VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

URČENÍ VLIVU UŽIVATELŮ NA SOCIÁLNÍCH SÍTÍCH

USER IMPACT IN SOCIAL NETWORKS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. PETR JIROUT**
AUTHOR

VEDOUCÍ PRÁCE — **Doc. RNDr. PAVEL SMRŽ, Ph.D.**
SUPERVISOR

BRNO 2017

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav počítačové grafiky a multimédií Akademický rok 2016/2017

Zadání diplomové práce

Řešitel: **Jirout Petr, Bc.**
Obor: Bioinformatika a biocomputing
Téma: **Určení vlivu uživatelů na sociálních sítích**
User Impact in Social Networks
Kategorie: Algoritmy a datové struktury

Pokyny:

1. Prostudujte rozhraní služby Twitter a dalších sociálních sítí a identifikujte potenciální problémy při analýze vlivu uživatele.
2. Navrhněte systém, který dokáže pravidelně získávat a analyzovat stahovaná data z vybraných sociálních sítí.
3. Implementujte systém pro určení vlivu jedinců na vybraných sociálních sítích.
4. Zhodnoťte dosažené výsledky a další možné pokračování tohoto projektu.
5. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- dle dohody s vedoucím

Při obhajobě semestrální části projektu je požadováno:

- Funkční prototyp

Podrobné závazné pokyny pro vypracování diplomové práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva diplomové práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap, které byly vyřešeny v rámci dřívějších projektů (30 až 40% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Smrž Pavel, doc. RNDr., Ph.D.**, UPGM FIT VUT
Datum zadání: 1. listopadu 2016
Datum odevzdání: 24. května 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav počítačové grafiky a multimédií
612 66 Brno, Božetěchova 2

L.S.

doc. Dr. Ing. Jan Černocký
vedoucí ústavu

Abstrakt

Tato práce pojednává o vytvoření systému pro analýzu sociálních sítí, umožňující určení vlivu vybraných uživatelů těchto sítí. Tento systém byl zpřístupněn veřejnosti pod MIT licencí a umožňuje jednoduchou rozšiřitelnost pro vlastní analytické přístupy. Použití tohoto systému bylo demonstrováno na příkladu analýzy vybraných profilů z české politické scény na sociální síti Facebook. Tato práce porovnala jejich vliv a aktivitu uživatelů na těchto profilech. Zároveň byl navržen nový přístup pro predikci aktivity a vlivu na těchto profilech a to pomocí identifikování tzv. oddaných fanoušků.

Abstract

This thesis describes the design and implementation of a system for social media analysis. This system provides a way of identifying social media user's influence. The system has been open sourced under the MIT license and is designed to be easily extensible. Example usage of this system is demonstrated for a chosen use case of analysing several selected Czech individuals and political parties which are active on the Facebook social network. The thesis compared their influence and activity. A new way of activity and influence prediction has been proposed, based on the identification of dedicated users.

Klíčová slova

Analýza sociálních sítí, Vliv uživatele, Facebook, Nástroj pro analýzu sociálních sítí

Keywords

Social Network Analysis, User Influence, Facebook, Tool For Social Media Analysis

Citace

JIROUT, Petr. *Určení vlivu uživatelů na sociálních sítích.* Brno, 2017. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Smrž Pavel.

Určení vlivu uživatelů na sociálních sítích

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana doc. RNDr. Pavla Smrže, Ph.D. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Petr Jirout
23. května 2017

Poděkování

Za rady a odbornou pomoc bych chtěl poděkovat mému vedoucímu doc. RNDr. Pavlovi Smržovi, Ph.D. Dále jsem vděčný všem těm, kteří mě během tvorby této práce jakkoliv podporovali.

Obsah

Kapitola 1

Úvod

V dnešní době, kdy sociální sítě prostupují do běžného života všech lidí a pomalu nahrazují tradiční média v oblasti distribuce zpráv a informací, je přirozenou otázku se ptát, kdo je ve skutečnosti vlivným členem těchto sítí. Těmto vlivným jedincům pak nabízí sociální sítě nenahraditelný prostředek, jak přímo komunikovat s jejich cílovými skupinami.

Identifikace těchto jedinců však není jednoduchá. Sociální sítě sice zveřejňují některé základní ukazatele vlivu, jako třeba počet odběratelů či fanoušků — tyto však nejsou dostatečně vypovídající a často s reálným vlivem nekorespondují [22]. Je tedy žádoucí poskytnout široké veřejnosti takový nástroj, který by umožňoval analýzu těchto vlivných jedinců. Tento nástroj nám pomůže pochopit, jací lidé na nás mají vliv ve světě sociálních sítí.

1.1 Cíl a postup práce

Cílem této práce je vytvořit systém pro analýzu sociálních sítí, umožňující porovnání vlivu vybraných profilů. Za tímto účelem byla nastudována problematika vlivu a sociálních sítí. Na tomto základě pak byla vytvořena architektura systému, který byl následně na základě vybraných požadavků implementován. Následně byla funkcionalita tohoto systému prezentována na analýze vybraných profilu ze sociální sítě Facebook.

V této práci se lze setkat s použitím některých anglických termínů, zejména v tabulkách. Anglické verze termínů jsou zvoleny proto, aby korespondovaly s obsahem souborů, které jsou generovány v rámci analýzy. Tyto soubory obsahují anglický text z důvodu zpřístupnění vyvinutého systému lidem z celého světa.

Problematika vlivu a návrhu architektury systému byla řešena v rámci předcházejícího Semestrálního projektu.

1.2 Členění práce

Ve druhé kapitole je rozebrána problematika vlivu z hlediska psychologie a sociálních sítí. Dále jsou v rámci této kapitoly popsány sociální sítě, jejich vlastnosti, reprezentace a metody analýzy.

Ve třetí kapitole jsou nejprve představeny požadavky na implementovaný systém a následně je představena jeho architektura, včetně popisu využitých nástrojů. Poté je popsán použitý formát dat a každá komponenta systému, včetně jejich vstupů, výstupů a principu funkce.

Ve čtvrté kapitole je pak představen vybraný ukázkový případ použití vyvinutého systému. Následně je popsán a jeho analytické výstupy jsou interpretovány a vyhodnoceny.

Kapitola 2

Rozbor řešené problematiky

Tato kapitola tvoří teoretickou část této diplomové práce. Je v ní nejprve rozebrána obecná problematika vlivu, dále pak stručný přehled sociálních sítí. Další část se pak věnuje vlivu v sociálních sítích a existujícím studiím na toto téma. Poslední část pak představuje některé přístupy k analýze sociálních sítí.

2.1 Vliv

Sociální vliv [1] (*Social Influence*) je proces změny chování, emocí nebo názorů jedince. Aby se jednalo o vliv, musí být tento proces způsoben jiným jedincem (nebo skupinou lidí). Kromě sociálního vlivu existují i další vlivy, které ovlivňují chování, emoce nebo názory jedince, např. vliv prostředí — tyto však nejsou sociálního charakteru a tudíž se jimi nebudeme dále zabývat (nejsou pro nás relevantní).

Za první milník v oblasti sociálního vlivu je považován experiment Herberta Kelmana z roku 1958 [34]. Kelman navázal na práci Solomona Asche, který experimentálně potvrdil [16, 17] vliv skupiny na rozhodování jedince — konkrétně jak se člověk ve frontě nechá přesvědčit, že ta jeho fronta je nejkratší, i když je ve skutečnosti nejdelší. Kelman svým experimentem položil základy budoucímu rozvoji této disciplíny, když v něm popsal tři oblasti vlivu dle míry a procesu ovlivnění.

Oblasti vlivu dle Kelmana

1. **Vyhovění** (*Compliance*): zdánlivý souhlas s ostatními, při zachování nezměněného vnitřního přesvědčení.
2. **Identifikace** (*Identification*): ovlivnění někým, koho jedinec respektuje a/nebo ho má rád. Člověk se snaží napodobit svůj model (se kterým se *identifikuje*).
3. **Internalizace** (*Internalisation*): úplný souhlas, nedochází k rozporu mezi vnějším a vnitřním přesvědčením. Jedinec názor/přesvědčení plně přejal a považuje ho za vlastní.

Dalším možným dělením vlivu je dle vnímání člověka, který je ovlivňován a musí změnit nebo přizpůsobit své přesvědčení či chování.

[1] Dále budeme pod pojmem „vliv“ vždy uvažovat „sociální vliv“.

Oblasti vlivu dle vnímání přesvědčovaného

1. **Konformita** (*Conformity*): člověk není přímo zvenčí ovlivňován, nicméně sám mění své chování tak, aby bylo v souladu se skupinou — například proto, aby byl člověk vnímán jako sympatičtější, protože se chová podobně jako ostatní.

2. **Vyhovění** (*Compliance*): člověk je o něco požádán a cítí, že se může sám, po zvážení všech faktorů, rozhodnout, zda žádosti vyhoví či ne.

3. **Poslušnost** (*Obedience*): člověk je o něco požádán, nicméně cítí, že prakticky nemá na výběr — je mu dána pouze zdánlivá možnost volby. Vyskytuje se často v armádních složkách.

2.1.1 Teorie sociálního vlivu

Pro téma této práce je klíčovou teoretickou základnou, z jejíchž principů tato práce vychází, Teorie sociálního vlivu (*Social Impact Theory*). Tuto teorii poprvé publikoval Bibb Latané v roce 1981 [40]. Je v ní ukázáno, jak se dá vliv kvantifikovat, čím je tvořen a jaké prvky ho ovlivňují.

Základem jsou tři složky, které identifikují, jak moc je jedinec ovlivňován. Stejně tak je možné je využít ke zjištění obrácené informace — který jedinec je vlivným. Tyto složky jsou: Sociální vlivy (*Social forces*), Psycho–sociální zákon (*Psychosocial law*) a Zesílení/rozmělnění vlivu (*Multiplication/division of impact*).

Sociální vliv

Tvrdí, že vliv lze vyjádřit vztahem $Impact = f(S \cdot N \cdot I)$. Tento vztah říká, že velikost vlivu je funkcí Síly (*Strength*), Počtu zdrojů (*Number*) a „blízkosti“ (*Immediacy*, ve smyslu obrácené vzdálenosti, tzn. čím blíže, tím větší vliv).

Nyní rozebereme jednotlivé činitele této funkce. Síla (ve smyslu intenzity) určuje, jak moc je daný zdroj vnímán jako vlivný (*influential*). Tato síla se odvíjí od mnoha faktorů a je silně závislá na kontextu jedince, z jehož pohledu je posuzována. Síla (intenzita) zdroje může vyvěrat z jeho postavení („autoritativnosti“, např. představitel státní instituce), sociálního postavení (např. milionář) nebo z historie vzájemných vztahů.

Počet zdrojů vyjadřuje počet zdrojů vlivu, které v daný okamžik působí na jedince. Například: pokud některou činnost/vzorec chování (třeba kouření) provádí 90 % vašich přátel, kombinovaný vliv na vás je větší, než když tutéž činnost provádí pouze 10 % přátel.

Blízkost vyjadřuje jak bezprostřední daný vliv je — z hlediska „rušení“ zdroje, tedy zda daný zdroj působí sám nebo působí několik zdrojů naráz, a z hlediska časové dimenze, tedy před jakou dobou zdroj působil. Například vliv z prohry vašeho oblíbeného týmu je silnější těsně po zápase než o dva týdny později.

Tento vztah ukazuje, že na velikosti vlivu, který působí na jedince, se rovnoměrně (vztah je čistě lineární) podílejí všechny složky stejnou měrou.

Psycho–sociální zákon

Tento zákon tvrdí, že vliv se taktéž řídí podle vztahu $Impact = s \cdot N^t$. Tento vztah říká, že počet zdrojů (N) má s narůstajícím počtem stále menší a menší přírůstek vlivu. Jinými slovy, pokud budu stále přidávat zdroje, dodatečný vliv každého následujícího přidaného zdroje bude menší než vliv předcházejícího zdroje. Mocnina t vyjadřuje obecně vyjádřený

vliv a zůstává konstantní, stejně jako škálovací konstanta s (kterou si můžeme představit jako „poddajnost“ jedince k ovlivnění ostatními).

Příkladem reálné situace, popisující tento zákon, je když učitel kárá žáka za kázeňský přestupek. Největší nárůst působeného vlivu je při změně z $N = 0$ (žádný učitel, tedy žák není kárán) na $N = 1$ (žák je kárán právě jedním učitelem). Můžeme vidět, že při dalším zvyšování N (dva, tři, ..., N učitelů) již kárání žáka nepřináší o moc větší vliv. Nutno podotknout, že se stále pohybujeme v oblasti psychologie a tudíž mapování dané přesné matematické funkce nelze uplatnit doslovně. Funkce $s \cdot N^t$ při zvyšujícím se N stále roste, lidské vnímání však nemá lineární průběh. Jeho změna je nejvyšší při přechodu z 0 do 1 (tedy nekárán $\rightarrow$ kárán), dalším nárůstem se ovšem nárůst intenzity prožitku postupně snižuje.

Rozmělnění vlivu

Poslední složka vlivu je popsána vztahem $Impact = f(\frac{1}{S \cdot N \cdot I})$. Tento vztah říká, že pokud vliv působí na více jedinců současně ($N > 1$), vliv na každého jedince je menší, než kdyby ten samý vliv působil na něj samotného ($N = 1$). Ze vztahu je to jasně patrné — se zvětšujícím se jmenovatelem (větší N) se zmenšuje výsledný vliv.

Tento fenomén se nevyskytuje pouze v Teorii sociálního vlivu, ale je široce známý i v dalších oblastech psychologie. Příkladem může být tzv. Efekt přihlížejícího (*The Bystander Effect*) [26]. Ten říká, že čím více lidí je přítomno u nějakého incidentu (kupř. autonehoda), tím menší je pravděpodobnost, že některý z nich (nikoliv každý!) poskytne pomoc. Pomůckou, jak tento efekt v případě potřeby obejít, je adresovat požadavky/příkazy konkrétním lidem. Tedy ne „*Zavolejte někdo sanitku.*“, ale „*Vy v té zelené bundě, zavolejte sanitku.*“. Tehdy je adresovaný jedinec vytržen z kontextu skupiny, kdy se vnímá jako pouhý jeden článek. Reálný vliv, který na jedince působí, se tím rapidně zvýšil, ačkoliv vliv samotného sdělení je stále stejný. Pouze došlo k tomu, že jsme snížili faktor rozmělnění vlivu, pomocí změny velikosti N na $N = 1$, a tím efektivně zvýšili velikost vlivu.

2.1.2 Dopady vlivu na rozhodování

Je zjevné, že člověk je ovlivňován svým okolím, které má vliv na jeho proces rozhodování i jeho výsledek. Nejrůznější entity se pak za cílem ovlivnění našeho rozhodování snaží na nás působit, zpravidla v jejich prospěch. Tento vliv může být cílený zvnějšku (a často nevědomý), například v podobě reklamy a marketingu. Vliv ale můžeme vnímat i „opačně“, a to v případě, že ho sami vyhledáváme. Například tehdy, když se rozhodujeme o koupi nějakého produktu a sami se obracíme na osoby (autority), kterým věříme a žádáme je o radu. V tom případě jsme taktéž ovlivněni, nicméně tentokrát na naší žádost.

Je nutno podotknout, že v případě koupi produktu má názor a doporučení našich blízkých velikou váhu, jak například ukázala práce autorů Sinha a Swearingena [46]. Je v ní ukázáno, že lidé mají tendenci spíše důvěřovat doporučením od lidí, kterým věří (přátelé, rodina), než automatickým systémům na prodejních portálech.

2.2 Sociální sítě

V dnešní době se asi najde málokdo, kdo by o pojmu sociálních sítí, sociálních médií či tzv. nových médiích, nikdy neslyšel. Sociální sítě za dobu jejich existence (cca posledních 20 let) [23] zvládly proniknout do mnoha aspektů našeho každodenního života. Část populace,

zastoupená především příslušníky generací Y a Z, je dokonce považuje za neodlučitelnou součást svého života.

Collins English Dictionary definuje [1] sociální síť následovně:

> *Webová stránka, která uživatelům umožňuje interagovat a komunikovat, typicky žádáním ostatních o přidání do svého seznamu kontaktů, tvořením či přidáním se do skupin zaměřených na jejich zájmy, či publikováním obsahu, ke kterému mají ostatní uživatelé přístup.*

Společnými znaky většiny internetových sociálních sítí je důraz na obsah tvořený uživateli, možnost vytváření profilů, vytváření komunit a propojování jednotlivých uživatelů. Jsou tedy přirozeným rozšířením vrozených lidských vlastností, díky nimž lidstvo od pradávna inklinuje ke tvoření komunit, do nového prostoru internetu. Lidé v reálném světě tvoří nejrůznější komunity, kdy jedinec je často členem několika odlišných komunit (například rodinné, pracovní, sportovní či zájmové) současně. Tato tendence se pak stejnou měrou přenáší do online sociálních sítí v té podobě, že uživatel může být členem několika komunit nejen v rámci jedné sítě, ale i v rámci několika odlišných sítí. Sociální sítě pak často nabízejí podpůrné prostředky pro vyjádření sociálních charakteristik, které lidé znají z běžného světa. Příkladem může být hodnocení, skóre, či karma uživatele, která může vyjadřovat popularitu, důvěryhodnost nebo třeba odbornost — záleží pouze na konkrétní síti.

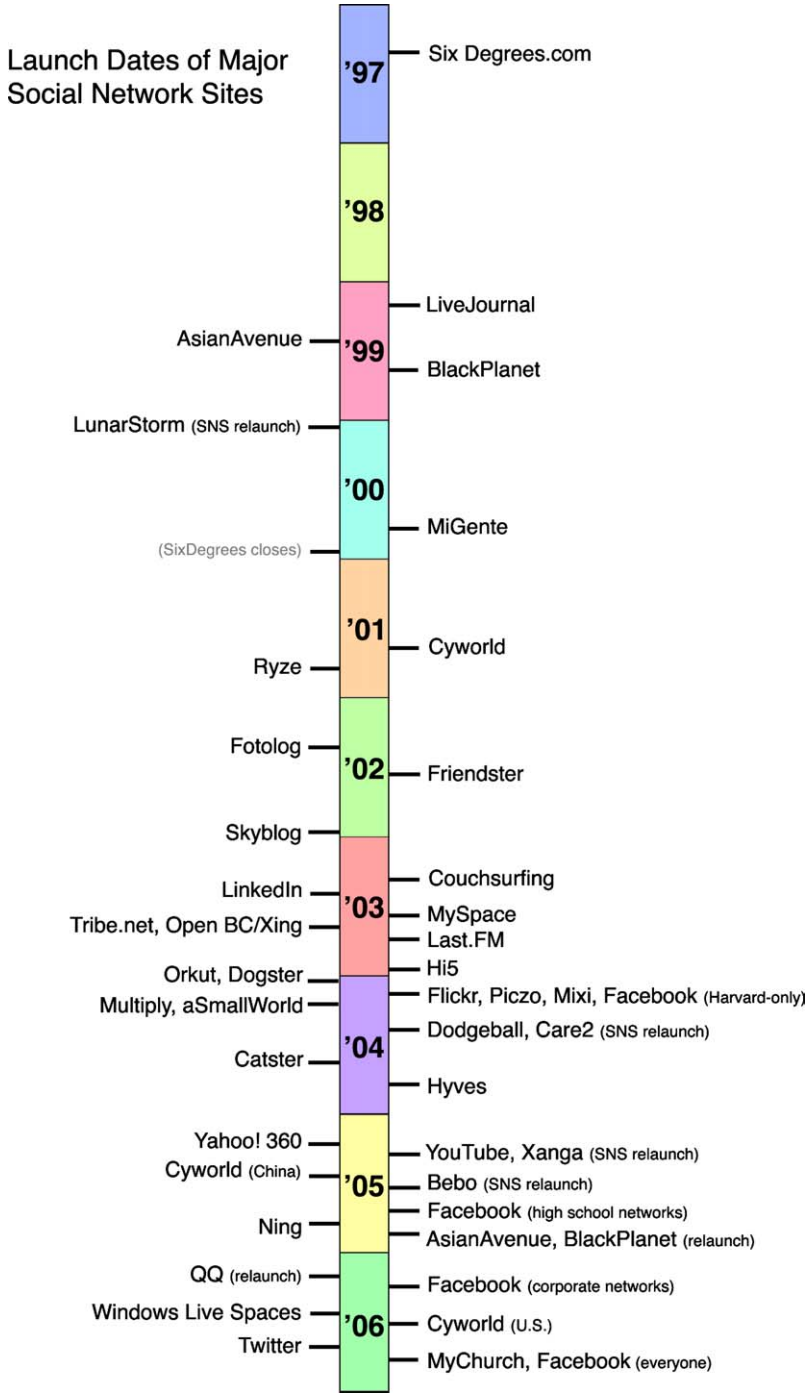
2.2.1 Historie

Za první sociální síť dnešní typu je považována *SixDegrees*, jejíž název vycházel z praktického experimentu [2] Stanleyho Milgrama [41]. Tato síť, vzniklá v roce 1997, umožňovala uživatelům posílat si navzájem zprávy a publikovat příspěvky na „nástěnky“ (*bulletin boards*) ostatních uživatelů, až do vzdálenosti 3 prostředníků. Interakce tudíž vyžadovala buď společné „známé“ nebo přímé spojení. Na svém vrcholu měla 3,5 milionu uživatelů a v roce 1999 byla prodána za 125 milionů dolarů. O dva roky později, v roce 2001, tato sociální síť zanikla.

Později vzniklo a postupně zaniklo ještě mnoho dalších sociálních sítích, z nejznámějších lze zmínit třeba Friendster, MySpace, LinkedIn, Twitter, Facebook a v poslední době Instagram a Snapchat [3]. Časová osa vzniku nejznámějších sociálních sítí je zobrazena na Obrázku 2.1.

[2] V tomto experimentu Stanley Milgram požádal účastníky, aby se pokusili dostat pohlednici k uvedeným adresátům pouze pomocí osobního předání další osobě. Průměrný počet prostředníků, přes který tato pohlednice prošla, byl 6. Pozdější práce Backstroma et al. [18] pomocí analýzy celé sociální sítě Facebook ($\approx$721 milionů uživatelů a $\approx$69 miliard spojení) zjistila, že průměrná vzdálenost (počet prostředníků) libovolných dvou osob je 4.

[3] Instagram a Snapchat jsou sociální sítě sloužící převážně ke sdílení fotografií.



Obrázek 2.1: Časová osa spuštění nejznámějších sociálních sítí do roku 2007. Převzato od Ellison, N.B. et al., 2007 [23].

2.2.2 Současnost

Dnes již není pochyb, že sociální sítě tvoří neoddělitelnou součást života mnoha lidí. Andrew Perrin v roce 2015 ukázal [42], že přibližně 65 % dospělých Američanů používá sociální sítě, oproti pouhým 7 % v roce 2005. Tento podíl uživatelů mezi obyvatelstvem je

ještě markantnější ve věkovém segmentu 18–29, kde plných 90 % populace používá sociální sítě. Zastoupení obou pohlaví mezi uživateli je rovnoměrné, stejně jako jejich etnicita. Lze očekávat, že uvedená čísla za uplynulé dva roky ještě trochu narostla.

V tabulce 2.1 jsou uvedeny přibližné počty uživatelů a hodnota nejznámějších sociálních sítí současnosti.

Jméno sítě	**Počet uživatelů**[4] [mil.]	**Přibližná valuace** [mld. USD]
Facebook	1 860	350
Snapchat	301	25
Twitter	284	14
Instagram	200	1 (cena akvizice v roce 2012)
LinkedIn	106	26 (cena akvizice v roce 2016)

Tabulka 2.1: Počet uživatelů a přibližná hodnota vybraných sociálních sítí.

Zároveň se objevují indicie, že sociální sítě mají i negativní důsledky na jedince. Jednou z hypotéz je, že jejich používání snižuje výkonnost při zpracování úkolů a obecně činí jedince méně šťastnými [21]. Další studie [25] ukázala vliv Facebooku na nálady žen a jejich vnímání svého těla. Pobyt na Facebooku měl negativní dopad na jejich náladu a následně byly více nespokojené se svým tělem a vzhledem obecně.

Nicméně, tyto trendy ještě nejsou dostatečně prozkoumány a není k dispozici přesvědčivé množství důkazů, které by je podpořily. Mimo jiné i proto, že tato oblast zkoumání je nová a můžeme očekávat, že v příštích letech budou studie na toto téma přibývat.

Twitter

Twitter[5] je sociální síť určená pro sdílení krátkých zpráv (nazývaných *tweets*) do 140 znaků. Uživatelé mohou tvořit, sdílet a přeposílat obsah, často tématicky zařazený, pomocí tzv. *hashtags*[6]. Tato síť je charakteristická rychlým šířením informací, které i přes svojí omezenou délku mohou mít značný dopad, viz dále. Zároveň se jedná o jednu z nejvíce navštěvovaných stránek na internetu.

Facebook

Facebook[7] je nejrozšířenější (viz 2.1) sociální síť na světě. Zprvu síť omezená pouze pro studenty Harvard University, byla zpřístupněna celému světu v roce 2006. Od té doby je tato síť jedním z nejvýznamnějších fenoménů a symbolů 21. století. Tato společnost již dávno překonala původní zaměření propojování lidí a v současné době přímo ovlivňuje světové dění. Jednak z pozice šíření informací a zpráv, kdy je Facebook dle některých spekulací u mladší generací mnohem vlivnější, než tradiční tištěná média a televize. Druhak i skrze přímé aktivity společnosti, kdy se jako člen *The Alliance for Affordable Internet* snaží zpřístupnit internet i do oblastí s jeho horší dostupností.

Uživatel Facebooku si po vytvoření profilu může přidávat další uživatele do svého okruhu, čímž je mu umožněno odebírání obsahu, který tito uživatelé produkují, sdílí či

[4] Počet měsíčně aktivních uživatelů.
[5] Dostupné na `www.twitter.com`
[6] *Hashtag* je slovo uvozené symbolem '#'.
[7] Dostupné na `www.facebook.com`

s ním interagují. Dále se může zapojit do nejrůznějších komunit, které pokrývají všechny oblasti lidské činnosti.

2.2.3 Dopad sociálních sítí

Sociální sítě taktéž poskytují velké množství dat pro oblast analýzy lidského chování. Tato data umožňují například předpovědět důvěryhodnost a množství spojení jednotlivých uživatelů na základě informací, které o sobě na této síti vyplní [39]. Dále je možné studovat vzory komunikace mezi jednotlivými uživateli [27]. V neposlední řadě je možné studovat strukturu sítě z hlediska typů uživatelů, například na uživatele pasivní, na uživatele, kteří se aktivně snaží do sítě zapojit nové jedince (*inviters*) a na uživatele, kteří slouží jako zprostředkovatelé pro ostatní (*linkers*) [38].

V posledních letech hrála sociální média a sociální sítě klíčovou roli v několika státotvorných událostech, známé pod názvem Arabské jaro (*Arab Spring*). Tehdy sociální sítě ovlivňovaly a prakticky tvořily politické debaty, které napomohly k šíření demokracie. Výraznou měrou se podílely na výměně vlád v Tunisku a Egyptě a na vypuknutí občanské války v Libyi [35, 32].

Této síly jsou si vědomy i autoritářské režimy, které často přikračují k permanentní či dočasné blokaci sociálních sítí, aby nemohlo dojít k šíření informací, které si režim nepřeje — a tím získání kontroly nad informovaností obyvatelstva, kdy se jedná prakticky o cenzuru. Příkladem může být zablokování Google Earth v Bahrajnu při anexování půdy královskou rodinou [45]. Je proto v nejvyšším zájmu demokracie a svobodného světa, aby sociální sítě zůstaly neregulované a necenzurované.

Bohužel, sociální sítě mají i své stinné stránky. Příkladem může být případ bývalé ředitelky komunikace společnosti IAC, Justine Sacco [44]. Ta v roce 2013 sdílela na Twitteru nevhodný vtip před nástupem do letadla a na konci jejího 11 hodinového letu se stala „hitem“ této sociální sítě. Tisíce uživatelů ji odsuzovaly, proklínaly a urážely za její vtip, přičemž tato reakce na sociální síti vedla k jejímu propuštění ze zaměstnání — a to vše ještě předtím, než její letadlo dosedlo na zem. Tento akt byl jedním z nejvážnějších případů kruté zlomyslnosti a samozvaného vykonávání „spravedlnosti“ anonymními uživateli internetu, které má ovšem dopad na zcela reálné osoby a jejich osudy. Justine se stala obětí tzv. veřejného zostuzení (*Public Shaming*), které bylo zesíleno „stádovým chováním“ (*Mob Mentality*) ostatních uživatelů. Tento případ ukazuje, že lze sociální sítě cíleně použít i k veřejné likvidaci jedinců.

Studie Nicka Hajli ukázala [31], že sociální sítě jsou cíleně používány prodejci ke zvýšení jejich důvěryhodnosti a podporování nákupních tendencí. Tato snaha o sociální interakci se navíc prodejcům vyplácí a vede ke zvýšeným prodejům.

Dalším příkladem možného využití sociálních sítí je vytváření znalostních „databází“, kdy uživatelé tvoří informačně hodnotný obsah [24]. Tento obsah pak komunitě slouží k uchování a šíření vybraných znalostí.

2.3 Vliv v sociálních sítích

Lze snadno ukázat, že Lataného zákony Teorie sociálního vlivu, popsané v sekci 2.1.1, se přirozeně vyskytují a uplatňují i v sociálních sítích na internetu. Zákony této teorie totiž definují vliv obecně, nezávisle na médiu, přes které je tento vliv uplatňován (osobně, či v našem případně po internetu). Díky tomu můžeme použít zákony Sociální, Psycho–sociální i zákon Zesílení/rozmělnění vlivu.

Práce autorů Kim a Srivastava [36] ukázala, že sociální sítě (kromě jiného) mají vliv na chování zákazníků při nákupu, v tomto případě v e-shopech. Jedním z vlivů je například sociální tlak, který mohou uživatelé sociální sítě zažívat při sledování příspěvků svých přátel či vzorů. Pokud vidí, že jejich idol si pořídil například nový telefon či auto, kterým se zde chlubí, mohou cítit vnitřní napětí například kvůli tomu, že jejich telefon je starší nebo auto méně výkonné. Toto napětí je pak může podnítit k tomu, aby si daný (či lepší) objekt koupili sami.

Dalším příkladem ovlivnění v sociálních sítích je záměrné čekání na recenze a zkušenosti prvních uživatelů (*early adopters*). Pokud si jedinec chce koupit nějaký nově uvedený produkt, může čekat na zhodnocení od prvních uživatelů, které sleduje na sociálních sítích. Až na základě těchto informací se pak rozhodne, zda nákup učiní či nikoliv.

Tento efekt je dále násoben faktem, jak moc věříme danému člověku. Míra této důvěry je ovlivněna několika faktory. Například ho můžeme znát osobně nebo ho sledujeme delší dobu a jeho názory rezonují s našimi, či se můžeme spolehnout na externí ukazatele „důvěryhodnosti“. Příkladem může být systém hodnocení na stránkách Amazon [8]. Zde může uživatel obdržet různé „odznáčky“ (například *Top Reviewer*), které jsou spojeny s určitými požadavky na kvalifikaci a které zvyšují jeho důvěryhodnost. Dále mohou ostatní uživatelé hodnotit recenzi daného produktu, zda ji shledávají nápomocnou či nikoliv. V závislosti na počtu uživatelů, kteří již shledali danou recenzi užitečnou, pak vnímáme míru důvěry k dané recenzi — a tím umožňujeme recenzentovi uplatit jeho vliv.

2.3.1 Existující studie

Oblast analýzy sociálních sítí, vlivu či šíření informací je v posledních letech značně aktivní. Publikované studie se zaměřují na různé aspekty sociálních sítí. Lze tedy najít práce z oblasti marketingu, psychologie, sociologie nebo informatiky. Pro téma této práce jsou zajímavé především studie z oblasti šíření informace v těchto sítích či určování vlivu. Tyto poskytly důkazy o přítomnosti vlivných jedinců, kteří mají významný vliv na šíření a distribuci informací [28, 14]. Jejich vliv často vycházel z faktu, že tito jedinci byli propojeni s velkým množstvím ostatních uživatelů dané sítě.

Zároveň bylo na vzorku bloggerů ukázáno, že ti s největším vlivem nemusí být ti nejaktivnější [14]. Cha et al. [22] analyzoval data ze sociální sítě Twitter a porovnal účinek tří uživatelských statistik na vliv a schopnost šíření informace skze tuto síť. Tyto statistiky byly následující: 1) počet spojení uživatele (*indegree*), 2) počet zmínek o něm (*user mentions*) a 3) počet sdílení jeho příspěvků (*retweets*). Objevil korelaci mezi počtem zmínění a přeposláním příspěvků, nikoliv však s počtem spojení. Na tomto případě tudíž můžeme ukázat, že počet následovníků (*followers*) není nejlepším měřítkem reálného vlivu daného jedince.

Zároveň je třeba vzít v potaz, že analýza dat ze sociálních sítí není vždy jednoznačná a různé přístupy mohou z podobných dat vyvodit odlišné závěry. Ukázkou mohou být studie Wenga et al. [47] a Cha et al. [22]. První studie tvrdí, že chování komunit na síti Twitter vykazuje velkou dávku reciprocity, tzn. příslušníci komunity sdílí a komentují příspěvky ostatních členů komunity, kteří jim to poté oplácí. Druhá studie však žádné podobné tendence neodhalila.

Jednou z nejzdařilejších a nejvlivnějších studií je práce Romera et al. [43], která odhalila několik charakteristik propagace vlivu na sociální síti Twitter. Tato studie ukázala,

[8] Více na stránkách Amazon (v angličtině): www.amazon.com/gp/help/customer/display.html?nodeId=201967050

že většina uživatelů funguje pouze jako pasivní konzument obsahu a tedy ho nikam dále nešíří. Vlivný jedinec tak musí nejen dosáhnout (ve smyslu že jsou vidět jeho příspěvky) na ostatní uživatele, ale musí navíc překonat i jejich pasivitu. Dále musí jeho příspěvky být dostatečně originální, kvalitní a musí být publikovány s určitou frekvencí. V této studii byl taktéž představen *Influence–Passivity* algoritmus, který identifikuje vlivné jedince na základě toho, jak aktivně lidé přeposílají jejich obsah.

2.4 Analýza sociálních sítí

Techopedia definuje [11] analýzu sociálních sítí následovně.

> *Analýza sociálních sítí je proces kvantitativní a kvalitativní analýzy sociální sítě. Měří a mapuje informační tok a změny vztahů mezi entitami, které disponují znalostmi. Entity mohou být jednoduché nebo komplexní a jsou tvořeny například webovými stránkami, počítači, skupinami, organizacemi či národy.*

Sociální síť si můžeme představit jako graf uzlů a hran, kde uzly jsou tvořeny uživateli této sítě a hrany reprezentují vztahy mezi těmito uživateli. Tento graf je navíc dynamický a v čase se mění, díky tomu, jak do sítě přibývají noví uživatelé, či jak uživatelé navazují nové vztahy nebo ruší stávající. Díky tomu můžeme studovat nejen jeho statické vlastnosti, ale i vývoj a dynamické trendy. Dále v této části kapitoly rozebereme reprezentaci a různé metody analýzy, které se v současné době používají.

Takto vzniklé analýzy neslouží pouze pro vědecké a výzkumné účely, ale lze je použít k ryze praktickým účelům. Příkladem může být jejich uplatnění v oblasti marketingu, kde se pomocí nich identifikují tzv. *Alpha Users*, což jsou uživatelé, kteří mají vysoký počet spojení s ostatními uživateli a ostatními vlivnými jedinci. Marketing cílený na tyto jedince (případně dohodnutí se s nimi na propagaci) má velkou efektivitu, neboť tito jedinci ovlivní ještě masu dalších, na které marketingový vliv nebyl přímo uplatněn [15]. Dalším příkladem je využití analýz sociálních sítí v kriminologii, kde slouží k identifikování vzorců trestných činů. Základním teoretickým kamenem této aplikace je *Crime Pattern Theory* [19], která vychází z předpokladu, že trestné činy nejsou náhodné. Je možné analyzovat komunikační vzorce jedinců, u kterých je známo, že se dopouští trestné činnosti. Tímto lze odhalit jejich komplice.

Také je možné z informací dostupných na sociálních sítích těžit znalosti (*Social Media Mining*). Těžené znalosti mohou být nejrůznějšího druhu, závisí pouze na cíli, pro který je chceme použít. Takto se dají vytěžit například informace o strukturách komunit, jak se vyvíjí, jací členové je tvoří nebo propagace informací skrze síť. Také je možné analyzovat obsah příspěvků, například analýzou sentimentu. V neposlední řadě pak informace o topologii a struktuře sítě (pro podrobnější popis viz 2.4.2).

2.4.1 Reprezentace sociálních sítí

Jak bylo řečeno výše, nejpříhodnější a nejčastěji používanou reprezentací sociální sítě je graf (z matematické oblasti Teorie grafů). Tento objekt se využívá k reprezentaci množiny jiných objektů a jejich vzájemných vztahů (propojení). Objekt je reprezentován vrcholem a vztahy mezi objekty jsou vyjádřeny jako hrany mezi odpovídajícími vrcholy (objekty). Obecně můžeme graf vyjádřit následovně (2.1).

$$G = (V, H) \tag{2.1}$$

Kde graf G je tvořen neprázdnou množinou vrcholů V a množinou hran H, která je tvořena dvojicemi vrcholů (viz definice 2.2).

$$H \subseteq ((u, v)|u, v \in V, u \neq v) \tag{2.2}$$

V našem případě nejsou v grafu povoleny smyčky, tzn. vztah vrcholu se sebou samým.

Vrchol grafu reprezentuje obecnou entitu konkrétní sociální sítě, nejčastěji její uživatel, stránka nebo účet. Hrany pak představují vztahy mezi těmito entitami, například následování (*following*) nebo „přátelství“ (*friendship*). Tyto hrany nemusí být pouze neorientované (například „přátelství“), ale i orientované (např. následování), kdy vztah má jasně danou orientaci, kdy například uživatel odebírá (*subscribe*) příspěvky jiného uživatele, kterýžto ovšem s prvním uživatelem nemusí mít žádný vztah. Tento vztah je tedy jednostranný.

Z teorie grafů pak můžeme vzít některé vlastnosti obecného grafu a aplikovat je v reprezentaci sociální sítě.

Vlastnosti grafů využitelné v sociálních sítích

Stupeň vrcholu vyjadřuje počet hran, které náleží danému vrcholu. Pro náš účel je taktéž vhodné rozlišovat (pouze v orientovaných grafech) směr, kterým hrana danému vrcholu náleží — zda do něj vstupuje (tehdy se jedná o vstupní stupeň), či z něj vystupuje (výstupní stupeň). Obecná definice stupně vrcholu pro neorientovaný graf je popsána ve vztahu 2.3.

$$deg(u) = |(h \in H|u \in h)| \tag{2.3}$$

V perspektivě sociálních sítí stupeň vrcholu vyjadřuje počet spojení, které má entita (např. uživatel) s ostatními uživateli. Může pak vypovídat o jeho vlivu (pokud má velký vstupní stupeň). Vzhledem k definici stupně vrcholu je pak zřejmé, že tento vliv vyjadřuje pouze bezprostřední vliv na bázi přímé interakce, neboť postihuje pouze ostatní entity se vzdáleností 1, tedy ty, které mají přímé spojení s daným uživatelem.

Cesta je v teorii grafů posloupnost vrcholů, kde každý vrchol má hranu s vrcholem následujícím v posloupnosti. Žádný vrchol ani hrana se v posloupnosti nevyskytují více než jednou. Obecná definice je vyjádřená výrazem 2.4.

$$\begin{gathered} C = (v_0, h_1, \ldots, h_n, v_n) \\ \text{kde } h_i = (v_{i-1}, v_i) \\ \text{a } v_i \neq v_j \text{ pro } i \neq j \end{gathered} \tag{2.4}$$

Pro účel analýzy sociálních sítí je pak možné použít cestu pro sledování distribuce informací nebo vlivu. Pokud je možné vytvořit cestu mezi uživateli (vrcholy), kteří sdíleli (nikoliv nezávisle, ale například přeposláním) tu samou informaci (např. článek), lze takto určit původce této informace, směr a způsob distribuce. Studiem těchto vzorů pak lze vyvodit závěry ohledně vlivu daných uživatelů.

Homofilie (*Homophily* či *Assortativity*) je pojem, který vyjadřuje frekventovanější navazování vztahů mezi uživateli, kteří si jsou v určitých vlastnostech podobní. Příkladem těchto vlastností může být četnější navazování spojení u uživatelů ze stejné geografické oblasti, politického přesvědčení, sociálního statutu, rasy, věku nebo náboženského vyznání. Tento fenomén opět vychází z lidské psychologie, kdy máme větší tendenci někoho zahrnout do svého „kruhu“, pokud máme s dotyčným něco společného. V sociálních sítích pak toto může vést ke vzniku uzavřených komunit, které se brání vstupu odlišných jedinců

a nepřipouští názory odlišné od jejich přesvědčení. Toto chování pak může směřovat až k ostrakizaci jedinců, kteří nevykazují potřebnou míru konformity.

Centralita je metrikou, která slouží k identifikaci nejdůležitějších (nejvlivnějších) vrcholů v grafu. Této metriky existuje několik typů, které se liší v tom, jaké vlastnosti identifikují důležitý vrchol. Nejjednodušším typem je tzv. *Degree Centrality*, která míru důležitosti vrcholu odvozuje od stupně daného vrcholu (čím větší tím důležitější). Dalším typem je *Closeness Centrality*, která určuje důležitost dle průměrné vzdálenosti nejkratší cesty mezi vrcholem a všemi ostatními vrcholy. Nejdůležitější vrchol je pak ten, který má tuto průměrnou vzdálenost nejmenší, tedy je ke všem ostatním nejblíže. Dále je možné měřit *Betweeness Centrality*, která určuje důležitost vrcholu podle toho, jak často se nachází uvnitř nejkratší cesty mezi dvěma libovolnými vrcholy. Nejdůležitější vrchol je pak ten, který „propojuje" co nejvíce ostatních uzlů. Tím efektivně umožňuje komunikaci a v případě potřeby může kontrolovat tok informací.

Koeficient shlukování (*Clustering coefficient*) určuje míru, jak moc se jednotlivé vrcholy mají tendenci shlukovat. Vysoký koeficient znamená, že vrcholy tvoří shluky. Tyto shluky pak identifikují jednotlivé, značně provázané komunity. Tvorba těchto shluků je typická při výskytu homofilie. Existují dvě varianty toho koeficientu: globální, kdy se zkoumají všechny možné trojice vrcholů a lokální, kdy uvažujeme pouze sousedy daného uzlu.

2.4.2 Metody analýzy sociálních sítí

Existuje mnoho metod, jak analyzovat existující sociální sítě. Některé jsou inspirované přístupy z teorie grafů a jiné jsou unikátní pro tuto oblast. Tyto metody se liší především dle účelu, pro který chceme sociální síť analyzovat. Jiné jsou přístupy pro analýzu struktur komunit, propagace informací, analýzu sentimentu, důvěry či v neposlední řadě detekce původců spamu. Vybrané metody taktéž závisí na míře kontroly, kterou chceme mít. Pokud se chceme kontroly do určité míry vzdát, můžeme použít některou, již existující, komerční službu (viz dále). Nevýhodou těchto služeb je jejich cena a především to, že jsou nabízené jako uzavřené skříňky (*black box*) a tudíž nevíme jistě, jak přesně analýza probíhá, jaké faktory jsou do ní zahrnuty a často ani jaká data byla použita — známe pouze konečný výsledek. Tehdy může být lákavé provádět analýzu svépomocí, za využití nejrůznějších existujících nástrojů. Opět záleží na tom, na jakou oblast analýzy se chceme zaměřit — například pokud nám jde především o přesně zacílená data, ale už ne tolik o následnou extrakci znalostí, můžeme si vytvořit vlastní nástroj pro stahování dat a extrakci znalostí nechat na cizím nástroji.

PageRank

Algoritmus PageRank byl poprvé publikovaný v roce 1998 [20] jako jádro tehdy nového internetového vyhledávače Google. Algoritmus původně určený pro měření důležitosti webových stránek však našel mnoho uplatnění i v dalších oblastech — například pro doporučení uživatelů vhodných pro odebírání na Twitteru [29]. Jeho základní princip spočívá v počítání výskytů odkazů na danou stránku a určení jejich kvality (důležitosti). Pokud má stránka vysoký počet kvalitních odkazů, které na ni směřují, je tato stránka hodnocena ve vyhledávání jako důležitější a významnější. Tento algoritmus je iterativní, průběžně tedy upravuje zjištěné informace na základě nejaktuálnějšího stavu (počtu a kvality odkazů). Její aplikace do sféry určování vlivu jedinců typicky spočívá v následující adaptaci. Místo počítání odkazů na stránku se počítají odkazy nebo zmínky konkrétních jedinců. Například zmínění

jejich profilů na sociálních sítích nebo sdílení jejich veřejných příspěvků. Konkrétní aplikace podobného přístupu je v práci Romera et al. [43], viz dále v části o IP algoritmu.

HITS

HITS algoritmus, vyvinutý Jonem Kleinbergem [37], je založen na podobném iterativním přístupu jako PageRank. Hlavním rozdílem je existence tzv. *authorities* a *hubs*, které reflektují skutečnost, že mnoho stránek na internetu slouží jako pouhé agregátory obsahu a odkazů (*hubs*), samy však nemají dostatečnou důležitost. Stránky, které tuto důležitost mají a zpravidla bývají často odkazovány z několika agregátorů, jsou nazývány autoritativními (*authorities*). Každá stránka pak má přiřazeny dvě váhy — první určuje, jak moc je autoritativní a druhá jak moc je agregátorem.

IP algoritmus

IP (*Influence–Pasivity*) algoritmus z práce Romera et al. [43] určuje vliv uživatelů na základě jejich aktivity při přeposílání příspěvků ostatních uživatelů sociální sítě. Vychází z několika následujících předpokladů.

Předpoklady IP algoritmu

- Vliv uživatele závisí jednak na počtu ovlivněných uživatelů, tak i na jejich pasivitě.
- Vliv uživatele závisí na oddanosti ovlivněných uživatelů. Tato oddanost je vyjádřena poměrem času věnovanému obsahu sdíleného daným uživatelem a veškerým dalším obsahem.
- Pasivita uživatele závisí na vlivu uživatelů, kterým je daný uživatel vystaven, ale není jimi ovlivněn. Jinými slovy, čím vyšší je celkový vliv, kterému je uživatel vystaven a na který nereaguje, tím pasivnější tento uživatel je.
- Pasivita uživatele závisí na tom, jak často není ovlivněn ostatními uživateli v porovnání s průměrným uživatelem dané sítě.

Tento iterativní algoritmus pracuje nad orientovaným ohodnoceným grafem, kde uzly reprezentují jednotlivé uživatele, hrana vliv mezi dvěma uživateli A a B a váha hrany reprezentuje vliv uživatele A na uživatele B (pokud je hrana orientována od A k B). Tyto váhy (reprezentující vliv) jsou iterativně počítány nad celým grafem a vyjadřují poměr mezi „přijatým“ vlivem a celkovým „vyzářeným“ vlivem od jednoho uživatele k druhému.

Analýza sentimentu

Dalším aspektem, který lze na sociálních sítích zkoumat, je tzv. *sentiment*. Sentiment je vlastnost přirozeného jazyka, která vyjadřuje postoj, emoce či zabarvení daného úryvku (nejčastěji textu). Určení sentimentu daného textu může identifikovat postoj řečníka k danému kontextu. Například, sentiment komentáře pod článkem může vyjadřovat postoj komentujícího člověka k tématu článku nebo jeho autorovi. Vzhledem ke komplexitě zpracování přirozeného jazyka je často dostačující zařazení sentimentu do z jedné ze tří kategorií (polarita): pozitivní, negativní a neutrální. Pokročilé metody pak dokáží určit postoj a emoci ještě přesněji, a to např. vztek, smutek nebo radost.

Získávání dat

Jedním z nejdůležitějších a zároveň nejkritičtějších kroků analýzy sociálních sítí je pak získávání dat, která jsou následně analyzována. Tento krok je vysoce závislý na konkrétním cíli analýzy a tedy na tom, jaká data chceme získat. V současné době některé sociální sítě poskytují veřejné API[9], skrze které je možné některá data a informace o akcích na sociální síti získat. Tyto informace jsou zpravidla omezené, neboť plný přístup k datům sociální sítě by umožnil poskytnout konkurentům výhodu. Dalším důvodem omezení je pak respekt (či dokonce legislativní povinnost) k soukromí uživatelů dané sociální sítě.

Dalším způsobem, jakým lze data získat, je stáhnout si je z dané sociální sítě přímo. Tato disciplína se nazývá *web crawling*, kde program systematicky a automaticky prochází vybrané stránky a vykonává nad nimi určité akce. Jednou z těchto akcí je *web scraping*, který z dané stránky extrahuje data, která lze uložit a následně, při jejich větším množství, analyzovat. Je nutno podotknout, že velice často jsou tyto techniky v rozporu s podmínkami užití sociálních sítí a jejich praktikování může vést až k nevratnému zablokování veškerého přístupu k dané sociální síti. Nicméně, sociální sítě se proti podobným akcím nikdy nemohou dokonale bránit, neboť striktní omezení přístupu by mělo negativní efekt na běžné fungování dané sítě.

Komerční služby pro analýzu sociálních sítí

V oblasti analýzy sociálních sítí působí mnoho společností s nejrůznějšími produkty. Tyto produkty se liší svou komplexitou, cenou a zaměřením. Obecně tyto služby pomáhají v dosažení komerčních cílů v několika oblastech. Mohou například pomoci v marketingu, kdy tyto nástroje identifikují aktuální trendy, nejlepší čas a místo pro propagaci a nejefektivnější přístup. Dále lze analyzovat aktivitu a reakce u publikovaných příspěvků například pro zjištění, jaká skupina uživatelů na daném sociálním médiu s firmou komunikuje. Je taktéž možné sledovat průběh aktuální kampaně, její ohlas či reakce na uvedení nových produktů. Všechny tyto znalosti pak mohou významně pomoci při plánování budoucích aktivit. V dnešní době, kdy sociální sítě získávají stále větší a větší význam, je pro firmy takřka nutností využívat podobných služeb, pokud chtějí své rozpočty na propagaci utrácet efektivně. Bez těchto nástrojů lze efektivně publikovat obsah v online světě (tak, aby měl kýžený dopad) jen velmi obtížně a prakticky „naslepo“.

Vybrané komerční služby pro analýzu sociálních sítí

- **Keyhole**[10]: tato služba nabízí měření dopadu vybrané značky či klíčového slova na sítích Twitter, Facebook a Instagram. Umožňuje nastavit a sledovat metriky v reálném čase, jako například dosah příspěvků, sentiment a čas největší aktivity. Dále umožňuje filtrovat obsah vlivných uživatelů (*influencers*), včetně jimi podnícené aktivity u ostatních uživatelů. Dále umožňuje sledovat konkurenci a automaticky vytvářet přehledy, včetně historických dat.

- **BuzzSumo**[11]: tento produkt umožňuje najít klíčové vlivné uživatele (*key influencers*) a nejvíce sdílený obsah pro vybrané téma na všech významných sociálních sítích. Tato

[9] API (*Application Programming Interface*): rozhraní pro programovou interakci se systémem, které specifikuje dostupné akce, protokoly a data.

[10] Služba Keyhole je dostupná na www.keyhole.co.

[11] Služba BuzzSumo je dostupná na www.buzzsumo.com.

služba pomůže najít klíčová témata a uživatele, na které pak můžete cílit (nebo jim nabídnout spolupráci), pro co nejefektivnější kampaň.

- **SocialBakers**:[12]: Další z řady analytických nástrojů pro sledování efektivity online aktivit, placených reklamních kampaní a podobně, opět na nejvýznamnějších sociálních sítích. Taktéž umožňuje identifikaci vlivných uživatelů.

- **AgoraPulse**:[13]: služba zaměřená především na sledování angažovanosti (*engagement*) uživatelů. Dále nabízí identifikaci uživatelů, kteří nejčastěji sdílí vybrané příspěvky. Celkově je tato služba cílená na usnadnění interakce s komunitou na sociálních sítích.

- **Followerwonk**:[14]: na rozdíl od předchozích služeb, které jsou zaměřené na více sociálních sítí, Followerwonk je specializovaný pouze na síť Twitter. Nabízí přehledy o aktivitě uživatelů, jejich demografii, kdy a jak často se přihlašují, jaká témata diskutují a mnohem více. Taktéž umožňuje identifikaci vlivných uživatelů pro vybranou oblasti a to na základě tzv. „sociální autority“ (*social authority*).

Žádná z výše uvedených společností neuvádí, jakým způsobem určuje vliv uživatelů (pokud tuto metriku nabízí). Toto je pochopitelné, vzhledem k tomu, že podobné informace jsou součástí obchodního tajemství a jejich držení představuje konkurenční výhodu.

[12] Služba SocialBakers je dostupná na www.socialbakers.com.
[13] Služba AgoraPulse je dostupná na www.agorapulse.com.
[14] Služba FollowerWonk je dostupná na www.moz.com/followerwonk/.

Kapitola 3

Návrh a implementace systému

V této kapitole je popsán návrh a architektura vytvořeného systému pro analýzu sociálních sítí. V první části jsou popsány požadavky na tento systém, ve druhé pak jeho architektura, použité nástroje, datový formát a komponenty systému. V poslední části jsou pak popsány analytické metody, které tento systém využívá.

Většina existujících nástrojů pro analýzu sociálních sítí se zaměřuje především na analýzy grafových vlastností těchto sítí, případně na jejich vizualizaci. Tyto charakteristiky sociálních sítí jsou taktéž důležité, nicméně nejsou jedinými typy informací, které o těchto sítích lze zjistit. Zmíněné nástroje se zabývají analýzou v globálním či větším měřítku, s důrazem na interakce mezi uživateli. Je ovšem možné přistoupit k analýze více cíleně a analyzovat přímo jedince a obsah, který publikují. Nástroj, který by toto umožňoval, by tím otevřel možnost analyzovat sociální sítě s granularitou na úrovni jednotlivců.

Dalším podstatným faktorem, který se musí vzít v potaz, je cenová náročnost existujících komerčních řešení. Tyto drahé nástroje si mohou dovolit například velké mediální skupiny, pro jednotlivce a analytické nadšence však představují značný výdaj a tím efektivně omezují dostupnost informací, které lze podobnými analýzami získat. Další nevýhodou komerčních řešení je nedostupnost syrových originálních dat, ze kterých vznikají výsledné analýzy. Uživatel tak nemá možnost si ověřit prezentované informace vlastním přístupem k analýze nad těmito daty.

Nástroj, který by byl *Open Source*[1] a umožňoval snadnou analýzu sociálních sítí, v současné době chybí. Cílem praktické části této diplomové práce je tvorba právě takového nástroje. Tento nástroj poskytne všem zájemcům o analýzu sociálních sítí základní rozšiřitelnou platformu, pomocí které se mohou ponořit do oblasti analýzy sociálních sítí s minimálními předchozími znalostmi. Umožní analyzovat vybrané uživatele ve zvoleném časovém období, pro nějž zobrazí přehledné grafy a statistiky, které mohou sloužit jako podpora pro zasazení zjištěných informací do sociálního a společenského kontextu.

3.1 Požadavky na systém

Jelikož výše zmíněný cíl je poměrně obsáhlý a jeho plné naplnění by vyžadovalo množství práce mnohonásobně přesahující možnosti jednotlivce a prostor nabízený diplomovou prací, je třeba daný cíl prakticky vymezit a vybrat podstatné vlastnosti a prvky systému, na které se tato práce zaměří.

[1] *Open Source*: program, ke kterému je volně dostupný zdrojový kód, který lze studovat a často i volně modifikovat. Přesné podmínky užití závisí na konkrétně zvolené licenci.

Nejdůležitějším požadavkem je pak uživatelská přívětivost. Toto se odráží i v rozumném výchozím nastavení dostupných parametrů, stejně jako v minimálním počtu nutných parametrů — ideálně pouze jméno uživatele sociální sítě, kterého si uživatel přeje analyzovat. Dalším aspektem uživatelské přívětivosti je minimální počet externích závislostí v tom smyslu, kolik dalších programů je nutné nainstalovat pro normální běh systému. Zároveň je vhodné omezit počet nutných akcí a úkonů na minimum, například nutnost vkládat či stahovat data do systému pouze jednou. Další vlastností je pak možnost spustit tento systém na kterékoliv platformě.

Pro technicky vybavenější uživatele je pak důležitá možnost snadného rozšíření systému, což je podpořeno zveřejněním zdrojového kódu a dokumentací. Zároveň je vhodné systém koncipovat tak, aby bylo snadné použít pouze některý koncepční modul systému (např. pro stahování dat ze sociální sítě) s tím, že o zbytek procesu analýzy se postarají jiné nástroje, dle volby uživatele. Dále je žádoucí, aby systém byl patřičně škálovatelný. Pokud je třeba udělat analýzu nad větším počtem dat, musí existovat snadná cesta, jak posílit výpočetní výkon systému.

Z hlediska obsahu analýz je důležité, aby systém poskytoval, kromě analýzy jednoho uživatele, taktéž možnost pro současnou analýzu a porovnání několika různých uživatelů. Je žádoucí specifikovat rozsah analýzy, ať již například počtem analyzovaných příspěvků či časovým obdobím.

Požadavky na systém

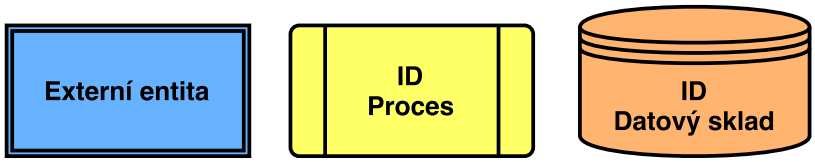
- Analýza všech typů interakce (sdílení, komentář, atd.).
- Možnost výběru časového období dat ke stažení.
- Minimum externích závislostí.
- Volně dostupný zdrojový kód (*Open Source*).
- Dokumentace vytvořeného systému a použitých formátů dat.
- Oddělené části systému zodpovědné za stahování dat, jejich načítání a analýzu.
- Analýza a vzájemné porovnání několika jedinců.
- Minimální správa a údržba systému.
- Škálovatelnost systému.

Cílovými skupinami uživatelů jsou (amatérští) datoví analytici, novináři nebo lidé analyzující aktuální dění. Vyvinutý nástroj předpokládá u uživatele alespoň mírně pokročilou znalost práce s operačním systémem a to na úrovni spouštění programů z příkazové řádky. Pro případné rozšiřování nástroje je pak potřeba i znalost programování.

Jelikož cílem je především vytvoření základní platformy pro analýzu sociálních sítí, bude dostupné uživatelské rozhraní omezeno na příkazovou řádku. Případné vytvoření prezentační vrstvy (například webové) je pak jednoduchou záležitostí, kdy lze napojit funkcionalitu a ovládání vyvinutého nástroje na prvky (například tlačítka) dané prezentační vrstvy a taktéž na stejném místě zobrazovat jeho výsledky, které jsou vždy vytvořeny na lokálním úložišti.

3.2 Architektura systému

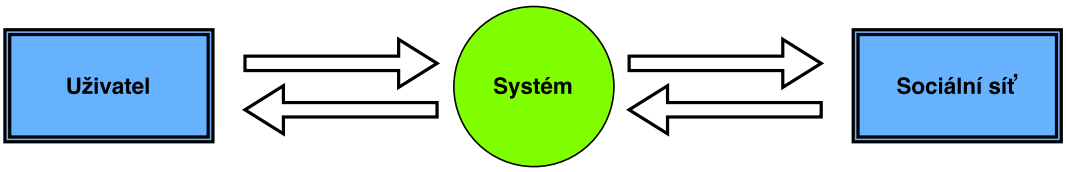
Architektura systému reflektuje požadavky popsané v části 3.1. Pro znázornění datových toků jsou použity diagramy, které nabízí především přehled o struktuře systému z hlediska vstupů a výstupů. Barvy v diagramech jsou zvoleny účelově a jsou konzistentní skrze celou práci. Jejich ukázka je zobrazena na Obrázku 3.1.



Obrázek 3.1: Ukázka použité notace v diagramech.

Obdélník s dvojitou hranou, vyplněný modrou barvou, představuje externí entitu, která stojí vně systému a interaguje s ním. Žlutý obdélník se zakulacenými rohy a dvěma svislými čarami reprezentuje proces, který představuje transformaci vstupů na odpovídající výstup. Nad jménem procesu se nachází identifikátor, který reprezentuje příslušnost v rámci hierarchie celého systému. Například proces s identifikátorem „3.2.1“ je součástí komponenty číslo 3, v rámci subkomponenty číslo 2 a jeho číslo je 1. Oranžový válec představuje datový sklad, kde se ukládají data pro pozdější interakci a analýzu.

Vzhledem k velikosti systému zde nebudou popsány implementační detaily, jako např. metody a členské proměnné tříd, ale bude kladen důraz na koncept celého systému. Z hlediska kontextu je vyvinutý systém zasazen mezi uživatele a jím vybranou sociální síť, jak je znázorněno na Obrázku 3.2. Uživatel dává systému příkazy (např. analýza vybraného uživatele), které jsou následně vykonány a je vyprodukován výsledek. Systém taktéž může, v závislosti na typu požadované akce, interagovat se sociální sítí — kupř. při stahování nových dat.



Obrázek 3.2: Kontextový diagram systému.

Vzhledem k požadavku na všeobecnou dostupnost vyvinutého systému je zdrojový kód publikován ve veřejném repozitáři pod MIT licencí [33].

Běhové požadavky

Pro spuštění a běh systému jsou nutné následující programy. Aktivní prostředí *Python* verze 3.4, do kterého je možné doinstalovat balíčky pomocí `pip` instalátoru (seznam balíčků je součástí systému). Dále je nutné mít nainstalované grafické prostředí pro *Python*, kvůli vykreslování grafů — například `tkinter`, což je standardní GUI balíček pro *Python*. A poslední požadavek je běžící instance *Elasticsearch* verze 1.6.

Návod na instalaci a nastavení systému se nachází v Příloze C.

3.2.1 Python

Jako hlavní implementační jazyk systému byl zvolen *Python 3*. Ten patří mezi nejpopulárnější programovací jazyky [10], ať již z hlediska počtu lidí, kteří ho ovládají, tak i z hlediska jeho oblíbenosti. Mezi datovými analytiky je často navíc první volbou mezi programovacími jazyky. Jeho zvolení tak bylo přirozené právě kvůli oblíbenosti mezi cílovou skupinou, tak i díky jeho rychlé učící křivce a zároveň jeho srozumitelnosti. Taktéž je tento jazyk multiplatformní, běh systému tudíž není omezen na konkrétní platformu. Jednou z mnoha dalších výhod je pak existence rozsáhlého a vyspělého ekosystému nástrojů pro datovou analýzu, zpracování dat a frameworků pro nejrůznější účely.

Verze 3 tohoto jazyka byla zvolena z několika následujících důvodů. Všechny operace nad řetězci a řetězce samotné jsou v této verzi, na rozdíl od verze 2, v kódování *Unicode* [2]. To je důležité především proto, že analyzované texty ze sociálních sítí mohou (a v případě češtiny obsahují) znaky, které jsou mimo standardní ASCII tabulku. Při nevyužití verze 3 by se pak mohly spontánně objevovat chyby způsobené nesprávným kódováním řetězců, ať již v systému samotném, nebo v komponentách třetích stran. Dalším důvodem je fakt, že podpora verze 2 končí v roce 2020 a následná migrace na verzi 3 by si vyžádala některé nutné úpravy. V současné době [3] tak není jediný důvod, proč zvolit *Python 2* místo *Python 3*.

V rámci implementace systému (tzn. ve zdrojovém kódu a v dokumentaci) je vždy výhradně použit anglický jazyk. Je použit z toho důvodu, že tento systém je zpřístupněn jako *Open Source*. Angličtina je de facto standardem v podobných projektech, neboť umožňuje použití systému lidem z celého světa. Všechny metody jsou přímo v kódu dokumentované ve formátu *reStructuredText* [4], který je následně automaticky zpracován programem *Sphinx* [9] a následně vytvořena dokumentace.

3.2.2 Elasticsearch

Jako technologie pro hlavní datový sklad, který slouží k uchovávání a interakci s daty získanými ze sociálních sítí, byl vybrán *Elasticsearch* [3]. *Elasticsearch* je distribuovaný *Open Source* nástroj pro práci s textem. Je to NoSQL databáze [5], nemá tudíž pevné schéma („tabulky“ v SQL terminologii) pro záznamy, které do ní lze vložit. Nabízí REST [6] rozhraní pro interakci a oficiální knihovny pro interakci z rozšířených programovacích jazyků. *Elasticsearch* je vyvíjen v jazyce Java a je schopen běhu na většině existujících platforem. Důvody pro zvolení této technologie hlavním datovým skladem jsou následující.

Elasticsearch nabízí pokročilou podporu pro vyhledávání, filtrování a ukládání obecných nestrukturovaných textových dat. Toto přesně odpovídá typu dat, který lze obecně ze sociálních sítí získávat. A právě tato podpora pro operace s těmito daty je důležitou

[2] *Unicode* je mezinárodní standard pro kódování, reprezentaci a zpracovávání textu většiny písem, které se na světě vyskytují. Nejčastější implementací tohoto standardu jsou kódovací formáty UTF-8 a UTF-16.

[3] Dříve se mohlo stát, že některé balíčky nebyly pro verzi 3 dostupné. Tento stav je již v současné době téměř minulostí, s několika výjimkami.

[4] *reStructuredText*: značkovací jazyk pro textová data s podporou automatického generování dokumentace, kdy některé znaky mají speciální sémantiku. Je ekvivalentem dokumentačního formátu *JavaDoc* (používaný v programovacím jazyce *Java*) pro *Python*.

[5] NoSQL databáze poskytují podobnou funkcionalitu jako SQL databáze, tedy ukládání a poskytování dat, nicméně data nejsou pevně modelována vztahy mezi tabulkami jako v relačních databázích, ale pouze volně samotnými daty.

[6] REST (*REpresentational State Transfer*): přístup pro interakci mezi systémy, který používá standardní HTTP metody a je bezstavový.

součástí vyvíjeného systému, neboť umožňuje delegovat operace na specializovaný a široce podporovaný nástroj. Tím je značně limitován možný výskyt chyb, které by se neodvratně vyskytly při implementaci podobné funkcionality ručně. Další výhodou je značná škálovatelnost tohoto řešení, která je vestavěnou součástí *Elasticsearch*. Je možné provozovat jednu velkou instanci (výpočetní cluster) na několika strojích — to vše naprosto transparentně pro koncového uživatele. V případě, že by uživatel systému chtěl analyzovat velké množství dat v kratším čase, stačí, pokud přesune tuto instanci na výkonnější stroj nebo cluster.

Vyčlenění zpracování a vyhledávání textových dat do *Elasticsearch* má ještě jednu podstatnou výhodu. Tím, že tato instance může běžet vzdáleně na jiném stroji, může ji používat více uživatelů současně a také lze její údržbu a správu na někoho delegovat, přičemž tento člověk nepotřebuje přístup k počítači běžného uživatele vyvinutého systému. Tento systém pak slouží jako tenký klient pro interakci a veškeré náročné operace jsou delegovány. Efektivním omezením výkonu systému je pak kromě instance *Elasticsearch* pouze kvalita a šířka pásma internetového připojení.

I přesto, že zdánlivě může jít nutnost použití *Elasticsearch* proti požadavku na minimum externích závislostí, je tento požadavek více než dobře vykoupen, a to výše zmíněnými výhodami.

Jedinou malou nevýhodou *Elasticsearch* je dokumentace. Ačkoliv pokrývá všechny podporované operace, je značně stručná, je v ní málo praktických příkladů a chybí v ní například ukázka komplexnějšího využití více podporovaných operátorů najednou, které jsou vždy popsány pouze jednotlivě. Na druhou stranu je komunita dostatečně velká a aktivní a tak není problém najít praktické návody či řešení problémů online. Stejné problémy se pak vztahují i na oficiální *Python* knihovny pro interakci (`elasticsearch` a `elasticsearch-dsl`), které mají pouze minimální množství příkladů a zároveň nejsou ani dokumentovány všechny podporované metody. Nutno podotknout, že se tento stav průběžně s každou verzí zlepšuje.

V rámci implementace byla použita verze 1.6. Tato verze byla zvolena místo nejnovější pouze proto, že výkonná instance právě této verze je dostupná v rámci výzkumné skupiny, pod níž tato diplomová práce vzniká. Nicméně migrace na nejnovější verzi by neměla vyžadovat mnoho úsilí — v zásadě stačí nainstalovat příslušné verze *Python* knihoven pro interakci.

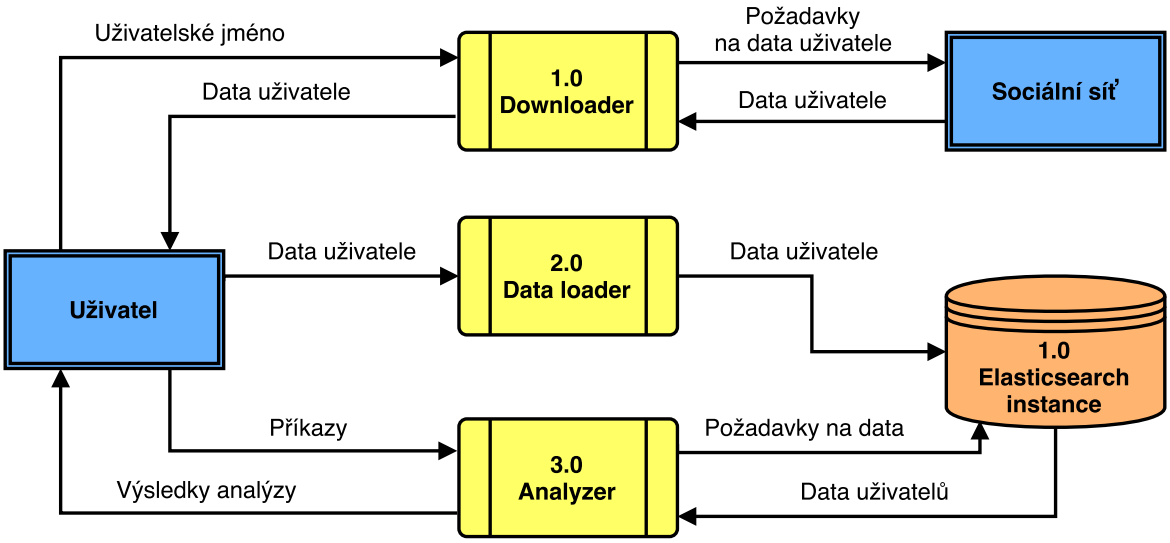
3.2.3 Schéma systému

Na obrázku 3.3 je graficky znázorněno schéma celého systému včetně datových toků. Jak bylo ukázáno na obrázku 3.2, systém je zvnějšku ohraničen dvěma externími entitami — samotným uživatelem systému a poté vybranou sociální sítí. Vnitřek systému se skládá ze čtyř částí, z čehož jedna je datový sklad, implementovaný pomocí *Elasticsearch* (více viz sekci 3.2.2). Další tři části, *Downloader*, *Data loader* a *Analyzer* odpovídají třem krokům klasického analytického procesu, které jsou společné v rámci všech analýz. Tyto části systému jsou uvedeny anglickým názvem z důvodu usnadnění jejich identifikace v kódu.

Kroky analytického procesu

1. Získání dat, v našem případě stažení ze sociální sítě — komponenta *Downloader*.
2. Nahrání dat do systému, v našem případě *Elasticsearch* a komponenta *Data loader*.
3. Analýza: vybrané operace nad daty — komponenta *Analyzer*.

Z architektury je na první pohled vidět důraz na vzájemné oddělení a nezávislost jednotlivých komponent. Toto je v souladu s požadavky na systém, neboť uživatel může pro jednotlivé fáze využít jiné prostředky. Jako příklad lze uvést situaci, kdy si uživatel opatřil data sám a tudíž komponentu pro stahování nevyužije. V případě, že uživatel chce využít všechny komponenty, je posloupnost akcí následující. Uživatel poskytne jméno uživatele, jehož chce analyzovat, komponentě *Downloader*. Ta ze sociální sítě stáhne všechna relevantní data a uloží je na uživatelův lokální disk. Poté tato data uživatel předá komponentě *Data loader*, která je obohatí o některé další informace a poté je vloží do datového skladu *Elasticsearch*. Jako poslední krok pak uživatel zavolá komponentu *Analyzer*, ze které získá výsledek analýzy.



Obrázek 3.3: Schéma architektury systému s datovými toky.

Tato architektura taktéž zajišťuje snadnou rozšiřitelnost. Pokud například uživatel přidá novou datovou položku nad rámec systému, stačí ji pouze zahrnout do dat, nahrát do systému a následně implementovat *Python* metodu, která nová data využívá. Pokud uživatel chce využít pouze již existující datové položky, stačí mu pouze poslední krok — implementovat patřičnou metodu.

3.2.4 Formát dat

Zde je popsán formát dat, použitý v celém systému. Tento formát dat je generován komponentou *Downloader* a následně zpracován komponentou *Data loader*, která data nahraje do datového skladu. Pokud chce uživatel využít vlastní nástroj pro stahování dat ze sociálních sítí, je nutné, aby poté výsledná data převedl do zde dokumentovaného formátu.

Vybraný formát pro data je otevřený standard JSON [7], který data uchovává v textové podobě, jsou tedy zpracovatelná bez další nutnosti dekódování. Tento standard je široce rozšířený a knihovny pro jeho zpracování existují pro většinu programovacích jazyků. V rámci všech dat lze rozlišit čtyři základní typy: interakce, příspěvek, uživatel a veřejný profil. Typ interakce má ještě několik podtypů, které rozebereme dále. V rámci jednoho souboru

[7] JSON *(JavaScript Object Notation)*: otevřený textový formát pro přenos a serializaci datových objektů. Podporuje základní datové typy, jako pole, řetězec, atd.

na disku mohou být elementy pouze jednoho typu, nelze je tak kombinovat (např. jeden soubor, který obsahuje interakce a příspěvky). Veškeré klíče musí být v kódování UTF-8.

U každého elementu všech těchto typů pak musí být přítomny klíče uvedené v Tabulce 3.1. Nutnost poskytnout unikátní (v rámci sociální sítě) identifikační řetězec `id` existuje z důvodu usnadnění interakce s daty. Jednak může uživatel přímo přistoupit na element, který ho zajímá a druhak se to pozitivně projeví na výkonu *Elasticsearch*. Pokud by uživatel `id` neposkytl, byl by vytvořen unikátní hash, který by sloužil jako klíč. Toto by však komplikovalo vyhledávání mezi elementy a porušilo vztahy mezi datovými elementy, kdy například jeden element odkazuje na jiný podle jeho `id` v rámci sociální sítě. Nutno podotknout, že většina sociálních sítí tento klíč automaticky generuje a jejich API ho poskytuje — toto omezení tudíž na uživatele neklade žádné zvláštní a náročné požadavky (jako by byla nutnost toto `id` generovat).

Přesné názvy klíčů v rámci API jednotlivých sociálních sítí se mohou lišit, jejich sémantika by však měla být dostatečně výmluvná. Právě proto je vždy třeba stažená data transformovat na zde použitou terminologii.

Jméno	Datový typ	Popis
`id`	`string`	Identifikační řetězec unikátní v rámci dané sociální sítě.
`origin`	`string`	Název sítě, ze který daný element pochází. Např. „`facebook`“.

Tabulka 3.1: Povinné klíče pro všechny datové typy.

Datový typ interakce reprezentuje interakce uživatelů sociální sítě s příspěvkem. Tato interakce má několik podtypů: komentář (`comment`), sdílení (`share`) a „líbí se“ (`like`). Každý tento podtyp vyjadřuje charakter interakce. v Tabulce 3.2 jsou uvedeny možné klíče pro datový typ interakce. Každý datový soubor s interakcemi musí mít v názvu jeden z následujících řetězců: „`ints`“, „`interactions`“ nebo „`interaction`“.

Datový typ příspěvek představuje příspěvek, který sledovaný uživatel publikoval na svém veřejném profilu. V Tabulce 3.3 jsou uvedeny podporované klíče pro tento typ. Každý datový soubor s příspěvky musí mít v názvu buď „`posts`“ nebo „`post`“.

Datový typ uživatel vyjadřuje informace o uživateli, který uskutečnil interakci. V Tabulce 3.4 jsou uvedeny podporované klíče pro tento typ. Každý datový soubor s uživateli musí mít v názvu řetězec „`user`“.

Datový typ veřejný profil vyjadřuje informace o veřejném profilu, jehož data a příspěvky chce uživatel našeho systému analyzovat. V Tabulce 3.5 jsou uvedeny podporované klíče pro tento typ. Každý datový soubor s veřejnými profily musí mít v názvu řetězec „`user_page_info`“.

Ukázky elementů těchto datových typů jsou v Příloze B.

Jméno	Datový typ	Popis
`message`	`string`	Textový popis této interakce. Např. text komentáře. Nepovinné.
`status_id`	`string`	`Id` příspěvku, jehož se tato interakci týká.
`status_author`	`string`	`Id` autora, který daný příspěvek publikoval.
`author`	`string`	`Id` autora této interakce.
`type`	`string`	Typ interakce. Jeden z `comment`, `like`, `share`.
`created_time`	`string`	Čas publikování této interakce ve formátu ISO. Např. „`2017-04-25T12:25:28+0000`“.
`like_count`	`int`	Počet „líbí se“ na této interakci. Nepovinné.

Tabulka 3.2: Možné klíče pro datový typ „interakce“.

Jméno	Datový typ	Popis
`message`	`string`	Text příspěvku.
`link`	`string`	Odkaz, který byl v příspěvku. Nepovinné.
`author`	`string`	`Id` autora příspěvku.
`status_type`	`string`	Typ příspěvku. Nepovinné.
`created_time`	`string`	Čas publikování příspěvku ve formátu ISO. Např. „`2017-04-25T12:25:28+0000`“.
`share_count`	`int`	Počet sdílení tohoto příspěvku. Nepovinné.

Tabulka 3.3: Možné klíče pro datový typ „příspěvek“.

Jméno	Datový typ	Popis
`name`	`string`	Celé jméno.
`first_name`	`string`	Křestní jméno uživatele.
`last_name`	`string`	Rodné jméno uživatele.
`link`	`string`	Odkaz na profil uživatele.

Tabulka 3.4: Možné klíče pro datový typ „uživatel“.

Jméno	Datový typ	Popis
`name`	`string`	Jméno profilu.
`link`	`string`	Odkaz na profil uživatele.
`fan_count`	`int`	Počet uživatelů kteří dali „líbí se“ tomuto profilu.
`talking_about_count`	`int`	Počet zmínek o profilu v poslední době. Nepovinné.
`is_author`	`bool`	Zde je profil autorem některých příspěvků. Musí být `true`.

Tabulka 3.5: Možné klíče pro datový typ „veřejný profil“.

3.2.5 Downloader

Tato komponenta zajišťuje stahování dat vybraných uživatelů z dané sociální sítě. Tato data jsou stažena přímo k uživateli na jeho lokální disk a to z několika důvodů. Prvním důvodem je zajištění oddělení jednotlivých komponent, kdy (jak je zmíněno výše) uživatel může získat data jiným způsobem, než pomocí této komponenty. Tento princip funguje i obráceně, kdy uživatel může ze systému využít pouze tuto komponentu ke stažení dat, která ho zajímají a která poté analyzuje vlastními prostředky. Dalším benefitem je fakt, že takto se vytvoří záloha dat na lokálním disku a data nejsou uložena pouze v datovém skladu, v případě že by se do něj přímo vkládala. Toto urychluje i případnou obnovu datového skladu, neboť stažení dat je časově náročnou operací, která se navíc může potýkat s omezením na straně sociální sítě, kdy je například omezen počet požadavků na časový rámec.

Vzhledem k poslání této komponenty je to zároveň jediná část systému, která je provázaná s konkrétní sociální sítí. Tento fakt plyne již z definice, neboť každá sociální síť má svá data jinak uspořádaná, pojmenovaná a dostupná. Proto je vždy nutné tyto rozdíly mezi sítěmi reflektovat a v rámci stahování dat používat jednotný formát a terminologii. Tento formát je podrobněji popsán v sekci 3.2.4. Proto je tato součást jedinou variabilní složkou systému, která se musí upravit či vytvořit pro každou novou sociální síť. Všechny ostatní části systému pak jsou od konkrétních sociálních sítí abstrahovány.

V rámci této diplomové práce byla tato komponenta implementována pro sociální síť Facebook, z důvodů uvedených v další kapitole. Facebook poskytuje *Open Graph API* [5], které umožňuje číst data z této sociální sítě. Pro přístup je třeba vytvořit aplikaci na Facebook Platform a získat autorizační token. Přístup i vytvoření aplikace je bezplatné, jedinou podmínkou je vytvoření účtu na této sociální síti. Ačkoliv má *Graph API* vestavěné limity požadavků na data, nejsou nikde přesně zdokumentované a během vývoje a stahování dat v rámci této práce nebyly nikdy překročeny.

Ačkoliv existují Python knihovny pro přímou interakci s *Graph API* (např. balíček `facebook-sdk`), žádná z nich neposkytovala takový přístup k API, který by bezproblémově zapadal do koncepce systému. Z toho důvodu byly vytvořeny metody pro přímou interakci s *Graph API*, které zjednodušují stahování potřebných dat. Příkladem může být automatické transparentní procházení stránek s výsledky (*paging*), v případě, že požadovaná data jsou příliš velká pro přenos v rámci jedné odpovědi a API vrátí kurzor na další stránku s daty.

Uživatel této komponentě předá pouze jméno či ID uživatele, jehož data chce stáhnout. Nejprve jsou staženy informace o tomto profilu (pro pozdější analýzu) a poté jsou průběžně stahovány všechny příspěvky tohoto uživatele. Pro každý příspěvek jsou pak staženy všechny jeho „líbí se“, komentáře a sdílení. V průběhu stahování jsou pak data průběžně ukládána, aby výsledná velikost jednotlivých datových souborů nepřesáhla nastavenou velikost (výchozí hodnota je `20 MB`, lze konfigurovat). Tímto je zajištěna nízká paměťová náročnost běhu programu, kdy není třeba všechna data držet v paměti. Zároveň je anulováno riziko náhlého zatížení disku v případě, že by bylo třeba najednou uložit velké množství dat až na konci běhu.

Tato část analytického procesu je nejzdlouhavější, pro příklad uveďme časovou náročnost stažení všech dat k 1 000 příspěvkům dvou subjektů. Jeden populárnější (107 279 odběratelů) měl celkem 376 746 interakcí ke stažení (324 322 „líbí se“, 50 522 komentářů, 1 902 sdílení). Stažení tohoto souboru dat trvalo celkem 90 minut. Druhý subjekt, méně populární (21 571 odběratelů) měl na stejný počet příspěvků 85 195 interakcí (47 017 „líbí se“, 18 671 komentářů, 19 507 sdílení). Stažení tohoto souboru dat trvalo celkem 45 minut.

I přesto, že počet interakcí byl více jak čtyřnásobně menší, čas byl menší pouze dvojnásobně. Částečným vysvětlením může být potřeba velkého počtu navázání spojení s API, neboť toto API dle REST principů správně neposkytuje jeden bod, na kterém by šlo obsloužit všechny požadavky a tím by se ušetřila režie častého navazování spojení.

Jediné omezení je, že analyzovaný profil musí být veřejný, případně musí mít uživatel takový token, který má dostatečné oprávnění k přístupu a čtení dat z profilu — lze tak analyzovat například svůj osobní profil nebo stránky, jejichž je uživatel administrátorem.

Praktickou ukázku použití této komponenty lze najít v části 4.1.1.

3.2.6 Data loader

Tato komponenta slouží k nahrání stažených dat do datového skladu. Tato data, která mají formát popsaný v části 3.2.4, jsou při nahrávání obohacena o další odvozené položky. Příkladem je přidání klíče `message_len` pro elementy, které obsahují klíč `message`. Ačkoliv tato odvozená data nijak nezvyšují informační hodnotu datového souboru, značně usnadňují některé analýzy. Toto je ukázka přístupu „něco za něco“, kdy vyměníme menší nárůst velikost dat za urychlení výpočtu.

Pokud se v datech nacházejí jiné klíče kromě těch popsaných výše, jsou do datového skladu vloženy taktéž. Toto usnadňuje rozšiřitelnost systému, kdy data nejsou striktně limitována a stačí pouze splnit základní požadavky.

Určení pohlaví ze jména

Pokud je do systému vkládán záznam o uživateli, který obsahuje jeho jméno, je zároveň vloženo jeho pohlaví (pokud není poskytováno přímo sociální sítí). Toto pohlaví je určeno pomocí vytvořeného klasifikátoru, který funguje na principu prostého porovnání s předem vytvořeným slovníkem. Tento slovník byl vytvořen ručně z dostupného seznamu českých jmen a obsahuje 804 záznamů. Prosté porovnání záznamu se slovníkem v tomto případě postačuje, neboť je velice pravděpodobné, že lidé budou svoje jméno mít uvedeno správně a bez chyb, neboť jejich profil na sociální síti slouží zároveň jako jejich vizitka.

Určení sentimentu

Pokud je do systému vkládán textový element (např. komentář nebo příspěvek), je k němu zároveň vložena i informace o obsaženém sentimentu (pro informace o tom, co je to sentiment, viz 2.4.2). Vzhledem k nedostupnosti hotového klasifikátoru českých textů byl v rámci této práce vytvořen klasifikátor vlastní. Pro jeho učení bylo použito 10 000 manuálně anotovaných Facebook příspěvků z datového korpusu skupiny NLP na Západočeské univerzitě v Plzni [30].

Pro vytvoření klasifikátoru byl použit `scikit-learn`, což je knihovna pro strojové učení, datovou analýzu a zpracování textu. Při trénování klasifikátoru se nejvíce osvědčila metoda *Support Vector Machine* s lineárním kernelem. Pro vektorizaci byl použit `TfidVectorizer`. Toto učení bylo provedeno na 9 000 náhodně vybraných elementů z korpusu, přičemž zbylých 1 000 elementů sloužilo k testování přesnosti klasifikátoru. Klasifikace probíhá do čtyř kategorií: `neutral` (neutrální sentiment), `bipolar` (obojetný sentiment), `negative` (negativní sentiment) a `positive` (pozitivní sentiment). Ve vzorcích popisující jednotlivé charakteristiky je použita následující notace: *tp* je počet *true positives*, *fp* je počet *false positives* a *fn* je počet *false negatives*.

Precision udává přesnost vytvořeného klasifikátoru pro testovací soubor (vzorec 3.1). *Recall* udává schopnost klasifikátoru správně určit správné výsledky (vzorec 3.2). *F1-score* lze interpretovat jako vážený průměr *precision* a *recall*, kde 1 je nejlepší hodnota a 0 nejhorší (vztah 3.3). *Support* udává počet elementů odpovídajících dané kategorii v testovacím souboru.

$$Precision = \frac{tp}{tp + fp} \tag{3.1}$$

$$Recall = \frac{tp}{tp + fn} \tag{3.2}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.3}$$

Výsledky učení jsou uvedeny v Tabulce 3.6. Z výsledků je vidět, že klasifikátor je dobrý v určování pozitivního a neutrálního sentimentu (*F1-score* je 0.76 a 0.72 respektive). Naprosto špatné výsledky má pro odhalování obojetného sentimentu (*F1-score* je 0.09). Vzhledem k tomu, že analýza sentimentu je velice komplexní problematikou, která není jádrem této diplomové práce, lze tento klasifikátor použít pro naše účely jako přibližnou aproximaci. Při interpretaci jeho výsledků je však třeba vzít v potaz, že v některých oblastech (obojetný sentiment) klasifikátor není spolehlivý. Pro orientační výsledky tento klasifikátor lze využít, především v oblasti určování pozitivního sentimentu.

Sentiment	Precision	Recall	F1-score	Support
neutral	0.65	0.81	0.72	427
bipolar	0.50	0.05	0.09	41
negative	0.50	0.37	0.42	150
positive	0.80	0.73	0.76	382
Average	0.68	0.68	0.67	1000

Tabulka 3.6: Výsledek učení klasifikátoru sentimentu.

Praktickou ukázku použití této komponenty lze najít v části 4.1.1.

3.2.7 Analyzer

Komponenta Analyzer zaštiťuje tu nejkomplexnější část procesu — datovou analýzu. Pro svou práci využívá data uložená v datovém skladu, která však nijak nemodifikuje. Na základě uživatelských příkazů vykonává analytické úlohy a vrací výsledky — ať již ve formě textového výstupu nebo grafu.

V rámci komponenty je podporován i tzv. komparativní výstup, kdy je možno v rámci jednoho analytického dotazu porovnat několik uživatelů. To je užitečné především v případě grafů, kdy pomocí vykreslení analytických hodnot pro více uživatelů do jednoho grafu lze snadno porovnat jejich charakteristiky. V rámci některých grafů je pak zároveň automaticky vykreslena křivka lineární regrese, aby uživatel mohl snadno zjistit směrovou tendenci dat. Zároveň je možné specifikovat, kolik posledních příspěvků chce uživatel analyzovat — není tudíž nutné provádět výpočetně náročné operace nad velkým počtem dat, pokud uživatele zajímá např. posledních 100 příspěvků.

Konkrétní využité analytické metody jsou podrobněji rozebrány v sekci 3.3. Obecný přístup implementace řešení analytických úloh by se dal popsat následovně.

Přístup k řešení analytických úloh

1. Stáhnutí vybraných dat — filtrování a transformace v rámci *Elasticsearch*.
2. Transformace dat uvnitř systému, jejich porovnání, určení charakteristik a vyvození analytických závěrů.
3. Zobrazení výsledků — v grafu nebo v textovém souboru.

Pro interakci s datovým skladem *Elasticsearch* byly využity *Python* balíčky `elasticsearch` a `elasticsearch-dsl`. Pomocí knihovny `matplotlib` pak byly generovány grafy.

Praktickou ukázku použití této komponenty lze najít v části 4.1.1.

3.3 Analytické metody

V rámci analytických metod byly využity následující přístupy. Tyto metody reflektují fakt, že dostupná data jsou velice omezená. Nejsou omezená množstvím profilů, které lze analyzovat, ale typem dat, která jsou poskytována. Například nyní již není možné zjistit prakticky žádné informace o uživatelích sítě Facebook, kromě jejich jména. Bohužel tak není možné zahrnout demografické údaje do obsahu analýzy, které by poodhalily například rozdíly v lidech aktivních na jednotlivých profilech či jací uživatelé jsou obecně nejaktivnější. Veškerá získaná data jsou tak pouze z veřejně přístupných profilů.

Toto omezení spočívá pouze v dostupnosti dat z veřejných API neprivilegovanému uživateli. Pokud je uživatel schopen opatřit si privilegovaný přístup, potenciální analytické možnosti systému se tím mnohonásobně zvyšují.

Z IP algoritmu 2.4.2 je vypůjčen koncept pasivity a aktivity uživatelů. Na tomto základě je následně určován vliv, který má profil na dané uživatele. Pokud jsou uživatelé aktivnější, má na ně profil větší vliv. Z tohoto algoritmu nebylo možné použít část s využitím poměru toho, co uživatel vidí a na co reaguje. Potřebná data bohužel nejsou k dispozici. Hodnocení aktivity a pasivity je pak založeno pouze na porovnávání mezi jednotlivými profily, kdy lze říci, že jeden profil je vlivnější a jak moc — nelze vliv však přesně kvantifikovat.

V některých generovaných grafech jsou vykresleny křivky lineární regrese s jejich koeficienty. Tyto umožňují snadné určení tendence vykreslených dat — zda rostou či klesají a jak rychle.

Ve statistických textových souborech se pak objevují tři typy informací. První jsou statistiky v poměru na jednoho fanouška daného profilu, druhé jsou v poměru na jeden příspěvek a třetím typem jsou informace o „oddaných“ (aktivních) uživatelích.

Dále je možné zjistit, kolik uživatelů je aktivních u více profilů. Např. jaká část „oddaných“ uživatelů profilu A je zároveň součástí „oddaných“ uživatelů profilu B. Tato data pak umožňují určit, jaké profily oslovují podobné cílové skupiny.

Dalším výstupem analýzy je seznam nejčastěji publikovaných domén. Ze všech odkazů sdílených uživatelem jsou vybrány domény, které jsou následně seřazeny dle frekvence výskytu. Je tak možné zjistit, jaké odkazy jsou na vybraném profilu nejčastěji sdíleny.

3.3.1 Typy interakcí dle míry vlivu

V části 2.1 byly představeny tři míry vlivu dle Kelmana. Tyto míry vlivu (*vyhovění*, *identifikace* a *internalizace*) lze promítnout do analyzovaných typů interakcí. Interakce typu „líbí se“ odpovídá míře vlivu *vyhovění*, která je prvním stupněm míry ovlivnění. Pokud někdo označí příspěvek jako „líbí se“, je tato interakce malým jednoduchým vyjádřením souhlasu, který však nemusí vycházet z hlubšího přesvědčení.

Interakce typu sdílení pak odpovídá druhé úrovni vlivu — *identifikace*. Pokud již uživatel něco sdílí, často tak činí proto, aby explicitně vyjádřil svůj souhlas s obsahem příspěvku a aby jeho okolí tento souhlas vidělo. Z tohoto důvodu lze tento typ interakce vnímat jako typ s mnohem větší vahou, než typ „líbí se“.

Poslední a největší míra vlivu pak není přímo mapována na konkrétní typ interakce, ale na jejich frekvenci. Pokud někdo interaguje (až již označením „líbí se“ či sdílením) u určitého množství příspěvků, lze u něj předpokládat, že se s obsahem publikovaným na daném profilu již plně ztotožnil a lze o něm říci, že odpovídá úrovni ovlivnění *internalizace*. V rámci ukázky byla tato hranice stanovena na 5 % všech příspěvků. Nad touto hranicí již uživatel obsah daného profilu internalizoval. Tuto hranici lze v rámci systému snadno konfigurovat.

Posledním typem interakce je komentář, který záměrně není přiřazen do žádné míry vlivu. Komentář totiž velice často může nést jak vyjádření souhlasné (tedy ovlivnění), tak i kritiku a vymezení se vůči obsahu. Nelze tak jednoznačně určit, zda je komentář projevem souhlasným či nesouhlasným.

Kapitola 4

Vyhodnocení a diskuse

Zatímco v kapitole 3 byl představen návrh a architektura systému, v této kapitole je rozebrán příklad použití tohoto systému pro analýzu vybraných uživatelů sociálně sítě Facebook z politického spektra, včetně následné interpretace analyzovaných dat. V poslední části jsou pak zmíněna potenciální omezení systému a této analýzy.

4.1 Praktické použití systému

Pro ukázku praktického použití implementovaného systému pro analýzu sociálních sítí byla jako cíl vybrána česká politická scéna na sociální síti Facebook. Sít Facebook byla vybrána z toho důvodu, že jsou na ní čeští političtí představitelé dostatečně aktivní, je populární mezi občany a zároveň nabízí dostatečně bohaté API pro získání dat o aktivitě na veřejných profilech a to o všech publikovaných příspěvcích (i historických).

Sociální síť Twitter byla zavržena z toho důvodu, že je v českém mediálním prostoru relativně málo využívaná, politici jsou na ní málo aktivní a navíc neposkytuje API pro přístup k historickým datům — například není možné zjistit, kteří všichni uživatelé sdíleli konkrétní status. Tato data je možná získat z tzv. *Streaming API*, kdy jsou dané informace poskytovány v reálném čase. To však klade značný nárok na stahování dat, které musí probíhat nepřetržitě, aby mělo patřičnou vypovídající hodnotu. V českém prostoru je pak problémem i malá aktivita — sběr dostatečného množství dat by trval neúměrně dlouho (řádově mnoho týdnů nepřetržitého monitorování).

Nicméně, jak bylo řečeno v části 3.2.4, vytvořený systém není omezený na jeden zdroj dat (zde konkrétně Facebook). Lze použít data z libovolného zdroje, pokud budou mít všechny potřebné datové položky.

Vybranými subjekty jsou veřejné profily českých, t.č. parlamentních, stran, plus profily jejich předsedů (jsou-li dostupné). Zároveň byly pro srovnání vybrány dvě neparlamentní strany. Seznam vybraných stran je v Tabulce 4.1. Dále byly vybrány profily několika výrazných osobností na Facebooku, jako je například prezident republiky (t.č. Miloš Zeman). Přehled vybraných osobností je v Tabulce 4.2. Mezi parlamentními stranami chybí strana Úsvit – Národní Koalice, která nemá na svém veřejném profilu dostatečnou aktivitu (průměrně pouze několik interakcí na jeden příspěvek). Mezi vybranými osobnostmi je pak její zakladatel Tomio Okamura se stranou SPD (Svoboda a přímá demokracie).

Pro všechny analyzované profily byla stanovena podmínka minimálního počtu fanoušků a to na 7 500. Toto omezení je z důvodu dostatečné výpovědní hodnoty výsledných analýz,

kdy se omezí potenciální výkyvy způsobené malým počtem fanoušků, které se postupně s vyšším počtem normalizují.

Zkratka strany	Jméno strany
ANO	ANO 2011
ČPS	Česká pirátská strana
ČSSD	Česká strana sociálně demokratická
KDU-ČSL	Křesťanská a demokratická unie – Československá strana lidová
KSČM	Komunistická strana Čech a Moravy
ODS	Občanská demokratická strana
SSO	Strana svobodných občanů
TOP 09	TOP 09

Tabulka 4.1: Vybrané politické strany pro analýzu.

Jméno	Funkce
Andrej Babiš	Předseda ANO
Bohuslav Sobotka	Předseda ČSSD
Miloš Zeman	Prezident ČR
Miroslav Kalousek	Předseda TOP 09
Petr Fiala	Předseda ODS
Petr Mach	Předseda SSO
Tomio Okamura	Předseda SPD
Václav Klaus ml.	Člen ODS

Tabulka 4.2: Vybrané osobnosti pro analýzu.

Pro každý profil bylo staženo 1000 posledních příspěvků. Stahování dat bylo ukončeno 13. 5. 2017 v 19:00, žádné novější příspěvky tudíž nejsou součástí analýzy.

4.1.1 Popis použití

Jak bylo řečeno v části 3.2.3, obecně má analytický proces tři fáze — stažení dat, jejich transformace a nahrání do datového skladu a následně provedení analytických operací nad těmito daty. Těmto třem fázím odpovídají tři komponenty vyvinutého systému. V této části je popsáno a ukázáno jejich praktické použití v odpovídající fázi.

Stažení dat

Komponenta pro stažení dat (popsaná v 3.2.5) je v našem případě omezená na sociální síť Facebook, z důvodů řečených výše. Pro její funkci je nutné poskytnout ID a tajný klíč vytvořené Facebook aplikace pro přístup ke *Graph API*. Tyto informace je možné buď poskytnout jako proměnné prostředí nebo pomocí parametrů skriptu. Dále je třeba poskytnout uživatelské jméno nebo ID uživatele [1], jehož data chceme stáhnout.

[1] Facebook ID uživatele lze získat z adresy jeho profilu například na stránce https://findmyfbid.com/.

Po spuštění se skript připojí ke *Graph API* a ověří, zda daný uživatel existuje a zda je jeho profil přístupný. Následně uloží informace o tomto profilu (např. počet fanoušků) a poté začne systematicky stahovat určený počet příspěvků (výchozí počet je 1000) od nejnovějších, včetně všech potřebných dat. Tato data jsou průběžně ukládána do cílové složky. Zároveň je zobrazován aktuální počet stažených příspěvků.

Ukázka zavolání skriptu pro stažení dat a následný výstup je ve Výstupu 4.1. Celkový objem stažených nekomprimovaných dat pro všechny subjekty je 1,7 GB. V tomto objemu je celkem 6 372 287 interakcí, z čehož je 5 520 540 „líbí se“, 797 229 komentářů a 54 518 sdílení. Počet sdílení nevyjadřuje celkový počet sdílení příspěvků, pouze veřejná sdílení, ke kterým byl poskytnut přístup.

```
1   # Download user data from Facebook
2   $ python3 fb_downloader.py --app-id IDKFA --app-secret IDDQD
        AndrejBabis
3   Downloading data from author:
4       AndrejBabis (id: 214827221987263)
5   Data will be saved into the following directory:
6       ../data/facebook/user_AndrejBabis
7
8   Started at 2017-05-12 18:01:12.263013
9
10  Posts downloaded: 10/1000
11  Posts downloaded: 20/1000
12  <...omitted lines...>
13  Posts downloaded: 990/1000
14  Posts downloaded: 1000/1000
15
16  Finished at 2017-05-12 21:44:53.154347
```

Výstup 4.1: Ukázka stažení dat.

Nahrání dat do datového skladu

Dalším krokem je pak nahrání stažených dat do datového skladu, který je implementován pomocí *Elasticsearch*. Parametry skriptu specifikují soubor či adresář, ze kterého se mají načíst datové elementy a dále pak adresu *Elasticsearch* a index pod který se data mají uložit.

Skript prochází předaný soubor či adresář a vkládá uložené datové elementy do lokální instance *Elasticsearch*, obohacené o několik položek. Ukázka nahrání dat jednoho uživatele je ve Výstupu 4.2.

```
1   # Load data into the data store
2   $ python3 data_loader.py --es-address localhost:9200 --es-index
        xjirou07 ../data/facebook/user_AndrejBabis/
3   Processing data file [1/24]: ../data/facebook/user_AndrejBabis/
        interaction_data_1.json
4       Inserted 76819 'interaction' elements
5   Processing data file [2/24]: ../data/facebook/user_AndrejBabis/
        interaction_data_2.json
6       Inserted 76819 'interaction' elements
```

```
7    <...omitted lines...>
8    Processing data file [24/24]: ../data/facebook/user_AndrejBabis/
         post_data_3.json
9        Inserted 121 'post' elements
```

Výstup 4.2: Ukázka nahrání dat do datového skladu.

Analýza dat

Poslední fází je pak analýza dat. Spuštěním skriptu jsou vykonány analytické operace a dotazy nad datovým skladem. Skript přijímá adresu *Elasticsearch* a jméno uživatele, kterého chceme analyzovat. Pokud nejsou specifikovány žádné přepínače, je spuštěna kompletní analýza v plném rozsahu. Ukázka analýzy dat jednoho uživatele je pak ve Výstupu 4.3. Plná analýza posledních 1000 příspěvků pro jednoho uživatele trvala cca tři minuty na obyčejném počítači (Intel Core i5 2.4 GHz, 4 GB RAM). Podobná analýza ve srovnávacím režimu pro tři uživatele zabrala cca osm minut.

Kompletní analýza byla provedena pro všechny testovací subjekty.

```
1    # Analyze user
2    $ python3 analyze.py --es-address localhost:9200 --es-index xjirou07 -u
         mach.svobodni
3    Results will be stored in the 'stats_mach.svobodni' directory
4    Saving statistics for author mach.svobodni... done
5    Saving plots for specified author(s)... done
6    Saving fan activity statistics for specified author(s)... done
```

Výstup 4.3: Ukázka analýzy uživatele.

4.1.2 Získaná data

V rámci analytické fáze systému vznikly obecně dva typy souborů pro každého analyzovaného uživatele. Prvním typem jsou textové statistiky (například průměrný počet sdílení příspěvku). Druhým typem pak jsou grafy s nejrůznějšími charakteristikami aktivity uživatele na dané sociální síti (např. vývoj sentimentu v komentářích). V této diplomové práci je uveden pouze výběr z těchto dat, kompletní data z analýzy lze nalézt na přiloženém CD.

V následujících tabulkách jsou uvedena nejzajímavější získaná data. Tyto tabulky záměrně obsahují anglické termíny, které korespondují s obsahem generovaných souborů a taktéž s terminologií použitou v části 3.2.4 o popisu datového formátu. První dvě tabulky obsahují průměrné počty různých typů interakcí na jednoho odběratele (fanouška) daného profilu — pro profily politických stran (Tabulka 4.3) a pro profily vybraných jednotlivců (Tabulka 4.4). Následující dvě tabulky obsahují průměrné počty různých typů interakcí na jeden příspěvek — opět pro profily politických stran (Tabulka 4.5) a pro jednotlivce (Tabulka 4.6).

Poslední dvě tabulky pak obsahují informace o počtu oddaných fanoušků.

Průměrný počet interakcí na fanouška

Z Tabulky 4.3 vyplývá několik zajímavých informací o profilech politických stran. Například strana ANO má 103 100 fanoušků, nicméně v průměru jsou tito fanoušci velmi málo aktivní — na jednoho vychází necelých jeden a půl interakce za posledních 1000 příspěvků. I počet

Name	Fan count	Interactions	Likes	Shares	Comments
ANO	103 100	1.48	1.22	0.11	0.26
ČPS	65 496	2.64	2.45	0.62	0.16
ČSSD	21 811	4.19	2.33	1.26	0.93
KDU-ČSL	17 173	4.51	3.67	0.53	0.75
KSČM	7 743	3.41	2.75	0.61	0.57
ODS	41 262	4.11	3.21	0.56	0.86
SSO	71 887	4.18	3.70	1.01	0.46
TOP 09	110 600	3.75	3.23	0.34	0.50

Tabulka 4.3: Průměrný počet různých typů interakcí na jednoho fanouška u profilů politických stran.

„líbí se“ je nejnižší ze všech, stejně tak počet sdílení. Průměrný počet komentářů je druhý nejnižší.

Mírným překvapením je pak profil strany ČSSD, která má druhý největší průměrný počet interakcí a zároveň největší počet sdílení a komentářů. KDU-ČSL má největší průměrný počet interakcí, který je však převážně tvořen typem „líbí se“. Strany SSO a ODS jsou na tom podobně s tím rozdílem, že SSO má větší podíl sdílení. KSČM a TOP 09 jsou na tom z hlediska angažovanosti podobně i přes diametrálně rozdílný počet fanoušků (7 743, respektive 110 600). Strana ČPS je v této tabulce na chvostu, především díky nízkému poměru komentářů a „líbí se“.

Name	Fan count	Interactions	Likes	Shares	Comments
Andrej Babiš (ANO)	121 174	10.18	8.79	0.59	1.37
Bohuslav Sobotka (ČSSD)	16 079	10.00	5.92	1.59	3.25
Miloš Zeman (Prezident ČR)	91 822	7.57	6.73	0.61	0.82
Miroslav Kalousek (TOP 09)	41 078	11.22	9.91	1.38	1.26
Petr Fiala (ODS)	21 823	11.57	9.88	1.58	1.60
Petr Mach (SSO)	26 106	7.89	7.15	1.95	0.71
Tomio Okamura (SPD)	257 017	5.08	4.45	2.85	0.61
Václav Klaus ml. (ODS)	64 639	10.12	9.45	1.36	0.66

Tabulka 4.4: Průměrný počet různých typů interakcí na jednoho fanouška u profilů jednotlivců.

Tabulka 4.4 obsahuje informace o jednotlivcích. Z hlediska průměrného počtu interakcí jsou na tom Andrej Babiš, Bohuslav Sobotka, Miroslav Kalousek, Petr Fiala a Václav Klaus ml. obdobně (10 až 11 interakcí na jednoho fanouška). Diametrální rozdíly jsou však mezi počty fanoušků. Naproti tomu Tomio Okamura má průměrný počet interakcí poloviční, i přes zdaleka největší počet fanoušků. Miloš Zeman a Petr Mach jsou pak ve středu mezi těmito dvěma tábory.

Pokud se však podíváme na skladbu těchto interakcí, zjistíme o něco odlišný příběh. Andrej Babiš má velký podíl typu „líbí se“, ale málo sdílení a střední počet komentářů. Bohuslav Sobotka má malý počet „líbí se“, střední počet sdílení a velký počet komentářů. Miloš Zeman má většinu interakcí z „líbí se“. Miroslav Kalousek a Petr Fiala jsou na tom podobně — velký počet „líbí se“, střední počet sdílení a střední počet komentářů. Petr Mach má střední počet „líbí se“, velký počet sdílení a málo komentářů. Václav Klaus ml. pak velký počet „líbí se“, střední počet sdílení a málo komentářů.

Nejzajímavější informace z této tabulky je složení interakcí Tomia Okamury. Ten, přestože má nejmenší počet „líbí se“, má naopak výrazně největší počet sdílení a málo komentářů.

Průměrný počet interakcí na příspěvek

Name	Fan count	Likes	Shares	Comments	Comment publ. [hours]
ANO	103 100	182.42	16.88	38.56	64.80
ČPS	65 496	157.61	40.01	10.12	12.24
ČSSD	21 811	51.50	27.86	20.47	66.04
KDU-ČSL	17 173	61.03	8.79	12.54	35.81
KSČM	7 743	25.58	5.67	5.28	38.45
ODS	41 262	126.28	22.10	33.84	45.72
SSO	71 887	260.48	71.11	32.44	13.00
TOP 09	110 600	340.39	36.21	52.34	28.86

Tabulka 4.5: Průměrný počet různých typů interakcí na jeden příspěvek u profilů politických stran.

Tabulka 4.5 obsahuje průměrné počty typů interakcí na jeden příspěvek a průměrnou dobu, po které je publikován komentář k tomuto příspěvku. Pro přehlednost je ke každému profilu opět uveden počet fanoušků.

Nejprve popíšeme situaci z hlediska počtu „líbí se“. Na prvním místě je TOP 09, což je očekávatelné, neboť má největší počet fanoušků. Avšak ANO, které má podobný počet fanoušků, má tento počet takřka poloviční. Taktéž se zde ukazuje efektivita jednotlivých příspěvků, která se dá vyjádřit jako počet fanoušků stránky nutných na jedno „líbí se“ u příspěvků. Efektivní strany si vystačí s poměrem okolo 300 na jedno „líbí se“ (SSO, KDU-ČSL, KSČM, TOP 09 a ODS), kdežto ty méně efektivní potřebují okolo 400 (ČPS a ČSSD) a nejvíce jich potřebuje ANO (566). Toto koresponduje s průměrnými hodnotami na fanouška z Tabulky 4.3.

Podobná situace je pak v poměru nutného počtu fanoušků na jedno sdílení, kde je rozdíl mezi nejhorší stranou (ANO – 6 107 fanoušků na jedno sdílení) a nejlepší (ČPS – 807 fanoušků) ještě propastnější. Zajímavá je pak korelace mezi těmito poměry u stran ČPS (807 fanoušků) a SSO (1000 fanoušků) a průměrnou dobou dobou publikace komentáře u příspěvků, kdy jsou téměř shodné.

Tabulka 4.6 pak obsahuje průměrné počty interakcí na jeden příspěvek u jednotlivců. Z hlediska počtu fanoušků potřebných pro jedno „líbí se“ jsou na tom všechny profily podobně (okolo 120 fanoušků na jedno „líbí se“), až na profil Tomia Okamury, který potřebuje

Name	Fan count	Likes	Shares	Comments	Comment publ. [hours]
Andrej Babiš (ANO)	121 174	979.00	65.74	153.04	17.32
Bohuslav Sobotka (ČSSD)	16 079	92.52	24.80	50.80	74.98
Miloš Zeman (Prezident ČR)	91 822	803.74	73.26	97.62	48.84
Miroslav Kalousek (TOP 09)	41 078	396.86	55.14	50.64	65.08
Petr Fiala (ODS)	21 823	209.18	33.41	33.93	26.08
Petr Mach (SSO)	26 106	183.97	50.29	18.23	17.99
Tomio Okamura (SPD)	257 017	1040.43	665.54	142.99	15.70
Václav Klaus ml. (ODS)	64 639	855.69	123.34	59.81	36.05

Tabulka 4.6: Průměrný počet různých typů interakcí na jeden příspěvek u profilů jednotlivců.

průměrně dvojnásobek. I z hlediska ostatních parametrů odpovídají naměřené hodnoty těm průměrným na jednoho fanouška z Tabulky 4.4.

Počet oddaných fanoušků

Name	Fan count	Dedicated fans	% of fan count
ANO	103 100	651	0.63 %
ČPS	65 496	426	0.65 %
ČSSD	21 811	276	1.27 %
KDU-ČSL	17 173	162	0.94 %
KSČM	7 743	120	1.55 %
ODS	41 262	337	0.82 %
SSO	71 887	885	1.23 %
TOP 09	110 600	759	0.69 %

Tabulka 4.7: Počet oddaných fanoušků u profilů politických stran.

Tabulky 4.7 a 4.8 obsahují informace o tzv. „oddaných“ fanoušcích. To jsou takoví fanoušci, kteří interagovali u více jak 5 % příspěvků daného uživatele. Hranice 5 % byla zvolena empiricky, nicméně v rámci vyvinutého systému je možné ji nastavit argumentem skriptu — není tak problém nastavit tuto hranici na libovolnou hodnotu. Tito lidé, kteří zároveň nemusí být přihlášeni k odběru příspěvků daného profilu, tvoří jádro aktivních lidí na vybraném profilu. Systém taktéž poskytuje informace o jednotlivých typech interakcí, tedy např. informace o tom, kolik je „oddaných“ fanoušků pouze v rámci sdílení.

Name	Fan count	Dedicated fans	% of fan count
Andrej Babiš (ANO)	121 174	3 502	2.89 %
Bohuslav Sobotka (ČSSD)	16 079	379	2.36 %
Miloš Zeman (Prezident ČR)	91 822	3 550	3.87 %
Miroslav Kalousek (TOP 09)	41 078	1 331	3.24 %
Petr Fiala (ODS)	21 823	816	3.74 %
Petr Mach (SSO)	26 106	675	2.59 %
Tomio Okamura (SPD)	257 017	4 244	1.65 %
Václav Klaus ml. (ODS)	64 639	3 713	5.74 %

Tabulka 4.8: Počet oddaných fanoušků u profilů jednotlivců.

Z těchto tabulek je opět jasně patrné, že počet fanoušků profilu nekoreluje s počtem aktivních uživatelů.

4.2 Diskuse

V předcházející části byla ukázána některá naměřená data. Tato a další naměřená data budou zasazena do kontextu a interpretována.

Prvním zjevným faktem, který získaná data prokazují, je že počet fanoušků není relevantní odhad vlivu a aktivity na daném profilu. Toto zjištění tak potvrzuje základní premisu této práce a stejné zjištění z práce Cha et al. [22], která to stejné potvrdila pro síť Twitter. Zde bylo to samé potvrzeno z několika úhlů — například u profilu strany ANO, která má vysoký počet fanoušků, ale malou průměrnou interakci (průměrně má dvakrát tolik fanoušků na jedno „líbí se“ a přes 6 100 fanoušků potřebných na jedno sdílení). Jiné profily s menším počtem fanoušků pak dosahují lepších hodnot interakcí – například SSO, která má lehce přes 1 000 fanoušků na jedno sdílení.

V této diskuzi vycházíme z přiřazení různé důležitosti a váhy jednotlivým typům interakcí, popsaného v části 3.3.1. Zjednodušeně, jedno „líbí se“ má menší váhu než jedno sdílení a komentář o vlivu jednoznačně nevypovídá (uživatel může obsah příspěvku schvalovat či ho kritizovat).

4.2.1 Interpretace dat

Z naměřených dat bylo zjištěno, že průměrná aktivita na profilech jednotlivců je několikanásobně vyšší než na profilech politických stran. Tento fakt může souviset se zaměřením a vnímáním sociálních sítí obecně, které slouží především pro kontakt mezi jednotlivci než mezi jednotlivcem a další entitou (např. politickou stranou).

Všechna naměřená data vykazovala silně pozitivní sentiment v komentářích u příspěvků. Tato skutečnost může být způsobena naučeným klasifikátorem, který je v identifikování pozitivního sentimentu o dost lepší (přesnost 80 %) než v identifikování negativního (přesnost 50 %).

Profily politických stran

Z profilů všech stran vychází nejhůře a jako nejméně vlivný profil strany **ANO**. Na velice nízkých úrovních pohybují počty všech typů interakcí v poměru na fanouška. V poměru na jeden příspěvek je pouze průměrná v počtu „líbí se“, ale podprůměrná v oblasti sdílení a průměrná v komentářích. I přes impozantní počet odběratelů tak není schopna uživatele zaujmout a vybídnout k větší aktivitě. Částečným vysvětlením může být nulová aktivita v období mezi volbami, kdy profil opět oživl až v druhé polovině roku 2016. I přesto si strana drží lehce nadprůměrný počet oddaných fanoušků: 651.

Poněkud překvapivě je na tom podobně profil **ČPS** (Piráti), kdy i přes cílení jejich programu na mladší generace (častým bojem proti regulaci internetu), taktéž dosahují malého vlivu a aktivity uživatelů. V průměrných počtech interakcí na příspěvek se již pohybují v lepší polovině vlivu, především díky velkému počtu sdílení. Nicméně počty komentářů a „líbí se“ jsou za poslední roky prakticky neměnné, straně se tudíž pravděpodobně nedaří oslovovat nové uživatele.

Profil **ČSSD** opět trpěl dlouhými výpadky v aktivitě, nicméně je zřejmá rostoucí tendence aktivity a interakcí. Na druhou stranu, rok 2015 byl z tohoto hlediska lepší než dosavadní aktivita okolo přelomu roku 2017. Je možné, že pauza způsobila pokles zájmu o interakci na profilu strany.

KDU-ČSL a **ODS** patří ke stranám, které mají relativně průměrný vliv založený na publikování líbivých příspěvků, nicméně počty sdílení jsou slabé. Slabé jsou taktéž počty oddaných fanoušků (162, respektive 337).

O profilu **KSČM** se toho nedá mnoho dedukovat z důvodu malé aktivity a dlouhé přestávky mezi publikací příspěvků. Toto je zjevně dáno demografií voličů, kteří se stranou komunikují jinými kanály. V poměru na počet fanoušků však tato strana vykazuje průměrný vliv.

Strana **SSO** (Svobodní) má v rámci vybrané sítě nadprůměrný vliv — především díky velkému počtu sdílení. Zároveň je však zde patrný mírný propad od poloviny roku 2016 oproti předchozímu půlroku. Tato strana má taktéž největší počet oddaných fanoušků: 885.

Strana **TOP 09** byla neaktivní od roku 2015, nicméně se začátkem roku 2017 začala publikovat příspěvky a aktivita na jejich profilu má od té doby vzrůstající tendenci. Nicméně podíl sdílení na celkových interakcích zůstává stále malý. I přes pauzu v publikaci má strana vysoký počet oddaných fanoušků: 759.

Je nutné dodat, že ačkoliv mezi profily stran jsou rozdíly, nejsou tak markantní a významné (kromě nejméně vlivného profilu ANO). U profilů stran je taktéž patrné, že jejich publikační aktivita často utichá na dlouhé roky a ožívají pouze v časech okolo voleb. Toto je dobře vidět s počátkem roku 2017, kdy všechny strany opět začínají publikovat a připravují se tak na nadcházející podzimní volby.

Z časů publikace příspěvků plyne, že všechny strany kromě ČPS, KDU a KSČM používají plánovače pro publikování, kdy většina příspěvků je zveřejněna v celou hodinu, menší množství pak v půl. Dalším společným elementem je skutečnost, že strany publikují většinu příspěvků během pracovního týdne a během víkendu pak aktivita značně poklesne.

Profily vybraných jednotlivců

Nejvýraznějším rozdílem mezi profily jednotlivců a těmi stranickými je ten, že ty stranické často omezují svou aktivitu na dlouhou dobu, zejména mezi volbami. Profily jednotlivců jsou na druhou stranu aktivní po celou dobu.

Andrej Babiš je středně vlivný jedinec. Ačkoliv jeho příspěvky mají mnoho „líbí se“ interakcí, jejich míra sdílení je podprůměrná a počet komentářů vysoký. Lze spekulovat, že sdílí líbivější obsah, se kterým se však mnoho lidí neidentifikuje natolik či ho nepokládá za tak důležitý, aby ho sdílelo dále. Počet oddaných fanoušků je vysoký: 3 502. Jeho trend je mírně rostoucí. V souvislosti s vládní krizí[2] došlo k mírnému navýšení průměrného počtu „líbí se“ na příspěvku a zároveň k rapidnímu nárůstu počtu komentářů pod příspěvky — z průměrné hodnoty 200 až k hodnotám okolo 3 000 u některých příspěvků.

Bohuslav Sobotka je nejméně vlivný jedinec na sociální síti Facebook. To je zčásti způsobeno jeho nízkým počtem fanoušků (16 079), ale i nízkým počtem sdílení a „líbí se“. V průměru na fanouška sice dosahuje průměrných hodnot, ale právě ve spojení s nízkým počtem fanoušků to jeho vliv významně omezuje. Počet oddaných fanoušků je velice nízký: 379. Jeho tendence je mírně rostoucí, s velikým nárůstem v poslední době, kdy jeho příspěvky najednou sbírají velké množství „líbí se“, komentářů a sdílení. Tento nárůst časově odpovídá vládní krizi.

Miloš Zeman patří mezi vlivnější jedince, jeho vliv a popularita příspěvků však cca od začátku roku 2016 klesá. Jeho vliv je však založen na typu interakce „líbí se“, s men-

[2] Premiér Bohuslav Sobotka 2. 5. 2017 oznámil, že kvůli kauzám ministra financí Andreje Babiše podá demisi, o tři dny později tuto demisi však stáhl a poté zaslal návrh na odvolání ministra financí. Novým ministrem by se měl stát Ivan Pilný, avšak 22. 5. tato vládní krize stále není vyřešena.

ším podílem sdílení. Jeho příspěvky jsou hojně komentované. Počet oddaných fanoušků je nadprůměrný: 3 550.

Miroslav Kalousek patří mezi průměrně vlivné jedince, který nepublikuje příspěvky frekventovaně. U jeho příspěvků převažují interakce „líbí se“. Jeho trend je mírně rostoucí, s lehce nadprůměrným počtem oddaných fanoušků: 1 331. Jeho příspěvky jsou středně komentované.

Petr Mach patří mezi méně vlivné jedince, a to především díky velkému počtu sdílení. Počet „líbí se“ a komentářů je však nízký. I počet oddaných fanoušků je průměrný (675), lze tak spekulovat, že oslovuje především své oddané fanoušky, mezi nimiž má vliv, ale tento nepřesahuje na další uživatele. Toto tvrzení podporuje i fakt, že za poslední dva roky není vidět žádný nárůst ve sledovaných statistikách.

Petr Fiala je srovnatelně vlivný jako Petr Mach, s podobnými charakteristikami, má však vzestupnou tendenci. Počet oddaných fanoušků je průměrný: 816.

Profil **Tomia Okamury** vykazuje několik zvláštností. Tou první z nich je velmi vysoký počet fanoušků: 257 017. To je více než dvojnásobek druhého nejvyššího počtu, který má Andrej Babiš. Zároveň ale má poloviční průměrný počet interakcí na jednoho fanouška a současně průměrný počet „líbí se“ a komentářů odpovídá hodnotám, které má právě Andrej Babiš. Nabízí se tak otázka, kolik z fanoušků tohoto profilu je skutečných a zda právě ten přebytek do dvojnásobku není tvořen falešnými profily. Dalším zajímavým faktem je množství sdílení, které uživatelé na tomto profilu průměrně produkují. Jak v počtu na fanouška, tak především v průměrném počtu na jeden příspěvek: 665. Z tohoto lze odvozovat velký vliv, který má tento profil. O tom svědčí i velké množství oddaných fanoušků: 4 244. Zároveň je ale třeba zmínit, že právě tento profil má největší klesající tendenci. Valnou většinu svých příspěvků publikuje v 7 a 8 hodin ráno nebo v 16 či 17 hodin odpoledne. V současné době se jedná o nejvlivnější analyzovaný profil. Pokud by se prokázala spekulace, že ne všichni fanoušci profilu jsou reální, pak by se statistiky jeho vlivu ještě zvýšily.

Václav Klaus ml. je vycházející hvězdou mezi analyzovanými profily. I přes začátek publikování teprve před rokem, patří mezi nejvlivnější jedince. S vysokým počtem „líbí se“ jak v přepočtu na fanouška, tak na status, tak i s nadprůměrným počtem sdílení. Jeho obsah je středně komentovaný. Nabízí se tak, že jím publikovaný obsah se často trefí do noty jeho fanouškům (velký počet „líbí se“ a sdílení). I velký počet aktivních fanoušků (3 713) značí o tom, že okolo sebe shromáždil velkou skupinu jedinců, které jeho obsah oslovuje a hodně se s ním identifikují. Zajímavostí je, že většinu svých příspěvků publikuje mezi 6 a 7 hodinou ranní.

Z analýzy jedinců lze usoudit na existenci dvou typů vlivných profilů. Prvním typem je profil, který publikuje velmi líbivé příspěvky, které však v lidech nerezonují natolik, aby je sdíleli. Příkladem jsou například profily Andreje Babiše a Miroslava Kalouska. Druhým typem jsou profily, které mají málo komentářů, střední počet „líbí se“ a vysoký počet sdílení. Tento typ vypovídá silném vlivu, neboť ostatní uživatelé se s příspěvky často identifikují a chtějí se o ně podělit dále. Příkladem takového profilu je Václav Klaus ml.

Ruční analýzou několika příspěvků, které jsou hojně komentované, bylo zjištěno, že často převažují negativní komentáře nad těmi pozitivními, odhadem v poměru 3:1. Komentátoři často kritizují a vymezují se proti obsahu příspěvku či jeho autorovi. Velký počet komentářů tak může signalizovat existenci silné opozice. Příkladem jsou profily Andreje Babiše a Tomia Okamury, které přes velké množství příznivců mají i značné množství opozice, které na těchto profilech působí.

Dále bylo experimentálně zjištěno, že počet oddaných fanoušků koreluje s průměrným počtem interakcí na jeden příspěvek. Tento poměr je cca 3:1, kdy na tři oddané fanoušky

připadá jedna „líbí se“ interakce na příspěvku. Proto může počet oddaných fanoušků sloužit jako přesnější odhad reálného vlivu než pouhý celkový počet fanoušků profilu.

Akvitita uživatelů napříč profily

Systém taktéž podporuje porovnání oddaných fanoušků napříč jednotlivými profily. Ukázalo se, že lidé aktivní (v rámci všech typů interakcí) na profilu strany jsou taktéž aktivní na profilu předsedy této strany. Nejvíce oddaných fanoušků společných pro profil strany je u ANO (Andrej Babiš): 467. Dále je pak ODS (Petr Fiala): 159.

Průnik aktivních jedinců mezi profily pak často dosahoval pouze řádu jednotek, s několika výjimkami. Těmi je průnik profilu Miroslava Kalouska a strany KDU-ČSL (15 lidí), který se dá vysvětlit jeho předchozím členstvím v této straně. Dále pak 23 lidí společných pro Petra Fialu a Miloše Zemana. Nejpřekvapivější skupinou je pak průnik 80 lidí aktivních jak na profilu strany Svobodných (SSO), tak i na profilu Tomia Okamury. 70 z těchto lidí označilo „líbí se“ u více jak 5 % analyzovaných příspěvků obou profilů.

Z hlediska komentářů je pak 11 společných komentátorů na profilu Miloše Zemana a Bohuslava Sobotky. Dalším zajímavým zjištěním je, že Václav Klaus ml. má společné oddané fanoušky (celkem 14) se stranou TOP 09, s jeho stranou ODS však nikoliv.

4.2.2 Porovnání s ostatními přístupy

I v rámci prostředí českých sociální sítí se již objevily přístupy, jak analyzovat vliv uživatelů. Příkladem je identifikace 77 nejvlivnějších Čechů, kterou provedl časopis Forbes [4]. Ten analyzuje profily osobností na několika sociálních sítích (Facebook, Twitter, Instagram a YouTube) a sleduje aktivitu dané osobnosti a jejích fanoušků, stejně jako jejich počet. Jejich přesná metodika bohužel není známá, není tudíž možné zjistit, jakou váhu mají jednotlivé faktory.

Další takovou službou je *Yeseter*, která byla například využita pro měření nesnášenlivosti na českém internetu [6]. Tato služba funguje především na principu monitorování českého internetu, kdy vytváří „katalog“ zmínek vybraných klíčových slov. Díky tomu je pak schopna vyhodnotit, jaká jsou aktuální témata na internetu, včetně jejich připojeného sentimentu. Zároveň je schopna tyto informace kategorizovat a tím zpřesňovat výslednou analýzu. Tato služba je opět uzavřená, uživatel tudíž nemá přesnou kontrolu nad vyhodnocením.

Systém vyvinutý v této práci je naopak plně otevřený a rozšiřitelný a v rámci analýzy poskytuje předzpracované charakteristiky jednotlivých profilů, které pak uživatel může interpretovat dle svého uvážení a jednotlivým elementům přiřadit libovolnou váhu. Další výhodou je, že stažená data jsou uživateli k dispozici a může je dále analyzovat dle své vlastní metodologie.

4.2.3 Limitace

Naměřená data a z nich vyvozené závěry mají pochopitelně několik omezení, která je třeba vzít v potaz. Nelze z nich například vyvozovat odhady volebních preferencí voličů. Svět sociálních sítí má svá demografická specifika, která nereflektují složení všech skupin voličů. Příkladem této diskrepance byla první veřejná volba prezidenta ČR v roce 2013. Tehdy byly sociální sítě výrazně nakloněny ve prospěch kandidáta Karla Schwarzenberga, který však ve volbě prohrál o téměř půl milionu hlasů.

Další potenciálním omezením je existence falešných profilů a tím pádem i falešných interakcí (např. označování „líbí se“), které jsou generovány automatickými systémy. Tyto falešné interakce jsou generovány především za účelem zvednutí popularity, kdy se například populární příspěvky mohou zobrazovat i uživatelům, ke kterým by se normálně nedostaly. Počet fanoušků Tomia Okamury by mohl být částečně způsoben právě falešnými profily, jak naznačuje nízká průměrná míra interakce.

Další limitací je vytvořený klasifikátor sentimentu, který je nerovnoměrně kvalitní pro identifikaci různých typů sentimentu — může tak ovlivňovat výsledné závěry.

Největší limitací jsou pak samotná data, která byla analyzována. Rozsah a typ dostupných dat bohužel v rámci této studie bohužel neumožnil provést některé zajímavé analýzy, týkající se například demografického složení interagujících uživatelů.

Kapitola 5

Závěr

V rámci této diplomové práce byl vytvořen systém pro analýzu sociálních sítí, umožňující určení vlivu vybraných uživatelů těchto sítí. Tento systém byl implementován v jazyce *Python 3* s využitím *Elasticsearch* a následně zveřejněn pod MIT licencí. Každý tak může tento systém studovat, rozšířit a používat. Tento systém je snadno rozšiřitelný, dokumentovaný a nevyžaduje žádnou speciální platformu pro běh.

Zároveň byla provedena ukázková studie vybraných subjektů ze sociální sítě Facebook, která představila praktické použití tohoto systému. Subjekty byly vybrány z české politické scény — jak z profilů politických stran, tak i z jejich jednotlivých představitelů. V rámci analýzy pak byly ukázány rozdíly v těchto profilech, aktivitě fanoušků a především vlivu, které tyto profily mají na ostatní uživatele sítě Facebook.

Dalším přínosem této práce je další nezávislé potvrzení faktu, že počet fanoušků nevypovídá o reálném vlivu daného profilu, ani jeho aktivitě či oblíbenosti mezi uživateli. Bylo ukázáno, že mnohem lepším způsobem predikce aktivity je identifikace tzv. oddaných fanoušků, jejichž počet pak přímo koreluje s popularitou jednotlivých příspěvků.

5.1 Možnosti pokračování

Vytvořený systém a přístup lze pochopitelně dále rozšířit. Možných oblastí potenciálních rozšíření je několik a netýkají se pouze limitací zmíněných v části 4.2.3.

První možností rozšíření je přidání další sociální sítě jako zdroje dat. Funkce tohoto systému byla ověřena na síti Facebook, požadovaná data je však možné získat i z dalších sítí (např. Twitter). Přidání nové sítě obnáší vytvoření či modifikace existující komponenty (více viz 3.2.5) pro stahování dat. Tato nová komponenta pak musí generovat formát dat popsaný v 3.2.4.

Vylepšit obsah generovaných analýz je možné i přidáním více dat. Ať již z hlediska přidání a porovnání více subjektů nebo z hlediska přidání nových polí do dat, například o interagujících fanoušcích. Tato nová data by pak umožnila analyzovat interakce z nového směru, například z pohledu demografie. Bohužel, tato data nejsou snadno dostupná a při jejich získávání je zároveň nutné dbát na podmínky užití a právo na soukromí uživatelů sociálních sítí.

Dalším potenciálním místem pro zlepšení je použití kvalitnějšího klasifikátoru sentimentu. Klasifikátor použitý v této práci je nestejnoměrně účinný pro detekci různých typů sentimentu a tudíž je výsledná analýza sentimentu zkreslená.

Jedním z dalších přístupů ke zpřesnění analýzy vlivu je analyzovat samotný text příspěvků a komentářů. Dalo by se tak například zjistit, jak klíčová slova spouští ve fanoušcích odezvu a zda například korelují výrazy použité v příspěvku s obsahem komentářů.

Literatura

[1] Collins English Dictionary - Complete & Unabridged 10th Edition. Únor 2017. Dostupné na: `http://www.dictionary.com/browse/social-networking-site`.

[2] Elasticsearch: Download and Installation. Květen 2017. Dostupné na: `https://www.elastic.co/downloads/elasticsearch`.

[3] Elasticsearch: RESTful, Distributed Search & Analytics. Květen 2017. Dostupné na: `https://www.elastic.co/products/elasticsearch`.

[4] Forbes – 77 nejvlivnějších Čechů na sociálních sítích 2016. Květen 2017. Dostupné na: `http://77.forbes.cz/`.

[5] The Graph API: The primary way to read and write to the Facebook social graph. Květen 2017. Dostupné na: `https://developers.facebook.com/docs/graph-api`.

[6] Jak měříme nesnášenlivost na českém internetu? Květen 2017. Dostupné na: `http://www.hatefree.cz/blog/790-jak-merime`.

[7] Pip: The PyPA recommended tool for installing Python packages. Květen 2017. Dostupné na: `https://pip.pypa.io/en/stable/`.

[8] Python 3 Setup and Usage. Květen 2017. Dostupné na: `https://docs.python.org/3/using/`.

[9] Sphinx: Python Documentation Generator. Květen 2017. Dostupné na: `http://www.sphinx-doc.org/en/stable/`.

[10] Stackoverflow Developer Survey 2017. Květen 2017. Dostupné na: `https://insights.stackoverflow.com/survey/2017#technology`.

[11] Techopedia.com. Únor 2017. Dostupné na: `https://www.techopedia.com/definition/3205/social-network-analysis-sna`.

[12] Tkinter: Standard Python GUI Package. Květen 2017. Dostupné na: `https://wiki.python.org/moin/TkInter`.

[13] Virtualenv: Tool To Create Isolated Python Environments. Květen 2017. Dostupné na: `https://virtualenv.pypa.io/en/stable/`.

[14] Agarwal, N., Liu, H., Tang, L. et al. Identifying the influential bloggers in a community. In ACM. *Proceedings of the 2008 international conference on web search and data mining*. 2008. S. 207–218.

[15] Ahonen, T. T. a Moore, A. Communities dominate brands. *London: Futuretext*. 2005.

[16] Asch, S. E. Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men. S*. 1951. S. 222–236.

[17] Asch, S. E. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and applied*. 1956, roč. 70, č. 9. S. 1.

[18] Backstrom, L., Boldi, P., Rosa, M. et al. Four degrees of separation. In ACM. *Proceedings of the 4th Annual ACM Web Science Conference*. 2012. S. 33–42.

[19] Brantingham, P. J. a Brantingham, P. L. Environment, routine and situation: Toward a pattern theory of crime. *Advances in criminological theory*. 1993, roč. 5, č. 2. S. 259–94.

[20] Brin, S. a Page, L. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*. 2012, roč. 56, č. 18. S. 3825–3833.

[21] Brooks, S. Does personal social media usage affect efficiency and well-being? *Computers in human behavior*. 2015, roč. 46. S. 26–37.

[22] Cha, M., Haddadi, H., Benevenuto, F. et al. Measuring user influence in Twitter: The million follower fallacy. *Icwsm*. 2010, roč. 10, 10-17. S. 30.

[23] Ellison, N. B. et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*. 2007, roč. 13, č. 1. S. 210–230.

[24] Evans, C., Professor Raymond Hackney, D., Wagner, D. et al. The impact of information technology on knowledge creation: An affordance approach to social media. *Journal of Enterprise Information Management*. 2014, roč. 27, č. 1. S. 31–44.

[25] Fardouly, J., Diedrichs, P. C., Vartanian, L. R. et al. Social comparisons on social media: The impact of Facebook on young women's body image concerns and mood. *Body Image*. 2015, roč. 13. S. 38–45.

[26] Fischer, P., Krueger, J. I., Greitemeyer, T. et al. The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological bulletin*. 2011, roč. 137, č. 4. S. 517.

[27] Golder, S. A., Wilkinson, D. M. a Huberman, B. A. Rhythms of social interaction: Messaging within a massive online network. In *Communities and technologies 2007*. [b.m.]: Springer, 2007. S. 41–66.

[28] Goyal, A., Bonchi, F. a Lakshmanan, L. V. Learning influence probabilities in social networks. In ACM. *Proceedings of the third ACM international conference on Web search and data mining*. 2010. S. 241–250.

[29] Gupta, P., Goel, A., Lin, J. et al. The Who to Follow Service at Twitter. In *Proceedings of the 22Nd International Conference on World Wide Web*. New York, NY, USA: ACM, 2013. S. 505–514. WWW '13. ISBN 978-1-4503-2035-1.

[30] Habernal, I., Ptáček, T. a Steinberger, J. *Facebook Data for Sentiment Analysis*. 2013. Dostupné na: http://hdl.handle.net/11858/00-097C-0000-0022-FE82-7.

[31] Hajli, M. N. A study of the impact of social media on consumers. *International Journal of Market Research*. 2014, roč. 56, č. 3. S. 387–404.

[32] Howard, P. N., Duffy, A., Freelon, D. et al. Opening closed regimes: what was the role of social media during the Arab Spring? 2011.

[33] Jirout, P. *Tool for social media analysis*. 2017. Dostupné na: https://github.com/pejirout/social-network-analysis.

[34] Kelman, H. C. Compliance, identification, and internalization three processes of attitude change. *Journal of conflict resolution*. 1958, roč. 2, č. 1. S. 51–60.

[35] Khondker, H. H. Role of the new media in the Arab Spring. *Globalizations*. 2011, roč. 8, č. 5. S. 675–679.

[36] Kim, Y. a Srivastava, J. Impact of social influence in e-commerce decision making. In ACM. *Proceedings of the ninth international conference on Electronic commerce*. 2007. S. 293–302.

[37] Kleinberg, J. M., Kumar, R., Raghavan, P. et al. The web as a graph: measurements, models, and methods. In Springer. *International Computing and Combinatorics Conference*. 1999. S. 1–17.

[38] Kumar, R., Novak, J. a Tomkins, A. Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*. [b.m.]: Springer, 2010. S. 337–357.

[39] Lampe, C. A., Ellison, N. a Steinfield, C. A familiar face (book): profile elements as signals in an online social network. In ACM. *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2007. S. 435–444.

[40] Latane, B. et al. The psychology of social impact. *American psychologist*. 1981, roč. 36, č. 4. S. 343–356.

[41] Milgram, S. a Gasteren, L. van. *Das Milgram-Experiment*. [b.m.]: Rowohlt Reinbek, 1974.

[42] Perrin, A. Social media usage: 2005-2015. *Pew Research Center*. 2015.

[43] Romero, D. M., Galuba, W., Asur, S. et al. Influence and passivity in social media. In Springer. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2011. S. 18–33.

[44] Ronson, J. How one stupid tweet blew up Justine Sacco's life. *New York Times*. 2015, roč. 12.

[45] Shirky, C. The political power of social media: Technology, the public sphere, and political change. *Foreign affairs*. 2011. S. 28–41.

[46] Sinha, R. R. a Swearingen, K. Comparing Recommendations Made by Online Systems and Friends. In *DELOS workshop: personalisation and recommender systems in digital libraries*. 2001.

[47] Weng, J., Lim, E.-P., Jiang, J. et al. Twitterrank: finding topic-sensitive influential twitterers. In ACM. *Proceedings of the third ACM international conference on Web search and data mining*. 2010. S. 261–270.

Přílohy

Příloha A

Obsah CD

Elektronická verze textu této diplomové práce se nachází v kořenovém adresáři přiloženého CD, kde je zároveň plakát. Na disku se také nacházejí následující adresáře:

Adresář `src`

- Zdrojový kód vyvinutého nástroje pro analýzu sociálních sítí.

Adresář `tex`

- Zdrojový kód textu této práce ve formátu LaTeX.

Adresář `data`

- Data z ukázkového použití: výstupy z analýzy (grafy, textové soubory).

Příloha B

Ukázky souborů

Zde jsou uvedeny ukázky jednotlivých datových typů a některých generovaných analytických souborů.

B.1 Datový typ interakce

```
{
    "type": "like",
    "id": "L_10154330398362233_43855944703_10154304317834704",
    "status_id": "43855944703_10154304317834704",
    "status_author": "43855944703",
    "author": "10154330398362233",
    "origin": "facebook"
}
```

Výstup B.1: Ukázka elementu datového typu interakce.

B.2 Datový typ příspěvek

```
{
    "id": "43855944703_10155281213979704",
    "status_type": "shared_story",
    "link": "<...omitted link...>",
    "message": "<...omitted lines...>",
    "share_count": 50,
    "author": "43855944703",
    "created_time": "2017-04-30T16:55:02+0000",
    "origin": "facebook"
}
```

Výstup B.2: Ukázka elementu datového typu příspěvek.

B.3 Datový typ uživatel

```
1  {
2      "last_name": "Novak",
3      "origin": "facebook",
4      "id": "1462571663763265",
5      "first_name": "Vladimir",
6      "link": "https://www.facebook.com/app_scoped_user_id
           /1462571663763265/",
7      "name": "Vladimir Novak"
8  }
```

Výstup B.3: Ukázka elementu datového typu uživatel.

B.4 Datový typ veřejný profil profil

```
1  {
2      "name": "Strana svobodnych obcanu",
3      "is_author": true,
4      "link": "https://www.facebook.com/svobodni/",
5      "about": "<...omitted lines...>",
6      "location": {
7          "country": "Czech Republic",
8          "city": "Prague",
9          "zip": "120 00",
10         "street": "Perucka 14"
11     },
12     "talking_about_count": 5617,
13     "fan_count": 71887,
14     "origin": "facebook",
15     "id": "43855944703",
16     "birthday": "02/14/2009",
17     "birthday_format": "MM/DD/YYYY"
18     "name_ascii": "svobodni",
19 }
```

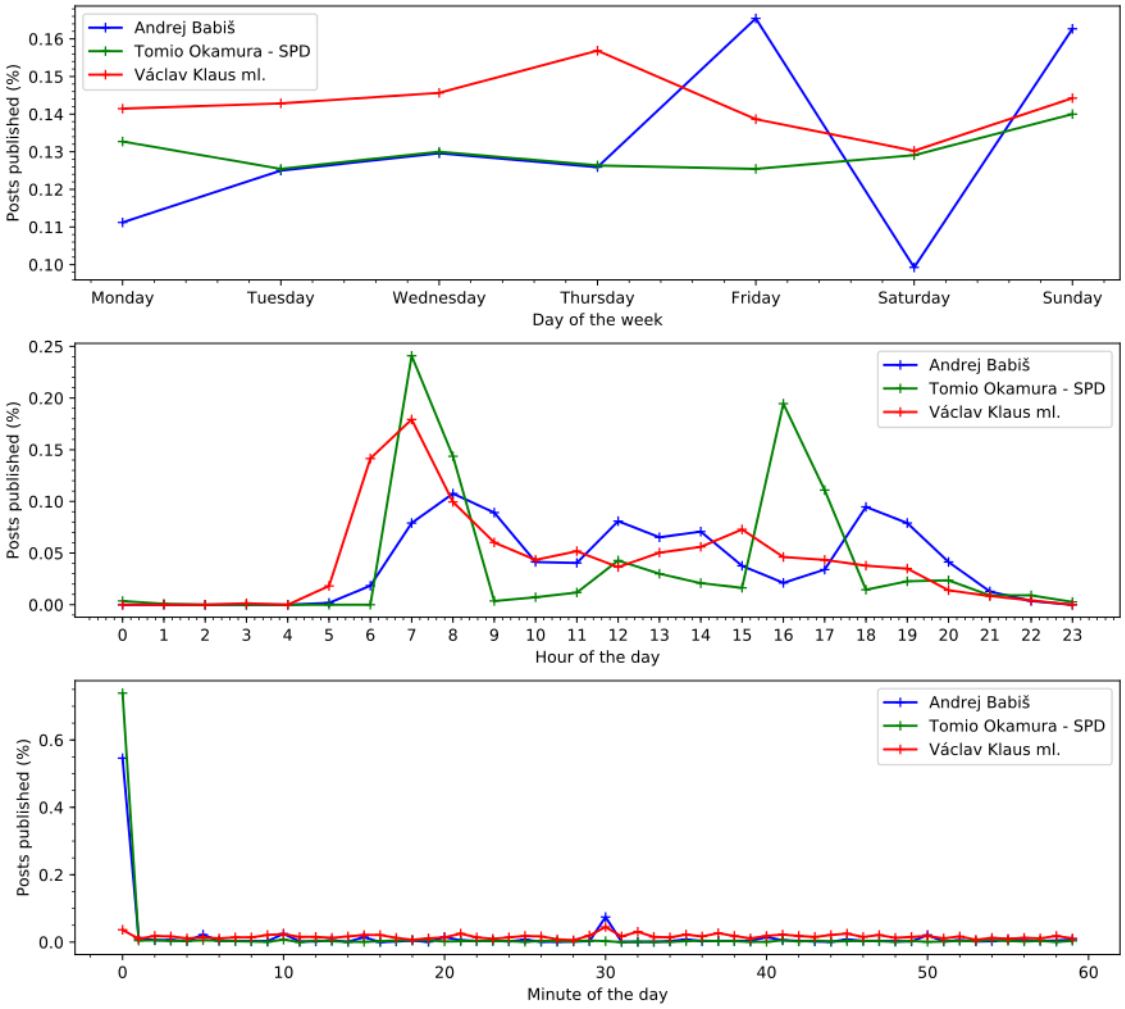
Výstup B.4: Ukázka elementu datového typu veřejný profil.

B.5 Generované statistiky uživatele

```
Author: svobodni
    Fan count: 71887
    Interaction count: 300597
    Interaction per fan: 4.18
    Likes per fan: 3.70
    Shares per fan: 1.01
    Comments per fan: 0.46

Stats of users who interacted on more than 5% of all author's posts
User count: 885 (1.23% of fans)

Liked >5%: 767 (1.07% of fans)
Commented >5%: 5 (0.01% of fans)
Shared >5%: 43 (0.06% of fans)

Liked and commented >5%: 13 (0.02% of fans)
Liked and shared >5%: 0 (0.00% of fans)
Shared and commented >5%: 0 (0.00% of fans)
Liked, commented and shared >5%: 0 (0.00% of fans)
```

Výstup B.5: Ukázka generovaných statistik uživatele.

B.6 Ukázka generovaného grafu



Obrázek B.1: Ukázka grafu distribuce času publikace příspěvků.

Příloha C

Instalace systému

V této části je popsáno prostředí potřebné pro běh systému a jeho instalace. Tento návod je sepsán pro prostředí Linuxu, nicméně postup je podobný i na jiných platformách.

C.1 Elasticsearch

Nejprve je nutné nainstalovat a spustit instanci Elasticsearch verze 1.6. Podrobný návod je k dispozici na oficiálních stránkách [2]. Tato instance slouží jako datový sklad systému.

C.2 Python

Dalším požadavkem je nutnost běhového prostředí Python 3, konkrétně verze 3.4. Postup instalace lze nalézt v oficiální dokumentaci [8]. Pro jednodušší instalaci a izolaci potřebných balíčků od existujícího Python prostředí je vhodné použít nástroj `virtualenv` [13].

Dále je potřeba nainstalovat grafický backend, který umožní vykreslování grafů. Nejjednodušší je použít `tkinter`, který je standardně v Pythonu podporován. Návod pro instalaci je k dispozici na stránkách Python organizace [12].

C.3 Systém

Nyní je potřeba nainstalovat samotný systém. Tento systém je šířený v podobě skriptů. Nejprve je nutno stáhnout celý veřejný GitHub repozitář [33]. Následně je nutné nainstalovat balíčky potřebné pro běh. Pro tento účel je v kořenovém adresáři repozitáře připraven soubor `requirements.txt`, který se předá nástroji `pip` [7] a ten nainstaluje všechny potřebné balíčky. Příkaz pro instalaci balíčků může vypadat následovně: `pip install -r requirements.txt`.

Nyní je systém možné používat. Ukázku použití lze najít v části 4.1.1.