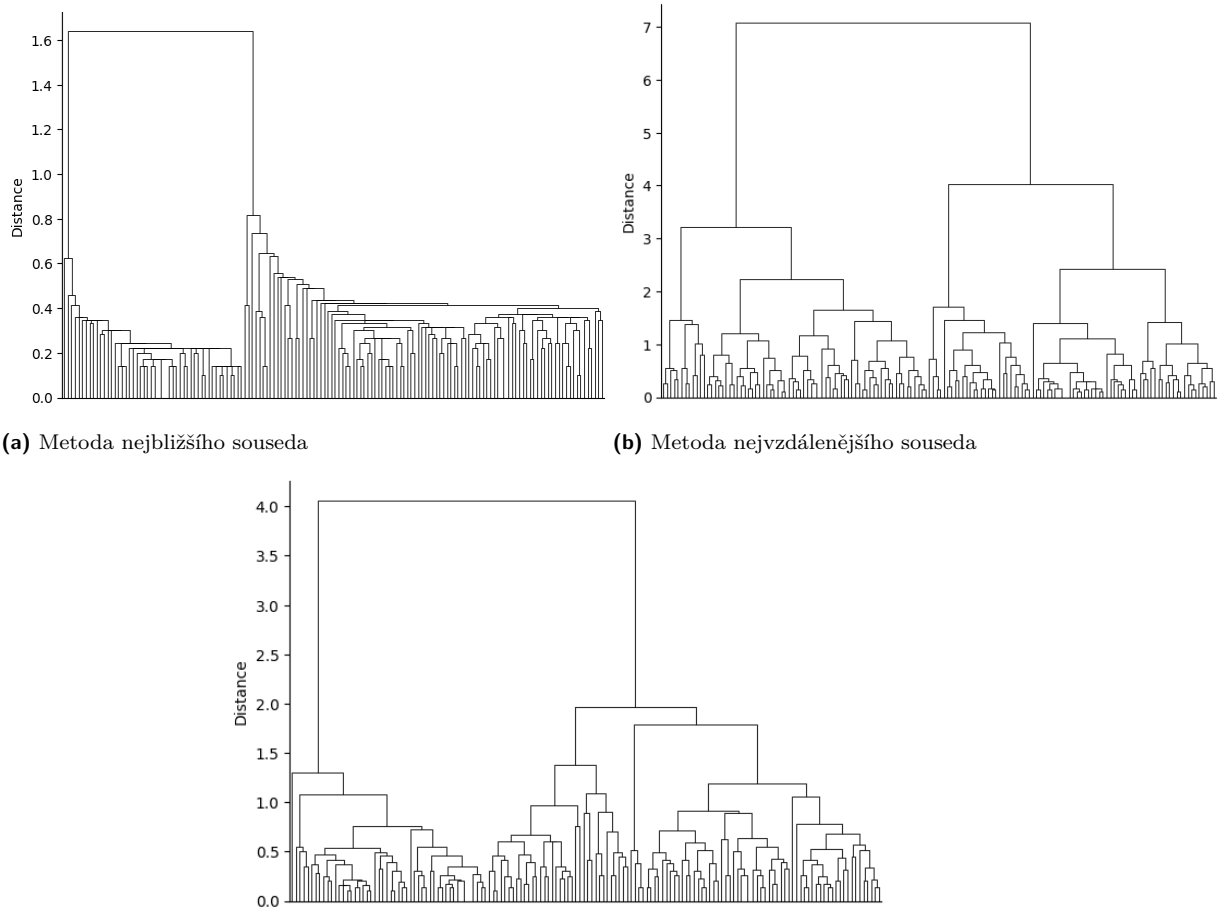
Kritériem pro sloučení (resp. rozdělení) shluků je nějaká míra odlišnosti. Autoři studie [26] pak uvádí, že většina hierarchických algoritmů používá některou z následujících měr[14]:

- Metoda nejbližšího souseda [15] – vzdálenost mezi shluky je rovna nejkratší vzdálenosti mezi dvěma body z rozdílných shluků.
- Metoda nejvzdálenějšího souseda [16] – analogicky je vzdálenost mezi dvěma shluky definována jako vzdálenost mezi dvěma nejodlehlejšími body ve dvou rozdílných shlucích.
- Metoda minimálního rozptylu [17] – Vzdálenost mezi dvěma shluky je definována jako průměr vzdáleností mezi všemi body z rozdílných shluků.

Nevýhodou metody nejbližšího souseda je tzv. „efekt zřetězení“, kdy existuje několik jednotlivých bodů ležících blízko u sebe, které propojí jinak vzdálené větší shluky. Nevýhodou metody minimálního rozptylu pak je, že může dojít k rozdělení podlouhlých shluků a ke sloučení částí sousedních podlouhlých shluků. [27, 23] Obecně lze říci, že při použití metody nejvzdálenějšího souseda vznikají kompaktnější shluky, zatímco při použití metody nejbližšího souseda vznikají shluky podlouhlé. Metoda nejvzdálenějšího souseda je tedy z těchto dvou metod obecně užitečnější. [26]

Výsledkem hierarchického shlukování je dendrogram. Dendrogram je stromový diagram, který zobrazuje postupné slučování (resp. rozdělování) shluků a hodnoty míry odlišnosti, při kterých sloučení (resp. rozdělení) nastalo (viz obrázek 1.6). [23, 28]

[14] popř. jejich varianty
[15] Single-link clustering
[16] Complete-link clustering
[17] Average-link clustering



(a) Metoda nejbližšího souseda

(b) Metoda nejvzdálenějšího souseda

(c) Metoda minimálního rozptylu

Obrázek 1.6 Dendrogramy pro hierarchická shlukování na Iris datasetu s využitím různým podobnostních metrik

1.2.4 Shlukování založené na hustotě

Metody shlukování založené na hustotě uvažují jako shluky oblasti s vysokou hustotou vzešlé z neznámého sdruženého rozdělení $p(x)$, které jsou od ostatních shluků odděleny oblastmi s řidším výskytem bodů. Tyto metody nepotřebují vstupní parametry jako počet shluků, nesnaží se zkoumat toto rozdělení, ani rozptyl shluků, které se v datech případně vyskytují. Shluky takto vzniklé nemusí být nutně skupiny bodů s nízkou vzájemnou nepodobností, tím pádem nemusí mít kulovitý [18] tvar, ale mohou nabývat v podstatě libovolné podoby. [14] Tyto metody uvažují speciální kategorii odlehlých bodů, které si nejsou podobné s žádnými jinými body a nehodí se tedy do žádného shluku. Takové body jsou běžně označovány jako šum. [29]

1.2.4.1 DBSCAN

Mějme vzdálenost r a minimální počet bodů k. Hustota bodu x_i je v tomto případě definována jako počet bodů k_i, které jsou od bodu vzdáleny méně než r. Pokud je $k_i > k$, bod x_i je označován jako *základní bod*[19]. Bod p je *přímo dosažitelný* z bodu q, pokud vzdálenost bodů je menší, než r a q je základní bod[20]. O bodu p pak řekneme, že je dosažitelný z bodu q, pokud existuje taková posloupnost bodů $p_1, \ldots, p_n$, kde $p_1 = q$, $p_n = p$ a platí, že každý bod p_{i+1} je přímo dosažitelný z bodu p_i. Poslední potřebnou definicí je spojitost. Dva body p a q jsou spojité, pokud existuje bod o, ze kterého jsou oba body dosažitelné. Shluky jsou pak maximální možné soubory spojitých bodů a jako šum jsou uvažovány body, které do žádného sluku nepatří. [29, 14]

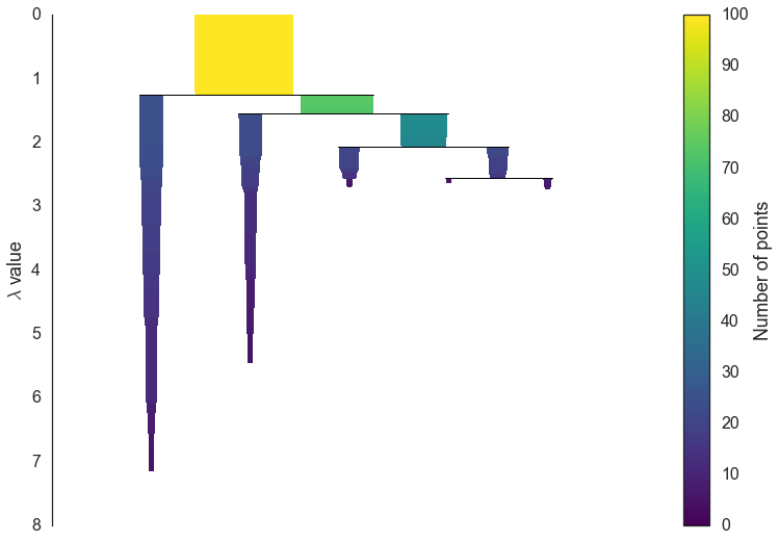
1.2.4.2 HDBSCAN

Tento algoritmus reaguje na některé nedostatky DBSCANu, zejména fakt, že DBSCAN pracuje se stejnou hustotou pro celý dataset a může tedy mít problém, pokud se v datasetu vyskytují shluky s různou hustotou, případně shluky vnořené. [30] HDBSCAN používá jako základ algoritmus DBSCAN*, který je původnímu algoritmu velmi podobný – liší se tím, že jako body do shluku patřící uvažuje pouze základní body, tedy pokud je bod dosažitelný a není základní, považuje ho DBSCAN* za šum. Dále nově definuje *základní vzdálenost* bodu p ($d_{core}(p)$) jako vzdálenost ke k-tému nejbližšímu bodu a vzájemně dosažitelnou vzdálenost bodů p a q jako $max\{d_{core}(p), d_{core}(q), d(p,q)\}$. Body je možné chápat jako vrcholy grafu G a vzájemně dosažitelné vzdálenosti bodů jako vážené hrany. Následně je na tomto grafu pomocí Jarníkova algoritmu vybudována kostra. Z této kostry jsou pak postupně odebírány hrany. Pokud při odebrání hrany vznikne spojitá komponenta o více než k bodech, je označena jako nově vzniklý shluk, v opačném případě se jedná o body, které se snižováním hustoty stávají šumem. Výsledek může být zobrazen jako zhuštěný dendrogram (viz 1.7), kde jsou zobrazeny pouze nově vzniklé shluky. Pro volbu shluků je používána převrácená vzdálenost $\lambda = \frac{1}{vzdalenost}$, kdy pro každý shluk platí že při určité hodnotě λ vzniká (λ_v) a při jiné hodnotě zaniká (λ_z), což platí i pro jednotlivé body p shluku, kde λ_p, je rovno hodnotě λ při kterém byl bod ze shluku vyřazen. V tu chvíli je možné definovat *stabilitu* shluku C jako

$$\sum_{p \in C} (\lambda_p - \lambda_v).$$

Zhuštěný dendrogram je poté procházen od nejmenších shluků ke kořeni, kde v každé úrovni je spočtena stabilita shluků a jejich rodičů. Pokud je součet stability potomků větší, než stabilita rodiče, přebírá tuto stabilitu rodič. V opačném případě je rodič označen jako finální shluk a všem jeho potomkům je toto označení odebráno. [31, 30]

[18] nebo obecně konvexní
[19] core point
[20] Žádný bod nemůže být přímo dosažitelný z bodu, který není základní
[21] postupném odebírání nejtěžších hran z kostry



Obrázek 1.7 Zhuštěný dendrogram, kde je zobrazeno štěpení shluků při postupném zvyšování hustoty[21]potřebné k utvoření shluku. Místo vzdálenosti je jako metrika použita její převrácená hodnota $\lambda = \frac{1}{vzdalenost}$. Počet bodů ve shluku je znázorněn pomocí odstínu a šířky shluku.

Kapitola 2
Experimenty a diskuze

V této kapitole bude představen použitý dataset. Dále proběhnou experimenty s výše popsanými metodami extrakce témat včetně vizualizací jejich průběhu a výsledků.

2.1 Průzkum a předzpracování dat

V této práci je používán dataset [32], který obsahuje 1305 knih básní z Korpusu českého verše. Korpus českého verše je lemmatizovaný, foneticky, morfologicky, metricky a stroficky anotovaný korpus české poezie 19. a počátku 20. století[33]. Každá kniha je dostupná jako json soubor, ve kterém je se kromě básní samotných vyskytuje veliké množství metadat (např. informace o autorovi, knize, verších, ...). Pro tuto práci je podstatná část, která pro každou báseň obsahuje lemmata jednotlivých slov. Lemmata bylo třeba z datasetu extrahovat a sestavit z nich jednotlivé básně a dále odstranit slova bez zvláštního významu (viz obrázek 2.1)[1].

2.2 LDA

K práci s metodou LDA byla využita její paralelizovaná implementace z knihovny gensim [34]. Extrakce byla provedena pro různé počty požadovaných témat (viz tabulka 2.1). Vizualizace modelu (viz obrázek 2.2) je vytvořena nástrojem LDAvis [35], který se skládá ze dvou částí. První část (obrázek 2.2a) zobrazuje dvoudimenzionální reprezentaci témat jako kruhy. Poloha jejich center odpovídá vzdálenosti mezi tématy. Jejich velikost se odvíjí od výskytu tématu v korpusu. Druhá část (obrázek 2.2b) zobrazuje nejdůležitější termy pro vybrané téma. Modré sloupce označují celkovou frekvenci termu a červené očekávanou frekvenci ve zvoleném tématu [35]. K vizualizaci byla použita implementace z knihovny pyLDAvis [2].

2.3 BERTopic

Již zmíněná modularita této metody umožňuje v každé její části volit odlišné přístupy k jejímu řešení. Experimentováno bylo s různými způsoby dimenzionální redukce vektorizovaných dokumentů (konkrétně pomocí metod PCA, t-SNE, UMAP a PaCMAP, viz obrázek 2.3) a s různými druhy shlukovacích algoritmů a jejich hyperparametry (K-means, K-medoids a HDBSCAN).

[1] Top2Vec je navržen tak, odstranění běžných slov nebylo potřeba, nicméně, jak bude dále ukázáno i v jeho případě je toto odstranění prospěšné.

[2] https://github.com/bmabey/pyLDAvis

Tvá loď jde po vysokém moři, v ně brázdu jako stříbro reje, svou přídu v modré vlny noří a bok svůj pěnné do peřeje.	loď vysoký moře brázda stříbro rej přijít modrý vlna nořit bok pěnné peřej
Tvá lana sviští, plachty duní a třepe vlajka. V noční chvíli zříš magický svit mořských tůní, a ve snu, Albatros jak pílí.	lano sviští plachta dunět třepat vlajka noční chvíle zřít magický svit mořský tůň sen albatros píle
Já samotním, jsem na ostrově, ohýnek topím, rybku lově zasedám na břeh za večera.	samotním ostrov ohýnek topit rybka lovit zasedat břeh
Dým v kotoučích se modrých krade, kdes písklo ptáče, ještě mladé, tma na mne hrozí z pološera.	dým kotouč modrý krást kdes písknout ptáče mladý tma hrozit pološero

(a) Původní text básně sestavený z tokenů

(b) Lemmatizovaná báseň po odstranění slov bez zvláštního významu

Obrázek 2.1 Ukázka zpracování první básně z datasetu.

Veškeré shlukování probíhalo v prostorech redukovaných na pět dimenzí. Experimenty proběhly v několika fázích a s metodami s nejhoršími výsledky nebylo experimentováno dále.

V první fázi bylo cílem vybrat nejlepší metodu dimenzionální redukce.

Pro K-means a K-medoids je nutné zvolit požadovaný počet shluků. V prvním kole testů byla pro tento účel použita metoda *silhouette score* (viz obrázek 2.4) a na základě jejích výsledků byl určen jako $n = 118$.

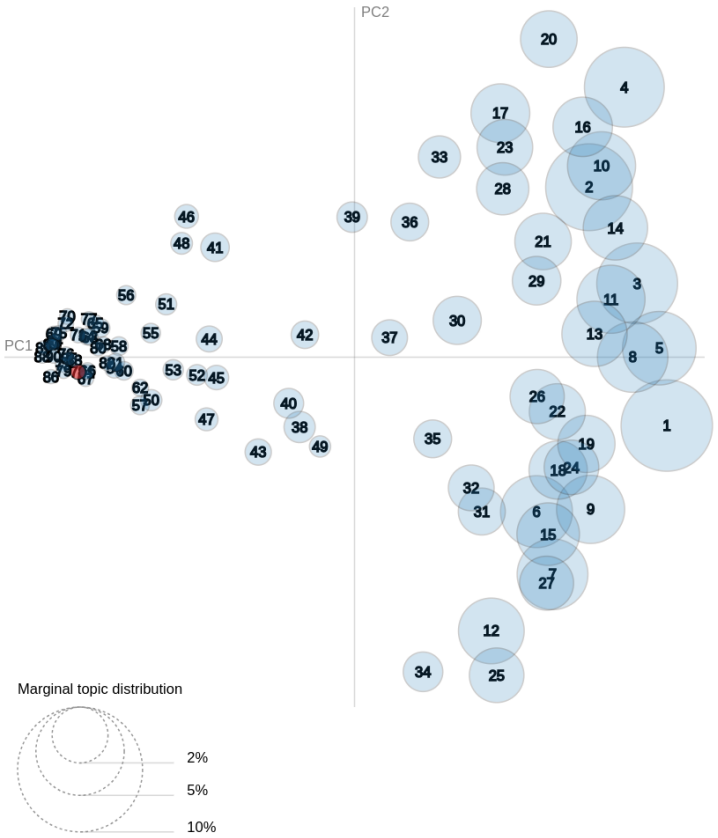
Koherence modelu byla pro oba shlukovací algoritmy nejvyšší, pokud byl pro redukci shlukovaných dat použit UMAP. Jak lze ovšem pozorovat na obrázcích 2.5 a 2.6, výsledné shluky se velmi liší.

Zásadním rozdílem je počet, HDBSCAN jich identifikoval téměř 300 oproti již dříve zvolenému číslu 118 u K-means. Na histogramech na obrázcích 2.5b a 2.6b pak lze vidět, že výrazně se liší i počty dokumentů v jednotlivých shlucích. V případě K-means počet klesá téměř lineárně, zatímco u HDBSCANu většina shluků obsahuje jen několik dokumentů. Vizualizace největších 100 shluků (obrázky 2.5a a 2.6a) jsou pak téměř komplementární, V případě K-means se největší shluky tvořily ve středu, u HDBSCANu pak po okrajích a dokumenty ve středu vizualizace byly naopak často označeny jako mimolehlé a do žádného shluku nebyly zařazeny.

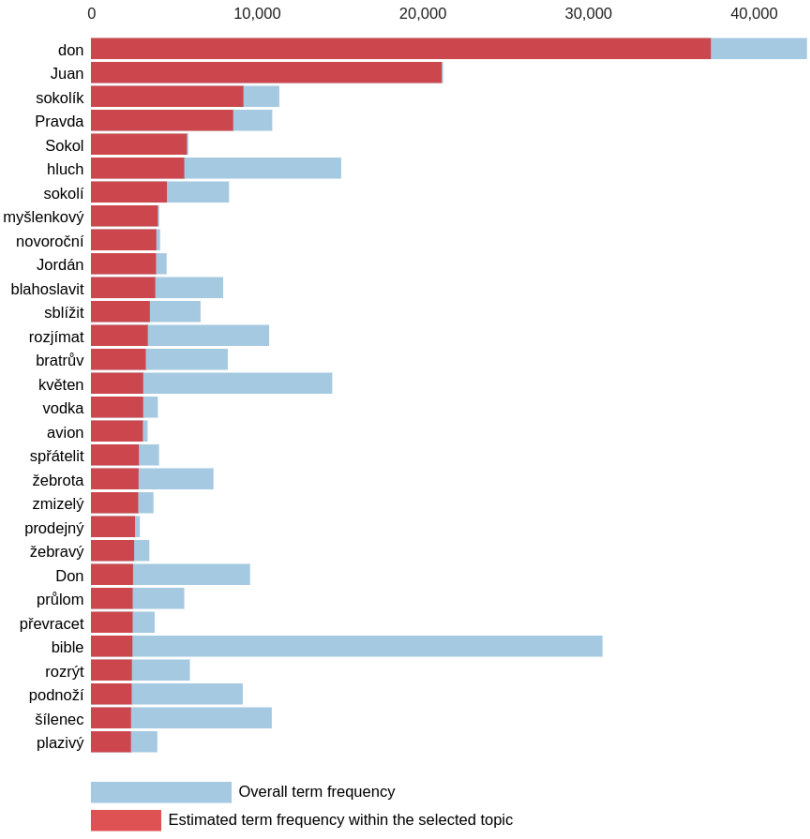
2.3.0.1 HDBSCAN

Další testování spočívalo ve volbě různé minimální velikosti shluků. Na základě výsledků z prvního kola byla využita pouze redukce vytvořená metodou UMAP. Výsledek (viz obrázek 2.7) ukázal, že nejvyšší koherence dosáhl model ve chvíli, kdy byl tento hyperparametr nastaven na hodnotu 5. Zároveň však oproti počáteční fázi[3] obsahuje mnohem více témat (téměř tisíc), kdy každé z nich obsahuje malé množství dokumentů. Otázkou tedy zůstává, zda zlepšení skóre v tomto případě indikuje to, že model lépe vystihl podstatu dat a tato se změna se na tématech a jejich názvech pozitivně projeví.

[3] Minimální počet dokumentů ve shluku byl nastaven na 5.

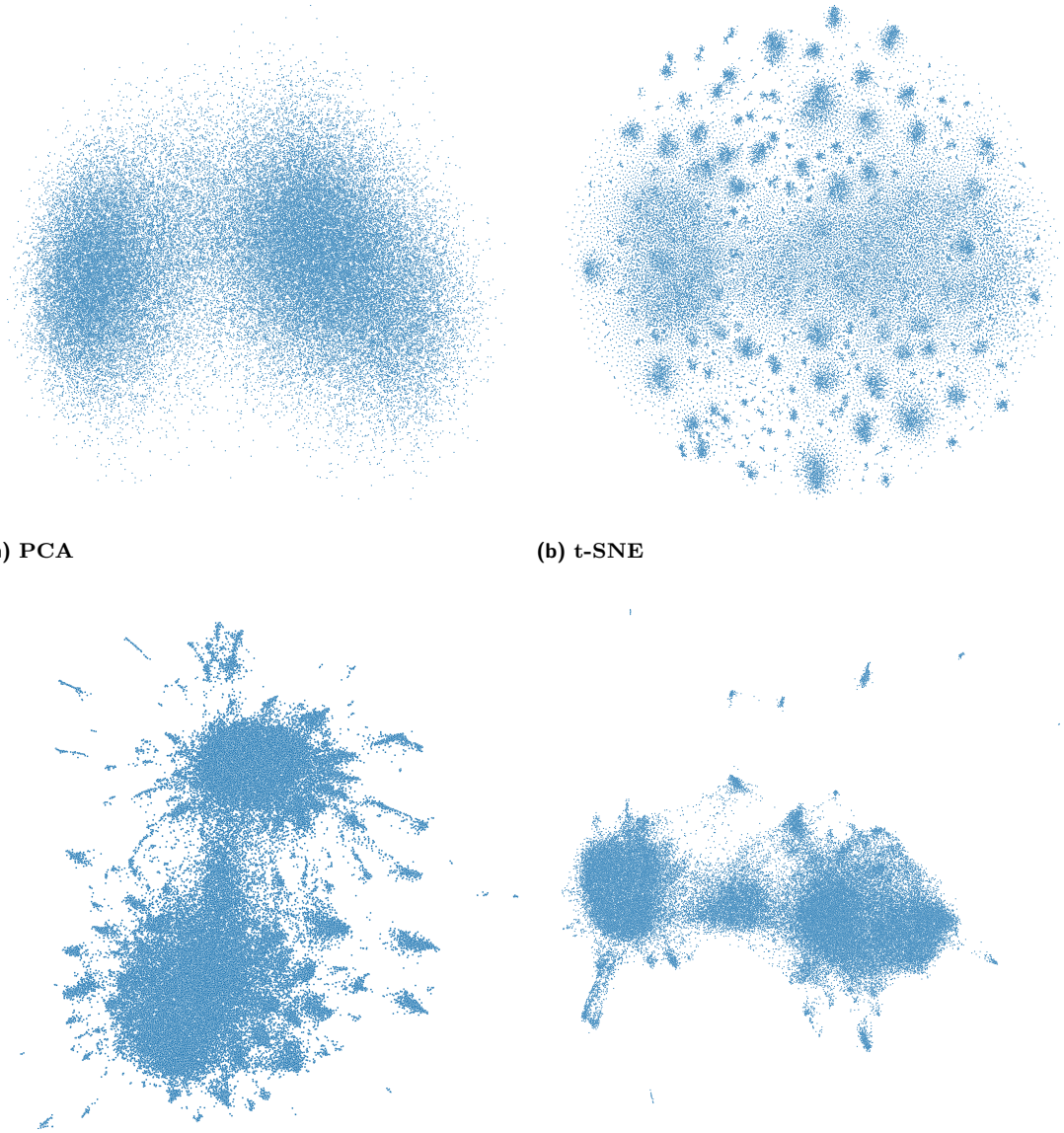


(a)



(b)

Obrázek 2.2 Vizualizace LDA pomocí pyLDAvis



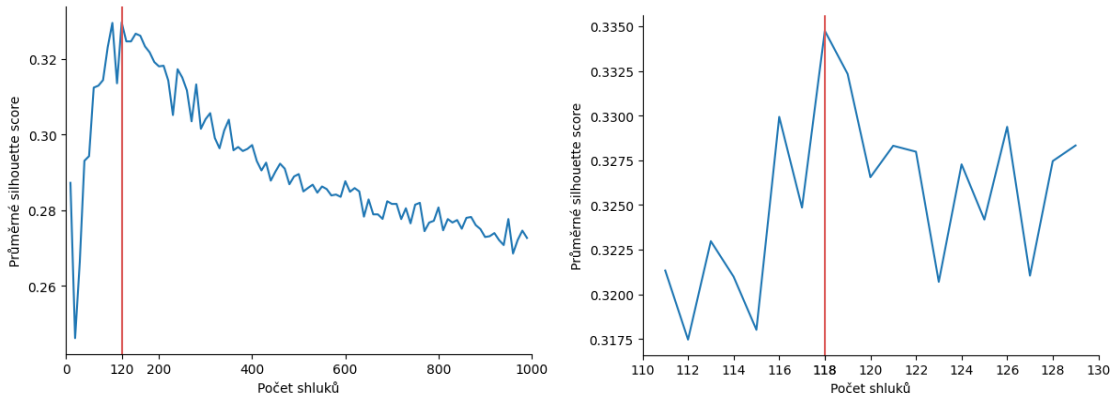
(a) PCA **(b) t-SNE**

(c) UMAP **(d) PaCMAP**

Obrázek 2.3 Zobrazení vektorizace dokumentů modelu BERT pomocí různých metod dimenzionální redukce.

počet témat	c_v koherence
10	0.386
30	0.435
50	0.451
70	0.438
90	**0.455**
110	0.434
130	0.444
150	0.432
170	0.432
190	0.431

Tabulka 2.1 C_v koherence modelu LDA pro různé počty témat.



(a) Průměrné silhouette score shluků pro $k = 10n$ kde $n \in \mathbb{N}$ a $k \in \langle 10, 1000)$.

(b) Průměrné silhouette score shluků pro k kde $k \in \mathbb{N}$ a $k \in (110, 130)$

Obrázek 2.4 Průměrné silhouette score shluků vytvořených algoritmem K-means pro jednotlivá k.

2.4 Top2Vec

Při testování extrakce témat pomocí metody top2vec byla využita implementace[4] od autora původní studie [11]. Extrakce probíhala se stejnými hyperparametry, jako v původní studii[5]. Z výsledků (viz obrázek 2.8) lze vidět, že počet dokumentů označených jako mimolehlé, je vyšší než u všech ostatních shlukovacích metod (přes 50 000). HDBSCAN v tomto případě identifikoval přes 200 shluků a opět je vidět, že většina shluků je malých a v modelu se vyskytuje několik velkých shluků obsahujících podstatnou část dokumentů. Oproti vizualizacím shlukování u BERTopicu popisky na obrázku 2.8a neoznačují názvy shluků, ale jedná se o jednotlivá slova, která leží nejblíže centrům shluků. C_v Koherence modelu je 0.425, nicméně autor argumentuje [36], že tato metoda není pro měření kvality tohoto modelu vhodná.

Výhodu popisovanou autorem – možnost neodstraňovat slova bez zvláštního významu se v tomto případě nepodařilo ověřit. Model se na datasetu s těmito slovy nepodařilo natrénovat, všechny dokumenty až na jeden byly označeny jako mimolehlé. Důvod tohoto jevu nebyl v rámci této práce objasněn. Po jejich odstranění proběhl trénink normálně.

[4] https://github.com/ddangelov/Top2Vec

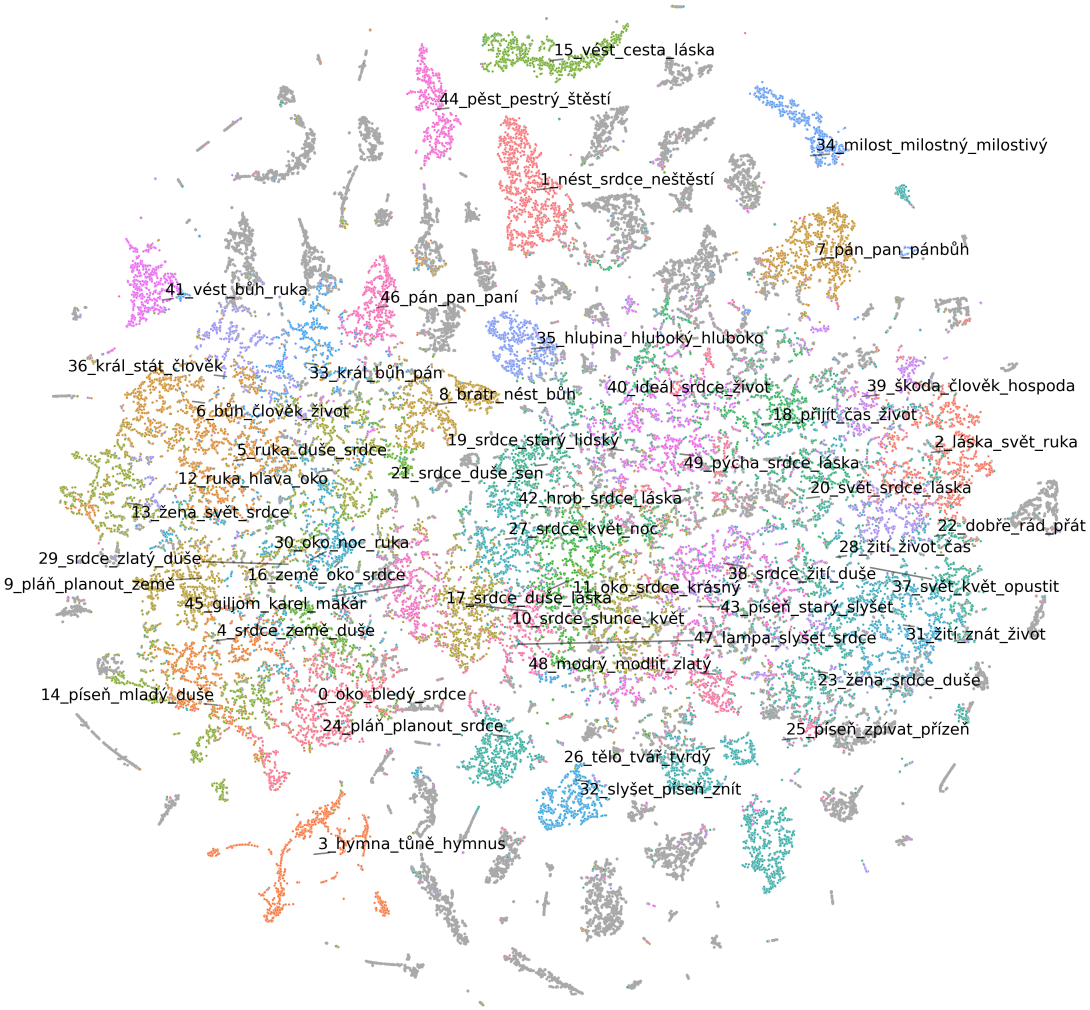
[5] Pro redukci byl využit UMAP, pro shlukování pak HDBSCAN s nejnižším počtem dokumentů ve shluku nastaveným na 10 a metrikou podobnosti pro shlukování byl kosinus úhlů vektorových reprezentací dokumentů (*cosine similarity*)

shlukovací algoritmus	metoda dim. redukce	c_v koherence
HDBSCAN	PCA	0.337
HDBSCAN	t-SNE	0.396
HDBSCAN	UMAP	**0.486**
HDBSCAN	PaCMAP	0.427
K-means 96	PCA	0.385
K-means 96	t-SNE	0.389
K-means 96	UMAP	**0.413**
K-means 96	PaCMAP	0.407

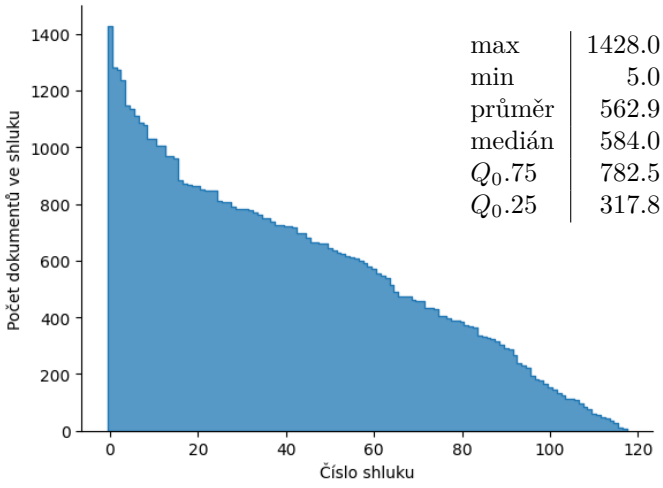
Tabulka 2.2 Výsledky první fáze experimentů s rozdílnými způsoby dimenzionální redukce.

2.5 Porovnání témat mezi jednotlivými metodami extrakce

Pro porovnání kvality extrakce témat byly z datasetu náhodně zvoleny 3 básně, které ani jeden z algoritmů neoznačil jako mimolehlé. První dvě navíc splňovaly podmínku, že se musí nacházet v jednom z padesáti největších shluků pro jednotlivé modely (tzn. typicky se jedná o dokumenty pocházejících z větších shluků), tedy lze očekávat, že popis vybraných témat bude obecnější. Třetí báseň je pak volena tak aby neležela v padesáti největších shlucích a bylo možné pozorovat, jak proběhne extrakce u menších shluků. Na obrázcích (2.9, 2.10 a 2.11) je kromě přiřazení témat také celý text básně a její wordcloud. V tématech extrahovaných BERTopicem je zejména v prvním případě (viz tabulka 2.9a) vidět dříve zmíněná nevýhoda – v názvu tématu se opakují velmi podobná slova (hlubina–hluboký–hluboko). Rozdíly ve výběru shlukovacího algoritmu se u BERTopicu na názvu tématu projevily, pokud byl dokument zařazen do menšího shluku (viz tabulka 2.11a). Kvalita extrahovaných témat na těchto příkladech se nejeví jako příliš dobrá. S básněmi mají názvy povětšinou společné jen některá ze slov, které v prvních dvou případech báseň příliš nevystihují. Ve třetím případě jsou názvy témat modelů top2vec a BERTopic lepší, nicméně u BERTopicu pouze v případě že shlukování probíhalo pomocí HDBSCANu.



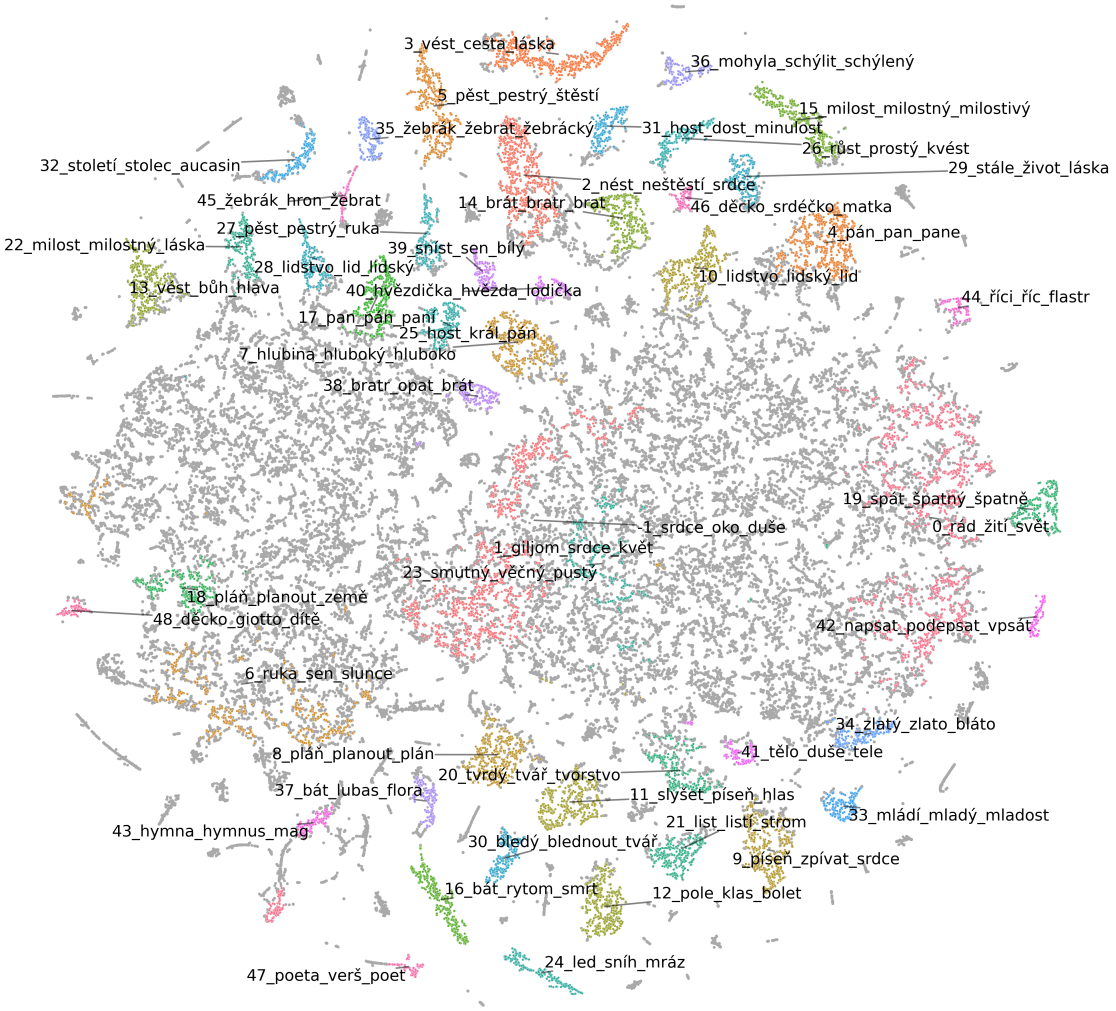
(a) Zobrazení shluků vytvořených algoritmem k-means pro $k = 118$ pomocí t-SNE společně s názvy témat. Zvýrazněno je 50 největších shluků.



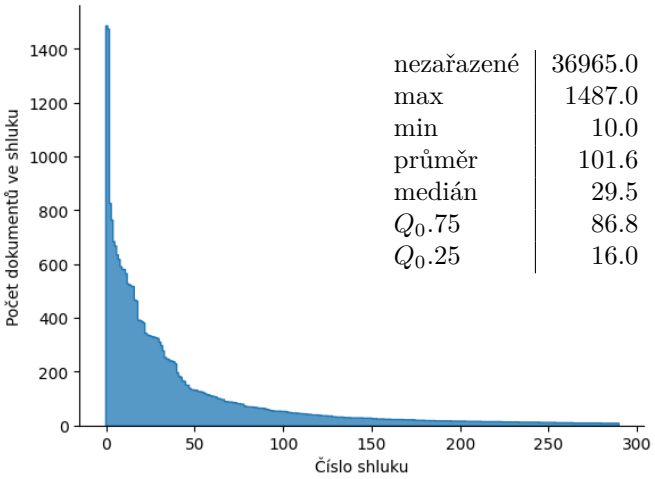
max	1428.0
min	5.0
průměr	562.9
medián	584.0
$Q_0.75$	782.5
$Q_0.25$	317.8

(b) Histogram pro počet dokumentů v jednotlivých shlucích spolu s tabulkou se základními statistikami o počtu dokumentů ve shlucích.

Obrázek 2.5 Vizualizace vytvořených shluků pro společně se základními statistikami o počtu dokumentů v jednotlivých shlucích.



(a) Zobrazení shluků vytvořených algoritmem HDBSCAN pomocí t-SNE společně s názvy témat. Zvýrazněno je 50 největších shluků.



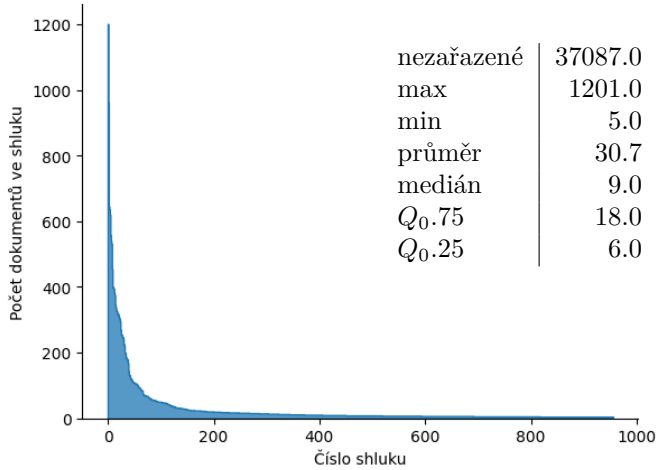
nezařazené	36965.0
max	1487.0
min	10.0
průměr	101.6
medián	29.5
$Q_0.75$	86.8
$Q_0.25$	16.0

(b) Histogram pro počet dokumentů v jednotlivých shlucích spolu s tabulkou se základními statistikami o počtu dokumentů ve shlucích.

Obrázek 2.6 Vizualizace vytvořených shluků pro společně se základními statistikami o počtu dokumentů v jednotlivých shlucích.

shlukovací algoritmus	min. velikost shluku.	c_v koherence
HDBSCAN	5	**0.544**
HDBSCAN	10	0.445
HDBSCAN	15	0.434
HDBSCAN	20	0.433
HDBSCAN	30	0.429
HDBSCAN	50	0.429
HDBSCAN	100	0.417

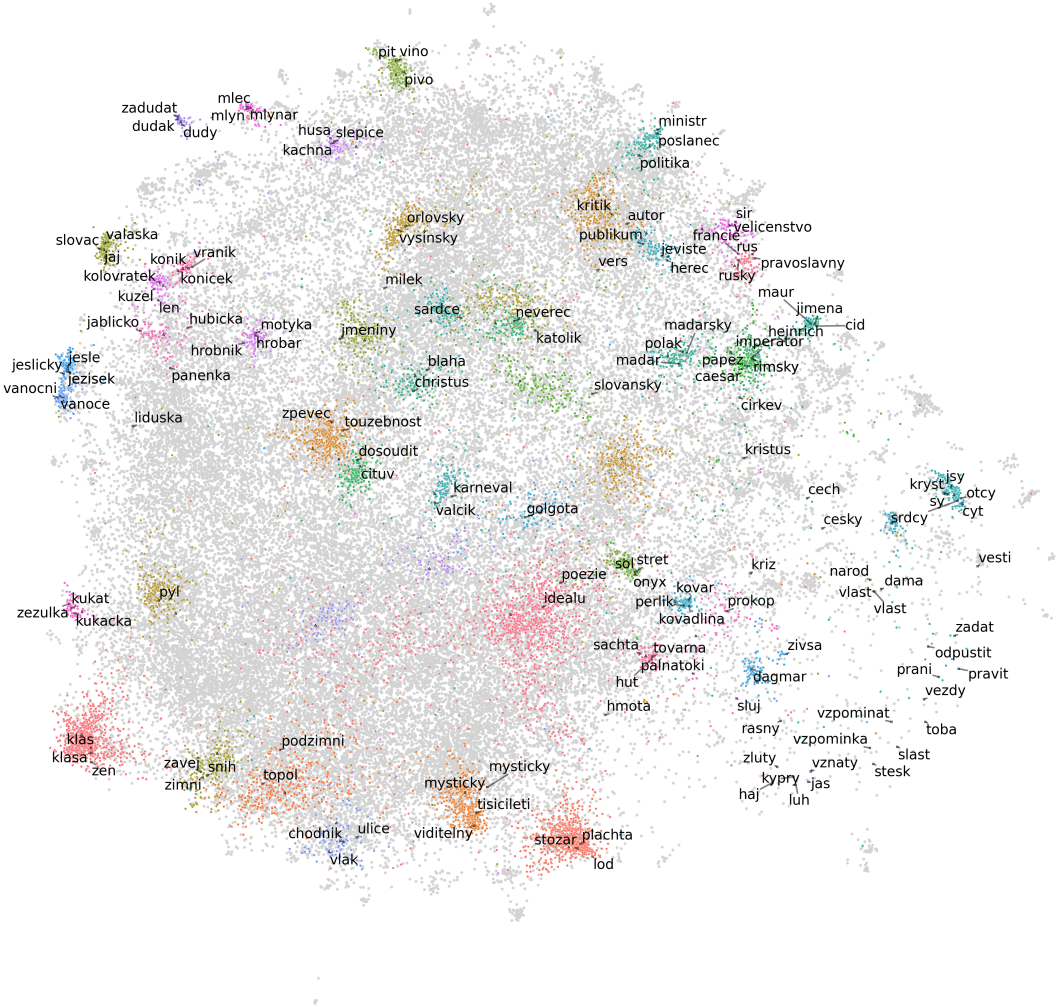
(a) Měřená c_v koherence modelu BERTopic pro shlukování s algoritmem HDBSCAN a rozdílnými minimálními velikostmi shluku.



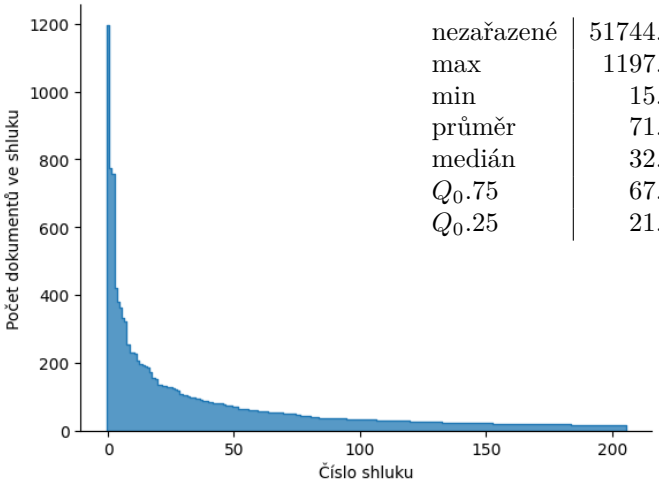
nezařazené	37087.0
max	1201.0
min	5.0
průměr	30.7
medián	9.0
$Q_0.75$	18.0
$Q_0.25$	6.0

(b) Histogram pro počet dokumentů v jednotlivých shlucích spolu s tabulkou se základními statistikami o počtu dokumentů ve shlucích.

Obrázek 2.7 Výsledky druhé fáze experimentů se shlukováním pomocí HDBSCAN pro různou minimální velikost shluků.



(a) Zobrazení shluků vytvořených algoritmem HDBSCAN. Shlukování probíhalo v pětidimenzionální redukci původních dat získaného pomocí metody UMAP, zobrazení je pak vytvořeno z této redukce pomocí t-SNE. Zvýrazněno je 50 největších shluků.



nezařazené	51744.0
max	1197.0
min	15.0
průměr	71.3
medián	32.5
$Q_0.75$	67.5
$Q_0.25$	21.0

(b) Histogram pro počet dokumentů v jednotlivých shlucích spolu s tabulkou se základními statistikami o počtu dokumentů ve shlucích.

Obrázek 2.8 Top2Vec: Vizualizace vytvořených shluků pro společně se základními statistikami o počtu dokumentů v jednotlivých shlucích.

Neznámou řečí, vím to, zpěv můj zní ti,
však oka tvého svit mi dobře známý,
tak v šero lesů mezi ruinami,
na severu v mé vlasti měsíc svítí.

Ó nech mne na tvých plných ňadrech sníti!
Jest sladko plout nad vášní hlubinami,
když víme, že jest lásky anděl s námi,
jenž v srázném pádu na křídlo nás chytí.

Nechť klesneme! Tvých očí ve azuru
můj duch si zdřimne zatím na tvé hrudi,
jak snící kondor na křídlech své touhy,

Až láskou naší pohnut Bůh se vzbudí,
své hvězdy strhá a z nich řetěz dlouhý
si uplete a vytáhne nás vzhůru!

model	téma
B K-means 118	hlubina_hluboký_hluboko_holubice_holub_duše_srdce_moře_holubička_les
B K-medoids 118	hlubina_hluboký_hluboko_holubice_holub_duše_moře_srdce_láska_holubička
B hdbscan 10	hlubina_hluboký_hluboko_holubice_holub_holubička_moře_duše_srdce_chlubit
B hdbscan 5	hlubina_hluboký_hluboko_holub_holubice_holubička_chlubit_dno_moře_holubičí
top2vec	leknin_hladina_vodni_rakosi_tune_rakos_tun_olse_siti_vlna
LDA 90	svět_duše_sen_život_žití_duch_krása_zřít_vědět_žít

(a) Extrahovaná témata pro jednotlivé modely.

Obrázek 2.9 První báseň pro porovnání extrakce témat.

Jdu plání zavlhlou... Jak ruka roztažená
do dáli přede mnou se táhne plocha polí,
a nad ní obloha se klene zachmuřená –
jen stromů několik tu krášlí úhor holý.

Když všecky dokola zde stromy pokáceli,
když seker železo se zarývalo všade
do mízy pojarní, – na ty jen zapomněli,
by mohly pozdravit svým šumem Jaro mladé.

V snech kdysi pohroužen já cítil, jak se žene
v les prvých na-
dějí bouř' mroucí v srdce kraji...
Vše nyní ztišeno... Teď v duši zachmuřené
snů stromy poslední se líně kolébají...

model	téma
B K-means 118	pláň_planout_srdce_plán_květ_planý_hvězda_nebe_slunce_duše
B K-medoids	pláň_planout_plán_srdce_květ_hvězda_planý_slunce_oko_duše
B hdbscan 10	pláň_planout_plán_planý_srdce_tma_hvězda_oko_světlo_nebe
B hdbscan 5	pláň_planout_plán_planý_provodiž_světlo_mrak_sníh_hvězda_tma
top2vec	podzimni_topol_zluty_alej_pesina_podzim_louka_chlad_zelen_mlcici
LDA 90	les_strom_tráva_pták_větev_mech_slunce_lesní_stráň_cesta

(a) Extrahovaná témata pro jednotlivé modely.

Obrázek 2.10 Druhá báseň pro porovnání extrakce témat.

Sedl orel na mečetu,
sivý orel Balkánu,
pod ním muezzíni světu,
hlásají zvěst koránu,
pod ním moslemín se béře,
k harému by touhy své,
vlíbal bujné do kadeře,
otrokyně tauridské.

Na mečetě na vysoké,
zdá se orlu starému,
o časech kdy smavooké,
prchnou děvy z harému,
o časech kdy umlkne tu,
tajemná zvěsť koránu,
a k hrdému peruť letu,
vznese orel Balkánu.

model	téma
B K-means 118	milost_láska_srdce_bůh_král_milostný_pius_stát_oko_svět
B K-medoids	srdce_ruka_hlava_svět_duše_láska_oko_stát_bůh_žena
B hdbscan 10	balkán_prochaska_turčín_korán_turek_puška_rája_marko_harém_pravoslavný
B hdbscan 5	balkán_prochaska_turčín_korán_rája_turek_marko_harém_urálu_pravoslavný
top2vec	turek_hajduk_marko_turcin_aga_junak_pasa_balkan_handzar_cernohorec
LDA 90	oko_krev_ruka_černý_tělo_tvář_oheň_plamen_hlava_žhavý

(a) Extrahovaná témata pro jednotlivé modely.

Obrázek 2.11 Třetí báseň pro porovnání extrakce témat.

Kapitola 3
Závěr

Cílem této práce bylo představit metody vizualizace shlukovacích algoritmů a prakticky demonstrovat jejich použití na Korpusu české poezie, nad kterým byla řešena úloha nesupervizovaného shlukování, včetně jejich porovnání, vyhodnocení výsledků a ucelené vizuální prezentace této úlohy. Tyto cíle byly převážně splněny. Určité nedostatky spočívaly v menším množství zkoumaných shlukovacích metod.

Bibliografie

1. BLEI, David M. Probabilistic topic models. *Communications of the ACM* [online]. 2012, roč. 55, č. 4, s. 77–84 [cit. 2023-03-15]. Dostupné z DOI: 10.1145/2133806.
2. MCAULIFFE, Jon; BLEI, David. Supervised topic models. *Advances in neural information processing systems* [online]. 2007, roč. 20 [cit. 2023-06-07]. Dostupné z: https://proceedings.neurips.cc/paper/2007/file/d56b9fc4b0f1be8871f5e1c40c0067e7-Paper.pdf.
3. BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* [online]. 2003, roč. 3, č. null, s. 993–1022 [cit. 2023-03-10]. ISSN 1532-4435. Dostupné z: https://dl.acm.org/doi/pdf/10.5555/944919.944937.
4. GROOTENDORST, Maarten. BERTopic: Neural topic modeling with a class-based TF-IDF procedure [online]. 2022 [cit. 2023-03-25]. Dostupné z DOI: 10.48550/arXiv.2203.05794.
5. DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* [online]. 2018 [cit. 2023-05-10]. Dostupné z: https://arxiv.org/pdf/1810.04805.pdf.
6. REIMERS, Nils; GUREVYCH, Iryna. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* [online]. 2019 [cit. 2023-05-10]. Dostupné z: https://arxiv.org/pdf/1908.10084.pdf.
7. AGGARWAL, Charu C; HINNEBURG, Alexander; KEIM, Daniel A. On the surprising behavior of distance metrics in high dimensional space. In: *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*. Springer, 2001, s. 420–434.
8. BEYER, Kevin; GOLDSTEIN, Jonathan; RAMAKRISHNAN, Raghu; SHAFT, Uri. When is "nearest neighbor" meaningful? In: *Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*. Springer, 1999, s. 217–235.
9. MCINNES, Leland; HEALY, John; MELVILLE, James. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. 2018.
10. JOACHIMS, Thorsten. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* 1996. Tech. zpr. Carnegie-mellon univ pittsburgh pa dept of computer science.
11. ANGELOV, Dimo. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470* [online]. 2020 [cit. 2023-06-05]. Dostupné z: https://arxiv.org/pdf/2008.09470.pdf.

12. LE, Quoc; MIKOLOV, Tomas. Distributed representations of sentences and documents. In: *International conference on machine learning* [online]. PMLR, 2014, s. 1188–1196 [cit. 2023-06-05]. Dostupné z: https://arxiv.org/pdf/1405.4053.pdf.

13. GUNOPULOS, Dimitrios. *Encyclopedia of Database Systems*. Clustering Overview and Applications [online]. Ed. LIU, LING; ÖZSU, M. TAMER. Boston, MA: Springer US, 2009 [cit. 2023-03-20]. ISBN 978-0-387-39940-9. Dostupné z DOI: 10.1007/978-0-387-39940-9_602.

14. KRIEGEL, Hans-Peter; KRÖGER, Peer; SANDER, Jörg; ZIMEK, Arthur. Density-based clustering. *Wiley interdisciplinary reviews: data mining and knowledge discovery* [online]. 2011, roč. 1, č. 3, s. 231–240 [cit. 2023-04-17]. Dostupné z: https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_605.

15. XU, Dongkuan; TIAN, Yingjie. A comprehensive survey of clustering algorithms. *Annals of Data Science* [online]. 2015, roč. 2, č. 2, s. 165–193 [cit. 2023-03-20]. Dostupné z DOI: 10.1007/s40745-015-0040-1.

16. HINNEBURG, Alexander. *Visualizing Clustering Results.* [online]. 2009. [cit. 2023-05-10]. Dostupné z: https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_617.

17. PEARSON, Karl. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*. 1901, roč. 2, č. 11, s. 559–572.

18. VAN DER MAATEN, Laurens; HINTON, Geoffrey. Visualizing data using t-SNE. *Journal of machine learning research*. 2008, roč. 9, č. 11.

19. WANG, Yingfan; HUANG, Haiyang; RUDIN, Cynthia; SHAPOSHNIK, Yaron. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. *The Journal of Machine Learning Research*. 2021, roč. 22, č. 1, s. 9129–9201.

20. INSELBERG, Alfred. Introduction. In: *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. New York, NY: Springer New York, 2009, s. 1–5. ISBN 978-0-387-68628-8. Dostupné z DOI: 10.1007/978-0-387-68628-8_1.

21. NANDA, Satyasai Jagannath; PANDA, Ganapati. A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm and Evolutionary computation* [online]. 2014, roč. 16, s. 1–18 [cit. 2023-04-15]. Dostupné z: https://www.sciencedirect.com/science/article/pii/S221065021300076X.

22. LEUNG, Yee; ZHANG, Jiang-She; XU, Zong-Ben. Clustering by scale-space filtering. *IEEE Transactions on pattern analysis and machine intelligence* [online]. 2000, roč. 22, č. 12, s. 1396–1410 [cit. 2023-04-15]. Dostupné z DOI: 10.1109/34.895974.

23. ROKACH, Lior; MAIMON, Oded. *Clustering methods.* [online]. 2005. [cit. 2023-04-15]. Dostupné z: https://www.researchgate.net/profile/Lior-Rokach/publication/226748490_Clustering_Methods/links/02e7e536dcb9b70ea3000000/Clustering-Methods.pdf.

24. OMRAN, Mahamed GH; ENGELBRECHT, Andries P; SALMAN, Ayed. An overview of clustering methods. *Intelligent Data Analysis* [online]. 2007, roč. 11, č. 6, s. 583–605 [cit. 2023-04-15]. Dostupné z: https://www.ire.pw.edu.pl/~arturp/Dydaktyka/PPO/pomoce/clustering.pdf.

25. SCHUBERT, Erich; ROUSSEEUW, Peter J. Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Information Systems* [online]. 2021, roč. 101, s. 101804 [cit. 2023-04-16]. ISSN 0306-4379. Dostupné z DOI: https://doi.org/10.1016/j.is.2021.101804.

26. JAIN, Anil K; MURTY, M Narasimha; FLYNN, Patrick J. Data clustering: a review. *ACM computing surveys (CSUR)* [online]. 1999, roč. 31, č. 3, s. 264–323 [cit. 2023-04-16]. Dostupné z: https://dl.acm.org/doi/pdf/10.1145/331499.331504.

27. GUHA, Sudipto; RASTOGI, Rajeev; SHIM, Kyuseok. CURE: An efficient clustering algorithm for large databases. *ACM Sigmod record* [online]. 1998, roč. 27, č. 2, s. 73–84 [cit. 2023-04-16]. Dostupné z: https://dl.acm.org/doi/pdf/10.1145/276305.276312.

28. EVERITT, Brian S. *Cambridge dictionary of statistics* [online]. Cambridge, England: Cambridge University Press, 1998 [cit. 2023-04-16]. ISBN 9780521593465. Dostupné z: https://archive.org/details/cambridgediction00ever_0/page/96/mode/2up.

29. ESTER, Martin; KRIEGEL, Hans-Peter; SANDER, Jörg; XU, Xiaowei et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *kdd* [online]. 1996, sv. 96, s. 226–231 [cit. 2023-05-20]. Č. 34. Dostupné z: http://www.cs.ecu.edu/~dingq/CSCI6905/readings/DBSCAN.pdf.

30. CAMPELLO, Ricardo JGB; MOULAVI, Davoud; SANDER, Jörg. Density-based clustering based on hierarchical density estimates. In: *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 17* [online]. Springer, 2013, s. 160–172 [cit. 2023-05-27]. Dostupné z: https://link.springer.com/chapter/10.1007/978-3-642-37456-2_14.

31. MCINNES, Leland; HEALY, John; ASTELS, Steve. *How HDBSCAN Works* [online]. [cit. 2023-05-31]. Dostupné z: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html.

32. PLECHAC, Petr. *versotym/corpusCzechVerse 1.0* [online]. Zenodo, 2021. Ver. 1.0 [cit. 2023-04-14]. Dostupné z DOI: 10.5281/zenodo.4569929.

33. Korpus českého verše. In: *Versologický tým* [online]. [B.r.] [cit. 2023-04-14]. Dostupné z: https://versologie.cz/v2/web_content/corpus.php?lang=cs.

34. ŘEHŮŘEK, Radim; SOJKA, Petr. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* [online]. Valletta, Malta: ELRA, 2010, s. 45–50 [cit. 2023-06-25]. Dostupné z: http://is.muni.cz/publication/884893/en.

35. SIEVERT, Carson; SHIRLEY, Kenneth. LDAvis: A method for visualizing and interpreting topics. In: *Proceedings of the workshop on interactive language learning, visualization, and interfaces* [online]. 2014, s. 63–70 [cit. 2023-06-26]. Dostupné z: https://aclanthology.org/W14-3110.pdf.

36. ANGELOV, Dim. *Mutual Information · Issue 30 · ddangelov/Top2Vec — github.com* [online]. [B.r.]. [cit. 2023-06-25]. Dostupné z: https://github.com/ddangelov/Top2Vec/issues/30#issuecomment-695785517.