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

ANALÝZA POSTOJŮ V OBLASTI AUTOMOBILOVÉHO PRŮMYSLU

SENTIMENT ANALYSIS IN AUTOMOTIVE INDUSTRY

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **ADAM BEZÁK**
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 bakalářské práce

Řešitel: **Bezák Adam**
Obor: Informační technologie
Téma: **Analýza postojů v oblasti automobilového průmyslu**
Sentiment Analysis in Automotive Industry
Kategorie: Algoritmy a datové struktury

Pokyny:

1. Seznamte se s přístupy a metodami počítačové analýzy postojů a se způsoby získávání dat ze sociálních sítí a dalšího, uživateli generovaného obsahu.
2. Shromážděte diskusní příspěvky ze serverů, věnujících se tématům v oblasti automobilového průmyslu, a z dalších zdrojů a zpracujte je do podoby vhodné pro průběžné testování systému.
3. Navrhněte a implementujte systém, který dokáže indexovat a analyzovat stahovaná data.
4. Vytvořte systém pro automatickou klasifikaci shromažďovaných dat, analýzu trendů a vizualizaci výsledků.
5. Demonstrujte vytvořený systém na vhodně zvolených příkladech ze zpracované datové sady.
6. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- Manning, C. D., Schütze, H., Foundations of Statistical Natural Language Processing, MIT Press, 1999, ISBN 0-262-13360-1.

Pro udělení zápočtu za první semestr je požadováno:

- funkční prototyp řešení

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 2016
Datum odevzdání: 17. května 2017

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

Cieľom tejto práce je oboznámiť sa so základnými metódami analýzy postojov na sociálnych sieťach. Téma práce je zameraná na automobilový priemysel, avšak princíp práce je možné použiť na akékoľvek iné skúmané odvetie. Podstatou praktickej časti je získanie dát zo sociálnych sietí, ich analýza a následná indexácia do ElasticSearch databáze. Ďalším cieľom práce je tieto dáta vizualizovať prostredníctvom portálu. Vytvorený webový portál poskytuje rôzne štatistiky popredných automobilových značiek, prehľad nových trendov alebo vizualizáciu názorov na konkrétne aspekty jednotlivých automobilov.

Abstract

The main theme of this thesis is to familiarize with the basic methods of sentiment analysis on social networks. Thesis's theme is aimed on the automotive industry, although this prinicipal can be used in any different examined branch. The basis of the practical part is to obtain data from the social networks, analyze them and then index them into ElasticSearch database. Another goal of the thesis is to visualize these data by means of a web portal. Created web portal provides various statistics of the leading automobile brands, an overview of new trends or the aspect visualization of the individual cars.

Kľúčové slová

Twitter, Facebook, extrakcia dát, analýza postojov, vizualizácia, strojové učenie, lexikón

Keywords

Twitter, Facebook, data extraction, sentiment analysis, visualization, machine learning, lexicon

Citácia

BEZÁK, Adam. *Analýza postojů v oblasti automobilového průmyslu.* Brno, 2017. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Smrž Pavel.

Analýza postojů v oblasti automobilového průmyslu

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Doc. RNDr. Pavla Smrža Ph.D. Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

.....
Adam Bezák
15. mája 2017

Poďakovanie

Týmto by som chcel poďakovať vedúcemu práce pánovi Doc. RNDr. Pavlovi Smržovi Ph.D. za odbornú pomoc, cenné rady a konzultácie.

Obsah

Kapitola 1

Úvod

Sociálne siete sú všade okolo nás. V dnešnej dobe ich využíva takmer každý. Ich používatelia môžu verejne zdielať svoje názory, nálady a vďaka tomu sa otvára nový rozmer pre analýzu dát. Najväčšou výhodou sociálnych sietí, z hľadiska analýzy dát, je získanie veľkého množstva názorov jednoducho a rýchlo. Istým typom sociálnej siete je možné nazvať aj niektoré webové portály, ktoré spájajú informácie z rôznych zdrojov jednotným spôsobom. Napríklad také, kde obsah tvoria príspevky napísané rôznymi užívateľmi.

Hlavnými cieľmi praktickej časti práce je vytvoriť všeobecný systém pre automatické sťahovanie, analyzovanie a vizualizáciu dát. Analýzou sa myslí predovšetkým určenie jedného z troch typov postoja príspevku (pozitívny, negatívny, neutrálny). Daný systém je inštanciovaný na podporu analýzy dát v automobilovom priemysle a je zameraný na vizualizáciu dát z viacerých zdrojov, predovšetkým na sociálnu sieť Twitter. Rozhranie medzi dátami a užívateľom je tvorené jednoduchým a prehľadným webovým portálom, ktorý uľahčí prácu ľudom, ktorých zaujímajú aktuálne automobilové trendy na sociálnych sieťach, chcú mať prehľad o najviac diskutovaných automobilových značkách alebo ich zaujíma názor na konkrétny model auta.

Predmetom teoretickej časti práce je poskytnúť čitateľovi prehľad predovšetkým o dvoch populárnych sociálnych sieťach (Twitter a Facebook) a jednom z najväčších automobilových portálov[1]. Zameriava sa taktiež aj na analýzu postojov a zohľadňovanie aspektov (kapitola 2). V 3. kapitole je navrhnutá koncepcia celého systému a analýza požiadaviek na tento systém. Implementácia je popísaná v kapitole 4, rozdelená na vnútornú štruktúru a logiku systému, implementáciu webového portálu a vizualizáciu výsledkov. Kapitola 5 opisuje príklady použitia a taktiež sa zaoberá experimentami práce užívateľov s webovým portálom. V kapitole 6 sa porovnáva presnosť a rýchlosť natrénovaného klasifikátora s použitou knižnicou na analýzu postoja a vyhodnocuje sa užívateľské rozhranie.

[1] recenzný portál http://www.parkers.co.uk

Kapitola 2

Rozbor riešenej problematiky

Táto kapitola sa venuje všeobecnému prehľadu o sociálnych sieťach, možnostiach získavania dát, ďalej opisu použitých sociálnych sietí. Taktiež i vysvetleniu základnych pojmov a teórií z oblasti analýzy dát, ktoré sa v práci ďalej používajú.

2.1 Sociálne siete

Sociálna sieť je termín používaný na opis webových služieb, ktoré umožňujú jednotlivcom vytvoriť verejný profil. Následne môžu komunikovať s ostatnými používateľmi prípadne zdielať svoje názory a myšlienky[2].

Sociálne siete získali významnú pozornosť v poslednom desaťročí. Stránky sociálnych sietí sú bežne známe pre šírenie nových informácií, zdielaní osobných aktivít, recenzií produktov, verejnému zdieľaniu obrázkov, reklám alebo vyjadrovaniu svojich postojov a stanovísk k určitým záležitostiam. Zdieľané novinky, najnovšie aktuality, politické debaty sú taktiež zverejnené a analyzované na sociálnych sieťach. Bolo zistené, že oveľa viac ľudí sa začína spoliehať na sociálne siete. Užívatelia sa niekedy rozhodnú na základe informácií zverejnených od neznámych jednotlivcov a spoliehajú sa na ne[14].

2.1.1 Extrakcia dát zo sociálnych sietí

Pomerne veľká závislosť na sociálnych sieťach vedie k vytváraniu objemných dát charakteristických tromi výpočtovými problémami a to:

- veľkosť
- šum
- dynamika

Dané problémy často zapríčiňujú, že dáta zo sociálnych sietí sú veľmi komplexné a je ich zložité analyzovať ručne. Z toho plynie potreba využiť príslušné výpočtové prostriedky k ich analýze. Dolovanie dát poskytuje celú radu techník pre detekciu užitočných vedomostí z masívnych dátových súborov. Tieto techniky využívajú pred-spracovanie dát, analýzu dát a interpretáciu konečných poznatkov[2].

Uplatnenie týchto techník na objemné dáta zo sociálnych sietí má potenciál pomôcť zlepšiť výsledky vyhľadávania pre bežné vyhľadávače, pre inštitúcie realizovať cielený marketing, pomôcť pochopiť správanie ľudí alebo personalizovať webové služby pre spotrebiteľov. Dokonca je možné odhaliť a zabrániť šíreniu nevyžiadanej pošty pre všetkých z nás. Kedže sú väčšinou dáta verejné, tak je možné tieto techniky skúmať a zlepšovať ich výkonnosť[3].

2.1.2 Dolovanie dát a automobilový priemysel

Automobilový priemysel je jedným z najväčších ekonomických sektorov na svete s viac než 90 miliónmi dopravných prostriedkov. V tomto sektore vládne veľká rivalita a to vyžaduje, aby predajcovia a automobilové spoločnosti starostlivo analyzovali a zaujímali sa o názory spotrebiteľov pre dosiahnutie konkurenčnej výhody na trhu. Analýza názorov spotrebiteľov na sociálnych sietach je dobrý spôsob ako zlepšiť svoje marketingové ciele a zvýšiť predaj automobilov.

Ako bolo spomenuté v 2.1, sociálne siete vedú k neuveriteľne vysokej úrovni komunikácie medzi užívateľom a predávajúcim. Podľa (2014 CMO council report) 23% kupcov áut konzultovalo kúpu s druhými užívateľmi a 38% spotrebiteľov povedalo, že by využili sociálne siete pred ďalšou kúpou[17].

2.2 Analýza postojov

Analýza postojov je dôležitý typ analýzy textu, ktorý sa zameriava na podporu rozhodovania extrakciou a analýzou názorovo orientovaného textu. Rozhodovací proces ľudí je ovplyvnený taktiež názormi formovanými ich okolím. Ak má osoba v pláne kúpiť nejaký produkt on-line, zvyčajne začne hľadaním recenzií a názorov druhých používateľov. To je dôvod, že analýze postojov, ako oblasti výzkumu, je venovaná značná pozornosť v posledných rokoch. Má široké spektrum využiteľnosti v rôznych aplikáciach, kde zistené informácie môžu pomôcť ľudom alebo spoločnostiam pri rozhodovaní[5].

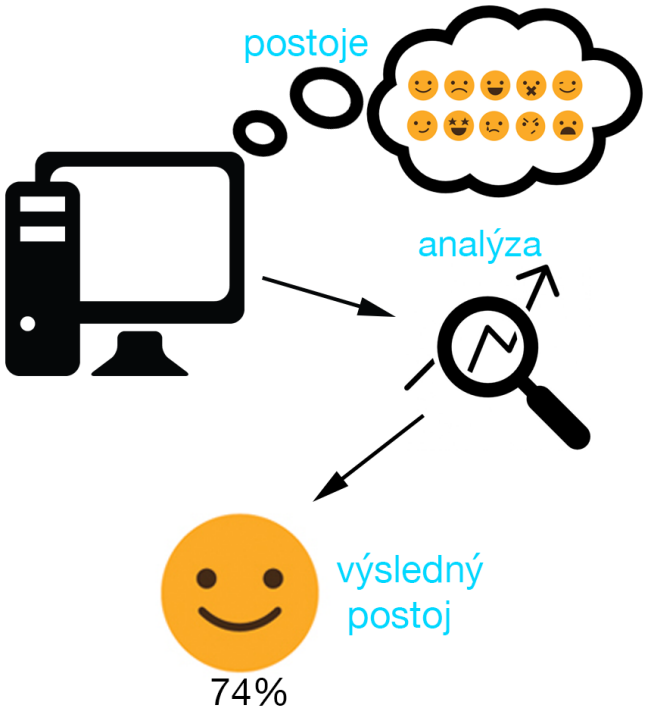
Základnou úlohou je klasifikácia polarity (škála od pozitívnych, cez neutrálne až po negatívne) daného textu na základe vyjadreného názoru. Väčšinou polaritu delíme na tri základné typy:

- pozitívna
- negatívna
- neutrálna

Avšak v skutočnosti sa v praxi môžeme stretnúť aj s rozdelením iba na dva typy (pozitívny a negatívny) alebo s využitím ohodnotenej stupnice, ktorá udáva do akej miery je postoj autora pozitívny (alebo negatívny).

Postoj emotikonov

Emotikon, ako je ":-)", je skratkou pre výraz tváre, ktorý umožňuje autorovi vyjadriť jeho nálady, emócie. Pomáha zapôsobiť na čitateľa a zlepšuje pochopenie významu postoja textu. Emotikony sú krok vpred a vyvíjajú sa spoločne s modernými komunikačnými prostriedkami. Rozdeľujeme veľa druhov emotikonov, ako napríklad, mimika tváre, pocity, počasie, dopravné prostriedky, jedlo, nápoje, zvieratá atď... Postupom času sa stávajú čoraz populárnejšie, a to vzbudzuje potrebu využiť ich význam a zahrnúť ich do analýzy postojov na sociálnych sieťach[12].



Obr. 2.1: Diagram analýzy postojov.

2.2.1 Aspekty

Pri analýze postojov užívateľovi často nevyhovuje rozdelenie len na základné triedy postoja (pozitívny, neutrálny, negatívny), ale zaujímajú ho, ktoré aspekty boli pozitívne alebo negatívne. Aspekt je teda zvyčajne entita reálneho sveta. Internetové stránky obsahujú značné množstvo recenzií na rôzne produkty. V týchto recenziách, ľudia chvália a kritizujú rozličné aspekty skúmaného tovaru. Je teda potrebné zahrnúť tieto skúmané aspekty k zisteniu celkovej polarity a určeniu pozitívneho (alebo negatívneho) postoja. Nie všetky analyzované dokumenty sú jednoznačné na určenie aspektu a jeho postoja. Majme recenziu:

I bought an iPhone a few days ago. The touch screen was really cool!

V tejto recenzii *iPhone* je entita a *touch screen* je aspekt. *Really cool* je postoj, ktorý cieli na spomenuté entity a aspekty[16]. Konečný postoj je ľahko určiteľný:

```
touch screen - positive
```

Naopak v recenzii:

I like spoiler design at my BMW but interior is really ugly.

V danej recenzii musíme určiť dva aspekty a to *spoiler* a *interior*. Výsledný postoj teda bude:

```
spoiler - positive
interior - negative
```

Prvé pokusy o detekciu aspektov boli postavené na klasickej extrakcií informácií, ktorá využívala štatistiky výskytov podstatných mien. Takéto prístupy fungujú pomerne úspešne

s prvým príkladom recenzie, v ktorej je spomenuté len jedno rozhodujúce podstatné meno. V prípade, že aspekty zahŕňajú málo vyskytujúce sa termíny (napr. v oblasti gastronómie – veľa druhov jedál) je úspešnosť horšia. Spoločné riešenia tohoto problému zahŕňajú zhlukovanie entít s pomocou ručne zhotovených pravidiel, metódy strojového učenia alebo spojenie oboch metód.

Stretávame sa aj s problémom, pri ktorom môžu aspekty ovplyvniť polaritu postoja v rámci jednej domény. Napríklad, v gastronomickej oblasti, *"cheap"* zvyčajne znamená pozitívny postoj, avšak negatívny ak sa jedná o opis predmetu. Ďalej veľmi veľa neutrálnych termínov (napr. *warm, heavy, soft*) ovplyňuje polaritu postoja na základe kontextu spojeného s konkrétnym aspektom[4].

2.2.2 Existujúce metódy

Analýze postojov je venovaná značná časť výskumu. Existujúce metódy analýzy postojov môžeme rozdeliť na dva hlavné typy:

- využitie lexikóna slov
- metódy strojového učenia

V prvom type sa využíva list slov, z ktorého každé slovo je ohodnotené špecifickou hodnotou polarity. Z toho môžeme určiť priemer polarity v texte a vypočítať výsledný postoj. Lexikálne slovníky sa líšia podľa kontextu a môžu byť vytvorené pre rôzne účely. To znamená, že lexikón, ktorý je použitý na analýzu postojov v oblasti športu sa líši od lexikónu využitého na analýzu postojov v politike[2]. Alebo je možné použiť všeobecný slovník, ktorý obsahuje široké spektrum ohodnotených slov.

Výhodou je, že slovník môže byť ľahko rozšírený inými synonymami. Na druhú stranu rozšírenie lexikónu môže spôsobiť zhoršenie presnosti, pokiaľ sa synonymá opakujú. Presnosť analýzy je možné zlepšiť odstránením neutrálnych slov, ktoré nie sú označené ani ako pozitívne ani ako negatívne.

Metódy strojového učenia sa často spoliehajú na riadenú klasifikáciu prístupov, kde je analýza postoju koncipovaná binárne. Tento prístup vyžaduje ohodnotené dáta k natrénovaniu klasifikátora. Jedna z výhod je schopnosť prispôsobiť a vytvoriť klasifikátory na špecifické účely. Nevýhodou je nedostupnosť veľkého množstva ohodnotených dát. Ohodnotené dáta treba vytvoriť a tu vzniká problém, že toto ohodnocovanie môže byť pomerne výpočtovo zložité dokonca aj nemožné pre niektoré úlohy.

Jedným z najznámejších i najjednoduchších klasifikátorov pracujúcich na báze strojového učenia je *Naivný Bayesov* (NB) klasifikátor. Ako už je v názve uvedené, tak pracuje na báze Beyesovho teorému[18] [8]. Pre dokument d a triedu c sa teorém rovná

$$P\left(c|d\right) = \frac{P\left(d|c\right)P\left(c\right)}{P\left(d\right)} \tag{2.1}$$

NB klasifikátor sa potom rovná

$$c^* = arg\,max c\, P\left(c|d\right) \tag{2.2}$$

NB klasifikátor rozhoduje na základe tzv. *"posterior rule"*. Inými slovami, vyberie tú triedu postoja, ktorá je najpravdepodobnejšia[6].

Využíva sa v rôznych aplikáciach na detekciu nevyžiadanej pošty, pri triedení dokumentov alebo na rozpoznávanie dokumentov. NB klasifikátor je veľmi efektívny, čo sa týka využitia procesoru a pamäte a vystačí si aj s menšou sadou ohodnotených dát.

Ďalším možným klasifikátorom je *Support Vector Machine* (SVM). Jeho podstatou je vyhľadanie tzv. nadroviny, ktorá v priestore príznakov optimálne rozdeľuje trénovacie data. Typicky ide o vyhľadanie maximálnej vzdialenosti rozdeľujúcej nadroviny klasifikátora k bodom z trénovacej množiny.

Klasifikátor *Maximálnej Entropie* (MaxEnt) je alternatívny algoritmus, ktorý je účinný v rôznych aplikáciach, ktoré sa zaoberajú automatickým jazykovým spracovaním. Metóda sa tiež používa ku strojovým prekladom alebo k značkovaniu textu[15]. Odhad pravdepodobnosti $P(c|d)$ má exponenciálne rozdelenie, ktoré ukazuje nasledujúci vzorec:

$$P_{ME}\,(c|d) := \frac{1}{Z(d)} exp \sum_{i} \lambda_i f_i(d,c) \tag{2.3}$$

Kde $Z(d)$ je normalizačná funkcia, ktorá sa využíva na správne zistenie pravdepodobnosti. $f_i(d,c)$ je príznakový model a λ_i je odhadovaný parameter.

Všetky prezentované metódy sa využívajú taktiež pri analýze postojov orientovanej na aspekty. V praxi sa väčšinou kombinujú a dosahujú veľmi dobré výsledky[7][11].

2.3 Systémy na získanie dát

2.3.1 Twitter

Twitter sa stal veľmi populárnou sociálnou sieťou. Ale podľa odborníkov popularita Twitteru dnes pomaly upadá, avšak stále existujú milióny ľudí ochotných prispievať svojimi názormi a úvahami. Tieto názory sú považované za vhodné online zdroje k využitiu v analýze dát. Výsledkom je, že analýza postojov na Twitteri je rýchly a účinný spôsob merania verejnej mienky ohľadom marketingu alebo spoločenských vied. Vyhodnotením ľudských postojov a názorov môže firma získať skorú spätnú väzbu nového produktu na trhu[10].

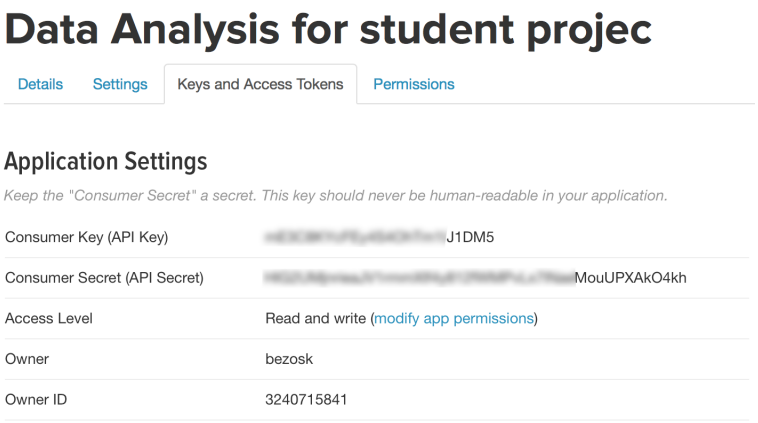
Príspevky na Twitteri sú obmedzené na 140 znakov, neformálne a je v nich použitých mnoho internetových slangov a emotikonov. Vďaka tomuto obmedzeniu na dĺžku sú správy na Twitteri ľahšie na analýzu, pretože autori zvyčajne píšu rovno k veci. Využitie špecifického slangu je hlavnou nevýhodou týchto správ.

Ďalším rysom príspevkov na Twitteri sú tzv. *hashtagy*. Hashtag je kľúčové slovo, ktoré začína znakom #. Na základe tohto slova je možné vyhľadať určité akcie, udalosti alebo akékoľvek iné témy.

Programové rozhranie

Twitter pre vývojárov pripravil skvelo zdokumentované *Twitter REST API*. Poskytuje im prístup k čítaniu a zapisovaniu dát. Aplikácie a užívatelia sa identifikujú podľa tzv. *OAuth* protokolu. Odpovede sú vo formáte *JSON*, ktorý je pomerne ľahko čitateľný. Za jednoznačnú výhodu môžeme považovať, že obsah, ktorý je možný vidieť cez webovú stránku je totožný s odpoveďou z API.

Pred tým, aby bolo možné komunikovať s API, sa musí vytvoriť Twitter účet a zaregistrovať používanie novej aplikácie. Po tomto kroku je už len potreba si vygenerovať tzv. `consumer` a `access` kľúče.



Obr. 2.2: Príklad vygenerovaných kľúčov pre Twitter API.

2.3.2 Facebook

Facebook je spolu s 1,86 miliardou aktívnych užívateľov jednou z najpopulárnejších a najznámejších sociálnych sietí[1]. Umožňuje im online komunikáciu s priateľmi a rodinou, zdieľanie fotografií, videí alebo príspevkov. Facebook hrá taktiež veľkú rolu v oblasti marketingu. Svetovo najznámejšie firmy doslova súperia o užívateľov a o počty palcov hore (tzv. páči sa mi to).

Z pohľadu množstva celkových dát je Facebook ideálnym zdrojom na analýzu postojov. Jeho užívatelia napíšu denne milóny príspevkov. Pokiaľ sa jedná o jeho bezplatné využitie tak hlavnou nevýhodou je jeho striktná ochrana súkromia užívateľa[9]. Na rozdiel od spomínaného Twitteru, kde všetky užívateľove informácie z profilu alebo jeho príspevky sú verejne dostupné, u Facebooku tomu tak nie je. Užívateľ Facebooku musí ručne odsúhlasiť a povoliť prístup aplikáciam tretích strán k jeho profilu. V praxi to znamená, že každý analyzovaný príspevok by musel byť odsúhlasený jeho autorom a to je z pohľadu veľkosti potrebných dát nemožné. I skrz toto drastické obmedzenie súkromia sa stále na Facebooku nachádza objemná časť verejne dostupných informácií vhodných na analýzu.

Programové rozhranie

Najznámejším programovým rozhraním Facebooku je *Graph API*. Pomáha vývojárom pri čítaní a zapisovaní dát. Je to dotazovacie rozhranie pracujúce s protokolom *Hyper Transfer Protocol (HTTP)*. Facebook *Graph API* poskytuje jednoduché a konzistentné zobrazenie sociálneho grafu Facebooku, jednotné reprezentácie objektov v grafe (napr. ľudí, fotiek, udalostí a stránok) a väzieb medzi nimi (napr. spoločných priateľov, zdieľaného obsahu, označených fotiek). Výhodou je, že v súčasnosti sa dá s *Graph API* pracovať vo viacerých programovacích jazykoch.

2.3.3 Web blogy

Webovými blogmi je možné nazvať užívateľmi zverejnené články na internete. Jednotlivé články pokrývajú širokú škálu tém od osobných záznamov po profesionálne žurnalistické správy o aktuálnych udalostiach. Podľa štúdií sú niektoré blogy považované za zdroje presnejších informácií na rozdiel od tých, ktoré poskytujú známe spravodajské servery. Blogy

sú zvyčajne verejne dostupné a ponúkajú čitateľovi možnosť komentovania ku každému príspevku samostatne[3].

Získavanie dát z blogov môže byť vykonané tzv. *crawling-om*, *scraping-om* alebo prístupom priamo k repozitárom, kde sú informácie uložené. Web crawler je automatický program, alebo skript, ktorý metodicky prehľadáva webové stránky a indexuje ich[13]. Stačí mu nastaviť počiatočnú adresu, podmienky (napr. o koľko úrovní sa maximálne zanoriť) a výsledkom je indexácia všetkých informácií spojených s počiatočnou adresou. Crawling využíva väčšina vyhľadávačov ako napr. Google, Yahoo alebo Bing. Scraping webových stránok čiastočne patrí pod crawling. Web scraper priamo využíva *HTTP* protokol a sťahuje obsah webovej stránky na dátové uložisko.

Kapitola 3

Návrh riešenia

Návrh a implementácia vychádza prevažne z teoretických poznatkov, ktoré sú popísané v kapitole č. 2. Pred samotným návrhom celého systému bol vytvorený online dotazník (príloha C.1), ktorého cieľom bolo spraviť menší prieskum a zistiť všeobecné požiadavky potencionálnych užívateľov na systém.

3.1 Analýza požiadaviek

Hlavnou úlohou práce je vytvoriť všeobecný systém na analýzu postojov. To znamená, že výsledné jadro implementácie systému by nemalo byť použiteľné len v oblasti automobilového priemyslu, ale i na iné odvetvia vhodné na analýzu dát. Prevažná väčšina dát bude pôvodom z Twitteru, ktorý je, ako bolo spomenuté v 2.3.1, najvhodnejším zdrojom dát a čím väčšie množstvo dát, tým lepšie. Webový portál Parkers[1] bude reprezentovať príklad sťahovania užívatelsky generovaného obsahu s úzkym doménovým zameraním. Z vyššie zadefinovaných podmienok vyplýva, že na analýzu postojov bude najefektívnejšie využiť všeobecný lexikón anglických slov.

Ďalšou úlohou je vhodným spôsobom vizualizovať výsledky užívateľovi. Z pohľadu užívateľa je webový portál najjednoduchšia voľba. Kedže sa jedná o informačný webový portál, tak by nemal byť veľmi zložito spracovaný. Jednoduchosť je prioritou pri návrhu vzhľadu webového portálu.

Webový portál bude dostupný online, z toho dôvodu je nutné premyslieť úložisko dát, ktoré sa použijú na vizualizáciu.

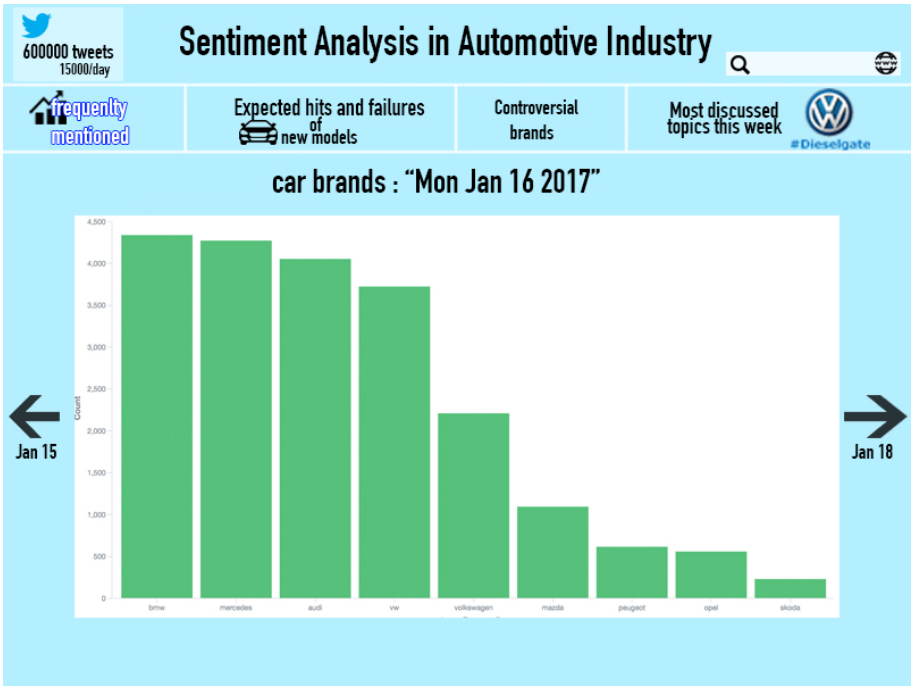
3.2 Použité technológie

V tejto podkapitole si popíšeme použité technológie jednotlivých komponentov systému. Od použitých významných knižníc pri analýze a spracovaní dát po vizualizáciu webovým portálom.

Analýza a získanie dát

V súčasnej dobe existuje veľké množstvo nástrojov vhodných na analýzu dát. Jedná sa o vývojové prostredia, ktoré využívajú grafické rozhranie. Tým pádom je ich používanie

[1] http://www.parkers.co.uk



Obr. 3.1: Prvotný mock-up webového portálu.

omnoho jednoduchšie, avšak prevažná väčšina je spoplatnených. Jednými najznámejšími nástrojmi využívajúce grafické rozhranie sú napr. Rapid Miner[2] alebo GATE[3].

Implementácia vlastného analyzátora, ktorý využíva funkcie z knižníc, je v tomto prípade omnoho efektívnejšia a prináša lepšie možnosti. Na internete je možné nájsť mnoho knižníc alebo frameworkov pre rôzne programovacie jazyky poskytujúcich nástroje na vytrénovanie vlastného analyzátora alebo využitie lexikónu slov. Časti systému, ktoré budú sťahovať, analyzovať a indexovať dáta budú implementované v jazyku Python. Ten ponúka množstvo skvelo zdokumentovaných knižníc, či už na analýzu textu (napr. Natural Language ToolKit[4]), alebo prístupu k využívaným API. Práve pre komunikáciu s Twitterom bola zvolená knižnica Tweepy[5]. Ta nám poskytuje prístup priamo k Twitter REST API, ktoré bolo spomenuté v 2.3.1. Pre sťahovanie príspevkov z Facebooku je vhodná knižnica Facebook Python SDK[6], ktorá je taktiež vyvinutá na komunikáciu s Facebook Graph API. Na extrakciu dát z web blogov je možné použiť buď scraper alebo crawler. Výsledný systém stiahne dáta len z jedného webového portálu, tým pádom je lepšie použiť scraper. Na túto úlohu budú použité dve knižnice a to Beautiful Soup[7] a Requests[8].

Indexácia dát

Najprv je potrebné si ujasniť celkový objem dát. Kedže sa dáta budú pravidelne aktualizovať a ich objem bude narastať, tak ukládanie dát lokálne priamo do jednej zložky

[2] https://rapidminer.com
[3] https://gate.ac.uk
[4] http://www.nltk.org
[5] http://www.tweepy.org
[6] https://facebook-sdk.readthedocs.io/en/latest/
[7] https://www.crummy.com/software/BeautifulSoup/
[8] http://docs.python-requests.org/en/master/

neprichádza do úvahy. Taktiež je potrebné dbať na rýchlosť vyhľadávania v zaindexovaných dátach. Najlepším riešením bude použiť nástroj ElasticSearch[9]. Je to serverový vyhľadávač vychádzajúci z Apache Lucene. Poskytuje distribuované multiužívateľské fulltextové vyhľadávanie. ElasticSearch je vyvíjaný v Jave a poskytuje podporu pre veľa programovacích jazykov. Na implementáciu celého systému budeme potrebovať klienta pre Python a JavaScript. I keď prioritne nám ElasticSearch poskytuje efektívne vyhľadávanie, dá sa použiť aj ako databáza na ukladanie analyzovaných dát. Spolu s nástrojom ElasticSearch bude vhodné použiť aj vizualizačný nástroj Kibana, ktorý nám zobrazí všetky uložené dáta a poskytuje nám jednoduché rozhranie pre vytvorenie rôznych grafov alebo histogramov.

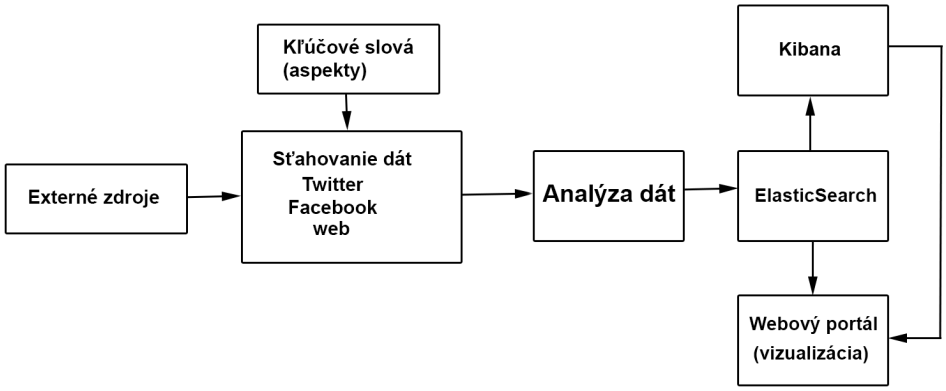
Tvorba webového portálu

Na implementáciu webového portálu bude použitý framework Bootstrap[10]. Je to jeden z najpopulárnejších HTML,CSS a JavaScriptových frameworkov. Zjednodušuje prácu pri tvorbe webových aplikácií a dodáva jej jednotný vzhľad. Na internete je možné nájsť mnoho šablón už priamo hotových webových aplikácií. Rozhodol som sa využiť HTML5 šablónu zo stránky https://bootstrapmade.com/ s názvom MyBiz[11].

Hlavným rysom webového portálu budú grafy. JavaScript nám poskytuje mnoho vizualizačných knižníc, z čoho najznámejšia a najpoužívanejšia je D3.js[12]. Výhodou je jej rýchlosť aj napriek použitiu s veľkým objemom dát.

3.3 Schéma systému

Na obrázku č. 3.2 je možné vidieť výslednú schému navrhovaného systému.



Obr. 3.2: Schéma navrhovaného systému.

Schéma systému bola navrhnutá ešte pred samotnou implementáciou. Je možné ju rozdeliť na samostatné moduly, ktoré na seba nadväzujú. Jednotlivé moduly si popíšeme v nasledujúcej kapitole.

[9] https://www.elastic.co
[10] http://getbootstrap.com
[11] https://bootstrapmade.com/mybiz-free-business-bootstrap-theme/
[12] https://d3js.org

Kapitola 4

Implementácia

Táto kapitola slúži na popis vnútornej štrukúry systému, teda základné informácie o zdrojových súboroch, použitých metód a tried alebo zaujímavých implementačných detailoch. Finálna podoba webového portálu je dostupná na školskom servery athena1[1].

4.1 Implementácia sťahovania dát

Nasledujúca podkapitola opisuje základnú podstatu implementačných detailov modulov slúžiacich na sťahovanie nových dát.

4.1.1 Twitter

Sťahovanie dát zabezpečuje súbor *twitter_down.py*. Po zadefinovaní potrebných kľúčov (2.3.1) je potreba inicializovať Tweepy objekt, prostredníctvom ktorého sa komunikuje s Twitter API. Tweety sa sťahujú na základe listu, ktorý obsahuje kľúčové slová, ktoré sa vyskytujú v sťahovaných tweetoch. Pri presiahnutí počtu povolených dotazov na API sa musí sťahovanie na 15 minút pozastaviť. Pri analýze sa využívajú len priamo tweety od užívateľov. Z toho vyplýva, že sa pridala podmienka na preskočenie tzv. retweetu. Ďalej sa využijú moduly generovania aspektov a zistenia postoja z textu tweetu popísaných v podkapitole 4.2. Najpodstatnejšou metódou je funkcia `make_content`, ktorá nám z celého JSON výstupu z knižnice Tweepy zhotoví prehľadný JSON. Tento JSON obsahuje len podstatné hodnoty, ktoré budeme potrebovať na analýzu dát. Následne sa zaindexuje do databázy. V tabuľke 4.1 je vidieť jeho obsah i s príkladom príspevku.

[1] http://athena1.fit.vutbr.cz/xbezak01/WWW/

Tabuľka 4.1: Príklad indexovaných JSON dát z Twitteru znázornených v tabuľke.

názov	**popis**	**príklad**
_type	značka auta	honda
aspects	spomenuté aspekty	price
aspects_mentioned	spomenuté typy aspektov	value
car_type	spomenuté typy áut	hatchback
car_type_mentioned	spomenuté typy áut menovite	jazz
compound	výsledok polarity z analyzátora	0.758
created_at	dátum vytvorenia	2017-04-08
entities_hashtag.text	spomenuté hashtagy	[honda,JAZZ]
entities_user_mentions	spomenutí užívatelia	
id_str	identifikátor príspevku	850563135509594112
lang	jazyk	en
place_country	skratka krajiny	*unknown*
place_name	krajina	*unknown*
sentiment	výsledný postoj	positive
text	text príspevku	HONDA JAZZ 1.4i-DSi CVT-7 SE AIR CON 12MTH MOT GREAT VALUE #honda#JAZZ
user_created_at	dátum vytvorenia účtu užívateľa	2013-03-01
user_description	popis profilu	The first and most important thing in life - is to try to control myself.
user_followers_count	počet odoberateľov	392
user_friends_count	počet priateľov	135
user_id_str	identifikátor užívateľa	1229336402
user_lang	jazyk užívateľa	RU
user_location	bydlisko užívateľa	Illinois, USA
user_screen_name	meno užívateľa	x3Amyy
user_statuses_count	počet príspevkov	94427
user_verified	overený užívateľ	false

4.1.2 Facebook

Sťahovanie dát na Facebooku prebieha podobne ako na Twitteri. Využíva sa súbor *facebook_down.py*. Rozdielom je, že sa prehľadávajú Facebookove stránky obsahujúce jednotlivé kľúčové slová a sťahujú sa ich všetky príspevky a príspevky užívateľov, v ktorých boli tieto stránky označené. Facebookových príspevkov je podstatne menej, takže stačí stiahnuť vždy posledných 20 príspevkov a indexovať len tie, ktoré sa ešte nenachádzajú v databáze.

Tabuľka 4.2: Príklad indexovaných JSON dát z Facebooku znázornených v tabuľke.

názov	**popis**	**príklad**
_type	značka auta	bmw
about	popis stránky	Welcome to the official BMW Facebook page! Get all the latest information about Sheer Driving Pleasure!
aspects	spomenuté aspekty	engine
aspects_mentioned	spomenuté typy aspektov	engine, motor
car_type	spomenuté typy áut	sport
car_type_mentioned	spomenuté typy áut menovite	i8
category	kategória stránky	Cars
created_time	dátum vytvorenia príspevku	2017-02-14
fan_count	počet fanúšikov	19865520
id	identifikácia príspevku	14101740225_10158368161985226
message	text príspevku	The BMW i8 is no ordinary sports car. A TwinPower Turbo engine and all-electric motor make for an exhilarating driving experience.
name	názov stránky	BMW
sentiment	výsledný postoj	positive
website	web stránka	http://www.bmw.com

4.1.3 Parkers.co.uk

Tento portál neponúka žiadne API. Skript s názvom *parkers_down.py* zabezpečí stiahnutie všetkych recenzií z tohoto portálu. Využíva funkciu `download` z modulu *web_down.py*. V tejto funkcii sa pracuje s knižnicami spomenutými v 3.2. Podstatou metódy je nájsť všetky recenzie o danom modely a postupne ich prechádzať a sťahovať ich obsah. Výsledný JSON, ktorý sa indexuje, závisí od toho ako daný užívateľ vyplnil formulár pri pridávaní recenzie.

4.2 Indexácia a analýza dát

Pred prvým použitím databázy ElasticSearch je potreba vhodne namapovať nový index, do ktorého sa budú ukladať dáta. Je nutné rozlíšiť, ktoré hodnoty môžu byť analyzované ako celok, a ktoré bude lepšie analyzovať po každom slove. Každý text, z ktorého budeme chcieť zistiť jeho postoj, sa namapuje ako `analyzed`. To nám uľahčí prácu pri tvorbe webového portálu, kedy budeme môcť vytvoriť rôzne grafy najviac používaných slov v textoch atď.

Konkrétny príklad namapovania indexu je možné vidieť na obrázku 4.1. Jednotlivé atribúty sú popísané v tabuľke 4.1.

```
'bmw': {
  'properties': {
      'aspects': {'index': 'not_analyzed', 'type': 'string'},
      'aspects_mentioned': {'index': 'not_analyzed', 'type': 'string'},
      'car_type': {'index': 'not_analyzed', 'type': 'string'},
      'car_type_mentioned': {'index': 'not_analyzed', 'type': 'string'},
      'created_at': {'type': 'date'},
      'text': {'index': 'analyzed', 'type': 'string'},
      'place_country': {'index': 'not_analyzed', 'type': 'string'},
      'place_full_name': {'index': 'not_analyzed', 'type': 'string'},
      'place_name': {'index': 'not_analyzed', 'type': 'string'},
      'user_created_at': {'type': 'date'},
      'user_description': {'index': 'analyzed', 'type': 'string'},
      'user_name': {'index': 'not_analyzed', 'type': 'string'}
  }}
```

Obr. 4.1: Príklad namapovania dokumentu pre značku BMW.

Programovo indexácia a analýza dát prebieha v každom predchádzajúcom module takmer identicky. Na zistenie postoja textu sa využíva modul s názom *sentiment_analysis.py*. Z tohto modulu sa zavolá metóda `sentiment`. Jej parametrom je text, ktorý chceme analyzovať. Avšak predtým sa musí vytvoriť objekt triedy `SentimentIntensityAnalyzer` Vader analyzátora z knižnice NLTK. Výsledok polarity analyzátora vráti metóda `polarity_scores`. Jedná sa o číslo na stupnici od -1 po 1. Na základe tohto čísla sa dajú určiť tri intervaly:

- $< -1, 0)$ - negatívny
- $(0, 1 >$ - pozitívny
- 0 - neutrálny

Ak je výsledok z Vader analyzátora neutrálny, tak sa využije modul `sentiment_emoji.py`, z ktorého sa použije najpodstatnejšia funkcia `contains_emoji`. V tele tejto funcie sa načíta JSON súbor `new_emoji_sentiment.json`. Tento súbor obsahuje, podobne ako u Vader sentimentu, ohodnotené emoji smajlíky na základe ich postoja. Výsledný postoj sa určí z priemeru týchto smajlíkov taktiež podľa troch intervalov.

Modul `aspect_gen.py` zabezpečuje generovanie aspektov zo zadaného textu. Pracuje s dvomi JSON súbormi. Prvý JSON s názvom `car_types.json` obsahuje všetky modely áut konkrétnych automobilových značiek rozdelených do kategórií.

```
{
    "bugatti": {
        "luxury": [
            "veyron",
            "chiron"
        ]
    }
}
```

Obr. 4.2: Príklad obsahu súboru `car_types.json`.

Druhý JSON súbor s názvom `aspects.json` tvoria aspekty.

```
{
    "aspects" : {
        "wheels": [
            "wheel",
            "rim",
            "disk",
            "roller"
        ]
    ...
```

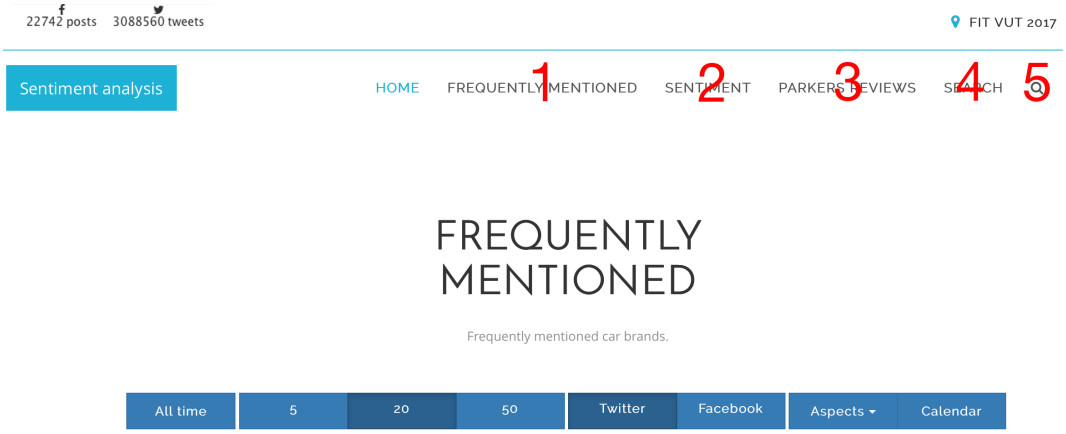
Obr. 4.3: Príklad obsahu súboru `aspects.json`.

Z daného modulu je najvýznamnejšia funkcia `aspect_generator`, ktorého parametrami sú daný text a kľúč, podľa ktorého sa vyhľadáva v prvom JSON súbore. Pre vyhľadanie aspektov a typov spomenutých jednotlivých modelov áut sa využívajú rôzne konštrukcie regulárnych výrazov.

Ako úložisko dát sa používa školský ElasticSearch dostupný na servery `athena1`. Aby bolo možné s ním komunikovať, tak sa musí inicializovať ElasticSearch objekt a pripojiť sa na port 9200. Ďalej sa využije triedna metóda `index`, ktorej argumentami sú automobilová značka, ktorá udáva typ dokumentu a telo výsledného slovníka, ktorý reprezentuje formát JSON.

4.3 Implementácia webového portálu a vizualizácia

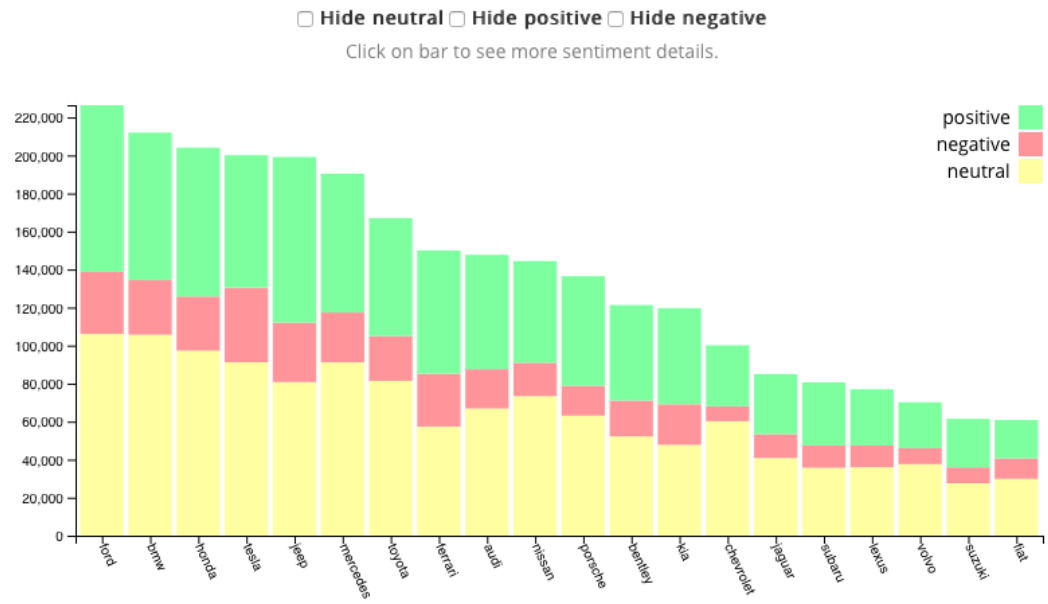
V nasledujúcej podkapitole sa stručne priblíži webový portál z pohľadu užívateľa, teda jeho užívateľského rozhrania. Taktiež podkapitola môže slúžiť ako príručka k používaniu. Implementácia webového portálu prebiehala ako posledná. Podstatou celej implementácie je odoslať dotaz na ElasticSearch, spracovať prijaté dáta a tie vizualizovať. Celý obsah sa nachádza na jednej stránke a je rozdelený do 4 častí.



Obr. 4.4: Záhlavie stránky, ktoré obsahuje menu.

1 - Frequently mentioned

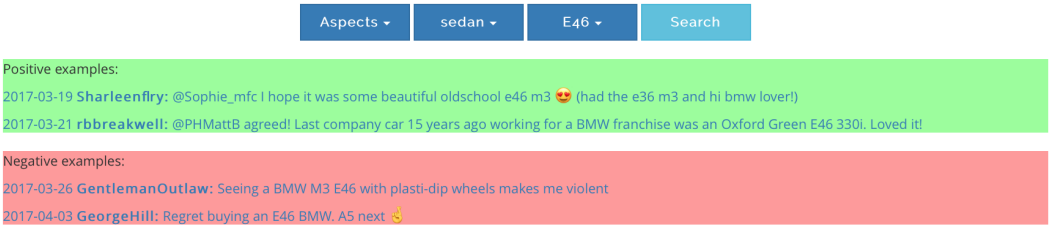
V tejto časti stránky sa nachádzajú štyri druhy grafov. V prvom stĺpcovom grafe je možné vidieť všetky príspevky rozdelené podľa názvu automobilových značiek a postoja. Ďalej sa tu nachádza koláčový graf, ktorý zobrazuje Top 10 prispievateľov na Twitteri. V tomto grafe je potrebné odfiltrovať všetky účty, ktoré sa využívajú len na zdieľanie automobilov na predaj. Teda účtov, ktoré obsahujú v názve kľúčové slovo `sale` alebo príspevkov, ktoré su označené aspektom `sale`. Ďalej si užívateľ môže pozrieť mapu sveta/USA, zobrazujúcu počet príspevkov z daných regiónov, alebo si môže zobraziť najviac vyskytované hashtagy za rôzne časové obdobia. Po kliknutí na automobilovú značku z prvého grafu sa animáciou dostaneme na dalšiu časť webového portálu.



Obr. 4.5: Príklad ukážky stĺpcového grafu.

2 - Sentiment

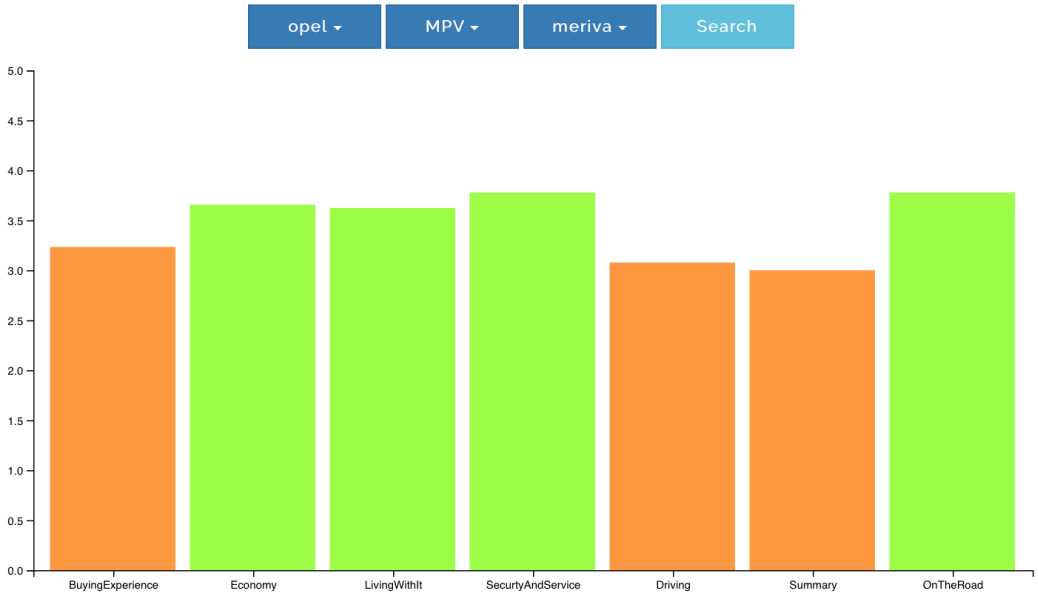
V tejto časti si užívateľ môže pozrieť štatistiky konkrétnej automobilovej značky. Napríklad celkový počet príspevkov za posledný týždeň, najviac vyskytujúce sa hashtagy alebo zobraziť na základe jednotlivých aspektov konkrétne príspevky. Obsah tlačidiel je vytvorený na základe dvoch JSON súborov spomenutých v podkapitole 4.2.



Obr. 4.6: Príklad ukážky vyhľadania postoja konkrétneho modelu auta.

3 - Parkers reviews

V tejto sekcii si užívateľ môže zobraziť stĺpcový graf na základe hodnotení na recenznom portále Parkers. Stupnica hodnotenia je v intervale od <0,5> a vypočíta sa ako priemer všetkých ohodnotených recenzií.



Obr. 4.7: Príklad ukážky grafu užívateľských hodnotení na recenznom portále Parkers.

4,5 - Search

Do tejto sekcie sa dá dostať viacerými spôsobmi:

- použitím vyhľadávača v záhlaví stránky
- kliknutím na kľúčové slovo v grafe najviac použitých hashtagov
- kliknutím na meno používateľa

Ďalej sa vyhľadanie rozdeľuje na:

- vyhľadanie podľa mena
- vyhľadanie podľa hashtagu

Ak sa vyhľadá slovo, ktoré sa začína znakom #, tak sa použije na vyhľadanie hashtagu v príspevkoch na Twitteri. Ak sa slovo nezačína týmto znakom, berie sa ako meno užívateľa. Po vyhľadaní v databáze sa zobrazí stĺpcový graf podobný grafu v časti Frequently Mentioned (obr. 4.5), náhľad na najviac pozitívne a negatívne príspevky a časový graf počtu príspevkov.

Kapitola 5

Príklady použitia systému

Nasledujúca kapitola prezentuje prácu s výsledným webovým portálom. Príklady použitia sú rozdelené do dvoch skupín:

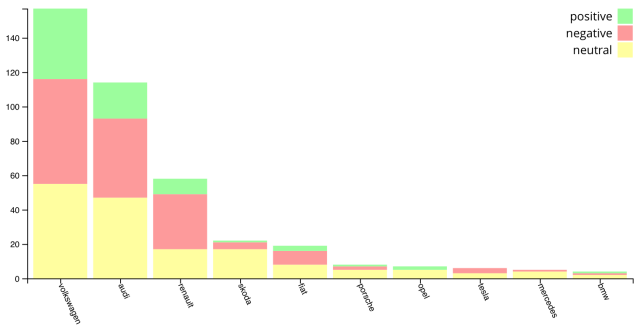
- Overenie funkčnosti na základe článkov a faktov
- Interakcia s užívateľmi

V prvých štyroch príkladoch sa overí predovšetkým reálna spojitosť medzi stiahnutými dátami a článkami/faktami z internetu, ktoré potvrdzujú zistené informácie. V ďalších príkladoch sa overí intuitívnosť používania užívateľského rozhrania. Zapojení boli užívatelia z cieľovej skupiny (automobiloví fanúšikovia), ktorí plnili rôzne úlohy s cieľom získať nové informácie. Po skončení práce bol týmto užívateľom poskytnutý dotazník, ktorého vyhodnotenie sa nachádza v podkapitole 6.3.

5.1 Prvý príklad

Prvý príklad sa venuje pokračovaniu kauzy s emisiami tzv. *dieselgate*. Tým pádom sa do položky vyhľadávača zadá kľúčové slovo `#dieselgate`. Vďaka rýchlosti nástroja ElasticSearch sa zobrazia grafy okamžite.

Vyhodnotenie



Obr. 5.1: Počet príspevkov jednotlivých automobilových značiek.

Z grafu na obrázku 5.1 je zrejmé, že najviac spomenutými značkami sú Volkswagen, Audi, Renault, Škoda a Fiat. Taktiež možno vidieť, že prevládajú príspevky s negatívnym postojom. Tieto výsledky nám potvrdzujú aj rôzne články na internete[1] [2].

Positive examples:

2017-04-04 **stevenbhoward:** Easy Come, Easy Go: Volkswagen Overtaken By Toyota, And Nissan/Renault https://t.co/FAhNZEKPNT #Dieselgate #VWscandal

2017-03-17 **pedroparralc:** It seems that Volkswagen and Audi also lied about Adblue clean Diesel solution #Dieselgate https://t.co/fV1N0ZwuM0 vía @focusonline

2017-03-28 **UrbanDwells:** A month in and my @subaru_usa #Outback is still the perfect replacement for my #vw golf 😍 #BuyBackMyTDI #dieselgate

Negative examples:

2017-03-16 **Karim_Khalaf:** But this whole #carbon thing is fake so why and how to blame cars manufacturers for cheating chips! #Dieselgate #Renault #Volkswagen #Audi

2017-03-22 **dpcarrington:** NOx emission tests by @WhichUK reveal Renault and Jeep are among worst offenders https://t.co/UEphuPjz8C #Dieselgate https://t.co/ujeD0KduyG

2017-03-22 **franzsch2:** More diesel fraud? German officials probing fraud charges vs. Mercedes' parent firm https://t.co/7nWlnAKdPn #dieselgate #fraud #cars

2017-03-28 **JemimaHoadley:** @CleanAirLondon @SadiqKhan @nicholascecil @SophiaSleigh am so angry we were sold diesel car by @Peugeot on #Dieselgate LIE

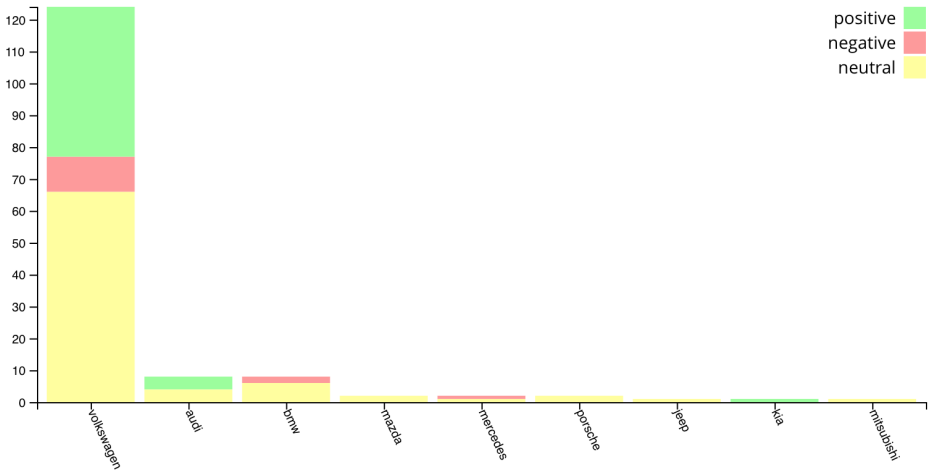
2017-03-23 **stephenweitzel:** #Volkswagen is a shameful company. Their #Dieselgate created many victims like me and they refuse to make it right. Time to boycott @VW

Obr. 5.2: Ukážka príkladov tweetov rozdelených podľa postoja.

5.1.1 Druhý príklad

Podstatou druhého príkladu je ukázať vplyv novovydaného modelu Volkswagen Arteon. Sleduje sa kľúčové slovo `#Arteon`.

Vyhodnotenie

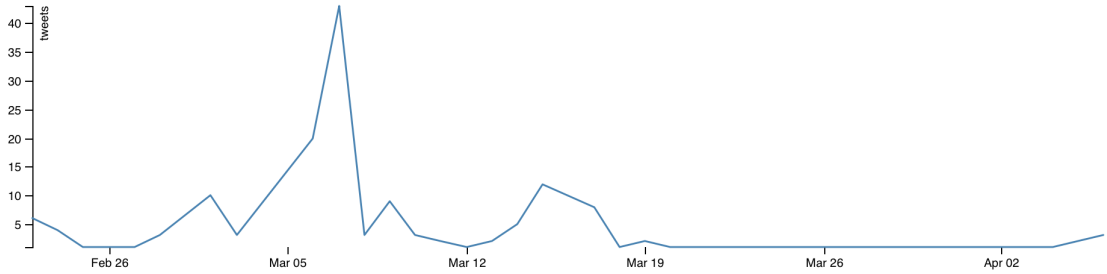


Obr. 5.3: Počet príspevkov jednotlivých automobilových značiek.

Ako je možné vidieť z grafu na obrázku 5.3, tak prevládajú pozitívne reakcie na nový model. Z časti je vidieť, že ďalšími spomenutými značkami s týmto modelom sú najviac konkurenčné Audi a BMW.

[1] http://www.rozhlas.cz/zpravy/svet/_zprava/pokracovani-dieselgate-francouzska-prokuratura-kvuli-emisim-vysetruje-fiat–1710180

[2] http://www.dw.com/en/audi-offices-raided-in-dieselgate-investigation/a-37941126



Obr. 5.4: Počet príspevkov za celkové obdobie.

Graf na obrázku 5.4 znázorňuje počet príspevkov za celkové obdobie chodu systému. Je zrejmé, že v okolí dňa 6.3.2017 bol najväčší výskyt príspevkov. Dôvodom bolo oficiálne predstavenie tohoto modelu na výstave v Geneve[3].

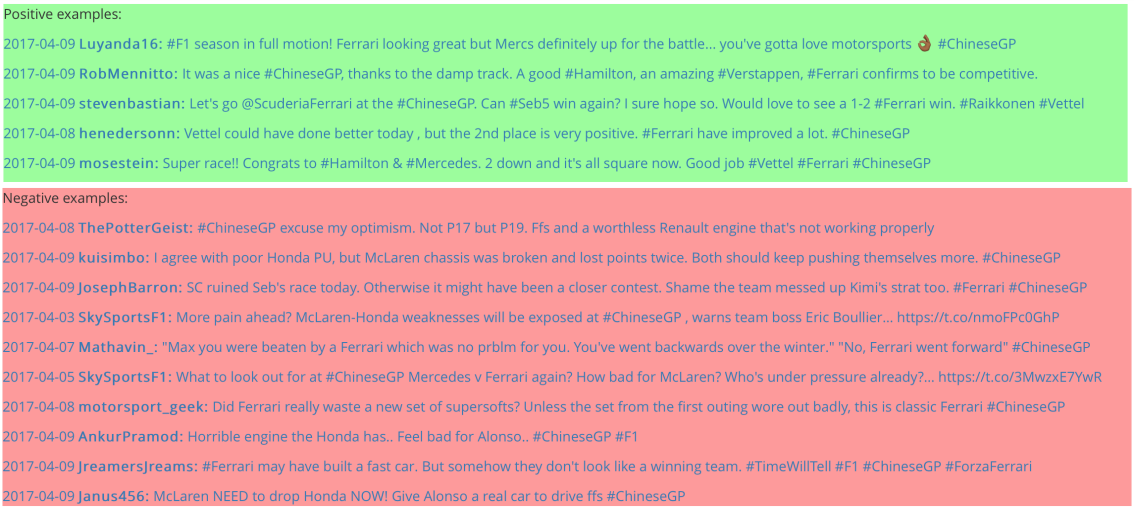
5.1.2 Tretí príklad

Tretí príklad sa prezentuje na základe grafu 5.5, ktorý vizualizuje najčastejšie používané hashtagy na Twitteri za určité časové obdobie. Ako je možné vidieť, medzi všeobecne používanými hashtagmi sa vyskytuje nové kľučové slovo a to *chinesegp*. Po kliknutí na toto slovo sa dostaneme do časti stránky, ktorá vizualizuje štatistiky pre jednotlivé hashtagy.

Obr. 5.5: Graf najčastejšie používaných hashtagov v dni 9.4.2017.

[3]http://www.carmagazine.co.uk/car-news/motor-shows-events/geneva/2017/vw-previews-new-2017-arteon-saloon/

Výsledkom je graf, ktorý poukazuje na najviac diskutované automobilové značky a to konkrétne Ferrari, Mercedes, Honda a Renault. Podstatnejšie informácie sú viditeľné na príkladoch príspevkoch s pozitívnym a negatívnym postojom 5.6.



Obr. 5.6: Ukážka príkladov tweetov rozdelených podľa postoja.

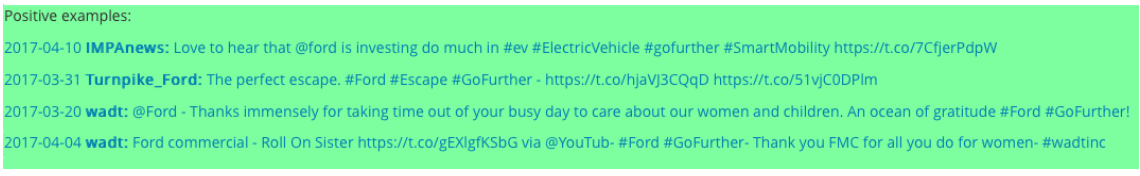
Z týchto príkladov je možné zistiť napr. výsledky pretekov a to konkrétne výhru Lawisa Hamiltona a neúspech Fernanda Alonsa a negatívny postoj k formuliam od Hondy.

5.1.3 Štvrtý príklad

V nasledujúcom príklade sa prezentuje vplyv reklamy danej automobilovej značky na sociálnej sieti Twitter. Prezentuje sa reklama od značky Ford s názvom Go Further. Tým pádom sa vyhľadá kľúčové slovo #GoFurther. V danej reklame Ford poukazuje na vývoj autopilotov, zdieľanie áut medzi ľuďmi a elektrické vozidlá.

Vyhodnotenie

Výsledkom je celkovo 51 príspevkov, v ktorých je spomenutý daný hashtag. Z toho 26 pozitívnych a 3 negatívne príspevky. Zaujímavejšie výsledky je možné vidieť v náhľade na niektoré pozitívne príspevky na obr. 5.7.



Obr. 5.7: Ukážka pozitívnych príspevkov.

Z príspevkov je možné vidieť, že reklamná kampaň zaujala predovšetkým v oblasti vývoja elektrických áut teda nasledujúca kampaň by sa mohla zamerať viac na túto oblasť a cieliť ju hlavne pre ženy.

5.1.4 Piaty príklad

Do tohto príkladu sa zapojili štyria potencionálni užívatelia webového portálu. Ich úlohou bolo pracovať s užívateľským rozhraním a plniť zadané úlohy. Následne im bol poskytnutý dotazník (C.2), prostredníctvom ktorého mohli zrecenzovať ako sa im pracovalo s webovým portálom, či získali nejaké nové informácie alebo aké funkcie by bolo potrebné zlepšiť, prípadne pridať. Nasledujúce úlohy boli zadané užívateľom a bolo pozorované ako a k akým odpovediam dospeli.

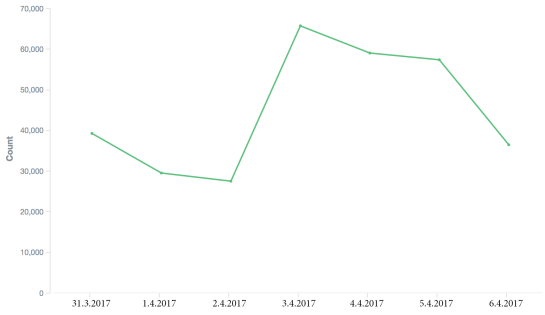
- Ktorá značka je celkovo najviac kritizovanou?
- V akom časovom období sa konala výstava Geneva motor show 2017?
- Vyhladanie recenzie auta Volkswagen Passat na recenznom portále Parkers a na Twitteri.
- Ktoré automobilové značky sú najčastejšie inzerované na Twitteri?
- Ktorá automobilová značka je najviac diskutovanou vzhľadom na jej cenu?

Kapitola 6

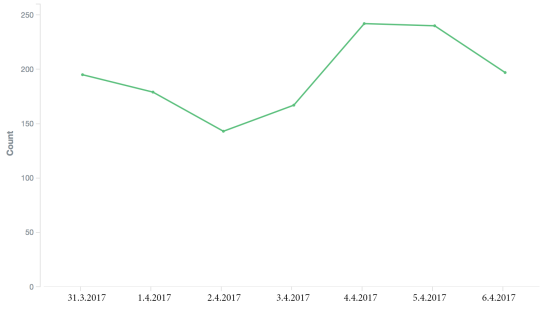
Vyhodnotenie systému

6.1 Množstvo dát

Prvým zámerom práce bolo indexovať dáta nepretržite, aj po odovzdaní práce. Z toho dôvodu boli v prvej finálnej verzii webového portálu aj grafy zobrazujúce štatistiky za posledný deň, týždeň atď. Ku koncu práce nastali problémy so školským ElasticSearch serverom, ktorý uchováva dáta. Tým pádom sa indexácia zastavila a finálny počet stiahnutých príspevkov je možné vidieť v tabuľke 6.1. Na pravidelnú aktualizáciu dátovej sady sa využíval nástroj `crontab`, ktorý pravidelne spúšťa skripty, ktoré zabezpečujú stiahnutie dát. Presne daná syntax je popísaná v prílohe B.1. Na zobrazenie rôznych grafov je možné využiť nástroj Kibana, pomocou ktorého môžeme ľahko vizualizovať ElasticSearch data. Grafy 6.1 a 6.2 znázorňujú množstvo príspevkov v denných intervaloch počas jedného týždňa na jednotlivých sociálnych sieťach.



Obr. 6.1: Množstvo dát na Twitteri za jeden týždeň (31.3 – 6.4).



Obr. 6.2: Množstvo dát na Facebooku za jeden týždeň (31.3 – 6.4).

Priemerne je denne na Twitteri pridaných okolo 40000 príspevkov o daných automobilových značkách. Na rozdiel od Facebooku, kde sa denne pridá niečo málo cez 180 príspevkov. Dáta sa začali indexovať začiatkom februára a momentálne sa v Twitter indexe nachádza viac než 3 milióny zanalyzovaných tweetov a vo Facebook indexe okolo 22000 dokumentov. To je dôvod, prečo väčšina vstupných dát je práve z Twitteru. Scraper, ktorý sťahoval všetky recenzie z recenzného portálu Parkers, stiahol a zaindexoval 22945 recenzií. V tabuľke 6.1 je možné vidieť objem zaindexovaných dát.

Tabuľka 6.1: Doposiaľ celkový objem dát.

	Veľkosť dát	Počet dokumentov
Twitter index	2.76GB	3088560
Facebook index	26.1MB	22742
Web Index	27.9MB	22945

6.2 Experimenty s klasifikátormi

Prvotná verzia analyzátora pracovala na princípe analýzy postoja natrénovaným klasifikátorom na báze Naivného Bayesovho algoritmu. Tento klasifikátor bol natrénovaný korpusom tweetov, ktorý obsahoval viac než 1000 ručne ohodnotených tweetov. V našom systéme je potrebná predovšetkým presnosť ako rýchlosť a to bol hlavný problém. V nasledujúcej tabuľke je možné vidieť porovnanie rýchlosti a presnosti jednotlivých analyzátorov na rôznych dátových sadách.

Tabuľka 6.2: Porovnanie presnosti VADER a NB analyzátora.

	VADER		NB	
Počet	Presnosť	Rýchlosť	Presnosť	Rýchlosť
10	60%	0.0198s	25%	2.1s
100	60%	0.317s	42.5%	2.22s
500	68%	1.495s	51.2%	2.7s
1000	68.8%	2.905s	59.4%	3.0s
5000	69.44%	13.25s	62.2%	7.32s
10000	69.61%	27.46s	61.03%	12.92
25000	70.37%	57.65s	60%	29.57s
50000	70.54%	82.05s	59.41%	54.49s

Ako možno vidieť, pri menších počtoch príspevkov u VADER analyzátora je presnosť konštantne okolo 68 až 70% a rýchlosť takmer identická. Z toho vyplýva, že použitie VADER analyzátora je vhodné a efektívnejšie.

6.3 Vyhodnotenie užívateľského rozhrania

Ktorá značka je celkovo najviac kritizovanou?

Na túto otázku vedeli celkom intuitívne odpovedať všetci užívatelia. Ich odpoveďou bola Tesla na Twitteri a Ford na Facebooku. K zisteniu výsledku využili graf v časti Frequently Mentioned. Z toho traja z nich použili prepínač, ktorý skryl počet všetkých pozitívnych a negatívnych odpovedí. Jeden z užívateľov ručne preklikal každý stĺpec a zistil najväčší počet negatívnych príspevkov.

V akom časovom období sa konala výstava Geneva motor show 2017?

K zisteniu odpovede na túto otázku bolo nutné u každého užívateľa, aby si najskôr prezreli celý portál a zistili celkovo aké možnosti ponúka. Každému užívateľovi trvalo 2 až 3 minúty, kým sa zorientoval a vedel sa pustiť do hľadania odpovede. Traja užívatelia použili vyhľadávač v záhlaví stránky. Z toho dvaja z nich zadali kľúčové slovo *#geneva* a na základe grafu,

ktorý zobrazuje počet príspevkov za jednotku času zistili, že najväčší počet bol v období od 5.3 do 9.3. Jeden užívateľ nesprávne zadal iba kľučové slovo bez hashtagu teda *geneva* a nezobrazili sa mu žiadne výsledky. Posledný užívateľ nepoužil vyhľadávač, ale intuitívne si zobrazil náhľad na sentiment značky Audi a na základe grafu počtu príspevkov za jednotku času a počtu najviac používaných hashtagov prišiel ku rovnakej odpovedi.

Vyhľadanie recenzie auta Volkswagen Passat na recenznom portále Parkers a na Twitteri.

Pri plnení tejto úlohy boli už užívatelia oboznámení s webovým portálom a už vedeli, akým spôsobom prísť k odpovedi. Všetci štyria správne použili sekciu Parkers reviews, v ktorej si vyhľadali štatistiku recenzií na daný model. Z tohto grafu všetci z nich zistili, že priemerné ohodnotenie tohto modelu je 4 z 5, teda prevládajú pozitívne postoje na daný model. Ďalšou pod-úlohou bolo vyhľadať daný model na Twitteri. Dvaja z nich použili sekciu Frequently mentioned, zobrazili si štatistiky na značku Volkswagen, potom štatistiky ku konkrétnemu modelu a potvrdili domnienku o prevládajúcich pozitívnych príspevkoch. Zvyšní dvaja užívatelia využili vyhľadávač a konkrétne hashtag *#passat*. Síce bolo vyhľadaných príspevkov takmer o polovicu menej, ale aj tak potvrdili svoje domnienky z prvej pod-úlohy.

Ktoré automobilové značky sa najviac predávajú prostredníctvom Twitteru?

Riešenie tejto úlohy bolo jednoduché. Ale aj napriek tomu dvaja užívatelia z počiatku nevedeli, ako sa dopracovať ku správnej odpovedi. Až po pár minútach si všimli menu `Aspects` a už boli na správnej ceste k riešeniu. Teda užívatelia využili sekciu Frequently mentioned, ktorá zobrazuje celkové štatistiky o všetkých značkách. Odfiltrovali príspevky v grafe len tie, ktoré majú spomenutý aspekt *sale*. Ich odpoveďou boli značky ako *BMW, Nissan, Honda, Ford, Toyota, Chevrolet a Jeep*. Tieto značky mali jednoznačnú prevahu v grafe.

Ktorá automobilová značka je najviac diskutovanou vzhľadom na jej cenu?

Táto úloha bola veľmi podobná predchádzajúcej, ale bola podaná až po vyplnení dotazníku a po konzultácií. V podstate išlo o test, či je webový portál ľahko zapamätateľný. Všetci užívatelia už vedeli že majú použiť menu `Aspects` v sekcii Frequently mentioned a dopracovali sa ku správnej odpovedi. Tou bola Tesla na Twitteri a Ford na Facebooku.

Vyhodnotenie a dotazník

Motiváciou testovania užívateľského rozhrania bolo predovšetkým zistiť názor potencionálnych užívateľov na vzhľad a funkčnosť celkového systému. Okrem toho sa vďaka tomuto testovaniu prišlo na pár podstatných programátorských chýb, ktoré boli následne odstránené. Napríklad dvojité zobrazovanie pozitívnych a negatívnych príspevkov. Taktiež sa ku grafom pridalo viacej vysvetliek, aby bolo jasnejšie, čo daný graf zobrazuje.

Dotazník B v prílohe C.2 je tvorený otázkami k celkovému zhodnoteniu práce. Nasleduje ukážka odpovedí.

Čo sa Vám páči na webovom portále? (vzhľad, rozloženie..)

- Najviac sa mi páči, že sa výsledky zobrazujú okamžite bez nejakého čakania (refreshovania stránky). Vzhľadovo pôsobí veľmi príjemným dojmom, pekné farby.
- Prehľadnosť, jednoduchosť, dá sa v ňom rýchlo zorientovať.
- Pekné grafické zobrazenie prvkov portálu.
- Jednoduchost, provázanost grafů a posloupný výpis informací.

Čo sa Vám nepáči na webovom portále?

- Absencia menej známych automobilových značiek.
- Mapa by mohla byť zarovnaná na stred.
- Zo začiatku je orientácia trochu chaotická, ale po chvíle práce s portálom bola práca prehľadnejšia.
- K některým funkcím přidat nápovědu, nebo je zvýraznit.

Aké výhody Vám prináša webový portál?

- Všetky odpovede na jednom prehľadnom mieste. Páči sa mi možnosť vyhľadať hashtagy v danom časovom úseku a vytvoriť si prehľad o tom, čo bola najviac diskutovaná téma v danom čase.
- Možnosť prehľadne si zistiť informácie od užívatelov, ich názory a hodnotenia.
- Možnosť zistenia názoru iných ľudí o určitých značkách, ktoré ma zaujímajú.
- Plno informací na jednom místě, hodnocení aut nejen podle značek.

Stretli ste sa už s podobným systémom na analýzu dat?

- Áno stretol, ale so všeobecným portálom na analýzu dát na Twitteri. Ten ale neponúkal takéto rozšírené štatistiky.
- 3x nie

Budete využívať v budúcnosti webový portál?

- Určite, keď sa nebudem vedieť rozhodnúť pri výbere auta.
- 3x áno

Kapitola 7

Záver

V úvode práce bol popísaný význam získavania dát zo sociálnych sietí a boli vysvetlené základné metódy ich analýzy. Ďalej bolo popísané, z akých sociálnych sietí sa budú informácie získavať a akým spôsobom. Taktiež bol popísaný celkový proces vytvárania vhodných dát na indexáciu a možnosti ako ich vizualizovať. Výsledkom práce je webový portál[1], ktorý tieto dáta vizualizuje do vhodných grafov a ponúka nám rôzne štatistiky, ktoré zaujmú hlavne automobilových fanúšikov. Tento portál je ku koncu práce prezentovaný na rôznych príkladoch, ktoré reprezentujú jeho možné využitie. Taktiež bolo otestované a vyhodnotené užívateľské rozhranie s potencionálnymi užívateľmi portálu.

Databáza obsahuje veľké množstvo dát, a tým pádom vývoj tohto portálu bude smerovať predovšetkým na rozšírenie užívateľského rozhrania. To znamená pridanie nových sekcií, grafov a predovšetkým rôznych prepínačov ku aktuálnym grafom. Ďalšie rozšírenie, ktoré by bolo vhodné, je pridať administrátorskú sekciu. Administrátor by mohol upravovať, pridávať aspekty alebo nové značky atď. Možným rozšírením je taktiež vypracovať webový portál do podoby mobilnej aplikácie.

[1] http://athena1.fit.vutbr.cz/xbezak01/WWW/index.html

Literatúra

[1] The Top 20 Valuable Facebook Statistics.
URL https://zephoria.com/top-15-valuable-facebook-statistics/

[2] Adedoyin-Olowe, M.; Gaber, M. M.; Stahl, F.: *A Survey of Data Mining Techniques for Social Network Analysis*. June 2014, [Online; navštívené 25.03.2017].
URL https://arxiv.org/pdf/1312.4617.pdf

[3] Barbier, G.; Liu, H.: Data Mining in Social Media. In *Social Network Data Analytics*, editace C. C. Aggarwal, kapitola Data Mining in Social Media, Springer US, 2011, s. 327–352.

[4] Brody, S.; Elhadad, N.: An unsupervised aspect-sentiment model for online reviews. *HLT '10 Human Language Technologies*, 2010.

[5] Hegde, V.; P, M. K.; Madhu, M.: Opinion Mining And Market Analysis. *International Journal of Applied Engineering Research*, 2015.

[6] Hercig, T.: *Aspects of Sentiment Analysis*. Dizertační práce, University of West Bohemia in Pilsen, 2015.

[7] Hercig, T.; Brychcín, T.; Svoboda, L.; aj.: UWB at SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, 2016, s. 342–349.

[8] Kumar, D.; Tyagi, A.; Tyagi, S.: *Sentiment analysis using naive bayes classifier*. October 2014, [Online; navštívené 27.03.2017].
URL https://www.slideshare.net/DevSahu2/sentiment-analysis-using-naive-bayes-classifier-39784368

[9] Kyriakou, A.: *Facebook's Graph API v2.0 kills data mining*. June 2014, [Online; navštívené 26.03.2017].
URL https://github.com/ptwobrussell/Mining-the-Social-Web-2nd-Edition/issues/205

[10] Lei, Z.; Riddhiman, G.; Mohamed, D.; aj.: Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. Technická zpráva, HP Laboratories, June 2011.

[11] Álvarez López, T.; Juncal-Martínez, J.; Fernández-Gavilanes, M.; aj.: GTI at SemEval-2016 Task 5: SVM and CRF for Aspect Detection and Unsupervised Aspect-Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California: Association for Computational Linguistics, 2016, s. 306–311.

[12] Novak, P. K.; Smailović, J.; Sluban, B.; aj.: Sentiment of Emojis. *PLoS ONE*, 2015.

[13] NT, B.: *Top 50 open source web crawlers for data mining*. January 2015, [Online; navštívené 26.03.2017].
URL http://bigdata-madesimple.com/top-50-open-source-web-crawlers-for-data-mining/

[14] Pang, B.; Lee, L.: Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, ročník 2, č. 1-2, 2008: s. 1–135.

[15] Pang, B.; Lee, L.; Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceeding EMNLP*, 2002: s. 79–86.

[16] Rana, T. A.; Cheah, Y.-N.: In *Artificial Intelligence Review*, 2016, str. 459–483.

[17] Shukri, S.; Yaghi, R. I.; Aljarah, I.; aj.: Twitter Sentiment Analysis: A Case Study in the Automotive Industry. In *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies*, Amman, Jordan, 2015.

[18] Vryniotis, V.: *Machine Learning Tutorial: The Naive Bayes Text Classifier*. October 2013, [Online; navštívené 25.03.2017].
URL http://blog.datumbox.com/machine-learning-tutorial-the-naive-bayes-text-classifier/

Prílohy

Zoznam príloh

Príloha A

Obsah CD

Priložené CD obsahuje nasledovné súbory a adresáre:

- */doc* zdrojové súbory technickej správy
- */src* zdrojové súbory pre prácu s dátami
- */web* zdrojové súbory webového portálu
- *xbezak01_BP.pdf* technická správa
- *plagat.pdf* plagát prezentujúci prácu a jej výsledky

Príloha B

Konfigurácia

B.1 Konfigurácia nástroja crontab

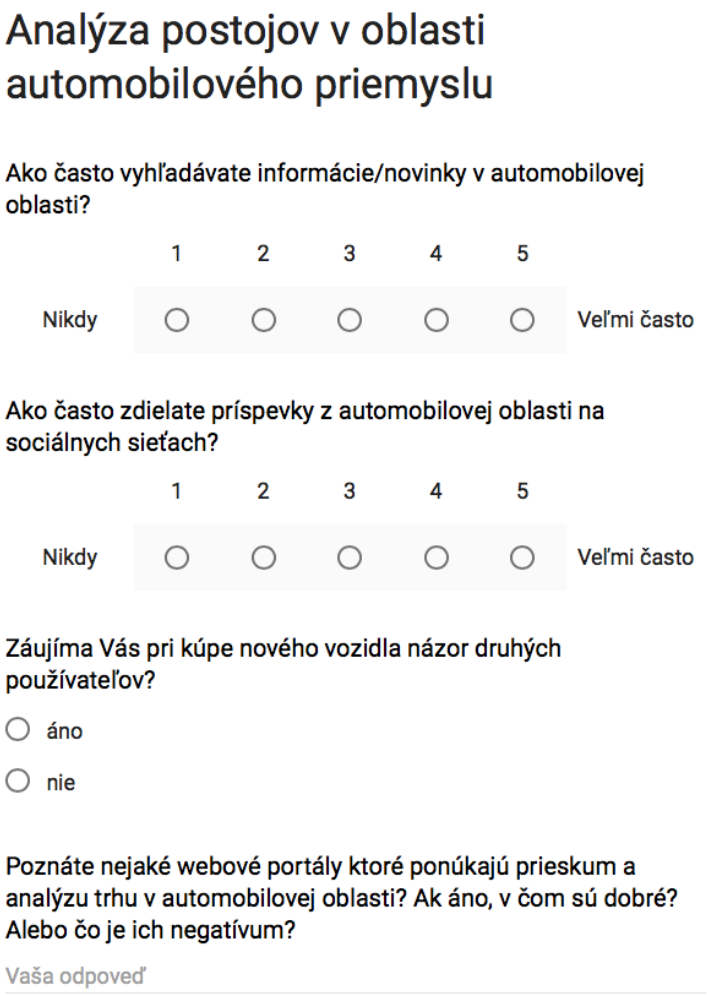
```
0 * * * * source twitter_down.py
0 * * * * source facebook_down.py
```

Obr. B.1: Konfigurácia nástroja crontab.

Príloha C

Dotazníky

C.1 Vzor dotazníku A



Analýza postojov v oblasti automobilového priemyslu

Ako často vyhľadávate informácie/novinky v automobilovej oblasti?

	1	2	3	4	5	
Nikdy	○	○	○	○	○	Veľmi často

Ako často zdielate príspevky z automobilovej oblasti na sociálnych sieťach?

	1	2	3	4	5	
Nikdy	○	○	○	○	○	Veľmi často

Záujíma Vás pri kúpe nového vozidla názor druhých používateľov?

○ áno

○ nie

Poznáte nejaké webové portály ktoré ponúkajú prieskum a analýzu trhu v automobilovej oblasti? Ak áno, v čom sú dobré? Alebo čo je ich negatívum?

Vaša odpoveď

Obr. C.1: Vzor dotazníku na zistenie požiadavkov užívateľov systému.

C.2 Vzor dotazníku B

Obr. C.2: Vzor dotazníku na zistenie vyhodnotenia práce so systémom.

Príloha D

Plagát

Obr. D.1: Plagát výsledného webového portálu.