Sem vložte zadání Vaší práce.

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra softwarového inženýrství

Bakalářská práce

Systém pro analýzu sociálních sítí

Alexander Poddubny

Vedoucí práce: Ing. Tomáš Bartoň

24. června 2015

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 24. června 2015 .....

Odkaz na tuto práci

Poddubny, Alexander. *Systém pro analýzu sociálních sítí.* Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2015.

Abstrakt

V dané práci se pojednává o shromažďování a statistické analýze údajů uživatelů sociálních sítí. Celkem bylo shromážděno a zpracováno desítky tisíc profilů uživatelů. Informace o uživatelích obsahuje široké spektrum jak soukromých, tak společensky významných položek.

Klíčová slova Webová služba, analyza, klasterizace, segmentace, skupiny uživateli, socialní síťe, Socialbakers, vKontakte, Python, Django, Pandas, Numpy, D3js

Abstract

This thesis deals with the collection and statistical analysis of social networks users. There were collected and processed tens of thousand of users profile. Information about the user contains a broad spectrum of both private and socially significant items.

Keywords Web service, analysis, clusterization, segmentation, user's group, social network, Socialbakers, vKontakte, Python, Django, Pandas, Numpy, D3js

Obsah

Seznam obrázků

Úvod

Úvod

Popularita analýzy sociálních dat nabírá na obrátkách po celém světě díky vzniku v 90. letech 20. stol. online služeb sociálních sítí (LiveJournal, Facebook, Twitter, YouTube, aj.). S tím je spojen fenomén socializace osobních údajů: veřejně přístupnými se stala fakta z biografií, elektronická komunikace, deníky, foto, video a audio materiály, záznamy o cestování, apod. Sociální sítě jsou tedy unikátním zdrojem informací o osobním životě a zájmech reálných lidí. Otvírá to bezprecedentní možnosti pro řešení výzkumných a podnikatelských úkolů (některé z nich nebylo možné efektivně řešit kvůli nedostatku dat), a také pro vytvoření pomocných služeb a aplikací pro uživatele sociálních sítí. Toto, mimo jiné, podmiňuje zvýšený zájem o shromažďování a analýzu sociálních dat společnostmi a výzkumnými centry.

V dané práci se pojednává o shromažďování a statistické analýze údajů uživatelů sociálních sítí. Celkem bylo shromážděno a zpracováno desítky tisíc profilů uživatelů. Informace o uživatelích obsahuje široké spektrum jak soukromých, tak společensky významných položek: pohlaví, datum narození, dosažené vzdělání, politické názory, vztah k alkoholu a kouření, rodinný stav účastníků, jejich zájmy, seznam oblíbené hudby.

Hlavní cíle práce:

- Srovnaní podobných řešení pro sběr a analýzu dat ze sociálních sítí
- Implementace webové aplikace, která umožní:
  - Sběr dat z vybrané sociální sítě za použití jejího API a to i v libovolném množství
  - Jednoduchou analýzu dat za účelem segmentaci uživatelů.
  - Vizualizaci výsledku

Struktura práce

První kapitola se zabývá stručným popsáním kontextu zadaní. Uvádějí se druhy sociálních sítí, existující skupiny uživatelů a také jsou popsáné možnosti využití informaci shromážděných ze sociálních síti.

Druhá kapitola popisuje populární webové aplikací pro shromažďování a analýzu dat ze sociálních sítí.

Třetí kapitola je věnována definicím a specifikacím požadavků, modelování scénářů. Popisuje výběr programovacího jazyka a webového frameworku pro realizaci.

Čtvrtá kapitola popisuje návrh důležitých části webové aplikaci: sběr, ukládaní dat, jejich správa a analýza.

Pátá kapitola popisuje samotnou aplikaci. Uvádějí se rozpracované scénáře, analýza dat, vizualizace.

Kapitola 1

Kontext

1.1 Pojem sociální síť

Pojem sociální síť je původně sociologickým, nikoliv informatickým, pojmem. Jak říká Velký sociologický slovník[1], jde o strukturovaný uzel, který je tvořen jedincem, skupinou nebo organizací. Uzly se propojují vzájemnými vazbami, které jsou vzájemně závislé a reciproční, a nabývají různých podob. Jde především o sociální vazby ve společnosti: například, přátelství nebo rodinná příslušnost, společné zájmy, vzájemné negativní vztahy – averze, nenávist, antipatie, etc. Jde tedy o něco, co sdružuje osoby jakožto jednotlivce nebo skupiny – leitmotiv, pojící prvek. Zde v této práci se zabýváme sociálními síti, které jsou v síti Internet, a poukazujeme na to, jak se sociální sítě přenášejí a existují ve virtuálním životě. Hlavním účelem internetové sociální sítě jako webové služby, je sdružování jednotlivců – uživatelů za účelem výměny informací, komunikace mezi nimi; sdružování se děje pod nějakou určitou hlavičkou. Každý z těchto uživatelů má svojí určitou identitu, kterou je uživatelský profil (určité místo na serveru, vyhrazené pro konkrétní osobu a data jí sdílená), pod volitelným jménem, povinnými nebo nepovinnými položkami k vyplnění, atd. Sociální sítě nabývají různých podob – od úzkoprofilových, tedy těch, které sdružují lidi pod jasně daným motivem (hudební, školní), nebo všeobecné, které jsou členěné do tématických podoborů. Jak již bylo řečeno výše, smyslem a tedy i definičním znakem sociálních sítí je sdružení lidí do určitých tématických, zájmových spolků – komunit za určitým účelem. Uživatelé sociálních sítí, jež jsou na nich zaregistrovány pod určitými profily, sdílejí mezi sebou data v textovém a multimediálním formátu (i.e. fotografie, videa, jiný obsah, umožňují-li to možnosti dané platformy, na které se sociální síť nachází). Taková komunikace probíhá přímo mezi dvěma uživateli – jako soukromé zprávy, které se zobrazují příjemci a odesílateli, nebo jako hromadné zprávy, které se zobrazují osobám, se kterými má sdílející zprávu uživatel vazbu (rozumí se tím v rámci sítě Facebook například status), nebo uživatelům jedné skupiny, je-li tato zpráva nasdílena do této skupiny.

V současné době ve světe hraje prim sociální síť Facebook, především v Evropě, Severní a Jižní Amerikách, Austrálii a Africe a části Eurázie. V Rusku a státech bývalého SNS dominují jiné sociální sítě – vKontakte a Odnoklassniki (v překladu znamená Spolužáci, mohlo by to připomenout taktéž i českou obdobu Spoluzaci.cz, leč tato síť nehledě na podobný záměr, má jiný vzhled, podobu a interface). Čína má svojí sociální síť – QZone, Facebook zde má podstatně menší podíl, stejně jako v Rusku. Mohli bychom se zmiňovat o dalších sítích (dnes považovaných za skoro mrtvé, ale přesto sehravších značnou úlohu v jejich vývoji), jako např. MySpace, Digg apod., profesně zaměřených Last.Fm a jí obdobných, budeme se však zabývat především velkými a populárními hráči, jakožto Facebook a vKontakte.

Jak již bylo řečeno výše – základním účelem a základním definičním rysem sociálních sítí je tvorba vztahů mezi uživateli za účelem sdílení informací v jakékoliv podobě. Jejich cílem je vytvořit co největší počet vzájemných vztahů mezi uživateli, využívaných jak pro jejich pohodlí (komunikace), tak i pro komerční účely propagace, podpory prodeje a distribuce výrobků. Jak ukážeme dále, sociální sítě jsou také využívány převážně jako vhodný doplněk k e-shopům, nebo začínajícími výrobci, umělci, výtvarníky etc. K propagaci svého zboží nebo uměleckých děl. Web 2.0 posloužil mocnou platformou pro tvorbu aplikací, sociální sítě nevyjímaje. Sdílený obsah se nemoderuje – tedy uživatel může sdílet vše, co se mu zlíbí, nicméně není tomu tak zdaleka pravda. Problematika sdílení obsahu je především záležitostí legislativy státu, jehož je uživatel občanem, resp. kyberprostoru, spadajícího pod teritoriální působnost konkrétního státu, ale z důvodu fake-identit, tedy profilů se lživými nebo zkreslenými informacemi, platí také i podmínky chování uživatelů v rámci jedné sítě, které jsou dány sítí samotnou. Navíc, také platí i mezinárodní úmluvy na potírání kriminality a trestné činnosti, je-li uživatelem sdíleným obsahem spáchán určitý delikt nebo zločin (stalking[2], pomluva, šíření dětské pornografie, etc.)

1.2 Skupiny uživatelů

Uživatelé sociálních sítí se dělí do 2 základních skupin – komunitních a virtuálních. Komunitní skupiny sdružují osoby se stejným zájmem nebo stejnou myšlenkou. Řadí se k nim skupiny, představující například politické nebo zájmové skupiny, spolky nebo organizace spolužáků, studentů, zahrádkářů, etc. Virtuální skupiny se zakládají na tom, že uživatelé označují jako přátele osoby, které ani nemusí znát, ale jsou jim z určitého důvodu blízcí; samozřejmě, jsou tam také i osoby, které znají osobně. Základem každé skupiny je identifikace uživatele a již nastíněný problém pravosti informací, uváděných na profilech. Identifikace uživatele se děje často prostřednictvím volby profilové fotografie, uživatelského jména (pravého nebo pozměněného, jde-li o touhu nějak přesněji vyjádřit svojí spojitost s určitou skupinou), a informací o uživateli. Příkladem

mohou posloužit např. profily příslušníků komunity „pejskářů“ na sociálních sítích, které mají jasně dané definiční znaky, a podle sdíleného obsahu lze konkrétního uživatele k takové komunitě přiřadit. Jak ukážeme dále, i profily v síti Facebook mohou splňovat známky skupin virtuálních a komunitních, poněvadž i uživatel může také na svém profilu sdružovat osoby se stejným zájmem – a to i ty, které nezná osobně, ale se kterými má navázaný určitý vztah. Zde se ale zabýváme spíše skupinou jako určitým kroužkem, který je uzavřený dovnitř, tudíž s okolím příliš nekomunikuje. Pokud však ano, pak se to děje prostřednictvím tzv. styčného uživatele, který se účastní několika takových skupin, a šíří informaci nejen ze skupiny, ale také i do ni, a to i prostřednictvím externích zdrojů. Mnohdy se informace do skupiny dostávají díky sociálním vazbám osob, a to buď z reálného nebo virtuálního života, především z osobní komunikace. V jakékoliv skupině existují 3 skupiny uživatelů: uživatelé-tvůrci obsahu, uživatelé, sdílející – distribuující obsah a pozorovatele, tedy ty, kteří ničím nepřispívají a pouze čtou.

1.3 Možnosti využívání

Na sociálních sítích jsou 3 druhy reklamy:

1. Tzv. reklama z doporučení, i.e. reklama, šířená mezi samotnými uživateli. Jedná se o nejstarší a nejúčinnější druh reklamy, kdy uživatele si nejvíce dávají na doporučení svých známých nebo přátel, případně sami takové kontakty vyhledávají. Statisticky se jedná o nejaktivnější a finančně nejúspornější způsob. Doporučení se může dít různou formou, především však sdílením, komentováním a hodnocením, někdy uživatelé sami zakládají skupiny, ve kterých doporučují určité zboží nebo výrobky, diskutují o nich a projednávají nedostatky. Nicméně, podobná reklama na sociálních sítích může nést i určitá rizika v podobě konkurence, která se může vydávat za určitého uživatele, který má s daným výrobkem zkušenosti, a tak obelhavat nebo záměrně zkreslovat údaje o výrobku.

2. Bannery, kontextová reklama. Jejich výhoda spočívá především v cílení na konkrétní uživatele v souladu s jejich zájmy nebo statistickým vyhodnocením jejich zájmů, vyhledávaných výrobků, nebo z informací z propojených účtů. Hlavními kritérií jsou však především věk uživatele, pohlaví nebo vzdělání – tedy přesné cílení reklamy na konkrétní uživatelské skupiny. Děje se to tak, že společnost, která chce získat nové klienty, potřebuje zacílit reklamu na osoby z určitého regionu, věku nebo s určitými zájmy. Následně se zadává investovaná částka, která se platí za určité období nebo počet přechodů na propagovanou stránku, a proto tedy platí pravidlo – čím delší reklamní období a čím větší počet zobrazení, tím větší částka se musí investovat. Totéž platí o záběru auditoria, které má být reklamou osloveno. Dost často se reklama zařizuje

prostřednictvím společností, zabývajících se tzv. SEO – Sear Engine Optimization, které zařizují reklamu pro firmy. Tyto společnosti vytvářejí určitý content-plan, tedy plán obsahu, který se bude sdílet v kontextové reklamě, detailně propracovávají veškeré rizikové faktory a snaží se optimalizovat celý proces tak, aby uživatelské stránky odpovídali přesně vyhledávaným klíčovým slovům a cílili, nehledě na existující konkurenci, přesto přivést konečného zákazníka na stránku uživatele.

3. Firemní stránky – dosti často se tento pojem používá ve spojení „firemní Facebook“, které označuje stránku firmy v této síti. Jedná se o v této práci již zmiňovaný druh propagace, kdy firma nebo jednotlivec založí vlastní stránku v sociální síti a sdílí určitý obsah, textový nebo multumedální, a vytváří tak okruhy osob, které tento obsah sdílejí dál nebo komentují a vytvářejí tak další spirálu kolem něj. Často se to používá především u internetových obchodů, které informují své zákazníky o různých událostech v režimu online – naskladnění zboží, nové výrobky, apod. Velice častým a efektivním zdrojem získání nových zákazníků je soutěž, která může přivést nové fanoušky pro firmu a získat nejen hlasující, ale nové zákazníky, kteří nehledě na výsledek výhry, budou buď nakupovat, nebo si tuto stránku alespoň zapamatují. Pomáhá to především novým firmám nebo značkám, o kterých spousta lidí předtím neměla ani ponětí.

Tyto trendy jsou dány především popularitou sociálních sítí – téměř každý má účet na Facebooku a aplikaci ve svém chytrém telefonu, díky čemuž je neustále v kontaktu se svými přáteli. Reklama se zobrazuje v tzv. newsfeedu pořád, ať již v podobě sponzorovaných odkazů (tedy nejdražší reklamní místa, na která je dosažena především SEO-společnostmi). I přesto její výhodou zůstává relativně nízkonákladnou reklamou, nehledě na to, že náklady na ni se mohou pohybovat okolo desítek tisíc korun.

1.4 Zákaznický servis a průzkum trhu na sociálních sítích

Servis pro stávající zákazníky hraje dosti velkou roli v hospodaření jakékoliv firmy, jelikož je velice důelžité si zákazníka nejen získat, ale i udržet ho. Špatná úroveň zákaznického servisu je důvodem k odchodu zákazníka, proto firma si musí uvědomovat síly sociálních sítí, které mohou zákazníky nejen dávat, ale take I odebírat, a to především díky jejich počínání. Neřeší-li firma požadavky a reklamace zákazníků náležitým způsobem, rychle se to rozšíří mezi ostatní uživatele, které budou takovou firmu spíše nedoporučovat, což je poměrně vražedné pro firmu, která zatím nemá stálou zákaznickou základnu. Sociální site nabízejí výbornou možnost k inovacím, a to díky zpětné vazbě od zákazníků. Pozorování a zkoumání chování ostatních firem, které působí na též sociální

site, hodnocení faktorů jejich chování, míry vyřízování reklamací nebo řešení ostatních požadavků, hodnocení úrovně prodeje konkurence, apod. Průzkum trhu, týkající se především dalšího rozvoje firmy, nemusí být jenom pasivní, ale také i aktivní, kdy konkrétní firma udělá anketu s volnou moderací, kam zákazníci mohou přispívat a vyjadřovat své názory.
Nejtriviálnějším příkladem je anketa „jaký produkt byste uvítali v nabídce“ nebo „chystáme se stát partnery X, uvítali byste to?“. Druhý případ je uveden také proto, aby se ukázalo, že díky takovým průzkumům firma nemusí riskovat zbytečně a zavádět do sortimentu zboží, o které zákazníci by neměli zájem. Ankety se mohou týkat také i otázek, spojených s chodem firmy, což je pro zákazníka poměrně velkou výhodou: znamená to, že si ho firma váží natolik, že ho připouští do dění ve firmě.

Kapitola 2

Rešerše současných řešení

Pro analýzu sociálních sítí pro různé druhy výzkumů jsou dnes využívány hlavně online aplikace, proto v této rešerši budu zkoumat v první řadě aplikace webové. Aplikace, již budu vyvýjet, bude též zaměřená na online využití širokou veřejností. Pro analýzu sociálních sítí existuje jen málo aplikací, nejpopulárnějšími jsou řešení od Social Bakers, Simply Measured, Popster, Media-vk V následujících podkapitolách se pokusím stručně popsat zmíněné aplikace, vyjmenovat jejich hlavní charakteristiky.

2.1 Socialbakers

Socialbakers[1] je společností pro analýzu sociálních a veřejných médií, která provádí správu služeb sociálních médií a hlubokou analýzu dat pro tisíce značek obchodujících na Facebook, Twitter, Google+, LinkedIn, YouTube, Instagram, and VK. Analytici Socialbakers dodávají zákazníkům přehled o monitorování profilů sociálních médií na Facebooku, Twitteru, LinkedInu, VK, and YouTube (s přidáním reportu pro Instagram). Jejích webová applikace nabízí přehled ukazatelů pro měření:

- Polularita
- track key ovlivnění
- Analýza souvisejících hodnot a interakcí
- Měřítko výkonnosti ve srovnání s konkurencí a průmyslovými standardy,
- Optimalizace stávajících sociální médií a generace grafických reportů.

Socialbakers Builder je nástrojem pro plánování a zveřejňování obsahu, stejně jako pro sběr konverzací o značce v sociálních médiích. Vedle hlavního

[1] `http://www.socialbakers.com/`, Socialbakers

produktu Socialbakers nabízí hned několik služeb, včetně EdgeRank[3] Checker, algoritmu hloubkového výzkumu, který, mimojiné, říká zákaznikům jaký kontent pro jejich značku funguje nejlépe na stránkách Facebook. Měří úroveň zvýšení popularity stránek, dostupnost, obsah a souvislost trendů, dokonce i zjištění jak uživatelé obsah vašich stránek přidávají do oblíbených, komentují a sdílí. Velké značky mohou analyzovat souhrnný výkon až 150 populárních stránek v jednom reportu a podávat zprávu o všech vašich hlavních stránkách zároveň. Mezi bezplatné služby se řadí také bezplatné marketingové reporty, které se dělí na 2 kategorie – regionální a odvětvové. Tyto reporty obsahují údaje o 15 odvětvích a 150 státech světa. Odvětví, která jsou analyzována SocialBakers: Telekomunikace, Média, Jídlo a pití, Finance, Móda a krása, Elektronika, E-Commerce, Auta, Alkohol, Aerolinie.

Byly získány 2 bezplatné reporty: 1 regionální pro Českou republiku a 1 odvětvový v oblasti elektroniky. Reporty jsou datovány květnem 2015. Nejdříve se podíváme na jeden odvětvový report, jenž obsahuje data:

- nejvíce se vyvíjející brandy2.1
- top 10 brandů podle počtu fanoušků2.2
- aj.

V regionálním reportu byla obsažena následující informace:

- top 5 průmyslech podle počtu fanoušků2.3
- celková statistika2.4
- top 5 brandů v síti Facebook podle koeficientu zapojení2.5
- top 5 nejvíce sociálních brandů2.6

2.2 SimplyMeasured

Simply Measured[2]Měří úroveň zvýšení popularity vašich stránek, dostupnost, obsah a souvislost trendů, dokonce i zjištění jak uživatelé obsah vašich stránek přidávají do oblíbených, komentují a sdílí. Velké značky mohou analyzovat souhrnný výkon až 150 populárních stránek v jednom reportu a podávat zprávu o všech vašich hlavních stránkách zároveň.

Nabízí ohodnotit stránku dle takových ukazatelů, jako:

- Počet přidání stránky do oblíbených
- Počet přijatých aktivit (spojitostí)

[2] `http://simplymeasured.com`, Simply Measured

- Počet unikátních lidí, kteří se zbývají vaší začkou v daném časovém období
- Počet click-průchodek na vaše odkazy
- Počet unikátních lidí, kteří by mohli vidět vaše příspěvky
- Počet shlédnutí lidmi vašich příspěvků
- Míra vaši účasti

Jimi nabízené bezplatné služby se nepodařilo získat kvůli problémům s jejich aplikací. Jako bezplatné reporty nabízejí analýzu pouze určité typy stránek na Facebooku – fan pages (fanouškovské stránky). Další conditio sinequa non je podmínka alespoň 2 milionů fanoušků. Například jsem chtěl analyzovat fanpage YouTube, která má více než 80 milionů fanoušků, ale následně jsem dostal chybovou hlášku, která poukazuje na nutnost autentifikace, kterou jsem provedl. Příkladů grafů je na webové stránce poměrně málo, na jednom z nichž jsem nenašel ani název analyzované stránky2.7, ale jak jsem pochopil z grafů, většina publika se nachází ve Spojených státech, a nejpopulárnějšími publikacemi na stránkách se ukázaly být příspěvky v latině.2.8

2.3 Popster

Dalším zdrojem je Popster[3], který poskytuje mimo analýzu sociální sítě Vkontakte také analýzy Facebooku, Instagramu a Twitteru. Služba zajišťuje zpracování každých 20 000 uživatelů během 60 vteřin. Služba je placená, a stojí $5 měsíčně za 1 sociální síť. Také existuje varianta objednávky služby na rok za $40, která zahrnuje podporu všech uvedených sociálních sítí. Prvních 5 reportů jsou zdarma. Existuje možnost analýzy jak jednoduchých stránek uživatelů, tak i skupin. Report obsahuje všechny publikace na stránce s možností třídění podle data a počtu komentářů/liků/sdílení. Report obsahuje 6 grafů, některé z nich rád představím:

- Histogram, který ukazuje aktivitu podle hodin a dnů v týdnu2.9
- Line graph (lineární graf), který znázorňuje změnu aktivit uživatelů podle takových parametrů, jako like, sdílení, komentáře.2.10

2.4 Media-VK

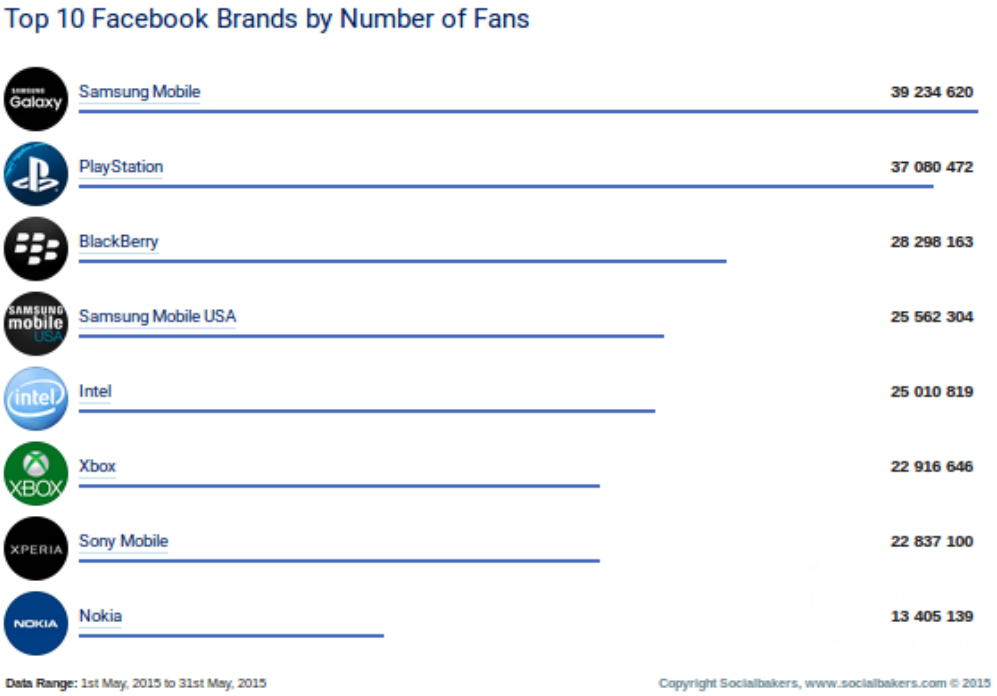
Dalším zdrojem je Media-VK[4] Služba nabízí analýzu jakéhokoliv community (neboli jinak označované skupiny) Vkontakte, hlavního zájmu fanoušků skupin

[3] `http://popsters.ru`, Popster
[4] `http://media-vk.ru/`, Media-VK

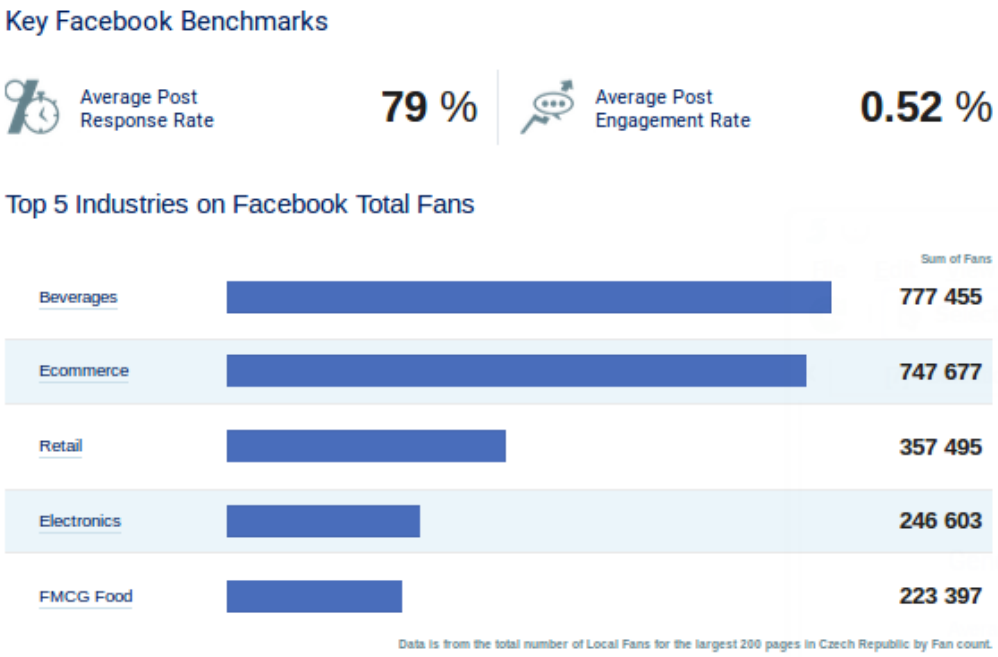
a dalších skupin, jejimiž jsou členy. Zdarma se poskytuje analýza 100-500 náhodně zvolených osob z té konkrétní skupiny. Podporuje analýzu do 30 tisíc uživatel, takový report přijde na $30. Report dané služby neobsahuje mnoho užitečných informací, nabízejí se pouze grafy typu PieChart, které znázorňují standardní atributy uživatelů.2.11 Reporty se posílají na emailovou adresu s menším zdržením.

V regionálním reportu byla obsažena následující informace:

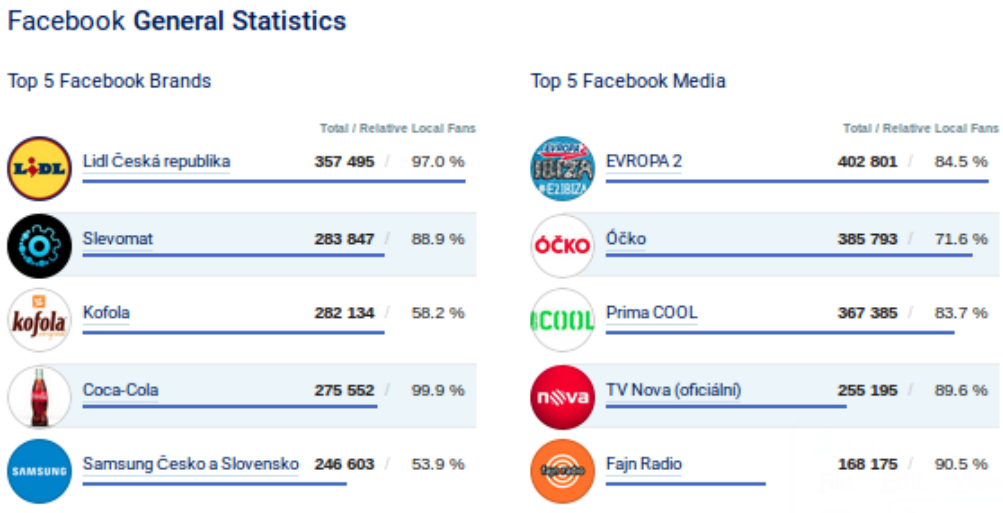
Obrázek 2.1: Socialbakers: Nejvíce se vyvíjející brandy



Obrázek 2.2: Socialbakers: Top 10 brandů podle počtu fanoušků



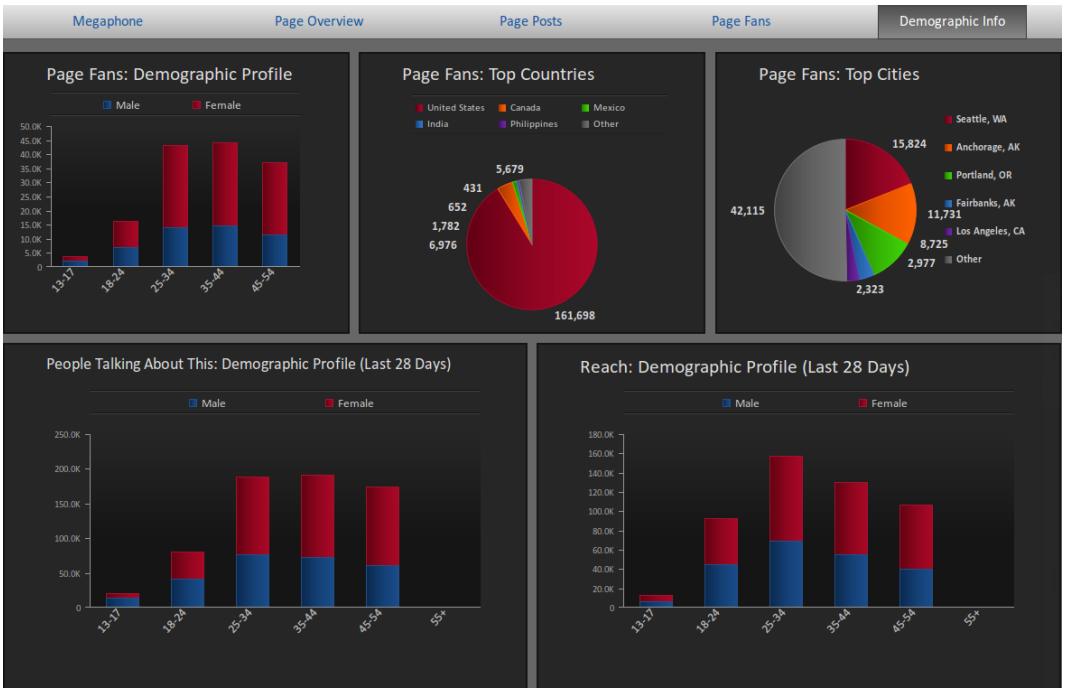
Obrázek 2.3: Socialbakers: Top 5 průmyslech podle počtu fanoušků



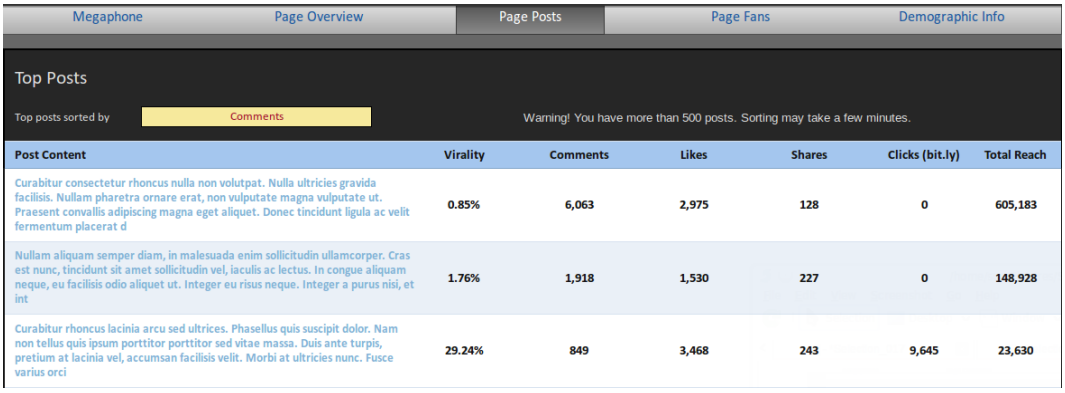
Obrázek 2.4: Socialbakers: Celková statistika

Obrázek 2.5: Socialbakers: Top 5 brandů v síti Facebook podle koeficientu zapojení

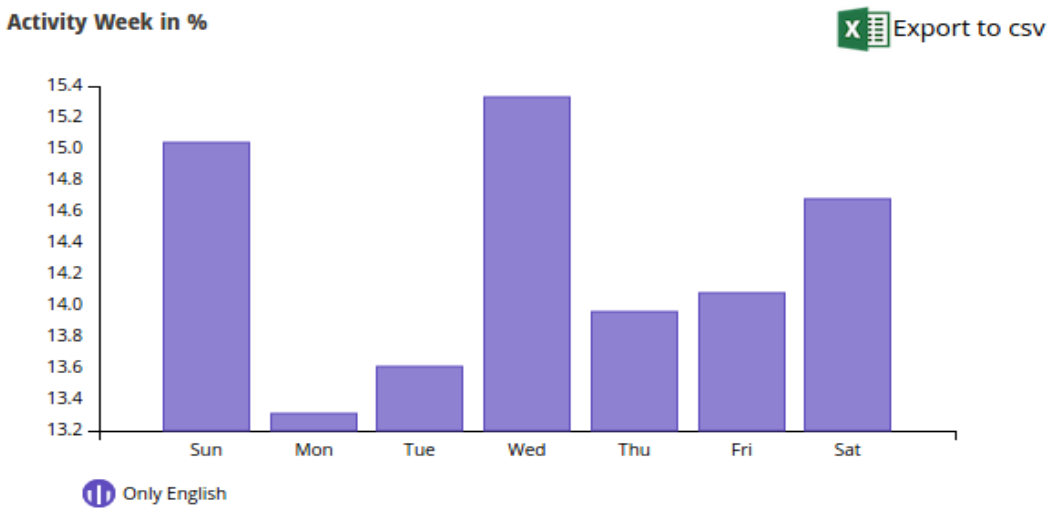
Obrázek 2.6: Socialbakers: Top 5 nejvíce sociálních brandů



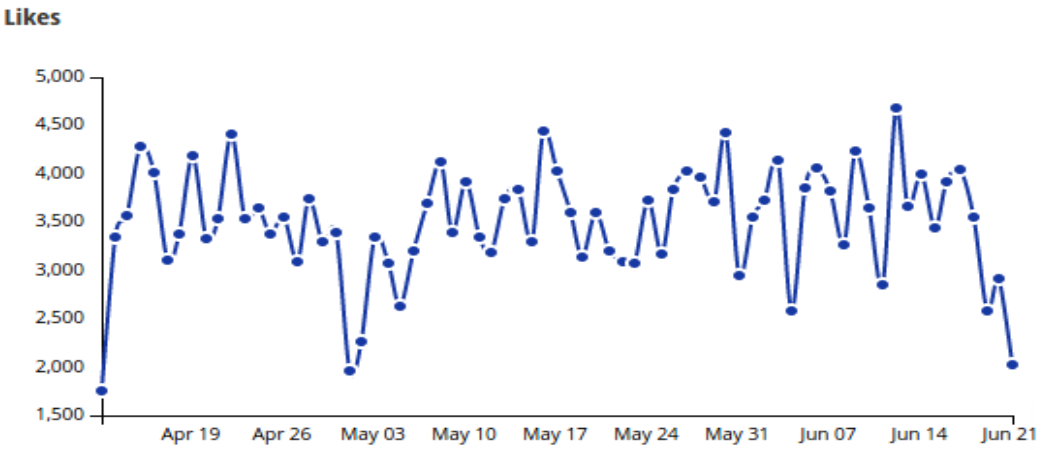
Obrázek 2.7: Simple Measure demografic report



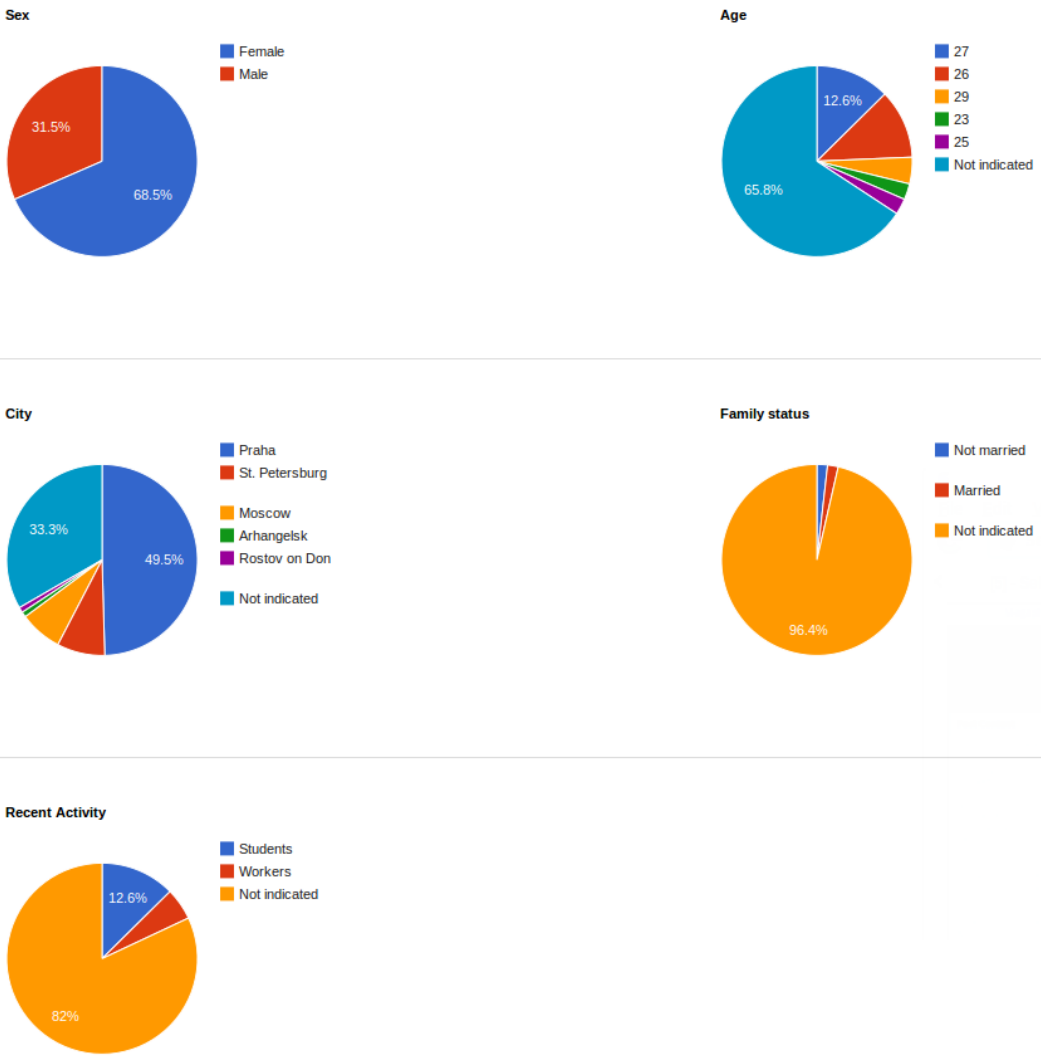
Obrázek 2.8: Simple Measure page posts report



Obrázek 2.9: Popster activity weeks



Obrázek 2.10: Popster likes



Obrázek 2.11: Media vk přehled

Kapitola 3

Analýza

3.1 Vyběr socialní sítí pro analyzu

Původně se plánovala realizace shromáždění a analýzy dat pro dvě sociální sítě: Facebook a vKontakte. Shromažďování dat se mělo uskutečnit pomocí poskytnutých API pro práci vnějších a vnitřních aplikací za účelem jejich integrace se sociálními síti.

Později však vyšla najevo jistá omezení, spojená s Facebook API (dále jen Use Graph API). Aplikace, vytvořené před 30.04.2014, používají Use Graph API v1.0 a mohou mít přístup k informacím, poskytovaným všemi uživateli, je-li taková informace poskytována pro veřejnost v nastavení soukromí. Později vytvořené aplikace mohou používat pouze Use Graph API v2.0, a mají přístup k informacím pouze těch uživatelů, kteří si takovou aplikaci nainstalují. Aplikace, vytvořené po 30.04.2015, povinně používají verzi 2.0.[4]

S přihlédnutím k této skutečnosti v této práci je popsána práce se shromážděním a analýzou pouze sociální sítě vKontakte, kterou bych rád představil.

- Jedná se o 23. nejnavštěvovanější web světa[5]
- 70 mil. unikátních uživatelů denně.
- Převážně rusky mluvící uživatelé.
- Uživatelé preferují ponechat otevřený přistup ke svým stŕankam

3.2 Požadavky

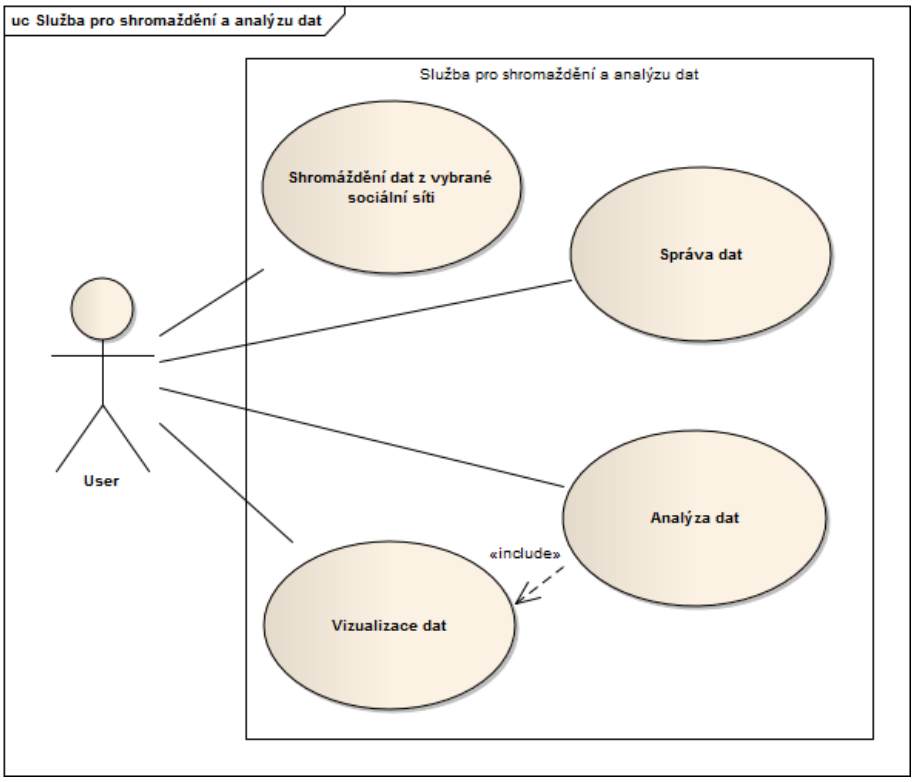
3.2.1 Funkční požadavky

- **Sběr základní informací uživateli skupiny**
  Aplikace musí získávat informace o účastnících skupiny a shromažďovat veškeré základní informace.
- **Sběr základní informací podle zadaných parametru**
  Aplikace musí poskytovat možnost vyhledávat uživatele na základě parametrů a shromažďovat jejich informace.
- **Sběr základní informací o přátelích uživatele sociální sítě**
  Aplikace musí shromažďovat unikátní identifikátory uživatelů, s nimiž je spojena zadaná skupina uživatelů
- **Sběr základní informací o skupinách uživatele sociální sítě**
  Aplikace musí shromažďovat unikátní identifikátory skupin, jichž je uživatel členem
- **Ukládaní dat do příslušné databázi**
  Aplikace bude ukládat veškeré shromážděné informace do lokální databáze počítače
- **Transformace dat do csv formátu**
  Veškeré informace se bude převádět do csv formátu pro zjednodušení zpracovávání dat.
- **Webová aplikace**
- **Vizualizace skupin uživatelů na základě zadaných parametru**
- **Klasterizace uživatelů**

3.2.2 Nefunkcni pozadavky

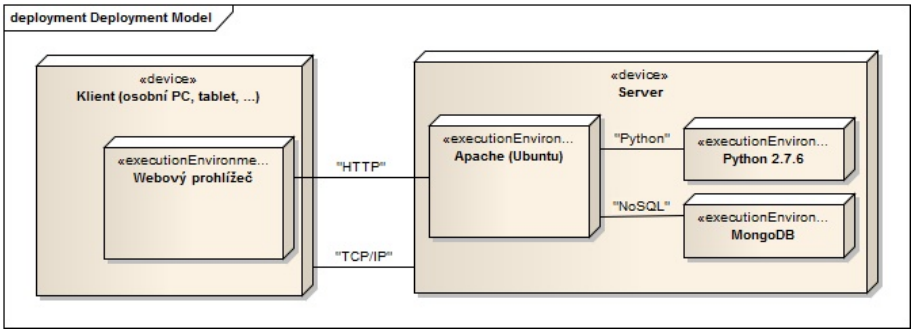
- Python Django framework
- Twitter Bootstrap
- Komunikace přes AJAX
- Intuitivni webové rozhraní
- Autentifikace a autorizace uživatelu

3.3 Případy užití



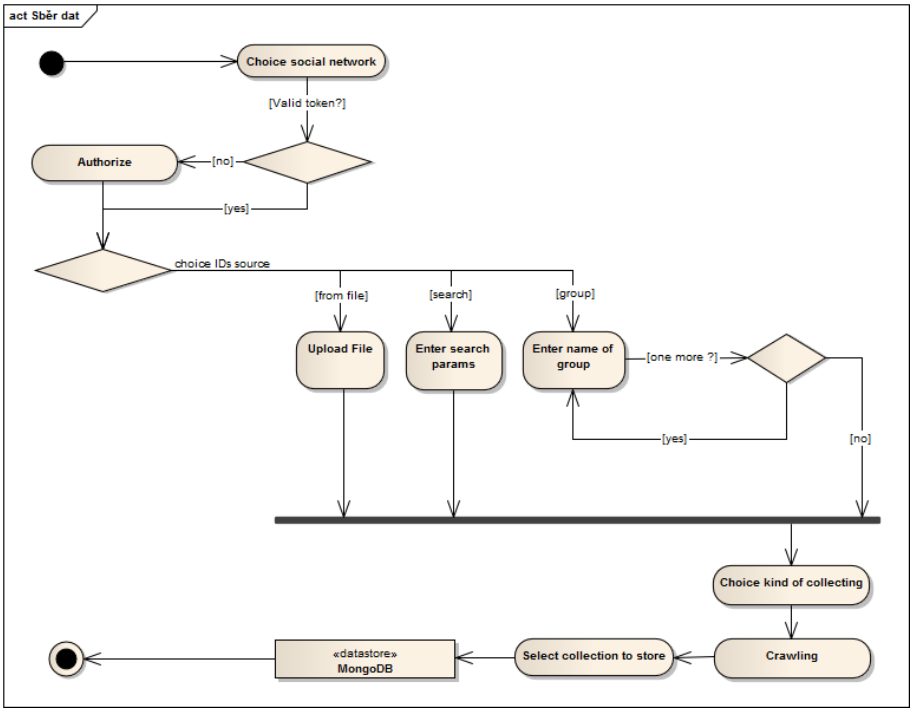
Obrázek 3.1: Diagram případu užití

3.4 Diagram nasazení

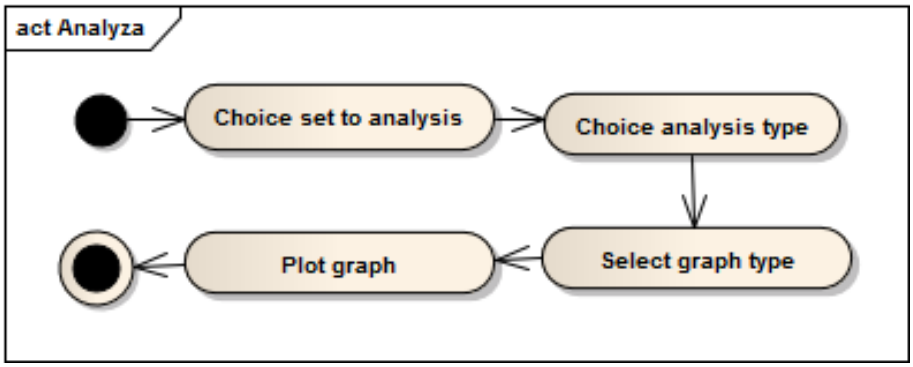


Obrázek 3.2: Diagram nasazení

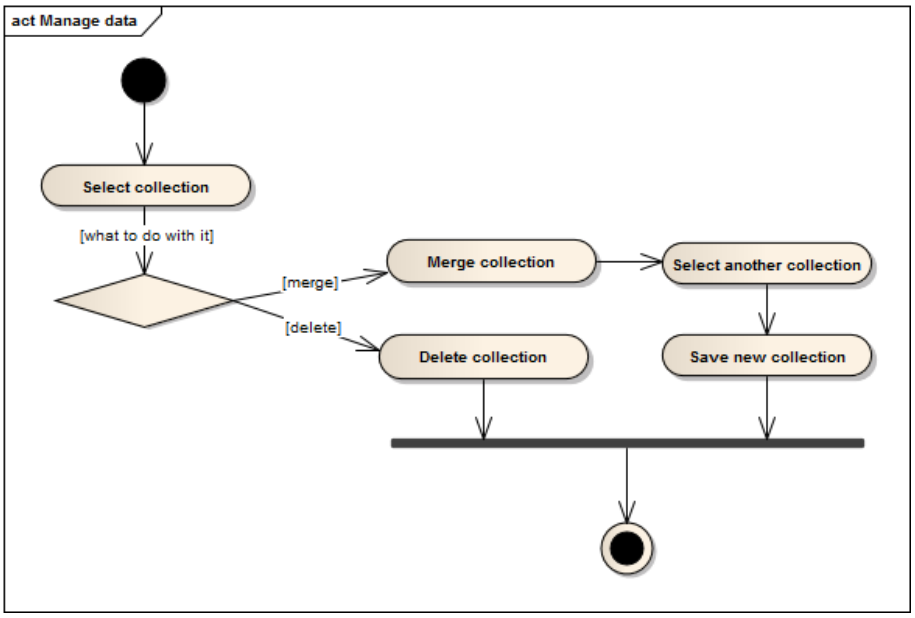
3.5 Diagramy aktivit



Obrázek 3.3: Diagram aktivit: Crawler



Obrázek 3.4: Diagram aktivit: Analýza dat



Obrázek 3.5: Diagram aktivit: Sprava dat

3.6 Volba programovacího jazyka

Moc bych chtěl realizovat tento projekt pomocí dynamicky typizovaného jazyka Ruby, jelikož existuje výborný framework Ruby On Rails[5] s excelentním API, který jsem již používal pro tvorbu webových aplikací, leč většina těch, kteří analyzují data, používají spíše Python. Vzhledem k neexistenci analogu takových knihoven pro analýzu dat, jako Numpy, Pandas, SciPy[6], rozhodl jsem se prostudovat webové frameworky Python.

3.7 Volba frameworku

Pro tvorbu webového rozhraní bylo prozkoumáno několik web-frameworků pro jazyk Python. Nejpopulárnějšími z nich jsou Flask[6], Bottle[7] a Django[8]. Pro ověření správnosti volby, s použitím každého z frameworků byla vytvořena stejná aplikace. Hodnotily se jak slabé, tak silné stránky kvality kodu a času, věnovaného tvorbě aplikace. Flask je mikroframeworkem, zaměřeným na malé aplikace s jednoduchými požadavky.[7] Django je vytvořen pro střední a velké aplikace, a má velkou škálu možností pro konfigurace a rozšíření.[8] Django používá vlastní ORM pro manipulace s databázemi, zatímco ve Flask a Bottle

[5] `http://rubyonrails.org/`, Ruby on Rails
[6] `http://flask.pocoo.org/`, Flask Framework
[7] `http://bottlepy.org`,Bottle Framework
[8] `https://www.djangoproject.com/`, Django Framework

můžeme zapojit jen ten nejvíce vyhovující. Nejpopulárnějším ORM pro Flask a Bottle je především SQL Alchemy, ale také existují DynamoDB, MongoDB, a velice jednoduché LevelDB a SQLite pro lokální použití. Zpočátku mě zaujal Flask, a psal jsem v něm, ale poté jsem narazil na omezení, spojené s komplikovaností struktury aplikace, nutné pro její rozšíření.

Django přichází s zabudovaným bootstrappingem.[9] Flask nezahrnuje nic jiného, nežli druh důvodu oddělení cílové skupiny a nesnaží se vybudovat rozlehlou MVP aplikaci.

Django odděluje projekt do individuální aplikace, kde Flask předpokládá projekt jako samostatnou «aplikaci» s několika pohledy a modely. To umožňuje replikaci projektu/aplikace rozdělením na Flasky, ve výchozím nastavení však tento pojem neexistuje.

Kapitola 4

Návrh

4.1 Shromažďování dat

4.1.1 Úvod

Je zřejmé, že sociální sítě poskytují přístup k osobním údajům uživatelů a umožňují shromažďovat tyto údaje pro určité účely. Shromažďování údajů uživatelů může poskytnout možnost pro analýzu uživatelských profilů a zakládat na tomto nějaké analytické modely, jako, např., cílení reklam, scoring[10], tvorba statistik, etc. Shromažďování uživatelských údajů se může uskutečňovat několika způsoby: pomocí curl parseru, který analyzuje HTML-stránky každého uživatele skupiny, ale sepsaní regulérních výrazů pro každý typ stránky je velice časově náročné. Sociální sítě poskytují svoje vlastní API, což usnadňuje shromažďování informací, odpověď přichází v XML nebo JSON formátu. Po analýze dat uživatelů sociálních sítí můžeme rozdělit tyto uživatele do různých skupin, podle pohlaví nebo podle zájmů, což nám umožní ve výsledku vylepšit přístup ke každé z výsledných skupin. Můžeme se také dozvědět, která skupina má největší zájem o daný produkt, a která volí spíše alternativní řešení. Abychom tohle všechno pochopili, musíme použit vedle rozdělení na skupiny podle určité kategorie také i klasterizaci a kontextní analýzu. Tohle všechno je jedním velkým komplexním problémem, kterým se zaobírá obor Data Mining. Data Mining[11] je analytickou metodologií získávání skrytých a potenciálně užitečných informací z dat.

4.1.2 Problematika

Sběr dat ze sociálních sítí Webové frameworky sociálních sítí jsou zdroji dat reálného času a jsou určeny k prohlížení a kooperaci se stránkami sociální sítě ve webovém prohlížeči, nebo pro použití uživatelských dat specializovanými aplikacemi. Jelikož scénáře využití frameworků sociálních sítí nepředpokládají automatické shromažďování dat velkého množství uživatelů, pak vzniká řada problémů:

- **Ochrana osobních údajů:** často přístup k údajům uživatelů je povolen pouze pro autorizované a registrované účastníky sítě, což vyžaduje podporu emulace uživatelského sezení pomocí accountů (účtů) určitého druhu

- **Slabá struktura dat:** ve mnoha případech programové frameworky (API) neumožňují získat údaje v potřebném formátu, proto musíme vybudovat jejich strukturovanou představu, využitelnou probudoucí automatické zpracování.

Při vyplňování svého profilu na sociální síti uživatelé často chybně nebo záměrně nevyplňují některá políčka nebo poskytují nepravdivé informace o své biografii, zájmech nebo zálibách. Mimo jiné, v kontentních sítích (Twitter, YouTube) uživatelský profil je často omezen na volbu základních atributů, nedostačující k řešení mnoha úkolů, jež předpokládají personalizaci výsledků. Tím pádem, jsou aktuální metody částečné identifikace autorů zpráv podle hodnot jejich demografických atributů. Zejména v systémech internetového marketingu a doporučení zvláštní význam hraje určení demografických atributů uživatele pro cílenou propagaci zboží a služeb ve skupinách uživatelů se stejnými hodnotami těchto atributů. Takové demografické charakteristiky nalézají své využití mimo internetové služby v různých disciplínách: sociologie, psychologie, kriminologie, ekonomika, řízení lidských zdrojů, etc.

Demografické atributy lze podmíněně rozdělit na:

- **Kategoriální:**pohlaví, národnost, rodinný stav, úroveň vzdělání, profese, zaměstnanost, náboženská a politická vzezření

- **Číselné:** věk, počet přátel, počet zveřejněných zprav.

- **Pořádkové:** vztah ke kouření a alkoholu.

Podmíněnost rozdělení je spojena s tím, že hodnoty číselného atributu lze zobrazit jako sadu kategorií, a zkoumat tento atribut jako kategoriální. Hodnoty věku se dají rozdělit zejména na několik věkových skupin, což se v praxi využívá poměrně často.

4.1.3 VK API

API pro práci s vKontakte nabízejí více než 300 funkcí a způsobů, ale v dané práci budou zohledněny pouze funkce následujících kategorií:

- Users (uživatelé)
- Wall (zeď)
- Audio (audiosoubory)
- Friends (přátelé)
- Groups (skupiny)

Pro komunikaci se serverem se posílá GET/POST požadavek na adresu:
'https://api.vk.com/method/' + method_name + params

Odpovědí je JSON s požadovanými údaji či informace o chybě. Základními metodami API byly:

- **users.search(params)**
  params – parametry hledání
- **users.get(users_ids, fields)**
  users_ids = množství uživatelů s ID, nepřevyšující 1000
  fields – zpětná informace o uživateli, například:
  Jméno, příjmení, pohlaví, město, univerzita, fakulta, oblíbené filmy, počet přátel, vztah k životu, lidem, alkoholu, kouření, informace o náboženských a politických postojích, rodinné vztahy, skype aj.
- **audio.get (user_id)**
- **friends.get(user_id)**
- **wall.get(owner_id, offset, count)**
  owner_id – ID uživatele
  count – nejvýše 20 pro 1 požadavek
- **groups.get(user_id)**
  vrátit všechny skupiny, jichž se uživatel účastní
- **groups.getById(groups_id)**
  vrátit veškerou informaci o skupině a ID všech účastníků skupiny

Analyzujeme-li sociální síť, získáváme následující data:

- Celkové informace o preferencích jedince
- Detailní statistické informace o různých kategoriích skupin, s pohlavím, věkem a jinými daty účastníků.

4.1.4 Získání Tokenu

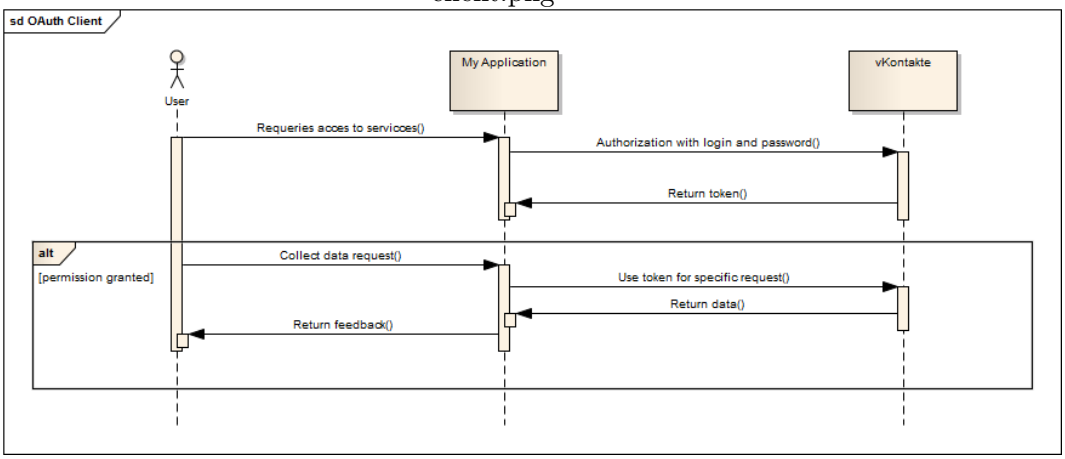
Mechanismus fungování V této části budou popsány všechny způsoby autorizace v této sociální síti, bude popsán formát datové výměny, poskytnuty příklady některých metod API, a také se ukážou některá omezení při práci s nimi. Pro získání přístupu k některým metodám API je nutné napřed získat token – určitý klíč, který autorizuje uživatele na stránce, jelikož přístup k některým metodám je možný pouze pomocí tohoto tokenu a příslušných přístupových práv. Autorizace probíhá pomocí protokolu Oauth 2.0.[12]

Oauth je autorizačním protokolem, který umožňuje jedné službě (aplikaci) vydání přístupových prav ke zdrojům uživatele u služby druhé. Protokol zbavuje nutnosti poskytnout aplikaci loginu a hesla, a také umožňuje vydání omezeného souboru práv, tudíž neposkytuje je všechny najednou.

Autentizace bývá dvou typů:

- **Klientská autentizace**
  - Klady:
    - Možnost vydání tokenu na neomezenou dobu (standardně na 24 hodin).
    - Je nutné registrovat vlastní aplikaci pro každého uživatele pro urychlení práce.
  - Zápory:
    - Maximálně 3 požadavky za sekundu za libovolného množství uživatelů

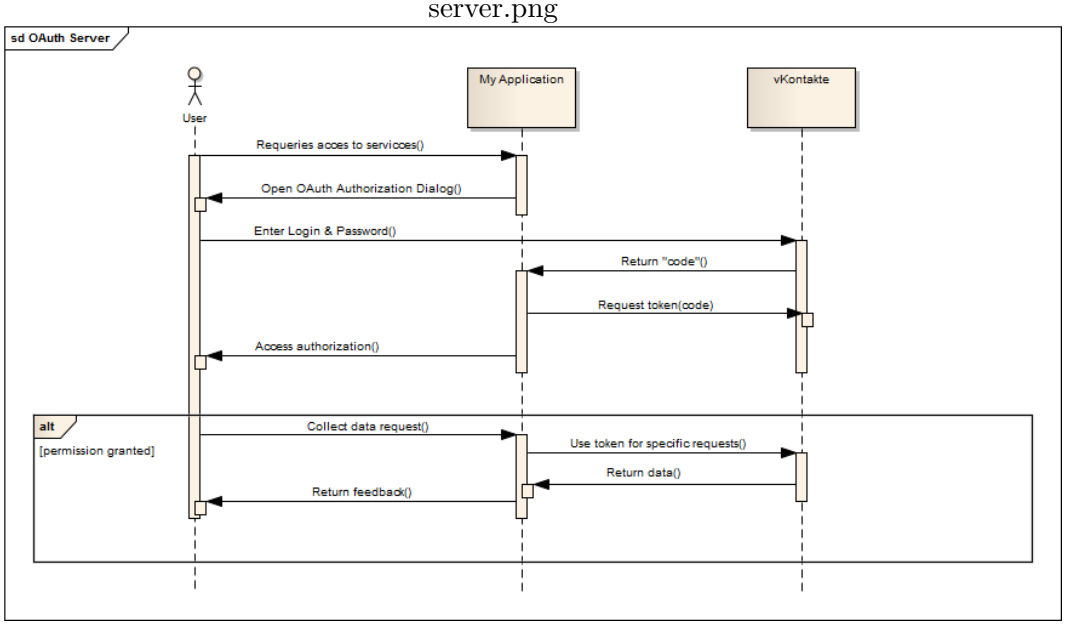
client.png



Obrázek 4.1: OAuth klientská autorizace

- **Serverová autentizace pro webové stránky**
  - Klady:
    - Pokud uživatel je již přihlášen do vKontakte, nepožadují se od něj přihlašovací jméno a heslo
    - 5 požadavků za sekundu, pokud přes naši aplikaci je přihlášeno méně než 10000 uživatelů
  - Zápory:
    - Token se vystavuje jen na 1 hodinu

server.png



Obrázek 4.2: OAuth serverova autorizace

Pro mou aplikaci byla vybrána klientská autentizace.

Autentizace za účelem získáni tokenu byla provedena pomocí knihovny Selenium.[9]

Selenium WebDriver — je programová knihovna pro řízení prohlížečů.

Existují některá ohraničení pro časté požadavky určitých typů a v tomto případě je nutné zadat Captcha[13] kód nebo vyčkat okolo 15 minut.

Při sběru informací o audionahrávkách načtených uživateli vKontakte byl často požadován Captcha a bylo rozhodnuto využít servis DeathByCaptcha[10], který v rozpoznávání captcha je úspěšný na 97 % a nabízí své služby za cenu $1,37 za 1000 úspěšně zpracovaných požadavků.

[9] `http://www.seleniumhq.org/projects/webdriver/`, Selenium WebDriver
[10] `http://www.deathbycaptcha.com`

Celkem při sběru informací o audio-nahrávkách k API vKontakte bylo odesláno okolo 18000 požadavků a za použití servisu deathbycaptcha bylo rozpoznáno 516 obrázků a získány informace o 7 miliónech audio-nahrávek.

Pro komunikaci s API vKontakte jsem vytvořil vlastní knihovnu, abych otdělil logiku práce sběru dat od webové aplikace, která zachycuje výjimky, zpracovává chyby a používá servis dathbycaptcha v případě potřeby.

4.2 Ukládání dat

4.2.1 Výběr správné databáze

Ve své aplikaci jsem použil 2 databáze. Jedna je relační, pro představení uživatelů mé služby a tokenů. Druhá je NoSQL databáze pro uchovávání databáze uživatelů sociálních sítí.

Během volby vhodné databáze pro uchovávání dat uživatelů sociálních sítí jsem se řídil především tím, aby formát ukládaných dat byl JSON. Dále jsem také prozkoumal různé varianty ukládání dat z profilů na sociálních sítích, a zvolil jsem NoSQL databázi MongoDB, protože NoSQL databáze jsou rychlejší v rozřazovacích systémech, zohledňujeme-li variantu s rozšířením vlastní služby a připojením dodatečných zdrojů.

SQLite[11] je integrovatelnou crossplatform-databází, která podporuje plnou sadu SQL-příkazů a je přístupná ve zdrojových kódech. Jejím hlavním plusem je jednoduchost. Tato databáze nevyžaduje přítomnost serveru, je zastoupena jen jedním souborem, a pro připojení k ní stačí zadání cesty a názvu tohoto souboru.

Nehledě na to, že Django podporuje široké využití databází, bylo se rozhodnuto používat modely pouze pro relační databázi.

Prvním důvodem je neustálá proměnlivost dat, flexibilní atributy, které se mohou měnit v čase, a proto není možné s jistotou říct, který z těchto atributů budeme potřebovat. Druhým důvodem je to, že oficiální větev Django nepodporuje NoSQL databáze, a v existujících řešeních (fork verze 1.6) se nepovedlo spojit modely mezi MongoDB a relační databází. Byly vyzkoušeny takové varianty, jako PostgreSQL, SQLite a další.

4.2.2 Rozdil klasické relační databáze i NoSQL

Relační databázový model:

- Rozseká do tabulek, mezi kterými jsou závislosti
- Plouživá řádky a sloupce
- Jednotlivé tabulky na sebe odkazuji pomoci cizích klíčů

[11] `https://www.sqlite.org/`, SQLite

- Při sběru informace musíme sáhnout do mnoha tabulek a zkombinovat výsledky dohromady
- Tabulky mají předem dáne schema, které je obtížně měnít

NoSQL

- Možnost vývoje bez návrhu schematy databázy
- Není předem definované schema tabulky
- Data jsou agregovaná dohromady
- Občas dochazí k duplicitě informací
- Lepší možnost distribuce databáze
- Možnost paralelního zápisu a čteni[14]

4.2.3 NoSQL - Not Only SQL

Mezi NoSQL databáze patří mnoho zástupců. Ti se mohou rozdělovat do kategorií podle toho, jaký datový model splňují. Mezi nejznámější datové modely patří sloupcově orientované databáze, databáze s klíčem a hodnotou, dokumentové databáze a objektové databáze. Modelů existuje samozřejmě více, např. grafové databáze a XML databáze, my se ale budeme věnovat pouze těm nejznámějším.

4.2.3.1 Sloupcově orientované databáze

Sloupcově orientované databáze ukládají data ve sloupcové formě a ne v řádkové, jak to známe z relačních databází. Sloupcově orientované databáze jsou výkonnější, když potřebujeme provéstnějakou agregační funkci na mnoha řádcích, ale na omezeném počtu sloupců. Výkonnost se také projeví, pokud chceme změnit hodnoty sloupců u všech řádek najednou.
Mezi nejznámější zástupce patří:

- HBase

4.2.3.2 Databáze s klíčem a hodnotou

Asociativní pole nebo HashMapa jsou nejjednodušší datové struktury, které dokážou uchovávat klíč s hodnotou. Klíčem je zde jedinečná hodnota, která může být jednoduše využita pro vyhledání dat.[15] Databáze s klíčem a hodnotou mohou být různých typů. Některé z nich ukládají data v paměti a jiné umožňují zaznamenání dat na disk
Mezi nejznámější zástupce patří:

- Redis
- MemcacheDB
- Cassandra

4.2.3.3 Dokumentové úložiště

Dokumentové databáze nejsou systémy pro správu dokumentů. Slovo „dokument“ v dokumentových databázích míní volně strukturovanou sadu klíčů-hodnot v dokumentech. Základním prvkem je zde dokument, do kterého se ukládají data a je označen jedinečným identifikátorem a množinou klíčů-hodnot. Pro ukládání dat se využívá většinou JSON.
Mezi dokumentovými databázemi jsou nejznámější:

- MongoDB
- CouchDB

4.2.4 MongoDB

Ze všech typu jsem vybral Dokumentové úložiště MongoDB[12], protože budu ukládat JSON dokumenty.

Data v MongoDB mají flexibilní schéma. Narozdíl od SQL-databází, v nichž se musí určovat a označit tabulkové schéma před vložením dat, soubory MongoDB nevynucují strukturu dokumentu. Tato flexibilita usnadňuje mapování dokumentů do subjektu nebo objektu. Každý dokument může porovnat datová pole představovaného subjektu, nehledě na to, že tato data mohou mít značné rozdíly. Avšak v praxi dokumenty v souboru mají podobnou strukturu. Klíčové rozhodnutí v tvorbě datových modelů pro aplikace MongoB spočívá ve struktuře dokumentů a jak aplikace ukazuje vztahy mezi daty. Toto jsou dva nástroje, které umožňují aplikacím ukazovat tyto vztahy:

- references (reference)
- embedded documents (vložené dokumenty)

4.2.4.1 Embedded documents

Embedded documents zachycují vztahy mezi daty ukládáním souvisejících dat do jediné struktury dokumentu. Dokumenty MongoDB umožňují vkládat struktury dokumenty do pole, nebo seřadit je uvnitř dokumentu. Tyto denormalizované (denormalized) datové modely umožňují aplikacím získávat a manipulovat se souvisejícími daty během jedné operace s databází. Modely vložených dat umožňují aplikacím uchovávat související kousky informací

[12] `https://www.mongodb.org/`, MongoDB

Obrázek 4.3: Denormalizovaná struktura MongoDB

ve stejném databázovém záznamu. V důsledku toho aplikace mohou vydávat méně dotazů a aktualizací k dokončení běžných operací.
Obecně se modely vložených dat používají v případech, kdy:

- mezi subjekty jsou „omezené“ vztahy
- mezi subjekty jsou vztahy „jeden-mnoho“. V těchto vztazích oné „mnoho“, dokumenty se vždy objeví nebo jsou prohlíženy v kontextu „jednoho“, neboli „rodičovského“ dokumentu.

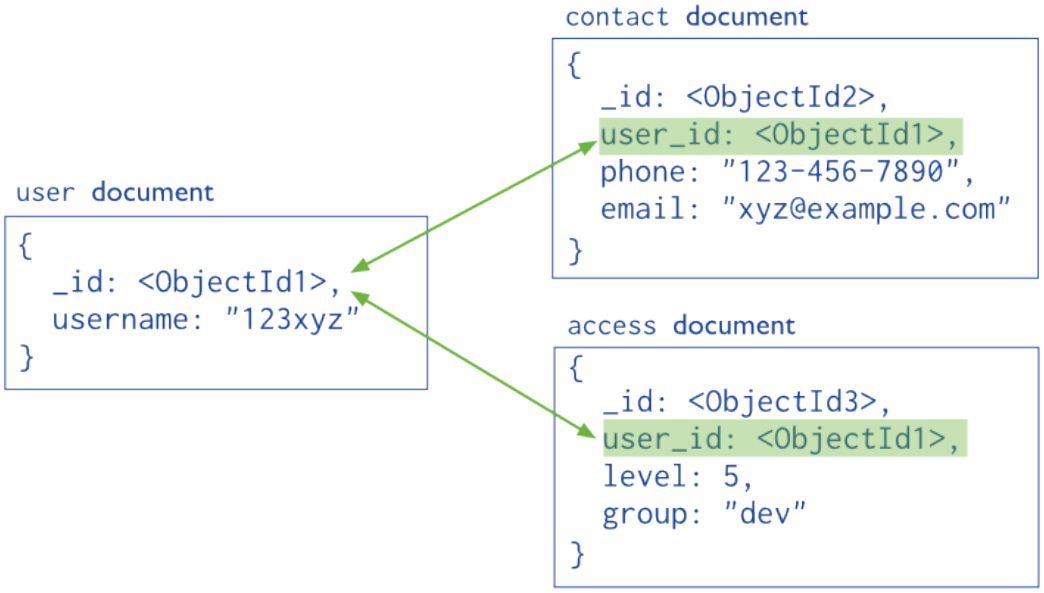
Vkládání nabízí lepší výkon pro operace čtení, stejně jako schopnost vyžadovat a získávat data během jedné databázové operace. Modely vložených dat umožňují aktualizovat související data během jedné operace. Nicméně, vkládání souvisejících dat do dokumentů může přivést k situacím, kdy dokumenty nabývají na objemu po jejich vytvoření.

4.2.4.2 References

Reference poskytují větší flexibilitu, narozdíl od vkládání. Nicméně, klientské aplikace musí vydávat navazující dotazy k vyřešení referencí. Jinými slovy, modely normalizovaných dat mohou vyžádat větší využití serverů. Obecně, modely normalizovaných dat se používají, když:

- vkládání vyústí v duplikaci dat, ale neposkytne dostatečný výkon při čtení, který má důsledky duplikace.

- se prezentují složitější vztahy typu „mnoho-mnoho“
- se modelují velké hierarchické soubory dat.



Obrázek 4.4: Normalizovaná struktura MongoDB

Pro představení souborů informací o uživatelích sociálních sítí volba padla na Embedded documents. Pro ovládání MongoDB z aplikace byl použit driver (ovládač) PyMongo. PyMongo pochází z distribuce Python, obsahuje nástroje pro práci s MongoDB, a je doporučeným způsobem práce s MongoDB.[16]

4.2.4.3 Domenova analyza

Diagram 4.5 popisuje fyzický datový model aplikace. Model Profile (profil) popisuje uživatele webové stránky, a je nástupcem stadardního modelu AbstractUser od Django, která je základem autentifikačního systému. Tento model představuje uživatele webové stránky a používá se pro ověření přístupových práv, registraci uživatelů a asociaci dat s uživateli. Pro představeni uživatelů v autentifikačním systému se používá pouze jedna třída.

Základní atributy profile:

- `username`
- `password`

Model Token obsahuje informace, nutné k autorizaci v sociálních sítích. Token samotný obsahuje pouze informaci o uplynutí jeho platnosti. Základní atributy tokenu:

- `social_network`
- `token`
- `expired_date`

Model Collection reprezentuje Kolekci MongoDB, do které se ukládají data ze socialních sítí. Obsahuje název kolekci a cestu k csv reprezentaci této kolekci. Atributy modelu:

- `name`
- `path`

4.3 Analýza dat

4.3.1 Použijte knihovny

4.3.1.1 Numpy

NumPy[13] je rozšířením jazyku Python, které přidává podporu velkých multidimenzionálních masivů a matic k velké knihovně vysokourovňových matematických funkcí pro operace s těmito masivy.

Vzhledem k tomu, že Python je interpretovaným jazykem, matematické algorytmy často fungují pomaleji, než v kompilovaných jazycích, jako např. C nebo Java. NumPy se snaží vyřešit tento problém pro velký počet výpočetních algorytmů zabezpečením podpory multidimenzionálních masivů, množstvím funkcí a operátorů pro práci s nimi. Tudíž jakýkoliv algorytmus, který může být vyjádřen v podstatě jako posloupnost operací s masivy a maticemi, funguje se stejnou rychlostí jako ekvivalentní kód v MATLABu, a po speciální optimalizaci rychlost může vyšplhat na rychlost kompilovaných jazyků.

4.3.1.2 Pandas

Tato sada činí z Pythonu velmi mocný nástroj datové analýzy. Pandas[14] poskytuje možnost tvorby souborných tabulek, seskupení a poskytuje pohodlný přístup k tabulkovým datám. Základními jsou Series a DataFrame. Series představuje indexovaný jednodimenzionální masiv hodnot. Je podobný jednoduchému slovníku typu dict, v němž název elementu bude odpovídat indexu,

[13] `http://www.numpy.org/`, Numpy
[14] `http://pandas.pydata.org/`, Pandas

a hodnota – hodnotě záznamu. DataFrame je indexovaným multidimenzionálním masivem hodnot, v němž každý sloupec DataFrame je strukturou Series.

Během práce se se soubory dat prováděly různé operace, jako např. Seřazování, filtrace, kontextní analýza a clusterizace.

4.3.2 Vizualizace

Pro tvorbu interaktivní webové aplikace se musel vyhledat dobrý nástroj pro její následnou vizualizaci, následně byla

provedena komparativní analýza takových nástrojů, jako GoogleCharts, HighSoft a D3.js.

	Google Charts Tools	D3.js	HighCharts
Formát	SVG + VML	SVG	SVG + VML
Formát dat	Javascript API	JSON, XML	JSON
Počet grafů	13	Neomezen	25+
Licence	Freeware pro všechny	BSD-3	Nekomerční využití: freeware.
Gradienty	Ne	Ano	Ne

[17]

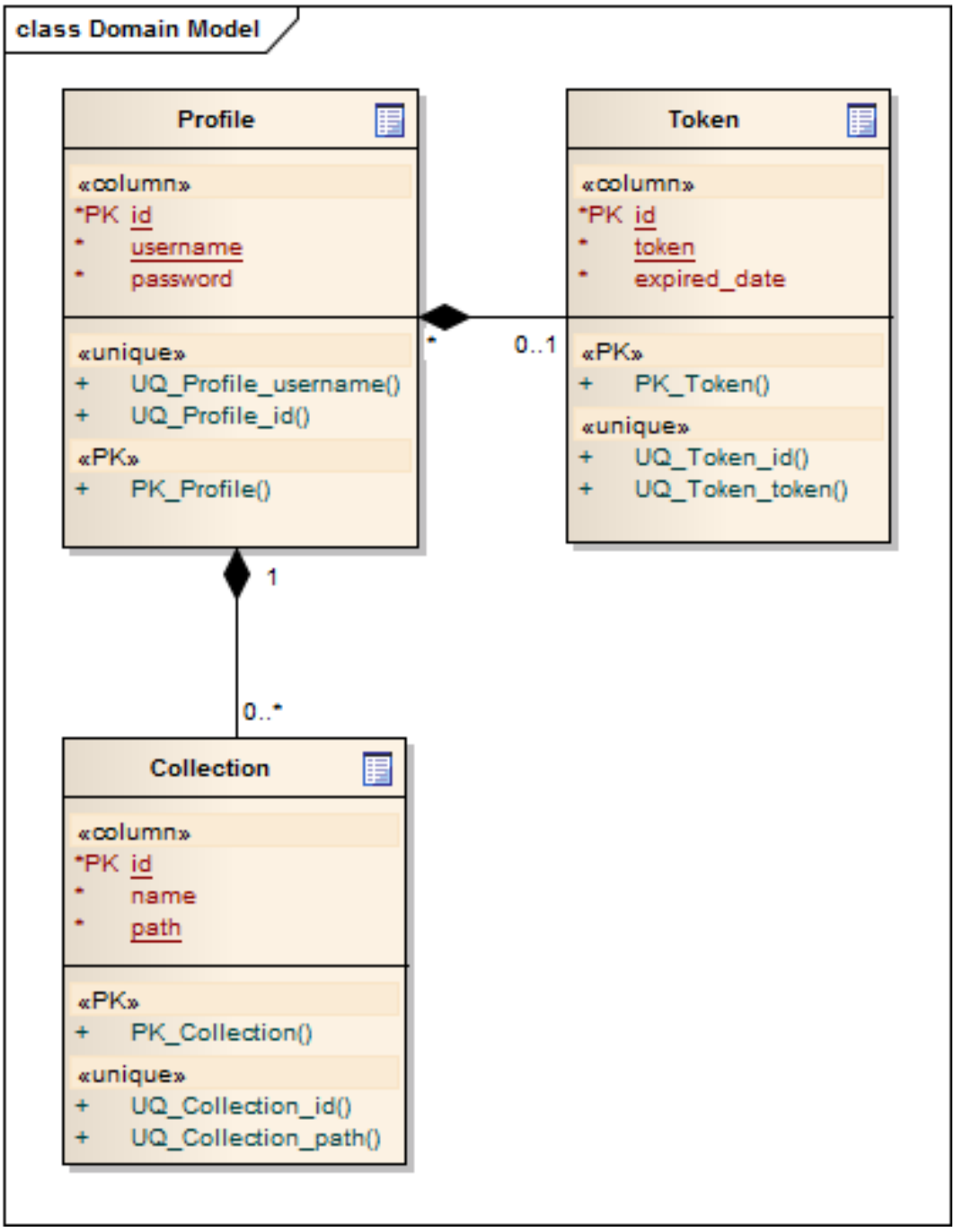
Nejjednodušší a nejefektivnější pro použití je knihovna HighCharts, jejímž jediným záporem ze mnou zpozorovaných je, podle meho názoru, absence gradientů, což neumožňuje plnohodnotné využití grafu BubbleCharts v této praci.

4.3.3 Clusterizace

Byla také provedena clusterizace uživatelů.

Clusterizace (neboli clusterová analýza) je metoda, pomocí níž se množství objektů rozděluje na skupiny, kterým se říká clustery. Uvnitř každé skupiny se musí nacházet „podobné“ objekty, objekty různých skupin se musí co nejvíce lišit. Hlavní odlišností clusterizace od klasifikace spočívá v tom, že seznam skupin není přesně určen – určuje se až během procesu fungování algoritmu.

Použil jsem metodu clusterizace k-means [18], kde jsem uvedl velký počet atributů, pomocí nichž se určovala míra podobnosti: vztah k alkoholu, vztah ke kouření.5.1 Metrika, kterou jsem použil pro nalezení vzdálenosti je Euklidova vzdálenost.[19]



Obrázek 4.5: Entity-relationship model

Kapitola 5

Vývoj webové aplikaci

5.1 Vývojová prostředí a další podpůrný software

5.1.1 Sublime Text

Sublime Text je editorem textových souborů a zdrojového kódu, který bych rád uvedl, jelikož představuje obsáhlou část mé aplikace. Narozdíl od IDE, Sublime Text je jednoduchým programem se spoustou kladů. Program obsahuje příkazový řádek, nabízí celou škálu volně stažitelných pluginů pro vylepšení a přizpůsobení základních funkcí, schopnost kontroly syntaxe různých jazyků, jako například JavaScript, Python, XML, CSS, HTML nebo schopnost automatického dokončování.

5.1.2 Mozilla Firefox a jeho doplňky

Hlavním prohlížečem pro vývoj své aplikace jsem zvolil svobodný multiplatformní prohlížeč Mozilla Firefox od společnosti Mozilla Corporation. Jeho předností je možnost volného stažení doplňků. Ve velkém množství však tyto doplňky mohou negativně ovlivnit funkční rychlost prohlížeče a zatížit operační paměť počítače. Dále zmíním jen několik nástrojů prohlížeče, které využívám pro svou práci, jako například Firebug nebo Web Developer.

Firebug
Firebug je velice populární doplňek umožňující snadnou kontrolu a edici HTML a CSS. Zároveň, prochází DOM, sleduje dotazy ze stránky na server, obsahuje rozšíření pro JavaScript a příkazovou konzoli.

Web Developer Ještě jedným nástrojem pro vývojáře je Web Developer s pohodlným a přehledným menu. Jednou z funkcí, která stojí za zmínku. je zobrazení generovaného zdroje, umožňující porovnání původního HTML kódu a výsledku po použití JavaScriptu. Samozřejmostí je velké množství nejen zákazových funkcí, umožňujících zákaz JavaScriptu, Javy, mezipaměti nebo vizuálních prvků na stránce, ale i dalších, které pracují s formuláři, cookies, kaskádovými styly, online validátory webových stránek a jejich propojením.

Nástroje vývojáře Základní verze prohlížeče Firefox obsahuje Nástoje vývojáře se spoustou prospěšných funkcí, jako například Editor stylů nebo Průzkumník, který umožňuje pohodlné sledování HTML prvků včetně jejich původu. Nabízí se zde i možnost zobrazení jednotlivých prvků s kaskádovými styly. Zmíněný Editor stylů je nápomocný pro okamžité vizuální zobrazení změn při edici kaskádových stylů.

5.1.3 Enterprise Architect

Tento nástroj hlavně vytváří UVL diagramy různých druhů, podporuje datové modelování a modelování obchodních procesů. Umožňuje forward a reverse engineering kódů, nabízí nástroje pro tvorbu verzí nebo různé druhy testování. Tento nástroj považuji za velice prospěšný pro celý proces vývoje aplikace, hlavně pro tvorbu diagramů případů užití.

5.2 Instalace a konfigurace aplikací

Instalační příručka k Django a ostatním nástrojům se nachází v přílozeA

5.3 Struktura adresářů

Struktura adresáře se nachází v přílozeC.

5.4 Ošetření vstupních dat a zabezpečení formuláře

Pro velké množství vstupních dat, které uživatel poskytuje aplikacím určeným pro správu dotazníků, je velmi důležité zabezpečit je před napadním a zneužitím. Nejběžnějším způsobem napadení je Cross-site request forgery, kdy útočník nenápadně podstrčí uživateli data, která pak uživatel odešle na server. Nastavení CSRF tokenu u formuláře dokáže před podobným útokem bránit. Každý takový formulář pomocí Django naváže na sebe uvedený řetězec, který se po odeslání validuje. Data mohou být také napadená pomocí jazyku JavaScript, kdy se škodlivý kód doručuje přes špatně ošetřené vstupy. Takovému způsobu útoku se říká Cross Site Scripting (XSS). Bránit se proti podobým útokům umožňuje šablonovací systém Django pomocí implicitního escapování kontextových proměnných pro zamezení výkonu potenciálně škodlivého kódu.

5.5 Uživatelské rozhraní

Menu se rozmístí v horní liště a bude obsahovat odkazy na úvodní stranku, kde bude nabizeno uživaeli se přihlasit. Též se do menu zařadí položky «Crawler»

a «Data Analysis» a «Collections».

5.5.1 Twitter Bootstrap.

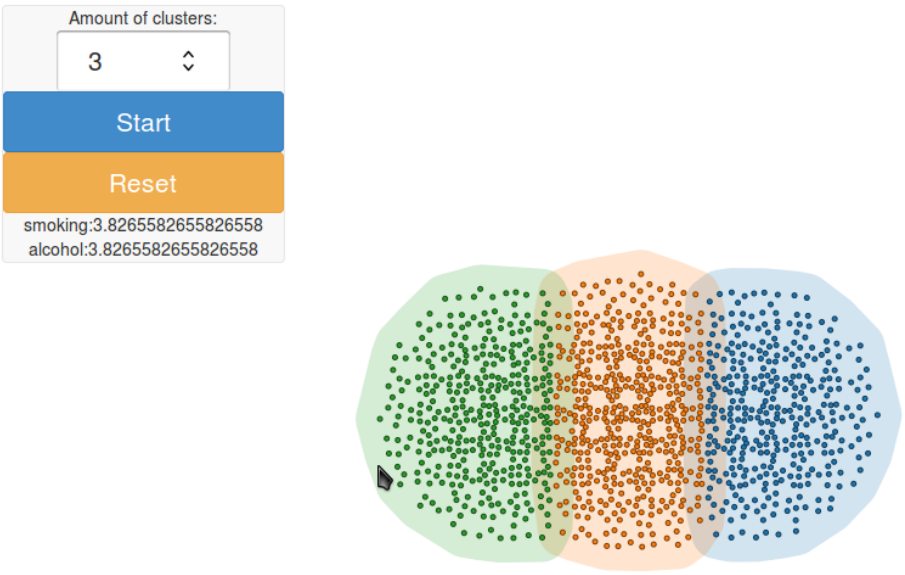
Twitter Bootstrap je velmi jednoduchý a volně dostupný soubor nástrojů pro vytváření moderního webu a webových aplikací. [1] Nabízí podporu nejrůznějších webových technologií HTML, CSS, JavaScript a mnoho prvků, které je možné snadno implementovat do své stránky. Interaktivní prvky jako jsou tlačítka, boxy, menu a další kompletně nastavené a graficky zpracované elementy je možné vložit pouze pomocí HTML a CSS. Výhodou tohoto souboru nástrojů je snadné zpracování jakéhokoliv uživatelského rozhraní ve webové aplikaci a nerozhoduje, zda to je například uživatelské rozhraní v administraci back-endových nebo front-endových aplikací[20]

5.5.2 jQuery

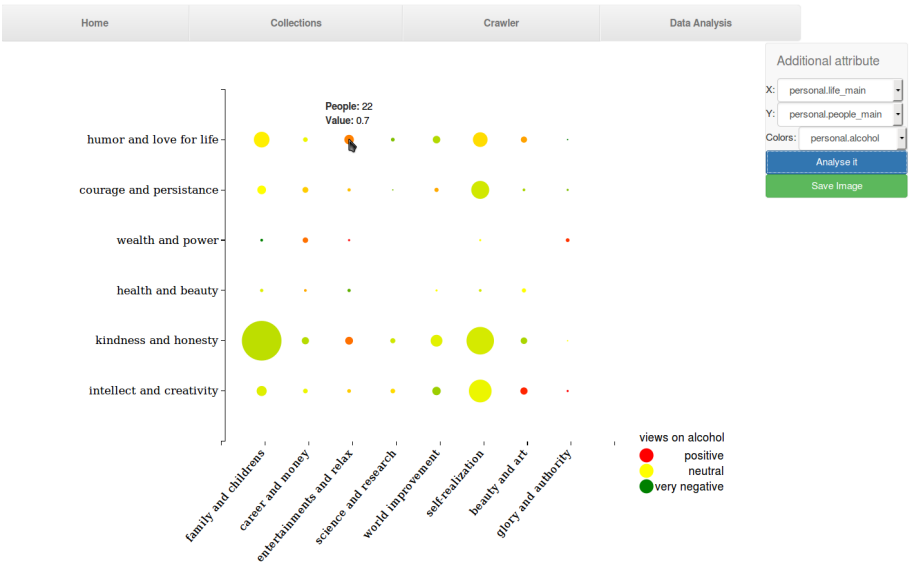
Pro snadnější implementaci pokročilých prvků a efektů do webové stránky, jsem využil volně dostupnou JavaScriptovou knihovnu JQuery.[21] Tato knihovna zahrnuje řadu funkcí usnadňujících procházení a změnu DOM, manipulaci s CSS, obsluhu událostí nebo tvorbu animací.

5.6 Ukázka aplikaci

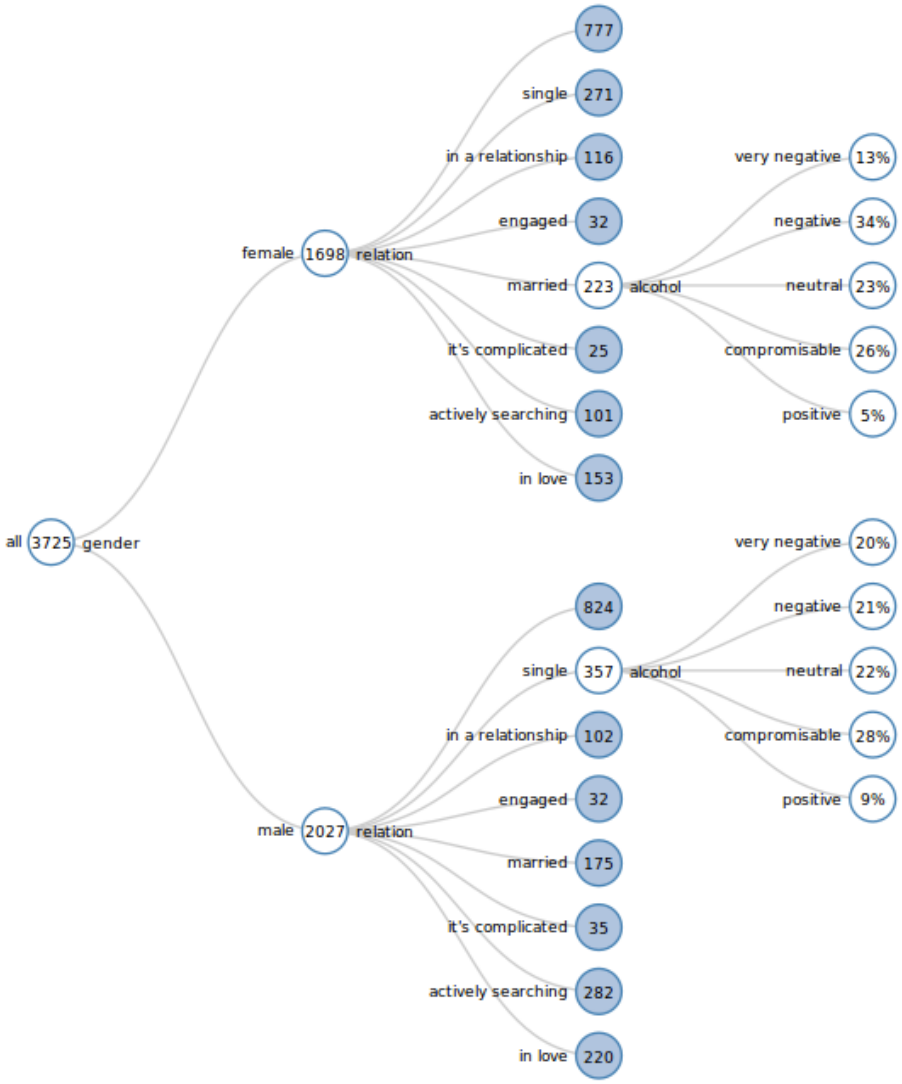
Ve mnou vyvinuté aplikaci můžeme provádět analýzu uživatel a výsledek promítat do grafů typů Histogram, PieChart (výsečový graf), Decision Tree (Rozhodovací strom) [22], Bubble Chart, aj. Uživatel mé aplikace má pro účely analýzy takové demografické atributy, jako: pohlaví, rodinný stav, jazyková vybavenost, vztah k životu, lidem, politické názory, město, stát, vysoká škola, fakulta, vztah ke kouření a alkoholu. Proanalyzoval jsem několik atributů zhruba 35 699 uživatelů, jež uvedli Českou republiku jako své současné bydliště. Počet lidí s vyplněným polem Rodinný stav: 5710 Počet lidí s vyplněným atributem „hlavní v životě“: 3274 Průsečík těchto množin činí 1881 osob. Zohledníme-li rodinný stav lidí5.4, jež za hlavní v životě označili osobní rozvoj, pak zjistíme, že počet osamělých osob procentuálně vzrostlo vůči vdaným/ženatým osobám.5.5 Většina analyzovaných atributů je kategoriální. Jako další příklad lze uvést diagram uživatelů, jež považují za hlavní v životě rodinu a děti, s rodinným stavem „ženatý/vdaná“, který také zobrazuje jejich vztah k lidem, a co tito uživatelé považují za podstatné v lidech. 5.6 V mé aplikaci jsou také přístupné rozhodovací stromy v fixovanou posloupností. Níže je představen jeden z nich:5.3 Také je umožněna clusterizace uživatelů, popsaná v sekci 4.3.3



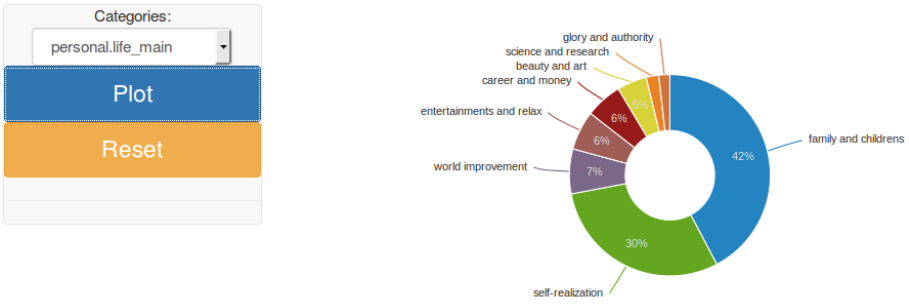
Obrázek 5.1: Klustery



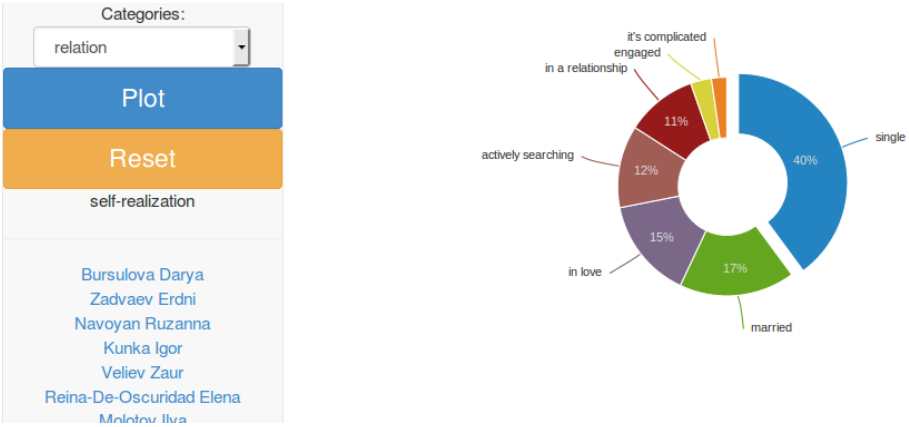
Obrázek 5.2: Vztah ke kouření



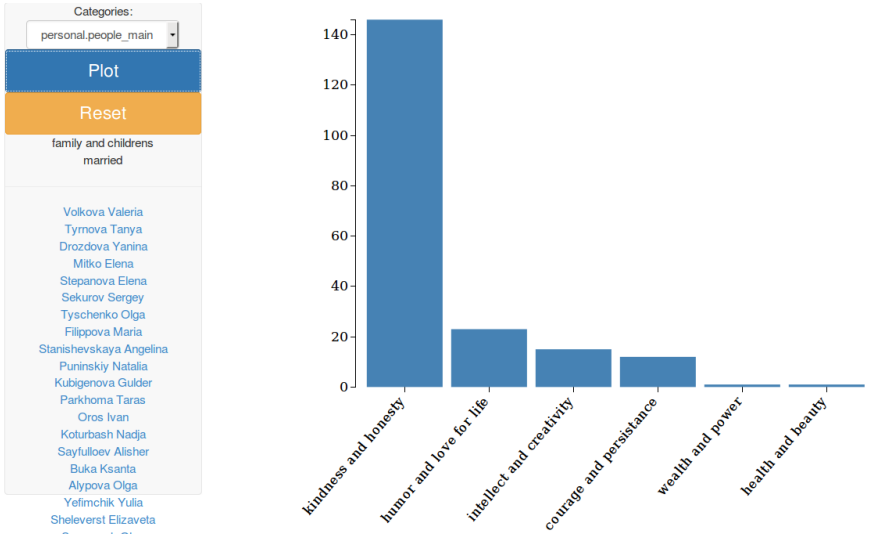
Obrázek 5.3: Rozhodovací strom



Obrázek 5.4: Diagram rodinného stav



Obrázek 5.5: Diagram rodinného stavu s podmínkou preferencí seběrealizaci v životě



Obrázek 5.6: Diagram vztahu k lidem

Závěr

V této práci byla provedena analýza online-aplikací pro shromažďování a analýzu dat, a byly popsány jejich charakteristické rysy. Byla vytvořena aplikace, umožňující ukládání a analýzu dat, shromážděné z vybrané sociální sítě. Bohužel, Facebook zakázal masový sběr dat aplikací bez publika k 30. dubna 2014. Byla představena sociální síť vKontakte, popsány charakteristické zvláštnosti shromáždění dat z této sociální sítě a také provedená statistická analýza, zaměřená na segmentaci uživatelů podle takových atributů, jako: pohlaví, politické názory, vztah k životu, lidem, vztah ke kouření a alkoholu, a jiné. Realizace online-aplikace pro shromáždění a analýzu dat byla provedena ve frameworku Django. Byly prozkoumány takové populární knihovny pro analýzu dat v Python, jako je Numpy, Pandas, a také knihovny Javascript a externí API pro vizualizace dat v internetovém prohlížeči: D3js, HighSoft a GoogleCharts. V této práci byla vyřešena otázka s rozpoznáváním Captcha, požadované v některých požadavcích, a také byla provedena asynchronní práce s postupnou analýzou dat pomocí session (sezení).

Literatura

[1] Winkler, J.; Petrusek, M.; aj.: *Velkỳ sociologickỳ slovník*. Karolinum Praha, 1997.

[2] Tjaden, P. G.: *The Crime of Stalking: How Big is the Problem?*. National Criminal Justice Reference Service, 1997.

[3] McGee, M.: EdgeRank is dead: Facebook's News Feed algorithm now has close to 100K weight factors. *Marketing Land*, 2013.

[4] Facebook Application Development FAQ. `https://developers.facebook.com/docs/apps/faq`, accessed June 10, 2015.

[5] Alexa Top 500 Global Sites. `http://www.alexa.com/topsites`, accessed June 10, 2015.

[6] Jones, E.; Oliphant, T.; Peterson, P.: {SciPy}: Open source scientific tools for {Python}. 2014.

[7] Ronacher, A.: Welcome| Flask (A Python Microframework). *URL: http://fla sk. pocoo. org/(v isited on 07/02/2013)*, 2010.

[8] Forcier, J.; Bissex, P.; Chun, W.: *Python web development with Django*. Addison-Wesley Professional, 2008.

[9] Sands, P.: World wide words: A rationale and preliminary report on a publishing project for an advanced writing workshop. *Academic Writing*, 2002.

[10] Schaefer, M.: *Return on influence: The revolutionary power of Klout, social scoring, and influence marketing*. McGraw-Hill, 2012.

[11] Han, J.; Kamber, M.; Pei, J.: *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

[12] Leiba, B.: Oauth web authorization protocol. *IEEE Internet Computing*, , č. 1, 2012: s. 74–77.

[13] Von Ahn, L.; Blum, M.; Hopper, N. J.; aj.: CAPTCHA: Using hard AI problems for security. In *Advances in Cryptology—EUROCRYPT 2003*, Springer, 2003, s. 294–311.

[14] Stonebraker, M.: SQL databases v. NoSQL databases. *Communications of the ACM*, ročník 53, č. 4, 2010: s. 10–11.

[15] Panyko, T.: *NoSQL databáze*. Diplomová práce, Jihočeská univerzita v Českých Budějovicích, 2013.

[16] Tsoukalos, M.: Using Django and MongoDB to build a blog. *Linux Journal*, ročník 2014, č. 238, 2014: str. 3.

[17] Chart js C3 D3 Highchart FusionCharts Google Chart | Comparison tables - SocialCompare. `http://socialcompare.com/en/comparison/chart-js-c3-d3-highchart-fusioncharts-google-chart-2i8mwb90`, accessed June 5, 2015.

[18] MacQueen, J. B.: Some Methods for Classification and Analysis of MultiVariate Observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, ročník 1, editace L. M. L. Cam; J. Neyman, University of California Press, 1967, s. 281–297. Dostupné z: `http://www.bibsonomy.org/bibtex/25dcdb8cd9fba78e0e791af619d61d66d/enitsirhc`

[19] Weisstein, E. W.: Euclidean Metric. MathWorld – A Wolfram Web Resource, 1999. Dostupné z: `http://mathworld.wolfram.com/EuclideanMetric.html`

[20] Otto, M.; Thornton, J.: Bootstrap. *Twitter Bootstrap*, 2013.

[21] De Volder, K.: JQuery: A generic code browser with a declarative configuration language. In *Practical Aspects of Declarative Languages*, Springer, 2006, s. 88–102.

[22] Friedl, M. A.; Brodley, C. E.: Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, ročník 61, č. 3, 1997: s. 399–409.

Příloha A

Instalační příručka

A.1 Krok 1: Instalace veškerých potřebných balíčků pro Python

Nejprve stáhneme Pip, systém správy balíčků v Pythonu:

```
sudo apt-get install pip
```

Pak provedeme instalaci knihoven pro analýzu dat:

```
sudo apt-get install python-numpy
sudo apt-get install python-pandas
```

A Selenium webdriver pro Python:

```
sudo pip install selenium
```

A.2 Krok 2: Instalace Django

Django je framework pro tuto aplikaci a Pipeline je knihovna pro Django, která poskytuje komprese jak pro CSS, tak pro JavaScript soubory:

```
sudo pip install django
sudo pip install django-pipeline
```

A.3 Krok 3: Instalace a konfigurace MongoDB

A.3.1 Instalace MongoDB

```
sudo apt-get install mongodb-server mongodb-clients
```

A.3.2 Spouštění mongodb serveru

Pro spouštění serveru je nutně ukazat zadat cestu k adresáři, kám chceme ukládat data:

```
sudo mongod --dbpath $db_path
```

A.4 Krok 4: Spouštění aplikací

Spouštím Django server jednoduchým příkazem:

```
python $app_path manage.py runserver
```

Příloha B

Seznam použitých zkratek

DOM Document Object Model

CSRF Cross-site request forgery

CSS Cascading Style Sheets

CSV Comma Seperated Values

XML Extensible markup language

HTML HyperText Markup Language

SEO Search Engine Optimization

SQL Structured Query Language

NoSQL Not Only SQL

MVP ModelView-Presenter

ORM Object-relational Mapping

API Application Programming Interface

AJAX Asynchronous Javascript and XML

JSON JavaScript Object Notation

VK vKontakte

Příloha C

Obsah přiloženého CD

- `text`
  - `src` ..... zdrojová forma práce ve formátu LaTeX
  - `BP_Poddubny_Alexander_2015.pdf` ..... text práce ve formátu PDF
- `vkapp` ..... Hlavní adresář
  - `crawler` ..... adresář s aplikací Crawler
    - `migrations` ..... Složka s migracemi
    - `templates` ..... HTML šablony pro prezentaci
    - `forms.py` ..... Formuláře
    - `models.py` ..... Modely aplikaci
    - `urls.py` ..... Routovaní pro Crawler
    - `views.py` ..... View interface pro praci s aplikaci Crawler
  - `data_analysis` ..... adresář s aplikací Data Analysis
    - `templates` ..... HTML šablony pro prezentaci
    - `forms.py` ..... Formuláře
    - `models.py` ..... Modely aplikaci
    - `urls.py` ..... Routovaní pro Data Analysis
    - `views.py` ..... View interface pro praci s aplikaci Data Analysis
  - `db_manager.adresář s aplikací Data Manager`
    - `templates` ..... HTML šablony pro prezentaci
    - `forms.py` ..... Formuláře
    - `models.py` ..... Modely aplikaci
    - `urls.py` ..... Routovaní pro Data Manager
    - `views.py` ..... View interface pro praci s aplikaci Data Manager
  - `mongosettings` .. Konfigurační soubor pro csv export dat z MongoDB databáze
  - `static` ..... Statické soubory: CSS, Javascript
  - `tmp` ..... Místo pro objekty session
  - `upload` ..... csv soubory uživatelů
  - `vkapp` ..... Společný adresář pro všechny aplikací
    - `__init__.py`

```
    ├── deathbycaptcha.py ..... DeathByCaptcha API
    ├── mongo_api.py ..... Mongo API
    ├── settings.py ..... Django nastavení
    ├── urls.py ..... Kořenové routovaní
    ├── vk_api.py ..... VK API
    └── wsgi.py
├── __init__.py
├── db.sqlite3 ..... SQLite databáze
└── manage.py ..... Django Kontroller
```