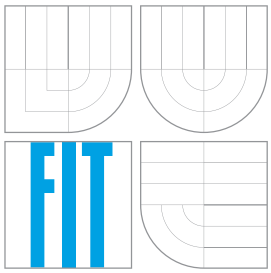
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

ANALÝZA DAT ZE SOCIÁLNÍCH SÍTÍ

ANALYSING DATA FROM SOCIAL NETWORKS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

PETR SKYVA

VEDOUCÍ PRÁCE
SUPERVISOR

Doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2016

Zadání bakalářské práce/18664/2015/xskyva02

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav počítačové grafiky a multimédií Akademický rok 2015/2016

Zadání bakalářské práce

Řešitel: **Skyva Petr**
Obor: Informační technologie
Téma: **Analýza dat ze sociálních sítí**
Analysing Data from Social Networks
Kategorie: Algoritmy a datové struktury

Pokyny:

1. Prostudujte rozhraní služby Twitter a dalších sociálních sítí
2. Navrhněte a implementujte systém, který dokáže pravidelně získávat, indexovat a analyzovat stahovaná data
3. Vytvořte systém pro automatickou klasifikaci shromažďovaných dat, analýzu trendů a vizualizaci výsledků
4. Demonstrujte vytvořený systém na vhodně zvolených příkladech.
5. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- dle dohody s vedoucím

Pro udělení zápočtu za první semestr je požadováno:

- Funkční prototyp

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

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% 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 2015
Datum odevzdání: 18. května 2016

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

doc. Dr. Ing. Jan Černocký
vedoucí ústavu

Abstrakt

Cílem této práce je prostudovat rozhraní sociálních sítí a následně navrhnout a implementovat systém, který dokáže pravidelně získávat, indexovat a analyzovat stahovaná data. Vytvořený systém je demonstrován na příkladech s počítačovými hrami a je implementován v jazyce Python, kde pro indexování dat využívá ElasticSearch.

Abstract

The goal of this bachelor thesis is to examine social networks, propound and implement the system, which manage acquire,index and analyze downloaded data. Created system is demonstrated on examples with computer games and it is implemented with Python programming language, where for indexing is using ElasticSearch.

Klíčová slova

Twitter, Facebook, získávání dat, analýza dat, indexování dat, automatická klasifikace, analýza trendů, vizualizace výsledků, Python, ElasticSearch, JSON.

Keywords

Twitter, Facebook, data downloading, data analysis, data indexing, automatic clasification, trend analysis, result visualisation, Python, ElasticSearch, JSON.

Citace

Petr Skyva: Analýza dat ze sociálních sítí, bakalářská práce, Brno, FIT VUT v Brně, 2016

Analýza dat ze sociálních sítí

Prohlášení

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

.....
Petr Skyva
17. května 2016

Poděkování

Tímto bych chtěl poděkovat vedoucímu své práce panu Doc. RNDr. Pavlu Smržovi Ph.D. za odbornou pomoc při vytváření práce a za veškeré cenné rady a konzultace.

Obsah

Kapitola 1

Úvod

V dnešní době se sociální sítě stávají pro většinu lidí neodmyslitelnou součástí života, proto se nabízí provádět analýzu volně dostupných dat a získávat názory velkého množství lidí na různá témata.

Z tohoto důvodu sociální sítě již jsou zajímavým a důležitým zdrojem informací pro malé a nadnárodní firmy, kterým se zde otevřela zcela nová možnost, jak kontaktovat zákazníky nebo zjišťovat co se jim líbí, anebo naopak nelíbí na jejich produktech. Sociální sítě jsou tedy v současné době velmi mocným nástrojem pro analýzu dnešní společnosti.

Velké množství lidí denně tweetuje a píše statusy na Facebooku o různých tématech; od nově vydané počítačové hry, až po nastávající volbu prezidenta. Díky tomu se nám otevírají možnosti získat data od velkého množství lidí z dané lokace za určitý čas, které následně analyzujeme.

Tato práce má za cíl objasnit pojem sociální síť. Charakterizovat a popsat různé, rozdílné sociální sítě. Následně navrhnout a implementovat systém, který dokáže pravidelně získávat, indexovat a analyzovat stahovaná data. Systém reaguje na univerzální vstupy dat a ani analýza není zaměřená na specifická témata. Data budou získávána ze dvou sociálních sítí: Twitteru a Facebooku. Dozvíme se něco o historii těchto sítí a zajímavé statistiky.

Ve 3. kapitole si ukážeme, jak a jaká data je možné získavat ze zmiňovaných sítí, a jaká jsou zde omezení při sběru dat.

4. kapitola ukazuje návrh řešení pro získávání a indexování dat. Poukážeme si na to, jak obejít případná omezení z předchozí kapitoly, a co přesně a jak budeme indexovat.

V 5. kapitole se dozvíme o analýze získaných dat a jejich vizualizaci.

V 6. kapitole si ukážeme implementovaný systém na příkladu s počítačovými hrami. Nahlédneme na nově vydávanou hru, na názory na ni a její aspekty. Dále ukážeme různé typy analýz od geolokace až po nejvíce zmiňovaná slova.

Kapitola 2

Sociální sítě

Sociální síť nebo společenská síť je internetová služba, která svým registrovaným uživatelům umožňuje vytvářet osobní profil. Díky tomuto profilu mohou uživatelé mezi sebou vzájemně komunikovat (chatovat), sdílet různé životní události a zážitky, fotky, videa nebo hrát různé hry. Mohou zde vznikat různé skupiny lidí na dané téma.

V dnešní době již existuje velké množství různých sociálních sítí na dané téma, od seznamování, sdílení videí a fotek, nebo navazování kontaktů se starými známými. Také existují odborné sociální sítě. O všech těchto typech se dozvíme v následujících sekcích.[8][10]

2.1 Nejznámější sociální sítě

V této sekci si popíšeme nejznámější sociální sítě dnešní doby, zařadíme si je a ukážeme, které budeme využívat v této práci.

Obrázek 2.1: Nejznámější sociální sítě

2.1.1 Facebook

Facebook je jedna z nejpopulárnějších sociálních sítí dnešní doby. Založen byl 1. února 2004. Facebook jako takový je akciová společnost s hodnotou přes 500 miliard dolarů, která

pořád stoupá. Umožňuje registrovaným užívatelům spoustu funkcí od vkládání fotografií až po odesílání instantních zpráv s přáteli. Facebook je dostupný v 84 jazycích. Je zde možné zakládat skupiny, události a stránky s různými tématy.

Název Facebook vznikl z letáků zvaných "Facebooks", které se rozdávají univerzitním studentům prvního ročníku. Facebook byl založen Markem Zuckerbergem, kde v jeho základech sloužil pouze pro studenty Harvardovy univerzity, jelikož jeho zakladatel byl jejím studentem. V té době fungoval na doméně thefacebook.com.

Jelikož se postupně stával Facebook úspešným, dostal se i na další univerzity. V Česku byl Facebook jako první zpřístupněn pro brněnskou Masarykovu univerzitu. Po univerzitách přišly nadnárodní společnosti a v budoucnu se Facebook vyvíjel dál.

Zlomové datum bylo 6. sprna 2006, kdy byl Facebook otevřen veřejnosti, kdy jediné omezení byla věková hranice třinácti let. V roce 2007 byl již mezi 10 nejnavštěvovanějšími webovými stránkami na světě.

Hlavní přijem Facebooku je jako u většiny neplacených služeb z reklamy. Zlomové období byl rok 2014, kdy Facebook odkoupil firmu WhatsApp.[12]

Statistiky a zajímavosti

- v dnešní době již přes 1,6 miliardy aktivních uživatelů,
- průměrně 480 000 nových uživatelů denně,
- velikost indexu pro vyhledávání je přes 200 Gigabytů,
- v současné době 3. nejnavštěvovanější internetová stránka na internetu,
- Facebook má uloženo přes 145 miliard fotografií, denně se jich nahraje až 200 milionů.

Bezpečnostní rizika

- sdílení příliš citlivých informací,
- důvěryhodnost uživatelů,
- nevhodné fotky a videa, které mohou poškodit uživatele,
- ukradení identity,
- zjednodušení práce pro "stalkery"
- až velké sebevědomí, zveřejňování kam uživatel jede na dovolenou, adresa, datum narození, jména dětí, manželky...

Facebook byl použit jako jedna ze dvou sociálních sítí v této práci, z důvodu velkého množství uživatelů a možností získání i osobních dat jako je jméno, datum narození, místo bydliště a veškeré příspěvky od daného uživatele. Také se zde nacházejí skupiny s velkým počtem uživatelů, kde je možné získávat veškeré příspěvky a data. Z těchto důvodů je Facebook vhodný pro analýzu různých témat od typů piv až po počítačové hry.

2.1.2 Twitter

Twitter není zdaleka tak robustní jako Facebook. Vyznačuje se především tzv. tweety, což jsou příspěvky dlouhé maximálně 140 znaků. Založen byl 21. března 2006 v San Franciscu v Kalifornii USA. Twitter je akciová společnost s hodnotou přes 532 milión dolarů.

Registrovaní uživatelé mají možnost odesílat nebo číst tweety, tí kteří nejsou registrovaní, je mohou pouze číst. Uživatelé mohou být buď sledovaní tudíž ostatní lidé vidí jejich tweety anebo mohou někoho sledovat. Takto je možné dávat například zprávy fanouškům anebo číst tweety oblíbených lidí.

Pro Twitter jsou důležité tzv. hashtagy, což jsou klíčová slova, které začínají #. Podle hashtagů je následně možné vyhledávat určité události, akce a jakákoliv jiné témata (refugee,Dota2). ódové jmeno bylo twttr, kde jako cíl bylo komunikovat s úzkou skupinou lidí pomocí SMS zpráv. Zlomový bod byl rok 2007, kde na konferenci SXSWi se zvýšil počet odeslaných tweetů za den z 20 000 na 60 000.

V současné době se Twitter využívá na vetšině konferencí, kde slouží jako zpětná vazba nejen pro pořadatele ale i účastníky konference. K velkým výkyvům počtu tweetů dochází většinou při velkých akcích jako je Mistrovství světa v hokeji nebo fotbalu. V současné době již Twitter neroste takovou rychlostí.[11]

Statistiky a zajímavosti

- na Twitteru se vyskytuje kolem 20 miliónů botů, což je méně než 5 procent všech uživatelů,
- i přesto jsou boti schopní ovlivnit názory na kulturu, produkty nebo různé politické názory,
- Twitter byl využit při organizování revolucí,
- Velká Británie vlastní boty pro zasílání automatických zpráv na Twitter, a získávání dat o uživatelích,
- nejpopulárnější účty na Twitteru jsou Katty Perry s 84 milióny followerů, nasledující jsou Justin Bieber,Taylor Swift,Barack Obama a Youtube

Bezpečnostní rizika

- sdílení příliš citlivých informací,
- ukradení identity,
- zjednodušení práce pro "stalkery", vyplněná lokace jak u profilu tak u příspěvků,

Twitter byl použit jako druhá sociální síť pro analýzu z důvodu výborného získávání dat v podobě tweetu a filtrování dle hashtagu. Také je mnohem jednodušší získat velké množství dat na dané téma oproti jiným sociálním sítím. U typu dat je i navíc oproti Facebooku možnost získání lokace čímž se opět zlepšuje možnost analýzy tématu i podle světových míst.

2.1.3 Instagram

Instagram je sociální síť, jejíž primární zaměření jsou fotografie a krátká videa. Registrovaným uživatelům umožňuje vkládat fotografie a videa, přidávat k nim komentáře a vyplnit si vlastní profil. Uživatel může mít volný anebo uzavřený profil, od toho se odvíjí, zda si cizí uživatelé mohou prohlížet jeho fotografie a videa. Pokud je profil uzavřený, je potřeba schválení sledování, pokud chceme prohlédnout cizí media.

Instagram byl založen 6. října 2010 v San Franciscu. V roce 2011 se podle Twitteru inspiroval hashtagy, které bylo možné dávat k fotkám a i následně podle nich vyhledávat. Instagram je známý pro své filtry, kterých nabízí nepřeberné množství. 12 dubna roku 2012 bylo uzavřeno odkoupení Instagramu Facebookem, kdy Instagram zůstane jako nezávislá firma.[3]

Statistiky a zajímavosti

- přes 160 milionů uživatelů měsíčně,
- slovo selfie se hlavně díky Instagramu stalo v roce 2013 jako slovo roku,
- přes 50 procent uživatelů přistupuje do služby přes produkty Apple,
- Instagram obsahuje 24 filtrů na fotky,
- v roce 2013 byl zaznamenán případ, kdy se na instagramu vyskytovaly fotky drog a následně probíhal prodej přes instantní zprávy.

Bezpečnostní rizika

- sdílení příliš citlivých informací,
- ukradení identity,
- zjednodušení práce pro "stalkery", vyplněná lokace jak u profilu tak u příspěvků,

Instagram nebyl použit v této práci z důvodu ne úplně vhodného hlavního sdělovacího prostředku, fotografií a videa. Jelikož budeme analyzovat především text, trendy a podobně, pouze komentáře k fotkám nejsou vhodný zdroj informací.

2.1.4 LinkedIn

LinkedIn řadíme opět mezi sociální sítě, ale řadíme jej do profesních sociálních sítí. Byl založen 14. července roku 2002 v Kalifornii. Registrovaní uživatelé zde opět mohou vyplnit vlastní profil, kde zadávají své vzdělání, minulou a současnou práci a zkušenosti, svůj životopis. V podstatě vše, co může zajímat budoucího zaměstnavatele. Uživatelé mezi sebou mohou navazovat kontakty, díky čemuž se mohou propojit i s potencionálními zaměstnavateli. Také zde mohou mezi sebou diskutovat o svých pracovních zájmech. Většina uživatelů se tedy řadí mezi odborníky v nejrůznějších oborech, manažery, ředitele a další. Jako uživatel se může registrovat i celá firma.

V současné době je LinkedIn největší profesní sociální sítí na světě. Díky tomu je možné vyhledávat své bývalé případně i současné kolegy a dostat se s nimi do kontaktu. LinkedIn je v současné době pro mnoho lidí jako vizitka na internetu, a právě z tohoto důvudo je hojně využíván personalisty a hledači talentů.

Dne 19. května roku 2011 vstoupil LinkedIn na burzu. Nicméně v současné době jeho hodnota klesá.[5]

Statistiky a zajímavosti

- LinkedIn má přes 400 milionů uživatelů z 200 států světa,
- každé 2 sekundy se registruje na LinkedIn nový uživatel,
- nejvíce registrovaných uživatelů je z USA a to 122 milionů,

Bezpečnostní rizika

- ukradení identity, poškozená vlastní značka se špatně napravuje,
- v červnu roku 2012 bylo ukradeno kolem 6 a půl milionu uživatelských hesel, které byly zveřejněny, i když zahashované na internet.

LinkedIn také nebyl použit v této práci, z důvodu příliš specifických dat, které se na této síti nalézají.

2.2 České sociální sítě

2.2.1 Lidé

Lidé je česká sociální síť, která se svým stylem podobá Facebooku, ale nemá takovou bázi uživatelů. Pro registrované uživatele se zde nachází seznamka, instatní chat a sdílení fotografií. Lidé byla vpodstatě první sociální síť na českém internetu. Vznikla v roce 1997 jako společný projekt Seznamu a Pinknetu.

V současné době zváště kvůli vlivu mezinárodních sociálních sítí jako Facebook a Twtitter návštěvnost a hodnota klesá.

2.2.2 Spolužáci

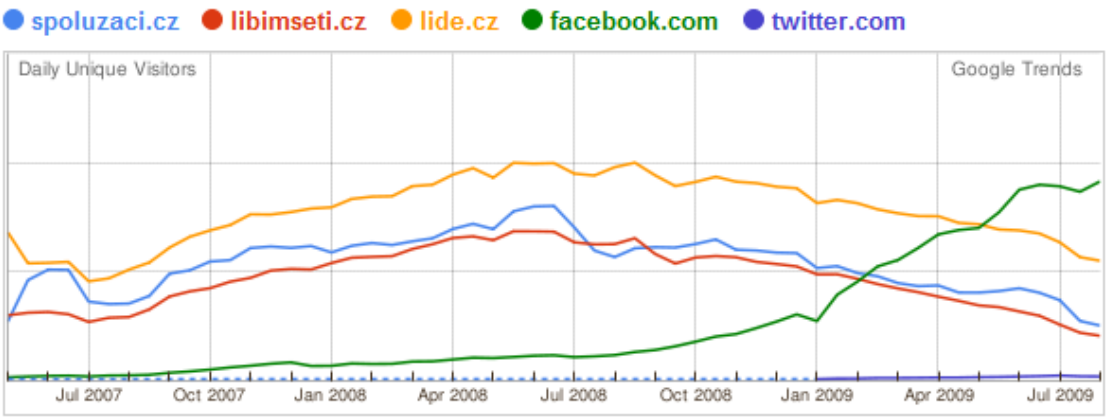
Spolužáci jsou dceřinou sociální sítí s Lidé. Jejich hlavní využití je hledání a navazování kontaktů s bývalými spolužáky. V současné době je to nejvíce navštěvovaná česká sociální síť.

2.2.3 Líbím se ti

Líbím se ti vznikl jako konkurent síťě Lidé. Poskytuje velmi podobnou funkcionalitu a snažil se napodobit již ne tolik známou sociální síť MySpace. V současné době tedy slouží především jako seznamka, sdílení fotografií a chatování s ostatními uživateli.

2.2.4 ČSFD

ČSFD je sociální síť zaměřená na filmy. Slouží jako místo, kde uživatelé diskutují nad filmy, hodnotí je a přidávají k nim zajímavosti, fotky a trailery. V současné době je to nejnavštěvovanější sociální síť (portál) pro filmové nadšence v České Republice.



Obrázek 2.2: Vývoj Českých sociálních sítí vůči Twitteru a Facebooku

Kapitola 3

Typy dat a jak je získat

V této kapitole si ukážeme jak a jaká data je možné získat ze dvou sociálních sítí Facebook a Twitter.

3.1 Proč získávat a analyzovat data ze sociálních sítí

Již víme co jsou sociální sítě a že se na nich nachází velké množství informací. V průběhu let se náš život změnil a nadále mění, sociální sítě nás ovlivňují ať v dobrém nebo špatném.

A proto je vhodné se podívat, co se na nich všechno nalézá a co z nich můžeme zjistit.

Co můžeme zjistit

- Kdo zná koho?
- Jak často určití lidé komunikují s jinými?
- Které sociální sítě generují největší hodnotu pro určité téma?
- Jak geografie ovlivňuje náš sociální život na internetu?
- Kteří lidé jsou nejpopulárnější na sociálních sítích?
- O čem se lidé baví? Jaké jsou současné trendy?
- Jaké nazory mají lidé na dané téma a ovlivňuje je například lokace?
- O co se lidé nejvíc zajímají na rozdílných sociálních sítích?
- Co je nejvíc populární pro určitá témata?

Odpovědi na tyto základní otázky často nabízejí cenné náhledy a prezentují lukrativní nabídky pro podnikatele, firmy, psychology a ostatní lidi, kteří chtějí vědět o čem se zrovna mluví.

3.2 Typy dat

V následující sekci si popíšeme, jaké různé typy dat můžeme získat a jak je získáme, ze dvou vybraných sociálních sítí, a také si vysvětlíme, proč jsme vybrali právě Facebook a Twitter.

3.2.1 Twitter

Twitter jako takový je velmi vhodný pro analýzu, jelikož obsahuje velké množství volně dostupných dat. Data můžeme získávat dvěma způsoby, sbíráním tweetů anebo získáváním volně dostupných informací vyplněných na uživatelských profilech.

Twitter je pro nás zajímavý, protože tisíce dat vznikají během sekundy a jsou volně dostupné ke stažení a analýze v reálném čase.

Tři hlavní důvody proč je Twitter vhodný

- Twitter API je elegantní, jednoduché a dobře dokumentované a je dostupé pro všechny,
- získaná data (tweety) jsou vhodně formátované a připravené pro analýzu,
- možnost získat všechna data, účty nejsou nijak zamknuté, díky tomu můžeme získávat data od kohokoliv.

Data tedy získáváme z volně dostupných tweetů. Tweety můžeme získávat dvěma způsoby, a to streamovaním tweetu nebo získaváním tweetu podle zadaných hashtagů.

Při streamování tweetu zadáme vstupní data ve formě slov, která chceme sledovat. Například názvy různých počítačových her. Následně v reálném čase získáváme veškeré tweety s klíčovými slovy, které uživatelé přidávaji na Twitter.

U získávání tweetu podle hashtagů vyhledáváme v minulosti a získáváme všechny tweety s daným hashtagem.

Tweety jsou esencí Twitteru, a i když to jsou jenom zprávy 140 znaků dlouhé se jménem uživatele a datum vytvoření, nachází se zde i velké množství metadat, která hned nevidíme. Metadata mají dvě další části a to entities a places. U entities jsou zmínění uživatelé, hashtagy, URL, a media spojená s tweetem. Místa jsou lokace v reálném světě, která mohou být připojena ke tweetu. My v této práci budeme využívat jen některé a to níže vyjmenované.[16]

Využité informace

- text tweetu,
- uživatel, který tweet napsal,
- lokace uživatele, kterou má vyplněnou na svém profilu,
- id uživatele,
- id tweetu,
- koordinace, kde byl tweet napsán,
- hashtagy použité v tweetu,
- url odkazy v tweetu,
- datum vytvoření tweetu (den,měsíc,rok).

3.2.2 Facebook

Facebook je v současné době největší sociální síť, už jen z tohoto důvodu je vhodné jej použít pro analýzu a získávání dat. Přesto nemá až takové výhody otevřenosti, kterými disponuje Twitter, jelikož má jedny z nejsofistikovanějších pravidel, jak chránit soukromí uživatelů. Data je možné získat pouze z otevřených skupin a profilů. Pokud chceme vidět všechna data od nějakého uživatele, musí nás přijmout za svého přítele. A i tak zde pořád existuje možnost sdílet fotky a příspěvky jen s vybranými uživateli.

Na druhou stranu existuje velké množství uživatelů, kteří mají svoje privátní data volně dostupná, jako věk, pohlaví, místo bydliště anebo rodinný stav. Dále se nabízí možnost využití lajků, vidíme o co se užívatelé zajimají a jaké jsou současné trendy. Možnost podívat se do minulosti nějaké skupiny či produktu. Jaké změny či novinky byly oblíbené a jaké se setkávaly spíše s kritikou. Na počtu komentářů je možnost vidět, zda daná skupina roste, stagnuje či upadá.

Facebook je tedy velice robustní, dobře zdokumentovaná brána do něčeho, co může být v současné době nejlépe organizovaný obchod s informacemi, za které nic neplatíte.[16]

Tři hlavní důvody proč je Facebook vhodný

- Facebook API je dobře dokumentované a je dostupné pro všechny, tudiž lze jednoduše a rychle najít řešení různých úkolů,
- získaná data nám zvětšují naše možnosti analýzy, jelikož jsou trochu odlišná od Twitteru,
- účty a skupiny mohou být zamknuté, ale ne všechny jsou, potom je možné získat citlivá data o uživatelích, což na Twitteru nelze docílit v takovém množství.

Data z Facebooku tedy získáváme z otevřených skupin/stránek a z otevřených profilů uživatelů. V naší práci budeme data získávat především z fan stránek vstupních témat.

Název stránky je vyhledán a zvolena první možnost, následně můžeme získat veškeré statusy a jejich komentáře. Celkový počet lajků na stránce a kolik lidí mluví o dané stránce. U každé stránky je opět velké množství informací a my využijeme pouze ty , které jsou pro nás nejdůležitější.

Využité informace

- název stránky,
- počet lajků stránky,
- počet lidí, kteří mluví o dané stránce,
- veškeré statusy a jejich komentáře,
- počet lajků u statusu a jejich komentářů,
- počet komentářů u statusu,
- datum vložení statusu (den,měsíc,rok).

Kapitola 4

Analýza a návrh řešení získávání a indexování dat

V této kapitole se podíváme, jakým způsobem budeme získávat data ze dvou zmíněných sociálních sítí a jak a jakým stylem je budeme indexovat. Nejdříve si ukážeme použité technologie a následně si popíšeme, jak jsme je využili při implementaci systému.

4.1 Používané pojmy v následující kapitole

4.1.1 API

API neboli Application Programming Interface je sbírka funkcí, tříd a protokolů, které můžeme jako vývojář využívat. Jsou to dané specifikace, které když programy dodržují, mohou komunikovat s jinými počítačovými programy. V podstatě je to rozhraní pro komunikaci s jinými počítačovými programy, které nám poskytuje stavební kameny a ty si můžeme poskládat jak potřebujeme.[1]

4.1.2 JSON

JSON neboli JavaScript Object Notation je odlehčený formát zápisu nějakého typu dat nezávislý na platformě. Je jednoduše čitelný i člověkem, ale především se snadno používá v počítačových programech. [4]

4.1.3 OAuth

OAuth je otevřený protokol vytvořený v listopadu roku 2006. Zaměřuje se na jednoduchost při poskytování specifické autorizace pro různé aplikace. Je to bezpečná cesta jak poskytnout ostatním lidem přístup k aplikaci a při tom nešířit jejich hesla po internetu. [6]

4.2 Použité technologie

Zde si popíšeme jednotlivě použité technologie. Zvolený programovací jazyk, ve kterém budeme implementovat v podstatě celý systém, použité volně dostupné knihovny a API. A jakou technologii použijeme pro indexování dat.

4.2.1 Programovací jazyk - Python

Celý systém budeme implementovat v programovacím jazyku Python. Python byl zvolen z důvodu jeho intuitivní syntaxe a jeho úžasného systému pro správu přídavných balíků, které velice usnadňuji práci s přístupem k API jednotlivých sociálních sítí a manipulaci s daty. Všechna získaná data jsou ve formátu JSON, což z nich dělá ideální vstup pro Python.

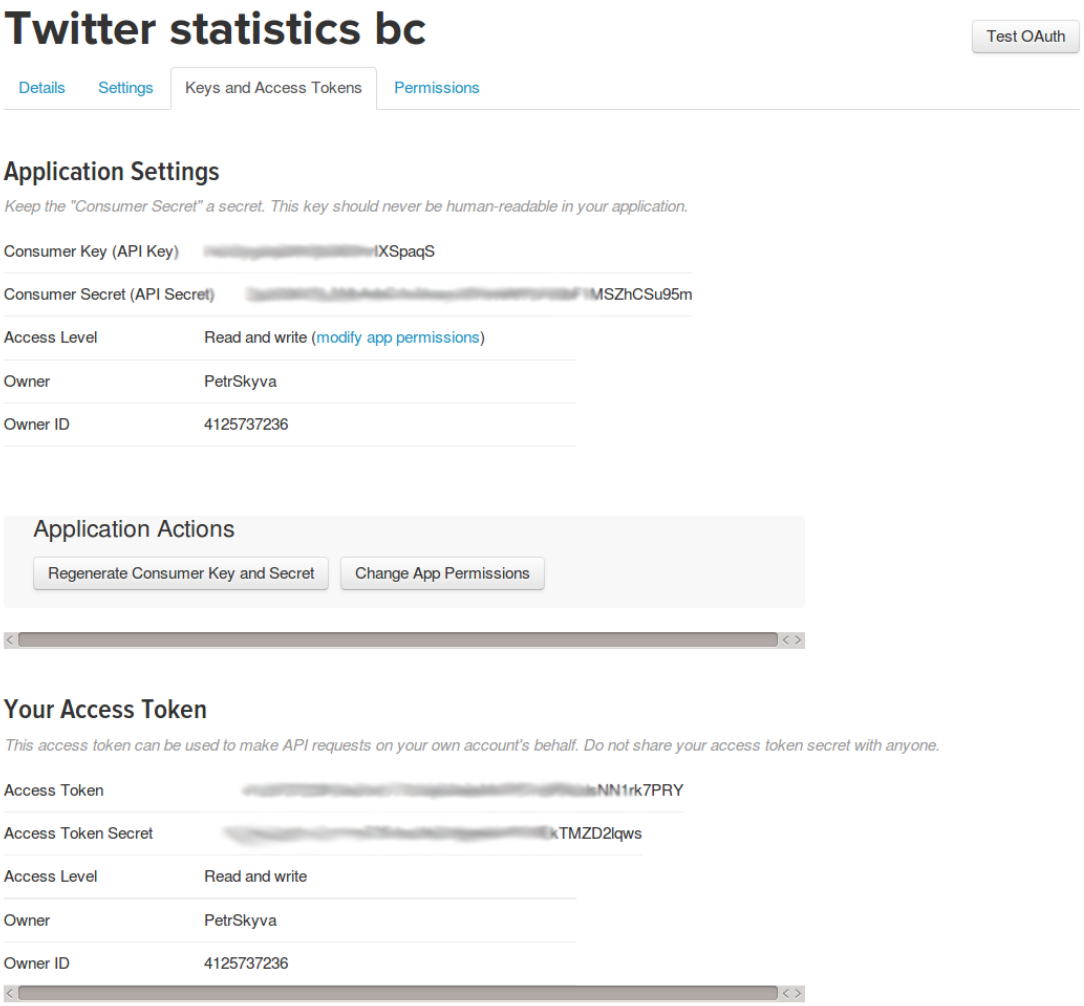
4.2.2 Python Twitter, Twiter API

Twitter poskytuje již zmíněné Twitter API pro získávání tweetů, údajů o uživatelích, přidávání nebo mazání uživatelů, které chceme sledovat. Možnost psaní tweetu a mnoho dalšího, v podstaté můžeme celý náš Twitter účet ovládat z terminálu.

My využijeme balík/knihovnu Pythonu, který se jmenuje Python Twitter. Python Twitter je v základě obal kolem Twitter API. Tato knihovna poskytuje čisté rozhraní v Pythonu pro Twitter API. Zjednoduší nám práci při získávání tweetu a ostatních dat. Výhoda této knihovny je, že obal nad API je v podstatě jedna ku jedné, tudíž vše co je v API je i v knihovně pod stejnými názvy a používá se i se stejnými parametry. Další výhodou této knihovny je výborná dokumentace, díky které neztrácíme čas dohledáváním jednoduchých informací.

Předtím než je možné Twitter API a tudíž i Python Twitter začít využívat, je zapotřebí se na Twitteru zaregistrovat. Poté můžeme vytvořit aplikaci na https://apps.twitter.com/. Dále je nutno provést autentifikaci s Twitterem pro naši nově vytvořenou aplikaci. Jelikož pro všechny operace je autentifikace zapotřebí. Twitter používá OAuth autentifikaci, z tohoto důvodu musíme pro naší aplikaci vygenerovat OAuth klíče, které budeme používat při jakémkoliv připojování na Twitter API.

V podstatě tedy vytvoříme aplikaci, která se bude autorizovat s účtem na Twitteru. Vypadá to jako zbytečná zátěž, proč jen nezadávat svoje přihlašovací jméno a heslo k přístupu k API? Pro nás by to bylo ideální, ale pro spolupracovníky nebo přátele je to nevhodné, protože by se musely sdílet jejich přihlašovací údaje a hesla, z tohoto důvodu Twitter a většina dalších sociálních sítí využívá OAuth.[9]



Obrázek 4.1: Potřebné klíče k připojení a práci s Twitter API

4.2.3 Facebook Python SDK, Facebook GraphAPI

Facebook stejně jako Twitter poskytuje vlastní API pro získávání dat, jako jsou údaje o lidech, stránky, různé akce a fotografie, případně propojení mezi nimi, přátelé přátel, označování fotografií. Facebook Python SDK slouží opět jako obal nad Facebook GraphAPI, ulehčuje a zrychluje nám tedy práci.

Abychom mohli Facebook GraphAPI využívat, je potřeba vytvořit účet na Facebooku. Po vytvoření účtu můžeme vytvořit naši novou aplikaci, přes kterou budeme přistupovat k Facebook GraphAPI a nebo využijeme defaultní aplikaci GraphAPI Explorer.

Facebook GraphAPI je tedy masivní množství strukturovaných dat, reprezentující sociální interakce skladájicí se z uzlů a propojení mezi těmito uzly.

Přístupový klíč, který budeme generovat v naší aplikaci je opět OAuth klíč, který budeme používat při každém dotazu na Facebook GraphAPI.[14]

Obrázek 4.2: Potřebný klíč k připojení a práci s Facebook GraphAPI

4.2.4 Elasticsearch

Elasticsearch je serverový vyhledávač vycházející z Apache Lucene. Poskytuje distribuované multiuživatelské fulltextové vyhledávání na zakladě RESTful rozhraní. Elasticsearch je vyvíjen v Javě a je dostupný jako open source pod licencí Apache. Elasticsearch je nejvíce populární podnikový vyhledávač.

Elasticsearch poskytuje tedy fulltextové vyhledávání v různých typech dokumentů. Poskytuje škálovaci výhledávání a běží téměř v reálném čase. Díky tomu je možné instantně filtrovat velké množství dat. Jelikož je Elasticsearch založen na Apache Lucene, snaží se poskytnout veškeré její možnosti skrz JSON a Java API. Pro náš systém je tedy nejdůležitější právě fulltextové vyhledávání. Díky němu můžeme prohledávat a filtrovat různé tweety, statusy a komentáře a nasledně tato data analyzovat. Veškerá získaná data tedy budeme indexovat do bežícího Elasticsearch. Jak budou vypadat indexy a co všechno budeme indexovat, se dozvíme v následující sekci.

Jěště je vhodné zmínit nástroje neboli balíčky, které nám usnadňují práci s Elasticsearch. Jedním z nejvíce využívaných při vývoji systému bude Elastic HQ. Je to nástroj naimplementovaný v JavaScriptu, tudíž běží na stejné adrese jako náš Elasticsearch. Po přistupu na adresu máme možnost graficky proklikávat jednotlivé clustery a uzly. Díky tomu nemusíme neustále dělat testovací dotazy přes RESTful.[2]

Firmy, které používají Elasticsearch

- Wikipedia,
- OpenTable,
- Tesco,
- SoundCloud,
- LinkedIn,
- WordPress,
- Uber.

4.3 Implementace indexování a získávání dat z Twitteru

Již víme, jaká data nám Twitter může poskytnout a jak je můžeme získat. Také víme, kam je budeme indexovat. Ukážeme si tedy implementační část získávání dat a jak tato data

budeme indexovat a především jaká data budeme indexovat, a která budou potřeba pro budouci analýzu.

Veškerá implementace probíhá v souboru *search.py*. Program může mít dva typy vstupních dat, podle toho, jestli data získáváme pomocí hashtagů anebo pomocí streamování. Po zvolení typu získávání dat musíme vytvořit ještě jeden vstupní soubor, ve kterém se budou nalézat uživatelem dané aspekty a lokace.

4.3.1 Získávání dat pomocí hashtagů

Pokud tedy chceme získávat data pomocí hashtagů, vytvoříme vstupní soubor, který se musí jmenovat *search_input.txt*. Soubor bude mít následující tvar:

```
název uzlu pro Elasticsearch
(,Jakým slovem budou výsledky indexovány,\#první hashtag,..,),téma=název tématu,..
```

Zde si ukážeme část vstupního souboru, na kterém budeme vytvářet testovací data.

```
games
(,Battleborn,#Battleborn,),genre=action
(,Counter Strike,#CSGO,#CounterStrike,),genre=action
```

Názvy se nemusí nutně shodovat s hashtagy, ale měly by popisovat celou skupinu, podle které budeme vyhledávat. Témat může být mnohem více než jedno, můžeme přidat například další *genre=rpg* k již existujícímu *genre=action*. Také můžeme vytvořit naprosto jíné téma jako např. *price=1500*.

Pokud máme vytvořený vstupní soubor, dojde k jeho načtení. Vstupní data se rozdělí na dvě části. První část jsou pouze hashtagy a pod jakým názvem chceme danou skupinu hashtagů v budoucnu indexovat. Tato data se uloží do proměné typu slovník v následujícím tvaru:

```
{'Battleborn': ['#Battleborn'],
'Counter Strike': ['#CSGO', '#CounterStrike', '#CS1.6']}
```

Druhá část dat jsou opět názvy, pod kterými budeme indexovat hashtagy a jejich témata a typy témat. Opět jsou uloženy ve tvaru slovníku v následujícím tvaru:

```
{'Battleborn': {'genre': ['action']},
'Counter Strike': {'genre': ['action']}}
```

Již jsme zmínili, že je potřeba vytvořit ještě druhý soubor s aspekty a lokacemi. Aspekty a lokace se budou využívat při zpracování přispěvku ještě před zaindexováním. Nejdříve zadáme, jaké typy aspektů chceme, například u počítačových her nás budou zajímat příspěvky, kde se mluví o příběhu, ceně, grafice anebo o hardwarové specifikaci. A následně zapíšeme slova, podle kterých se při výskytu v tweetu zařadí.

Po aspektech zbývají tedy už jen lokace, ty budou v podstatě ve stejném tvaru. Nejdříve část země nebo kontinentu, například SE Asia a všechny státy pro danou lokaci. Opět se to bude přiřazovat dle výskytu v tweetu. Vstupní soubor se musí jmenovat *search_aspects.txt*. Výsledný soubor může mít tedy následující tvar:

```
{
  "aspects":[
    {
      "hardware requirements":[
      "GTX980",
      "GTX Titan Z",
      ...
      ],
      ...
    }
  ],
  "tweet_location":[
    {
      "SE Asia":[
      "Vietnam",
      "Thailand",
      "Laos",
      ...
      ],
      ...
    }
  ]
}
```

Po načtení obou souborů, než začneme postupně pro všechny hashtagy z prvního souboru získávat data, musíme zjistit poslední id zaindexovaného tweetu. To provedeme dotazem na Elasticsearch pro daný uzel, kde v dotazu nás bude zajímat pole _timestamp, podle kterého seřadíme výsledky a následně pole _id. Veškerá data budeme indexovat za poslední uložený tweet.

Data budeme tedy získávat pomocí dotazů na Twitter API, kdy před každým dotazem je potřeba vyčkávat 1,5 sekundy, abychom nepřekročili limit dotazů z jednoho účtu. Jedním dotazem je možné získat až 100 tweetů pro daný hashtag. Což budeme využívat i my a získavat data od nejnovějšího tweetu s daným hashtagem až po nejstarší dostupný tweet. Do minulosti budeme klesat díky id tweetu, kdy si uložíme vždy poslední id tweetu a použijeme ho jako vstupní bod při dalším dotazu na Twitter API. Takto můžeme získat až týden staré tweety s daným hashtagem. Získat tweety starší než týden již není možné, je to jedno z mála omezení z Twitter API.

Po dotazu na Twitter API získáme tedy určitý počet tweetu, které postupně po jednom projdeme a zaindexujeme. Jako první zjistíme, zda tweet není už zaindexován. Tuto kontrolu provádíme také z důvodu znovuspuštění programu. Pokud je tweet již zaindexován, uložíme pouze jeho id a pokračujeme na další. V případě, že tweet ještě nemáme zaindexovaný, přiřadíme mu aspekty a lokaci podle druhého načteného souboru. Aspekty a lokaci přiřazujeme na základě výskytu daných slov. Například pokud máme tweet: *Today I went to vietnam, and start playing Battleborn on my GTX980.* Přiřadíme lokaci *SE Asia* a i přímo misto *Vietnam.* A co se týče aspektů přiřadíme *hardware requirements.* Díky tomuto můžeme v budoucnu získat například jen příspěvky týkající se ceny, příběhu nebo čehokoliv jiného. Případně se podívat kam lidé cestovali, ale o tom až v dalších kapitolách.

Po přiřazení lokace a aspektů přidáme základní údaje o tweetu, které budeme indexovat.

Veškerá data, která budeme indexovat

- aspekty - aspects,
- lokace v tweetu jako SE Asia - tweet_location,
- určitá lokace v tweetu jako Vietnam - tweet_location_country,
- uživatel, který napsal tweet - user,
- lokace uživatele - location,
- id uživatele - user_id,
- samotný text tweetu,
- název pod jakým bude tweet za indexován např. Battleborn nebo Counter Strike - value_name,
- koordinace - coordinates,
- url vyskytující se v tweetu - url,
- datum a čas vytvoření tweetu - day,monhth,year,time.

Jakmile jsou všechna data připravena, vytvoříme *PUT* dotaz na Elasticsearch a zaindexujeme daný tweet s daným *id*, které jsme zjistili na začátku.

Pokud program ukončíme předčasně, uloží si do souboru poslední použité *id* pro daný název tzv. *value_name*.

4.3.2 Získávání dat pomocí streamu

Druhou možností je získávat data je pomocí streamu. Pro použití této metody je opět potřeba vytvořit dva vstupní soubory. První soubor bude trochu odlišný, ale v zásadě má podobnou syntaxi. Druhý soubor s aspekty a lokacemi můžeme použit ten stejný jako v případě hashtagů.

Jaký má mít tedy vstupní soubor tvar? Mohou být dva typy. Buď použijeme jedno slovo, které nás bude zajímat, pokud o něm někdo promluví a i pod tímto názvem tweet zaindexujeme, a nebo budeme mít vic slov, ale zaindexování proběhne podle prvního v řadě. První soubor se musí jmenovat *search_input.txt*.

```
název uzlu pro Elasticsearch
jakým slovem budeme výsledky indexovat a streamovat,téma=název tématu,..
(,název pod jakým slovem budeme indexovat a streamovat,
slovo které využijeme pro streamování),téma=název tématu,..
```

Následně si opět ukážeme jak vypadá část vstupního souboru, na kterém budeme vytvářet testovací data.

```
games
Battleborn,genre=action
(,Counter Strike,csgo),genre=action
```

Po vytvoření vstupních souborů je opět načteme. Načítání vzniká úplně shodně jako u hashtagů, jen klíčová vyhledávaná slova nezačínají hashtagem.

Streamování začíná shodně jako hashtagy a to tím, že nalezneme poslední uložené id v Elasticsearch. Následně se chování liší, u hashtagů jsme získávali data postupně pro jeden hashtag. U streamování na vstup vložíme pole všech vyhledávaných slov a výsledky získáváme po jednom tweetu.

Výhodou streamování je, že není potřeba kontrolovat, zda tweet již je zaindexován, protože streamování získává tweety nově napsané v reálném čase. Nicméně po každém získaném tweetu je potřeba vyčkávat opět 1,5 sekundy abychom nepřekročili limit Twitter API.

Po získání tweetu tedy opět proběhne případné přiřazení aspektů a lokací a vyplní se zbytek dat pro indexování. V této chvíli se indexují totožná data jako u hashtagů zmíněná výše.

4.4 Implementace indexování a získávání dat z Facebooku

Ukázali jsme si, jaká možná data můžeme z Facebooku získat. Nyní si ukážeme, jaká data přesně budeme získávat a jak. Veškerá implementace probíhá v souboru *fb.py*. Program získává na vstup jeden soubor, který je velice podobný jako vstupní soubor u vyhledávání podle streamování. Jediný rozdíl je, že lze získat data pouze podle jednoho názvu. Název opět musí být pevně zadán a to *fb_input.txt*. Pro lepší pochopení si ukážeme příklad.

Lze

```
games
Battleborn,genre=action
Counter Strike,genre=action
```

Nelze

```
games
Battleborn,genre=action
(,Counter Strike,csgo),genre=action
```

Důvodem, proč lze zadat jen jeden název, je to, že budeme získávat data pouze ze skupin s daným názvem.

Po vytvoření vstupního souboru ho tedy načteme a zadané názvy *(Battleborn, Counter Strike)* uložíme do pole. Dále musíme opět načíst témata *(genre=action)* a uložíme je do slovníku stejně jako u vyhledávání a stahování dat z Twitteru. Než začneme získávat data, musíme opět zjistit stejně jako u Twitteru poslední uložené *id*. Zde nás nezajímá *id* tweetu nýbrž *id* poslední uložené stránky.

Po získání a nastavení vstupních dat začneme postupně procházet pole se zadanými názvy a pro každý název vytvoříme dotaz na Facebook API.

Zde se objevuje jeden problém, který nebyl nějak vhodně vyřešen a to, že nemusíme přímo získat stránku kterou chceme. U jasných věcí, jako je například počítačová hra Minecraft získáme od Facebook API odpověd s výsledky ze správné stránky, protože tento název jako takový je jedinečný. Je ale možné, že pro daný název existuje více stránek, ale nás zajímá ta největší, tedy nejpočetnější. Nicméně pro názvy jako StarWars dochází k problému, kdy je potřeba zadat specifičtější název. Jelikož pouze pro vstup StarWars

nám nemusí být vrácen jako výsledek hra, ale například film. Abychom to tedy shrnuli, je potřeba si uvědomit při vytváření vstupního souboru, jaká data opravdu chceme.

Po dotazu na Facebook API získáváme z dané stránky nepřeberné množství dat, ale přesto nás opět budou zajímat jen některá, pro nás důležitá. Jako první bude název stránky, přestože nakonec zvolíme řešení, že budeme indexovat dle názvu ve vstupním souboru a ne podle názvu stránky, protože se tyto dva údaje mohou poněkud lišit. Například právě pro zmíněný Counter Strike název stránky může být CounterStrike. Potom by při následné analýze mohlo dojít k problému, jakým je rozdílný název u Twitteru a Facebooku. Případně uživatel by již nemusel znát správný název. Dále to je celkový počet lajků na dané stránce. Následovaný počtem lidí, kteří mluví o dané stránce.

Poslední věc, než provedeme první zaindexování je položka istopic. Ta bude sloužit jako rozlišení při analýze, protože nás nemusí zajímat veškerá data z dané stránky jako jsou statusy atd., ale pouze kolik má daná stránka lajků a kolik o ní mluví lidí. V tom případě budeme využívat tuto položku.

Pokud tedy máme připravena první data, tak provedeme kontrolu, zda již daná stránka není zaindexovaná. Pokud ano, provedeme aktualizaci počtu lajků a počtu lidí, kteří o dané stránce mluví. V opačném případě provedeme pouze zaindexování nově získaných dat.

Po zaindexování tzv. náhledu stránky provedeme další dotaz na Facebook API, který nám poskytne veškeré statusy a komentáře k daným statusům pro danou stránku. Zde se vyskytuje další problém - co vše přidávat případně aktualizovat. My budeme pouze přidávat nové věci, tudíž pro každý status zjistíme, zda byl již zaindexován a pokud ano, tak půjdeme na další. Tímto se ochuzujeme o aktualizaci statusů, ale pro testovací účely to již necháme takto, protože indexujeme velké množství dat a v podstatě se nikdy nevracíme na danou stránku dvakrát. Úprava, aby bylo možné aktualizovat statusy, je ve své podstatě jednoduchá - zjistíme, zda se daný status nachazí v Elasticsearch a pokud ano, tak víme, jaké id tento záznám má, a je již jednoduché ho upravit.

Pokud se tedy status nevyskytuje v Elasticsearch, začneme získávat data, která chceme zaindexovat. První bude počet lajků pro daný status. Jelikož Facebook API neposkytuje počet lajků pro daný status, ale vrací uživatele, kteří olajkovali daný status, je potřeba je všechny sečíst. Problémem je, že v základu poskytne posledních dvacet pět lidí, kteří olajkovali status a ne všechny, ale také poskytne odkaz na další. My tento odkaz upravíme tak, aby nám ideálně Facebook API vrátilo všechny lidi v jednom dotazu. Uděláme to tak, že u poskytnutého obrazu upravíme položku limit=25 na mnohem vetší číslo. Tímto způsobem tedy nakonec získáme všechny lidi, kteří olajkovali daný status, sečteme je a výsledek si uložíme a později zaindexujeme.

Další položkou bude čas vytvoření statusu, který bude rozdělen na den, měsíc a rok. A jako poslední nás budou zajímat komentáře ke statusům, jejich texty, které budeme indexovat. U komentářů musíme vytvořit nový dotaz na Facebook API, kde parametry budou id daného statusu a opět zde bude jeden z parametrů limit. My ho nastavíme na 50, protože při větší hodnotě můžeme překročit limit pro Facebook API.

Po získání komentářů od statusu je po jednom projdeme a uložíme všechny texty do pole a také získáme jejich počet. Pokud máme všechny tyto informace připravené, zbývají poslední dvě položky. První položkou bude název (např. Počítačové hry), pro kterou daný status patří, a druhá položka bude nastavení istopic. Zde nastavíme *istopic* tak, abychom při budoucí analýze měli možnost získat nejen obraz stránky (Pouze název,počet lajků,počet lídí mluvících od dané stránce) ale i veškerá data dané stránky. Pokud máme vše připraveno, zaindexujeme naše nově získaná data do Elasticsearch.

Veškerá data, která budeme indexovat

- jméno stránky - topic_name,
- počet lajků stránky - likes,
- počet lidí mluvících o dané stránce - talking_about,
- parametr, zda se jedná o obra stránky - istopic,
- text statusu - status_msg,
- počet lajků statusu - likes,
- datum vytvoření statusu - day/month/year,
- komentáře ke statusu - comments,
- počet komentářů ke statusu - comments_counter,
- pro jaký název se statusy řadí - topic_name,
- pro jaké téma se statusy řadí - topic.

Kapitola 5

Analýza indexovaných dat a jejich vizualizace

V této kapitole si ukážeme, jaké máme možnosti analýzy naších již zaindexovaných dat. Nejprve si ukážeme opět vstupní soubor, a co všechno do něj můžeme přidat. Následně se podíváme, jaké možné výsledky analýzy je možné získat a jak je provést. Nakonec se podíváme na vizualizaci těchto dat nejen jako JSON, ale i vykreslování různých grafů.

5.1 Použité technologie při analýze

5.1.1 TextBlob

TextBlob je knihovna pro Python verzi 2 i 3, která slouží pro zpracování textových dat v angličtině. Poskytuje jednoduché a dobře dokumentované API pro ponoření se do základních úloh NLP (Natural language processing) jako označování řeči, extrakce jmenných frází, analýza sentimentu, třídění, překlad a mnohem dalšího. Při naší analýze budeme využívat pouze analýzu sentimentu, která je založena na používání algoritmu strojového učení Naive Bayes.

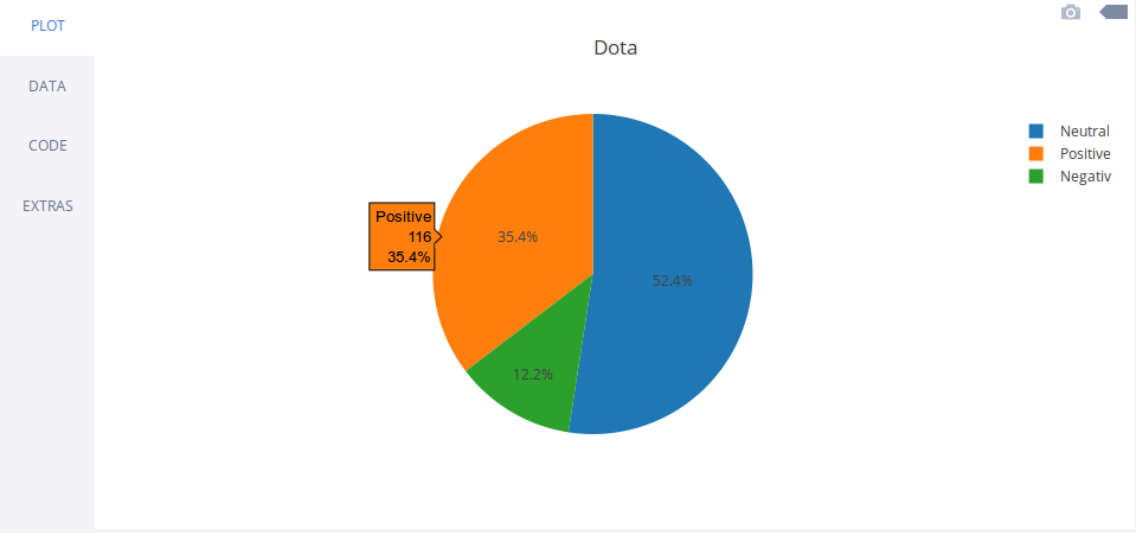
Naive Bayes je strojově vycvičen na datech z filmových recenzí, dále již nebyl cvičen, protože dosahuje použitelných výsledků. Jako výsledek vrací polaritu sentimentu, tedy menší než nula je negativní, vetší než nula je pozitivní a právě nula je neutrální.

Ukázky analýzy sentimentu

- Beer is good. - 0.7 pozitivní,
- Hangover is horrible - -1.0 negativní,
- Beer is good, but hangover is horrible. - -0.15 negativní,
- Game is kind of ok, but first 10 minutes are bad. - 0.05 pozitivní.

5.1.2 Plotly Python

Plotly Python je open-source grafová knihovna. Plotly Python poskytuje API pro vytváření jednoduchých, ale velice efektvních a krásných grafů. Výhodou je, že veškeré grafy se vytvářejí a zobrazují na webové strance plot.ly. Samozřejmostí je stáhnout si vytvořený graf a výhodou je i jeho možná editace po vytvoření, ať už změna popisků či hodnot.



Obrázek 5.1: Ukázka grafu vygenerovaného pomocí plotly

5.2 Vstupní data pro analýzu Twitteru

Pro spouštění analýzy je potřeba vytvořit vstupní soubor, kde si zvolíme co chceme analyzovat, nastavit různé filtry případně propojit indexovaná data z Facebooku a Twitteru.

Vstupní soubor je ve formátu JSON a můžeme zvolit několik parametrů, které si popíšeme dále na výše zmíněném přikladu s počítačovými hrami. Nejdříve si popíšeme jednotlivé položky a následně si ukážeme celkový zápis. Soubor musí mít název *input_analyze_data*.

5.2.1 dbname

Dbname slouží pro název uzlu v Elasticsearch. Pro náš příklad se bude tedy dbname rovnat games.

```
"dbname" : "games"
```

5.2.2 name

Name je název indexovaných dat, tedy pro počítačové hry to budou jednotlivé nazvy her. V našem příkladu to může být například Battleborn.

```
"name" : "Battleborn"
```

5.2.3 year,month,day

Rok, měsíc a nebo den, ze kterého chceme získat zaindexované tweety.

```
"year" : "2016",
"day" : "20",
"month" : "11"
```

5.2.4 user_number_of_posts

Počet tweetů pro daného uživatele. Zde se nám otevírá první možnost vytváření například různých skupin uživatelů. Pokud chceme pouze uživatele, kteří o daném názvu/tématu mluvili alespoň desetkrát, zde si to zvolíme.

```
"user_number_of_posts" : "1"
```

5.2.5 different_topic

Počet různých témat, o kterých daný uživatel mluvil. U počítačových her nás budou zajímat například uživatelé, kteří komentovali nejen akční hry ale i hry jiných žánrů.

```
"different_topic" : "1"
```

5.2.6 location

Lokace uživatele - tato lokace nesouvisí s lokací zmíněnou v příspěvku. Slouží například k tomu, pokud chceme znát názor na dané téma pouze z New Yorku.

```
"location" : "New York"
```

5.2.7 category

Zde si zvolíme kategorii - téma. Tím se nám nabízí další možnost filtrování dat. Mohou nás například zajímat veškeré hry žánru rpg v tomto měsíci. Většinou použijeme buď parametr name anebo category. U kategorie budeme vždy zapisovat ve složitějším tvaru. První bude parametr topic, kde bude název topicu a druhý hodnota value, což bude hodnota topicu. V našem příkladu to může být genre a action.

```
"category" :[{
    "topic" : "genre",
    "value" : "action"
}]
```

5.2.8 talked_about

Tímto parametrem specifikujeme uživatele, kteří se zmínili o jiných kategoriích. Opět je potřeba zápis ve složitějším tvaru a to stejně jako u předchozí kategorie: topic a value.

```
"talked_about" :[{
    "topic" : "genre",
    "value" : "action"
}]
```

5.2.9 talked_about_count

Počet, kolikrát daný uživatel mluvil o rozdílném tématu z talked about.

```
"talked_about_count" : "1"
```

5.2.10 talked_about_same

Tento parametr nastavujeme pouze v případě, že vyhledáváme dle parametru name. Pak tímto určíme, zda mluvení o uvedeném name se bude započítávat do počtu o mluvení o jiném žánru. Osvětleme si to na našem příkladu. V podstatě nás zajímají uživatelé, kteří zmínili určitou hru a zároveň se zmínili alespoň desetkrát o hře ze stejného žánru. Jelikož jsou žánry stejné, mohou se nacházet uživatelé, kteří desetkrát mluvili o hře z daného žánru, ale desetkrát to byla naše vyhledávaná hra. Proto nastavíme hodnotu na True a tím tomuto zamezíme.

```
"talked_about_same" : "false"
```

5.2.11 aspects

Tímto parametrem specifikujeme uživatele, kteří o daném tématu napsali příspěvek s daným aspektem.

```
"aspects":[
  {
    "aspect1" : "hardware requirements"
  }
]
```

Tímto jsme si uvedli všechny parametry pro analýzu Twitteru. Pokud chceme pouze analýzu Facebooku anebo propojit výsledek analýzy Twitteru a Facebooku do jednoho souboru, přidáme další níže uvedené parametry. Jak tedy může vypadat vstupní soubor pro analýzu dat z Twitteru si nyní ukážeme.

```
{
  "dbname" : "games",
  "name" : "Battleborn",
  "year" : "2016",
  "user_number_of_posts" : "10",
  "different_topic" : "2",
  "talked_about" :[
   {
    "topic" : "genre",
    "value" : "action"
   }
  ],
  "aspects":[
   {
    "aspect1" : "hardware requirements"
   }
  ],
  "talked_about_count" : "2",
  "talked_about_same" : "true"
}
```

5.3 Vstupní data pro analýzu Facebooku

Vstupní data pro analýzu Facebooku vkládame do stejného souboru jako u Twitteru a to *input_analyze_data*.

5.3.1 dbname

Dbname slouží stejně jako u Twitteru, pro název vyhledávaného uzlu v Elasticsearch.

```
"dbname" : "fbgames"
```

5.3.2 istopic

Tento parametr slouží pro vyhledání pouze obrazu indexované stránky z Facebooku, anebo získání veškerých dat včetně komentářů atd.

```
"istopic":"0"
```

5.3.3 name

Name je název indexovaných dat v Elasticsearch stejně jako u Twitteru.

```
"name" : "Battleborn"
```

5.3.4 category

Zde nastavujeme vyhledávanou kategorii - téma. Opět to stejné jako u Twitteru.

```
"category" :[{
    "topic" : "genre",
    "value" : "action"
}]
```

5.3.5 status - day,month,year

Zde nastavujeme statusy pouze pro určité datum.

```
"status" :[{
    "year" : "2016",
    "month" : "11",
    "day" : "22"
}]
```

5.3.6 message

Parametr message slouží k vyhledávání v komentářích pod statusy.

```
"message" : "Horrible"
```

Zde jsme si uvedli všechny možné parametry pro analýzu Facebookových dat. Jak by mohl vypadat vstupní soubor si nyní ukážeme.

```
{
  "facebook":[
   {
    "dbname" : "fbgames",
    "istopic":"1",
    "name" : "Battleborn"
   }
  ]
}
```

5.4 Analýza indexovaných dat

Díky tomutu velkému množství různě nastavitelných parametrů můžeme provádět mnoho zajímavých analýz. Otevíraji se zde možnosti rozdělit uživatele do skupin. V případě her to mohou být hráči, kteří hrají hodně, tudíž hodně mluví o mnoho různých hrách z různých žánrů. Anebo hráči, kteří moc nehrají hry, takže se o hrách tolik nezmiňují.

Dle analýzy aspektů můžeme zjistit, v čem je například problém daného produktu. Pokud bude hodně pozitivních či negativních příspěvků s daným aspektem.

Můžeme analýzu Twitteru propojit s Facebookem nebo se osamostatnit pouze na Facebook. Zjistit, která rpg hra má nejvíce lajků. Nebo zjistit zda nevyšel nějaký důležitý update podle rapidního nárůstu lajků na novém statusu oproti předchozím. Pokud ano, zjistit z komentářů, jaký na to mají hráči názor.

Nyní si tedy popíšeme implementační část, a jaké výsledky vlastně získáme. Po vytvoření vstupních dat spustíme program. Data se načtou do jednotlivých proměných a začnou se vyhledávat data v Elasticsearch. Při analýze je v podstatě možné využít pouze jednoho spojení mezi daty. A to mezi vyhledávaným jménem a tématech, o kterých daný uživatel mluvil. Tímto docílíme dotazů typu: zajímají nás uživatelé, kteří komentují hru Dota2, ale v minulosti již alespoň pětkrát mluvili o různých hrách z žánru moba.

Toto spojení tedy probíhá tak, že se vyhledají všechny tweety s určitým jménem *Battleborn*. A následně se získají veškeré příspěvky každého uživatele, i ty, které neobsahuji Battleborn. A podle nich probíhá filtrování uživatelů a výsledků podle vstupních dat. Po vyfiltrování anebo spojení těchto dvou vyhledávání získáváme nějaké uživatele a jejich tweety.

Tyto tweety poté začneme analyzovat. První probíhá analýza sentimentu pomoci TextBlobu pro každý tweet. A počítají se zvlášť pozitivní, negativní a neutrální příspěvky pro daný název. Dále se počítají slova v každém tweetu pro daný název, která budou nakonec seřazena. Počítáme slova, která nas pravděpodobně zajímají, tudíž ignorujeme slova jako: the,he,will,was,for... . Na konci jsou započítaná slova seřazena a ve výsledku je zobrazeno prvních dvacet slov. Toto počítání slov následně slouží pro první pohled na danou analýzu. V případě, že hledáme například jak si nová hra stoji co se týče hardwarové specifikace, z těchto slov můžeme na první pohled vidět, že je například problém s drivery a nebo právě naopak že hra běží plynule i na starších strojích.

Po tomto už se dostáváme skoro k výslednému JSONu. Předtím než ho pošleme na standardní výstup, si vypíšeme deset nejaktivnějších uživatelů, tedy deset uživatelů s nejvíce příspěvky pro daná vstupní data. Zde následuje malé rozdvojení, buď byl ve vstupních datech i Facebook a proto nejdřív zpracujeme Facebook a přidáme výsledek k současnému výsledku anebo výsledek vypíšeme rovnou.Nejdřív se tedy podíváme na případ, kdy Facebook byl ve vstupních datech, jelikož jsme prozatím popsali jen část analýzy Twitteru.

Začátek je stejný jako u Twitteru. Vstupní data se načtou a vyhledají v Elasticsearch. Poté se začnou přidávat k současnému výsledku nebo v případě, že analyzujeme pouze Facebook, se začne vytvářet nový výstup. U analýzy Facebooku především rozlišujeme zadaný parametr istopic. V jednom případě přidáváme k výsledku pouze základní údaje o dané stránce dle názvu. V druhém případě přidáváme veškeré statusy a komentáře k nim.

5.5 Obsah výsledného JSONu

Nejdůležitější částí je vždy název, a k tomu název uživatele a veškeré jeho tweety. Tweety vždy obsahují text, lokaci a analýzu sentimentu.

```
"Battleborn": {
  "CGNpromo": [
    {
      "comment": "Damn @Battleborn is so addicting!
      more streams to come soon!\nhttps://t.co/QYL4WGctHn\n
      \n#CGN ftw @Fr3nchKitti3",
      "location": "CGN",
      "sentiment": "positive"
    },{
      "comment": "Battleborn!!!! Come say Hi! https://t.co/6PbrgGVkpd
      via @Twitch @Fr3nchKitti3 #CGN #TheDivision @girlstreamers
      https://t.co/NqiQLPwKyF",
      "location": "CGN",
      "sentiment": "neutral"
    }
  ]
}
```

Další částí je tedy celkový počet pozitivních, negativních a neutrálních tweetů.

```
"Battleborn": {
  "positive": 65,
  "negative": 22,
  "neutral": 48
}
```

Následující výčet nejčastěji zmiňovaných slov ve všech typech sentimentu. Na ukázku si uvedeme pouze jeden typ sentimentu a výčet prvních tří slov.

```
"negative_text": [
  [
    "Ready",
    19
  ],[
    "Driver",
    10
  ],[
    "Nvidia",
```

```
    9
  ]
]
```

Dále výsledný JSON obsahuje Facebookovou část a to buď část, kde je pouze základní pohled na danou stránku v následujícím tvaru.

```
"facebook": {
  "topic": [
    [
      "action"
    ]
  ],
  "talking_about": 76320,
  "likes": 256695
}
```

Nebo se vyskytuje část se zmíněnými statusy informacemi k nim, a jejich komentáři. Opět si ukážeme pouze část tohoto výstupu.

```
"facebook": [
  {
    "date": "30-04-2016",
    "status": "Ready to rock on a whole new level! #Battleborn heroes
    will be in Rock Band 4!",
    "status_likes": 136,
    "comments": {
      "Benjamin Scott Firchau": {
        "message": "Hopefully I can have a band of different coloured
        miko's and call it shrooms",
        "likes": 2
       },
       "Matt Davies": {
         "message": "Any clue when the servers go live?
         Got my copy and eager to play :(",
         "likes": 0
       }
  }
]
```

5.6 Analýza lokace v tweetu, aneb kam uživatelé cestují

Zbývá nám ještě jedna část analýzy, a to kam uživatelé cestují. Při indexování jsme v tweetu rozeznávali aspekty, ale i lokace. A právě tahle část se zabývá zmíněnou lokací. Veškerá analýza je implementována v souboru *analyze_loc.py* . Cílem této analýzy bylo ukázat, kam lidé cestují. Rekněme tedy, že si zvolíme získávat data z naších zmiňovaných sociálních sítí, která se týkají různých nemocí. Tedy v této chvíli nás například zajímají uživatelé, kteří cestovali někam do jihovýchodní Asie a mluvili o nějaké nemoci.

Pro spuštění této analýzy je potřeba vytvořit vstupní soubor s různými parametry, kteý se musí jmenovat *input_loc_data*.

5.6.1 dbname

Opět stejně jako u předchozích analýz název uzlu v Elasticsearch.

```
"dbname": "games"
```

5.6.2 user

Tímto parametrem zvolíme přímo jednoho uživatele a zjistíme, kam všude cestoval.

```
"user": "PetrSkyva"
```

5.6.3 travel to

Zde nastavíme, kam lidé cestovali, a získáme veškeré dané uživatele, kteří cestovali do dané oblasti.

```
"travel to" : [
  "SE Asia"
]
```

Jakmile máme vstupní soubor, můžeme spustit daný program a získáme výsledky. Výsledky se zobrazují ve tvaru slovníku. Uveďme příklad, kdy se zajímáme, kteří uživatelé cestovali do jihovýchodní Asie. Výsledkem tedy bude vždy uživatel, daná lokace a státy v dané lokaci.

```
{u'PetrSkyva': {u'SE Asia': [u'Laos', u'Vietnam']},
u'BigxDeuce': {u'SE Asia': [u'Bali']}}
```

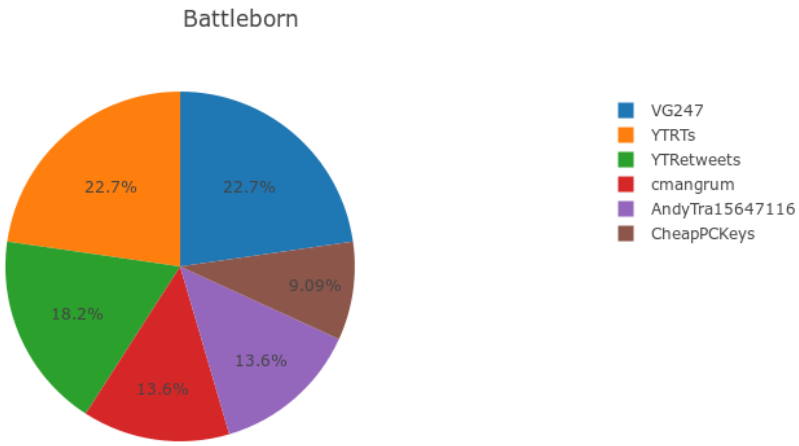
5.7 Vizualizace výsledku

Nyní již víme, jak jsou strukturované vstupní soubory, co tedy můžeme analyzovat. Také víme, jak vypadají výstupní data, a proto si nyní ukážeme, jak můžeme tato data vizualizovat pomocí Plotly. Ukážeme si, jaké grafy systém podporuje případně jak vytvořit z výstupního JSON souboru nové grafy v podstatě jakéhokoliv typu.

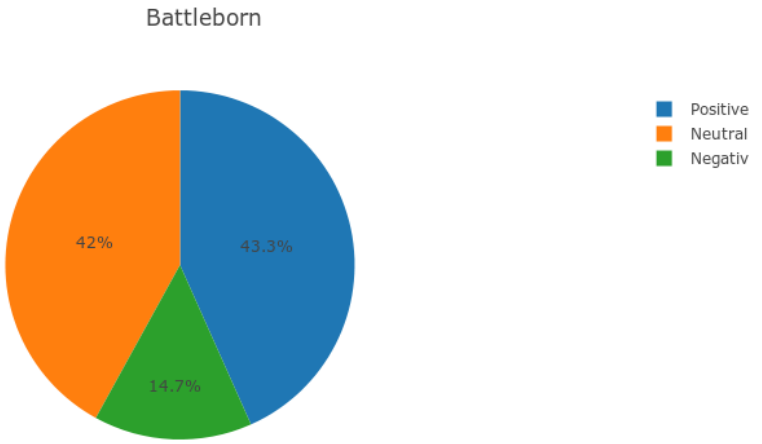
Jelikož PlotLy má výborně popsané API a mnoho základních příkladů, jak generovat různé typy grafů v Pythonu, je pak jen na uživateli, aby si vytvořil grafy, které potřebuje pro analyzovaná data. My v našem systému tedy generujeme grafy z výsledného JSONu a nejvíce aktivních uživatelů Twitteru. Co se Facebooku týče generujeme grafy díky celkovému počtu lajků všech statusů za jeden den. Systém na ukázku generuje čtyři druhy grafů. Dva pro data z Twitteru a dva pro data z Facebooku.

5.8 Grafy z dat Twitteru

Jak jsme zmínili, generujeme dva grafy. První graf znázorňuje uživatele s nejvíce příspěvky. Na druhém grafu znázorňujeme celkový počet pozitivních, negativních a neutrálních příspěvků.



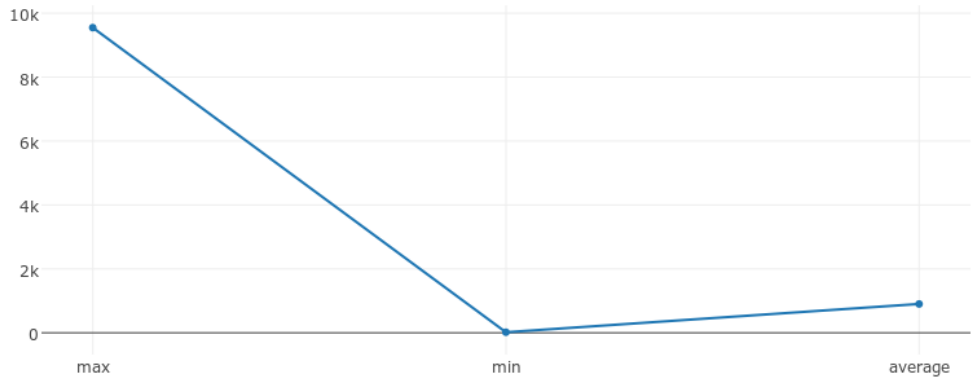
Obrázek 5.2: Prvních 10 uživatelů s nejvíce příspěvky.



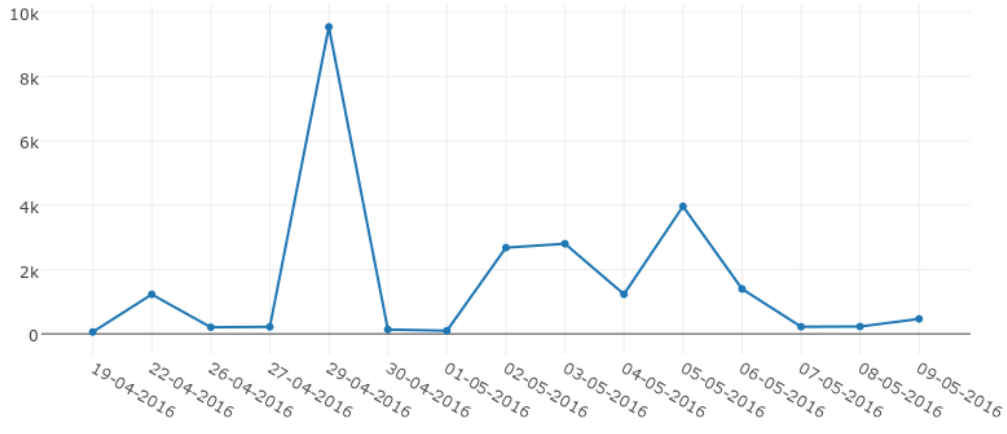
Obrázek 5.3: Celkový počet pozitivních, negativních a neutrálních tweetů.

5.9 Grafy z dat Facebooku

I pro Facebooková data generujeme dva grafy. První graf znázorňuje minimum, maximum a průměr lajků na všech statusech pro danou stránku. Na druhém grafu vidíme histogram, kde je vyobrazen počet lajků pro každý den, ve kterém byl napsán alespoň jeden status.



Obrázek 5.4: Minimum, maximum a průměr lajků pro danou stránku.



Obrázek 5.5: Histogram počtu lajků na den.

Kapitola 6

Příklad použití systému

V této kapitole si demonstrujeme vytvořený systém na různých příkladech. Všechny příklady se budou týkat počítačových her, protože sběr dat probíhal pouze pro data tohoto typu. V Elasticsearch je zaindexováno přes 200 000 tweetů týkajících se počítačových her. U každé hry jsme rozlišovali žánr. Některé hry mohou mít samozřejmě více žánrů. Dále vyplněné aspekty jsou hardwarovou specifikaci, příběh, grafiku a cenu. Ukážeme si jednoduché příklady, až po ty složitější.

U všech tweetů bohužel nejsou vyplněny aspekty a lokace, protože sběr dat probíhal již v době, kdy tato funkce nebyla prozatím naimplementována. Přeindexovat veškerá data by zabralo značnou dobu, z tohoto důvodu bylo rozhodnuto, že se data nechají jak jsou a začnou se jen přidávat již správná další data.

6.1 První příklad

V prvním příkladě se podíváme na hru Battleborn, jelikož byla chvíli v Betě a teprve nedávno vyšla. Proto nás pro začátek zajímají poze veškeré tweety o dané hře. Jak si v současné době tedy vede, jestli se mluví více pozitivně a nebo negativně. A podíváme se na nejaktivnější uživatele a nejpoužívanější slova. Vstupní data tedy budou nejjednodušší.

```
{
   "dbname" : "games",
   "name" : "Battleborn"
}
```

Výsledky analýzy

- Celkový běh systému než se zobrazil výsledek - 2 minuty 42 vteřin,
- celkem 52 985 tweetů,
- stoho 25 872 pozitivních,
- 19 822 neutrálních,
- 7 291 negativních,
- nejaktivnější uživatel NevadaBaseball - 170 tweetů,

- druhý BattlebornPAX - 123 tweetů,
- třetí Battleborn - 115 tweetů,
- nejčastější slova u pozitivních tweetů - live, more, Beta, win, chance, video, PC, Open, beta, PS4, awesome, good, fun,
- nejčastější slova u negativních tweetů - Tap, beta, Ready, 2K, Android, Overwatch, iOS, bad, animated, Launches, intro,
- nejčastější slova u neutrálních tweetů - beta, Overwatch, PC, today, launch, PS4, Open, playing, video, join, HDTV,
- Pozitivní tweet: RT @Battleborn: Dive into Battleborn with guns blazing! Or swords, or rockets The BadassBeta is now available on #PS4!,
- Pozitivní tweet: @Battleborn 15-0! I love the beta,
- Negativní tweet: RT @Jorraptor: The OPEN BETA for Battleborn only has 276 viewers on Twitch, this is bad, really bad.,
- Negativní tweet: Battleborn, to buy or not to buy! I mean it looks fun but I dunno.....,
- Neutrální tweet: Hmm. Its almost the end of the working day. Still no review code for Battleborn. Shall I buy it?,
- Neutrální tweet: We Can't Choose Between Battleborn or Overwatch.

Co jsme tedy zjistili? Výsledky Battlebornu jsou více než výborné, jelikož polovina tweetů je pozitivní a pouze jedna sedmina je negativní. Nicméně narazili jsme na jeden problém a tím je nejaktivnější uživatel. Očividně se slovo Battleborn vyskytovalo i v baseballu, což je taky hra, ale ne ta, která nás zajímá. Dále u nejčastějších slov vidíme, že lidé jsou pravděpodobně především pozitivní z toho, že je beta otevřená pro všechny. Pravděpodobně bude i více lidí hrát hru na PC než na konzolích a baví je. Z negativních slov můžeme zjistit, že se něco odehrává na mobilních zařízeních a nemá to úspěch. U neutrálních tweetů se nám jen potvrzuje domněnka, že více lidí spouští hru na PC.

6.2 Druhý příklad

V druhém příkladu se podíváme opět na základni dotaz a to na hru ze stejného žánru jako Battleborn, a to Counter Strike. Uvidíme tedy srovnání mezi dvěma současně hodně diskutovanými hrami.

```
{
   "dbname" : "games",
   "name" : "Counter Strike"
}
```

Výsledky analýzy

- Celkový běh systému než zobrazil výsledek - 2 minuty 37 vteřin,

- celkem 49 579 tweetů,
- z toho 18 960 pozitivních,
- 27 957 neutrálních,
- 2 662 negativních,
- nejaktivnější uživatel csgoreddit - 926,
- druhý playerscsgo - 479)
- třetí CSGOPlayerGE - 357
- nejčastější slova u pozitivních tweetů - CSGO, enter, Giveaway, AK-47, Daily, Hydroponic, Doppler, Karambit, Tiger, win,
- nejčastější slova u negativních tweetů - CSGO, video, liked, Wear, Minimal, AK-47, entered, vulcan, Black,Giveaway,
- nejčastější slova u neutrálních tweetů - CSGO, Enter, giveaway, skin, Fade, Serpent, AK-47, follow, asiimov, Buttlefly,
- Pozitivní tweet: I liked a @YouTube video https://t.co/6pJBulDhsk CS GO - CRUISE GLOBALS!! (Counter Strike Global Offensive Gameplay!),
- Pozitivní tweet: I liked a @YouTube video from @m3rkmus1c https://t.co/vjGJl3x5VG BOWIE KNIFE UNBOXING!? (CSGO Bowie Knife),
- Negativní tweet: RT @mattelmer_: When the rank was on point #csgo #riprank,
- Negativní tweet: Clayton wasted his csgo account just trying to beat me in a 1v1 by hacking and still lost all 3 games,
- Neutrální tweet: AK-47 Fire Serpent (FN) #CSGO skin giveaway!,
- Neutrální tweet: 1v1 csgo after school tomorrow @SteveChahkah.

Co jsme tedy zjistili? Výsledky jsou opět velice kladné. Negativní tweety se skoro nevyskytují, ale problém je, jak se o této hře mluví. Vetšina tweetů se týká rozdávání skinů. Proto je i analýza sentimentu ne úplně správná. Nicméně vidíme, že oproti Battlebornu se zde vyskytují mnohem aktivnější uživatelé. Zde nejčastější uživatel na skoro stejný celkový počet tweetů napsal sedmkrát víc tweetů než nejaktivnější uživatel u Battlebornu. Nejčastější slova nám jen potvrzují domněnku, že většina tweetů se týká nějaké reklamy a soutěží.

6.3 Třetí příklad

Nyní tedy zjistíme, jaký názor mají na Battleborn pouze zkušení hráči. Co budou pro nás zkušení hráči? Zmínili se o Battlebornu alespoň pětkrát, napsali alespoň o dvou různých hrách ze žánru jako je Battleborn tedy action. Dále k naší analýze přidáme i Facebook. Bude nás ale zajímat, kolik má Battleborn lajků na Facebooku a kolik o něm mluví lidí.

```
{
  "dbname" : "games",
  "name" : "Battleborn",
  "user_number_of_posts" : "5",
  "different_topic" : "2",
  "talked_about" :[{
    "topic" : "genre",
    "value" : "action"
  }],
  "talked_about_count" : "2",
  "talked_about_same" : "true",
  "facebook":[{
    "dbname" : "fbgames",
    "istopic":"1",
    "name" : "Battleborn"
  }]
}
```


Výsledky analýzy

- Celkový běh systému než zobrazil výsledek - 1 minuta 55 vteřin,
- celkem 2 277 tweetů,
- stoho 1 067 pozitivních,
- 973 neutrálních,
- 237 negativních,
- nejaktivnější uživatel TwitchGro - 76,
- druhý TwitchShare - 74
- třetí zimisss - 72
- nejčastější slova u pozitivních tweetů - live, full, come, more, new, win, PS4, chance, Beta, PC, Xbox, Follow,
- nejčastější slova u negativních tweetů - Deals, Daily, TV, max, mad,4K, tap, bad, 2K, iOS, live, few, gearbox,
- nejčastější slova u neutrálních tweetů - Come, beta, review, overwatch, progress, playing, week, PS4, Join,
- Pozitivní tweet: Battleborn is so amazing, the writing is so good. And all the different characters are super fun to play as,
- Pozitivní tweet: RT @Battleborn: Swords are sharp, rockets are mounted, queues are clear. Play the Battleborn BadassBeta today on PS4!,
- Negativní tweet: Wish I could bring y'all some battleborn videos but can't afford the game on XB1,

- Negativní tweet: @Battleborn Killed Guardian Vyn and it's still showing as an active quest and I can't go anywhere?,
- Neutrální tweet: Streaming some @Battleborn. Come hang out.,
- Neutrální tweet: Of to work. When I'm home Il be streaming some battleborn battleborn twitch xboxone,
- celkový počet lajků na Facebookové stránce 281 183,
- celkový počet lidí mluvící na Facebooku o Battlebornu 42 172.

Co nám ukázal tento příklad? Jakmile se zaměříme na hráče, kteří hrají víc, výsledky jsou mnohem méně obsáhlé. Aktivních hráčů z celkového počtu tweetů není ani jedna desetina. I přesto ale opět převládá pozitivní názor nad negativním. A podle pozitivních slov můžeme usoudit, že aktivnější hráči jsou zde více z playstationu než z PC. Bohužel se ale většina tweetů zaměřuje především na reklamu na svůj stream ze hry nebo něco podobného. Na konci můžeme vidět, jak si Battleborn vede na Facebooku. Má velký počet lajků, ale přesto i když nahlédneme na stránku, moc lidí nemá zájem o novinky. Více se na toto téma podíváme v následujícím příkladě.

6.4 Čtvrtý příklad

V tomto příkladě se podíváme na aspekty, a také se podíváme detailněji na Facebook. Opět budeme chtít alespoň hráče, kteří zmínili Battleborn alespoň pětkrát. A co se týče aspektů, zaměříme se na hardwarovou specifikaci.

```
{
  "dbname" : "games",
  "name" : "Battleborn",
  "user_number_of_posts" : "5",
  "aspects":[{
    "aspect1" : "hardware requirements"
  }],
  "facebook":[{
    "dbname" : "fbgames",
    "istopic":"0",
    "name" : "Battleborn"
  }]
}
```

Výsledky analýzy

- Celkový běh systému než zobrazil výsledek - 17 vteřin,
- celkem 1 564 tweetů,
- stoho 705 pozitivních,
- 552 neutrálních,

- 307 negativních,
- nejaktivnější uživatel 8rend - 6,
- druhý VarunV3rma - 6
- třetí IAmSp00n - 5
- nejčastější slova u pozitivních tweetů - drivers, geforce, latest, nvidia, prep, team, time, live, save,
- nejčastější slova u negativních tweetů - ready, drivers, nvidia, released, geforce, new, driver, PC,
- nejčastější slova u neutrálních tweetů - see, why, gave, watch, launch, tomorrow, trailer, giving,
- Pozitivní tweet: Battleborn Tap available now for iOS and Android (via @NewsfusionApps Gaming News),
- Pozitivní tweet: In love with Battleborn's Prologue Cinematic. Excellent animation from all involved. @Jeff_Lai's FX are to die for!,
- Negativní tweet: GearboxSoftware hey guys/girls, is there gonna be a patch for AMD cards for the performance issues with Battleborn? Games unplayable for me,
- Negativní tweet: @TeamRamTodd @Battleborn We think it may apply when you launch game and get to title screen. Let me know what happens.,
- Neutrální tweet: Nvidia Geforce 365.10 WHQL Driver Released,
- Neutrální tweet: Battleborn Review in Progress,
- maximální počet lajků na status na Facebookové stránce 9 548,
- minimální počet lajků na status na Facebookové stránce 18,
- průměrný počet lajků na status na Facebookové stránce 902.

Co nám ukázal tento příklad? Opět jsme se zaměřili na hráče, kteří hrají aktivněji tudíž výsledků je míň. Nicméně zaměřili jsme se na aspekt hardwarové specifikace a zde vidíme, že pozitivní příspěvky již až tak nepřevládají a z nejčastějších slov můžeme odvodit že byl nějaký problem s drivery. Také vidíme že Battleborn má i nějakou aplikaci na iOS i Android. Na konci statistiky se můžeme podívat jak si vede Battleborn na Facebooku co se týče komentářů. V minulém příkladu jsme viděli že stránku Battleborn má olajkovanou přes 280 tisíc lidí a i přesto průměrný počet lajků na komentář je extrémě nízký. Z JSON výsledku vidíme že nejzajímavější komentář je ten, kdy oznámili vypuštění hry Battleborn na určité datum.

Kapitola 7

Závěr

7.1 Vyhodnocení výsledků práce

V úvodu práce jsme se podívali na různé sociální sítě a popsali si je. Ukázali, které budou pro nás zajímavé, a které budeme využívat v naší práci. Dále jsme si vysvětlili principy sběru dat a ukázali, jaká data nás zajímají a jak je budeme zpracovávat.

V polovině práce jsme si ukázali, jak začít pracovat se systémem. Jak vytvářet vstupní soubory a jaké jsou možnosti systému. U toho jsme si i ukázali implementaci a použité technologie. Na konci práce jsme použili náš vytvořený systém na ruzných příkladech.

Všechny příklady se odehrávaly na tématu počítačových her, protože je to velice živé a diskutované téma na sociálních sítích a není nouze o data. Proběhla tedy řada testů použití systému. Prakticky jsem ukázali, jaká možná data jsou se systémem možná analyzovat a vytvářet.

Výsledkém práce je tedy systém implementovaný v jazyce Python, který je schopný pravidelně získávat, indexovat a analyzovat stahovaná data. Při analýze dat probíhá klasifikace dat podle různých zadaných aspektů. Podle počtu nejčastějších slov jsme schopní analyzovat současné trendy v daném analyzovaném odvětví. Tyto výsledky nakonec je možné vizualizovat pomocí grafů, případně pouze prohlédnout textový výsledek.

7.2 Co udělat do budoucna

Do budoucna by bylo vhodné, aby systém podporoval více jazyků než jen angličtinu. Bylo by tedy potřeba implementovat vlastní analyzátor sentimentu. Dále by bylo vhodné přidat další sociální sítě jako LinkedIn, Instagram či cokoliv podobného. Jako poslední by bylo vhodné refaktorovat spojování vyhledávání v Elasticsearch, abychom zajistili větší rychlost analýzy.

Literatura

[1] *API*. [online]. [cit. 2016-04-28]. Dostupné z: http://www.webopedia.com/TERM/A/API.html.

[2] *Elasticsearch*. [online]. [cit. 2016-05-01]. Dostupné z: https://www.elastic.co.

[3] *Instagram*. [online]. [cit. 2016-05-01]. Dostupné z: https://cs.wikipedia.org/wiki/Instagram.

[4] *JSON*. [online]. [cit. 2016-04-28]. Dostupné z: http://www.json.org/json-cz.html.

[5] *LinkedIn*. [online]. [cit. 2016-05-01]. Dostupné z: https://cs.wikipedia.org/wiki/LinkedIn.

[6] *OAuth*. [online]. [cit. 2016-05-01]. Dostupné z: http://oauth.net.

[7] *SIREn*. [online]. [cit. 2016-04-25]. Dostupné z: http://siren.solutions/searchplugins/join/.

[8] *Sociální síť*. [online]. [cit. 2016-04-25]. Dostupné z: https://cs.wikipedia.org/wiki/Sociáln%C3%AD_s%C3%ADť.

[9] Developers, P.-T.: *Python Twitter*. [online]. [cit. 2016-04-19]. Dostupné z: https://github.com/bear/python-twitter.

[10] Havlová, J.: *Sociální síť*. [online]. [cit. 2016-04-25]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000015947&local_base=KTD.

[11] Heppermann, C.: *Twitter: The Company and Its Founders*. Abdo Publishing Company, 2012, ISBN 1617833371.

[12] Kulhánková, H.: *Fenomén facebook*. BigOak, 2010, ISBN 978-80-90764-0-0.

[13] Loria, S.: *Textblob*. [online]. [cit. 2016-04-25]. Dostupné z: http://textblob.readthedocs.io/en/dev/.

[14] Mobolic: *Facebook-sdk*. [online]. [cit. 2016-05-01]. Dostupné z: https://github.com/mobolic/facebook-sdk.

[15] Postlethwaite, B.: *PlotLy*. [online]. [cit. 2016-04-22]. Dostupné z: https://plot.ly/python/.

[16] Russell, M. A.: *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More 2nd Edition*. O'Reilly Media, 2013, ISBN 1449367615.

Přílohy

Seznam příloh

Příloha A

Obsah CD

Přiložené CD obsahuje tyto adresáře:

- */doc* - zdrojové soubory technické zprávy,
- */src* - zdrojové soubory technické práce,
- */xskyva02_BP.pdf* - technická zpráva ve formátu pdf,
- */plakat.pdf* - plakát k technické práci ve formátu pdf,
- */tests* - ukázkový vstup/výstup aplikace, JSON a grafy,
- */elastic* - soubory nutné pro spuštění Elasticsearch.

Příloha B

Manual

B.1 Spuštění Elasticsearch

Jako první spustíme Elasticsearch. A to takto *./siren-elasticsearch-1.4-bin/example/bin/elasticsearch*. Po spuštění Elasticsearch je pořeba nakonfigurovat uzly které budeme využívat. První vytvoříme uzel zadáním následujícího příkazu do terminálu (slovo games nahradíme za námi zvolený nový uzel):

```
curl -XPUT "http://localhost:9200/games"
```

Dále ho namapujeme a nastavíme (opět slovo games změníme za námy vytvořený uzel v předchozím kroku).

```
curl -XPUT "http://localhost:9200/games/chargepoint/_mapping" -d '
{
  "chargepoint" : {
    "properties" : {
      "_siren_source" : {
        "analyzer" : "concise",
        "postings_format" : "Siren10AFor",
        "store" : "no",
        "type" : "string"
      }
     },
     "_siren" : {},
     "_timestamp": {
       "enabled": true
     }
   }
}'
```

Nyní již můžeme začít získavat případně analyzovat data.

B.2 Spouštění získávání dat

Pro spustění získávání dat z Twitteru musíme vytvořit vstupní soubory.

- *hashtags.txt* - vstupní soubor pro získávání dat pomocí hashtagů,

- *search_input.txt* - vstupní soubor pro získávání dat pomocí streamu,
- *search_aspects.txt* - vstupní soubor s aspekty a lokacemi, jak pro hashtagy tak pro stream.

Spouštění zahájíme spuštěním souboru *python search.py*. Na konci souboru se vyskytují dvě volání funkce. *mainHashTags()* a *streamerTrack()*. Pokud chceme spustit získávání dat pomocí hashtagů, zakomentujeme druhou funkci. Pokud chceme získávat data pomocí streamu, zakomentujeme první funkci.

Pro spuštění získávání dat z Facebooku musíme vytvořit vstupní soubor.

- *fb_input.txt* - vstupní soubor pro získávání dat z Facebooku.

Pro spuštění získávání dat z Facebooku musime v souboru *fb.py* odkomentovat poslední řádek *getGroupData()*. Následně spustíme získávání dat *python fb.py*. Pokud chceme analyzovat data z Facebooku, je potřeba tento řádek opět zakomentovat před spuštěním *analyze.py*.

B.3 Spouštění analýzy dat z Twitteru a Facebooku

Před spuštěním analýzy je potřeba vytvořit tento soubor.

- *input_analyze_data* - vstupní soubor pro analýzu dat ať pro Twitter nebo Facebook.

Analýzu spouštíme pomocí příkazu *python analyze.py > vystup*. Můžeme to zvolit i bez výstupu, ale to se veškeré výsledky vytisknou na standardní vstup.

B.4 Spouštění analýzy lokace, aneb kam lidé cestují

Před spuštěním analýzy je potřeba vytvořit tento soubor.

- *input_loc_data* - vstupní soubor pro analýzu dat cestování.

Analýzu spouštíme pomocí příkazu *python analyze_loc.py > vystup*. Výstup opět nemusíme zvolit, ale je to vhodné.