Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Katedra informačních technologií

Studijní program: Aplikovaná informatika
Obor: Informační systémy a technologie

Hodnocení e-Word Of Mouth českých bank na Facebooku a webových komentářích

DIPLOMOVÁ PRÁCE

Student : Bc. Petr Škola
Vedoucí : doc. Ing. Ota Novotný, Ph.D.
Oponent : Ing. Ivan Jelínek

2015

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracoval samostatně a že jsem uvedl všechny použité prameny a literaturu, ze které jsem čerpal.

V Praze dne 26.4.2015 . . . . . .

Bc. Petr Škola

Poděkování

Rád bych poděkoval doc. Ing. Otovi Novotnému, Ph.D za rady a vstřícnost, které mi věnoval při vedení této práce. Chtěl bych poděkovat také své přítelkyni Ivaně Jandové za velkou trpělivost, kterou měla, během doby vypracování této práce.

Abstrakt

Diplomová práce se zabývá analýzou příspěvků z internetových diskuzí a z facebookových stránek, které se týkají bank a bankovních produktů. Tato analýza je zpracována kvůli potenciální hodnotě pro marketingové účely, ale i pro další uživatele mimo oblast marketingu. Hlavním cílem je návrh pravidelně podávaného přehledu metrik a charakteristik, které čtenáři poskytnou informace o aktuálních tématech, která se probírala na sledovaných datových zdrojích.

V úvodu jsou popsány metody marketingového výzkumu s ohledem na možnost získávání datových podkladů z internetových diskuzí. Dále jsou zpracovány dílčí cíle. Základem je popis technik pro stahování dat ze jmenovaných zdrojů, což je popsáno konkrétněji na dvou vytipovaných webových stránkách a pěti facebookových profilech bank, které jsou zdrojem dat. Dalším dílčím cílem je vytvoření programů pro stahování těchto dat v Javě a jejich uložení v Elasticsearch. Tato data jsou obohacena o analýzu sentimentu příspěvků. Pro zpracování hlavního cíle jsou definovány metriky a charakteristiky, které se budou zobrazovat. Následně jsou data analyzována pomocí navržených vizualizací v aplikaci Kibana. Takto připravená data jsou interpretována a je navrhnuto zobrazení pro jejich distribuci, které je hlavním cílem této práce.

Přínosem této práce je popis zpracování dat, která lze získat z internetových stránek a veřejných facebookových profilů s důrazem na jejich obsah, a možnosti jejich dalšího obohacení a zobrazení uživatelům.

Klíčová slova

Marketing, Facebook, internetové diskuze, Elasticsearch, Kibana

Abstract

The diploma thesis analyzes internet discussions and Facebook sites that relate to banks and banking products. This analysis has been prepared for the potential value for marketing purposes, but also for other users outside the field of marketing. The main objective is to propose a regularly administered overview containing metrics and characteristics, which would provide information about the current status of the topics that were discussed at the monitored data sources.

At first are described the methods of marketing research regarding to the possibility of obtaining data based from Internet discussions. Next are described the intermediate objectives. The base is description of downloading data from the mentioned sources. There have been identified two web pages and Facebook profiles of five banks, which are the data sources. Another objective is to create Java programs for downloading data and storing them in Elasticsearch. This data is enriched by sentiment analysis of users' comments. The main objective is based on defined metrics and characteristics that will be displayed. Subsequently, the data are analyzed using the proposed visualization in the application Kibana. The resulting data are interpreted, and there is designed a form of their distribution, which is the main objective of this work.

The contribution of this work is the description of the processing of data that can be obtained from the website and public Facebook profiles with emphasis on their content, and their further enrichment and finally data visualization designed for wide range of audience.

Keywords

Marketing, Facebook, Internet discussion, Elasticsearch, Kibana

Obsah

1. Úvod

Veřejná hodnocení produktů a služeb jsou v dnešní době dostupnější než kdy dříve. Internet je zdrojem nepřeberného množství názorů uživatelů o produktech a službách. Skrývá potenciál, který je možné použít pro získání informací širšímu okruhu uživatelů. Tyto informace jsou na internetu k dispozici na různých webových stánkách, diskusích, fórech a sociálních sítích. Těžit z těchto zdrojů mohou firmy, odborná veřejnost i zpětně samotní uživatelé, kteří příspěvky napsali. Mým cílem je vybrat z těchto zdrojů malý vzorek a ukázat, jakým způsobem lze informace získat a prezentovat, aby byly použitelné pro širší publikum.

1.1. Vymezení tématu práce a důvod výběru tématu

Sociální sítě a internetové diskuse jsou pro firmy zdrojem informací, které mohou využít pro zlepšení svých produktů a pro zjištění zpětné vazby na své služby. Pomocí analýzy dat získaných z webu mohou firmy lépe poznat potřeby stávajících i potenciálních zákazníků. Tato práce si dává za cíl ukázat, jak je možné získat zpětnou vazbu a podněty pro rozvoj produktů v odvětví bankovnictví s využitím veřejně dostupných dat ze sociálních sítí a webových stránek, které se zabývají bankovními službami. Výstupem práce je popis metod a postupů zpracování veřejně dostupných dat z internetu s cílem vytvořit pravidelně podávaný přehled o aktuální situaci v daném odvětví z pohledu lidí, kteří přispívají na sociálních sítích na profilech bank nebo v diskusích pod články vybraných webů, které se bankovnictvím zabývají.

1.2. Cíle práce, metriky měření dosažení cílů a jejich cílové hodnoty

Na trhu existují služby, které dodávají firmám přehled o tom, kde se na webu o těchto firmách hovoří. Klasický monitoring medií firmám dává přehled o tom, kde byla ve vydaném textu zmínka o některém z klíčových slov pojících se k dané firmě a jejím produktům nebo sféře podnikání. Tento přehled je důležitý z pohledu mediálního obrazu firmy. Ale pokud jde o internetová média, nedává zpětnou vazbu, která by byla vyjádřená lidmi a jejich komentáři pod článkem nebo pod příspěvkem na sociálních sítích. Sociální sítě navíc poskytují další prostředky, které mohou podat přehled o mínění lidí o produktech a službách bank.

Hlavním cílem je vytvořit pravidelný přehled údajů, který hodnotí banky podle toho, v jakých souvislostech o nich uživatelé hovoří, jak často jsou zmiňovány a jak si obecně stojí v očích uživatelů. Pro splnění tohoto cíle je nutné identifikovat média na webu a stránky na sociálních sítích, kde lze získávat pravidelně informace od zákazníků a uživatelů bankovních služeb, které po zpracování budou dávat přehled o aktuálním chování zákazníků. Tato data jsou stažena a zpracována pomocí nástroje Elasticsearch.

Pro zpracování hlavního cíle je nutné splnit následující postup. Data ze všech zdrojů jsou stažena a uložena do Elasticsearch, dále jsou obohacena o kategorii sentimentu. Data jsou následně vizualizovány v nástroji Kibana. Tím je získán základní pohled na datové vstupy a ty jsou za jedno období interpretovány. Data je nutné dále zobrazit v distriubuovatelném přehledu, což je hlavním cílem této práce. Tento výstup slouží pro zveřejnění formou emailu nebo na internetu v pravidelných intervalech.

1.3. Způsob/metoda dosažení cíle

Cílového výstupu bude dosaženo vytvořením systému, který zabezpečí stažení potřebných dat z internetu, jejich uložení, následné zpracování a analýzu dat. Data jsou v kontextu této práce komentáře, které se týkají bankovnictví vyskytující se na českých internetových stránkách nebo facebookových (FB) profilech českých bank. Je navrženo několik metrik a charakteristik, které souhrnně za každou banku hodnotí, jak si stojí v očích uživatelů. Metriky slouží jen pro základní analýzu dat a jako prostředek pro získání přehledu o kvantitativní stránce témat, o kterých se v příspěvcích hovoří, a jejich složení v různých řezech napříč daty. Součástí analýzy je vytvoření modelu pro kategorizaci sentimentu příspěvků. Pro tuto analýzu jsou prozkoumány české zdroje a je navržena jednoduchá metoda pro její provedení. Základem je rešerše zdrojů pro sentiment analýzu a přístupů k ní.

Předposlední částí je návrh dashboardu, který metriky přehledně zobrazuje a dodává další podrobnější informace. Díky tomu lze data interpretovat. Výstupem práce je sada vizualizací, které slouží k distribuci uživatelům emailem nebo na internetu nebo i tištěnou formou.

1.4. Předpoklady a omezení práce

Pro stažení uložení dat jsou vytvořeny programy za použití knihoven v jazyce Java. Pro analýzu dat je použit software Elasticsearch, další pomocné pluginy a aplikace Kibana. Systém stahování a uložení dat je spuštěn na osobním notebooku.

Nelze v rozsahu této práce stahovat veškerý obsah internetu, který se týká bank. Proto je vybráno několik zdrojů, které jsou stahovány a analyzovány. Konstrukce systému umožňuje přidání dalších webů a facebookových profilů, které mohou být přidány jako rozšíření této práce. Jako vzorek dat jsou použity FB profily pěti největších bank v ČR, které poskytují služby univerzálního bankovnictví, kde lze předpokládat největší uživatelskou aktivitu. Výběr webových stránek byl na autorově uvážení a byly vybrány na základě vlastní zkušenosti. Před analýzou dat nelze říci, zda je zdroj dostatečně kvalitní. Lze ručně procházet jednotlivé stránky v rámci webu a ručně počítat relevantnost příspěvků k bankovní tématice. Tato analýza provedena nebyla, i proto že je relativně jednoduché přidat další webové stránky jako datový zdroj uživatelských příspěvků. Byly vybrány stránky, kde lze předpokládat silné zastoupení relevantních příspěvků a jejich kvalita je vyhodnocena v závěru.

Cílem není detailně popisovat použitou technologii a komentovat veškerý kód vytvořených programů pro stahování dat, které jsem vytvořil nebo použil z veřejně dostupných zdrojů s licencí, která mě k jejich nekomerčnímu použití opravňuje. Soupis použitých knihoven a autorů je uveden ve zdrojích. Tyto metody a technologie jsou dostatečně popsány v jiných zdrojích a dokumentaci. Důraz je kladen na výsledky analýzy získaných a zpracovávaných dat.

Může se stát, že množina stažených dat nebude dostatečně silná a hodnocení nebude vypovídající na dané periodě (například za týden). V tomto případně by bylo nutné přidat další datové zdroje - komentáře na dalších webových stránkách nebo profilech na Facebooku. Další možností je prodloužit sledované období na čtrnáctidenní nebo měsíční. Tento předpoklad se nepotvrdil a z týdenního vzorku dat je možné analýzu zpracovat.

Budou zpracovány jen české internetové stránky a tedy jen české texty, protože se zabývám českými bankami. Zpracování českého textu je pak také složitější při sentiment analýze, lze ji dále zpřesňovat úpravou technik použitých pro vyhodnocení.

Provedení sentiment analýzy je obecně složitý problém a o to víc v českém prostředí díky rozmanitosti Českého jazyka oproti jiným jazykům. Dílčím cílem této práce zjistit, jaké jsou možnosti české analýzy sentimentu pro hodnocení komentářů z vytipovaných zdrojů. Je vytvořen kategorizační model, jehož kvalita je zhodnocena v závěru. Jelikož není hlavním cílem vývoj modelu, je jen na vzorku dat vypočítána jeho úspěšnost. Není tedy mým cílem úspěšnost dalšími možnými postupy zvyšovat.

Nastavení systému, metriky a dashboard jsou aktuální v době zpracování diplomové práce. Předpokládá se, že pro dlouhodobé používání bude nutné systém zpracování a jeho dílčí celky aktualizovat.

Analyzovaná data nemusí vždy vyjadřovat názory jejich autorů, data může dále zkreslit cenzura, která funguje na Facebooku i u webových diskusí. Pokud převládnou účelně vytvořené komentáře a naopak budou chybět cenzurované, bude tímto způsobem zkreslena i analýza dat. Tento problém není v práci nijak ošetřen.

1.5. Struktura práce

Úvodní část práce je věnována marketingu a marketingovým nástrojům a principům, které pomáhají získat a hodnotit zákaznické požadavky. Jako výchozí bod je popsán marketingový výzkum, což je proces získání relevantních a kvalitních informací z různých zdrojů a různými způsoby, které manažeři potřebují pro svá podnikatelská rozhodnutí. Pak je zde popsán princip Word Of Mouth na internetu, tedy předávání zkušeností a názorů a ovlivňování ostatních uživatelů v on-line prostředí. Navazuje popis metody Voice Of Customer (Hlas zákazníka), která porovnává způsoby získávání zpětné vazby. Déle následuje popis metody Net Promoter Score, která v základu vyjadřuje loajalitu a spokojenost zákazníků díky bodovému ohodnocení odpovědi na jedinou otázku. A to zda by uživatelé produkt nebo službu doporučili jiným lidem. V závěru kapitoly je uvedena rešerše prací, které se zabývají analýzou sentimentu, jsou popsány metody pro provedení sentiment analýzy.

Následující kapitola 3 popisuje základní principy stahování dat z webových stánek pomocí takzvaných web crawlerů a získávání dat z Facebooku pomocí API. Jsou uvedeny obecné principy a požadavky, které vedou k získání dat z těchto zdrojů. V další části kapitoly je popsáno, které facebookové stánky a které weby byly vybrány pro analýzu. Navazuje popis vytvořených programů, které slouží pro stažení a obohacení dat o kategorii sentimentu. Pro oba datové zdroje jsou použity Java knihovny, které zjednodušují tuto činnost.

Další kapitola je o Elasticsearch a dalších použitých pluginech a aplikacích, pomocí kterých je analýza dat provedena. Popisuje způsob předzpracování a indexace dat, vyhledávání komentářů a jejich analýzu. Sumarizuje postup pro stažení údajů potřebných pro výpočet metrik a dalších charakteristik, které jsou navrženy v následující kapitole.

Způsob výpočtu metrik a dalších charakteristik je popsán v kapitole 0. Jsou zde definovány metriky, které lze z komentářů na Facebooku a na webu získat a dále je navržen způsob jak

tyto informace agregovat do dashboardů tak, aby vyjadřoval, jak se o bankách v daném období hovoří a jaká je jejich pozice ve sledovaných metrikách oproti jiným bankám. Kromě metrik jsou zde navrženy další nekvantitativní charakteristiky dat, které ukazují, jakým tématům se ve stažených příspěvcích uživatelé věnovali.

Předposlední kapitola obsahuje návrh dashboardu. Ten má sloužit širšímu publiku, které chce poznat, jaký je obraz bank na webu a Facebooku v očích veřejnosti. Jsou zde zobrazeny definované metriky. Dále je popsána možnost zobrazit témata, o kterých se hovoří pomocí nástroje pro vytváření clusterů carrot2. Následně je v kapitole 8 navrženo konsolidované řešení, které by dokázalo zobrazit všechna agregovaná a konsolidovaná data na jednom místě.

Závěrečná kapitola 9 shrnuje výsledky práce a dává doporučení pro další rozvoj analýzy dat z internetových diskusí a Facebooku. Obsahuje také zjištěná omezení a možnosti pro další rozvoj.

1.6. Výstupy práce a očekávané přínosy

Hlavním výstupem práce je vizualizace hodnocení jednotlivých aspektů vnímání služeb zákazníků bank. To je zpracováno pomocí tabulek, které splňují požadavek na zobrazení, kterým lze data jednoduše Pro distribuovat různými kanály. Složitější, ale variabilnější prezentaci dat je vytvořen návrh dashboardu, který je dílčím cílem této práce. Oba zmíněné výstupy mají širokou škálu konzumentů. Může být jedním z ukazatelů marketingových oddělení bank, jimž zhodnotí, jak si stojí oproti jiným bankám. Zobrazují, jaká témata se daný týden probírala nejvíce. Může tedy také sloužit jako zpětná vazba při zavedení nového produktu, přehledu o konkurenci anebo pro objevení přání zákazníků.

Klientům napoví, jaká banka je k zákazníkům vstřícná a o které lidé hovoří ve spojitosti s jakými tématy. Dlouhodobější sledování výstupu tedy může napovědět, kde žádat bankovní produkty. Případně naznačí, jaké problémy řeší zákazníci bank.

1.7. Rešerše prací s podobným zaměřením

Tématem tvorby a hodnocení marketingové kampaně na Facebooku se zabývá mnoho vysokoškolských kvalifikačních prací. Většina prací vytváří nebo hodnotí komunikaci konkrétní firmy. Následující výčet je výběrem prací na podobné téma.

Viktor Gleich se zabývá v diplomové práci s názvem Marketing na sociálních sítích definicemi metrik pro vyhodnocování marketingových akcí na Facebooku. Popsal

marketingovou strategii pro společnost Módní guru. Dále porovnává nástroje a aplikace pro monitoring metrik a ukazuje výstupy těchto nástrojů. Výsledkem jsou doporučení pro společnost Módní guru. (1)

Bakalářská práce Ljuby Kotáskové s názvem Budování značky na sociálních sítích porovnává komunikaci čtyř českých bank na jejich facebookových profilech. Pro porovnání používá nástroj Analytics PRO od Socialbakers, kde sleduje tři metriky: počet fanoušků, Engagement rate (počet interakcí fanoušků vůči počtu fanoušků) a Respose rate (počet příspěvků na stránce banky vytvořených fanoušky). Další výzkum probíhal pomocí dotazníkového šetření, což není pro mou práci relevantní. (2)

V diplomové práci Ondřeje Linharta jsou popsány přístupy, které umožní použít data ze sociálních sítí, zejména z Facebooku a Twitteru, jako součást Business Intelligence řešení podniku. Autor získává data pomocí Graph API z Facebooku. Tato práce dává hlavně detailní přehled o datech dostupných na Facebooku. (3)

Získání dat z Facebooku a jejich uložení je také popsáno v bakalářské práci Martiny Hradské. Autorka vytvořila aplikaci pro pravidelné stahování dat z této sítě pomocí vlastní aplikace s uložením dat do Apache Solr. Analýze dat se dále nevěnuje. (4)

Diplomová práce Martina Šveráka Analýza nestrukturovaného obsahu z veřejně dostupných sociálních médií za pomocí nástroje Watson společnosti IBM pokrývá tři datové zdroje a pomocí nástroje IBM analyzuje data týkající se společnosti Vodafone CZ z několika profilů na Facebooku, Twitterových účtů a webových diskusí. (5) Tématem je tato práce nejblíže mé práci. Rozdíl je v použité technologii a v oblasti, kterou zpracovávám, a také to, že mým cílem není věnovat se jedné značce ale odvětví.

2. Marketingový význam analýzy

Internetové diskuse a sociální sítě mohou být zdrojem dat pro marketingový výzkum, jehož charakteristika je popsána v první části této kapitoly. Tato část je podkladem pro identifikaci oblastí marketingového výzkumu, kterým může analýza uživateli generovaných dat na internetu poskytnout informace. Další část kapitoly shrnuje základní marketingové přístupy používané pro získání zpětné vazby a její hodnocení. Následuje část popisující metodu Net Promoter Score, která hodnotí službu nebo produkt otázkou, zda by ho zákazníci doporučili. Význam marketingu je v uspokojování potřeb zákazníka. Marketéři tedy potřebují vědět, co si o jejich produktu zákazníci myslí a co by měli zlepšit, aby byl lepší a žádanější než konkurenční. K tomu slouží marketingový výzkum. (6)

2.1. Marketingový výzkum

Marketingový výzkum je dlouhodobá aktivita, která má za cíl zjistit tržní potenciál, kapacitu trhu, verifikovat produkt, jeho cenu, zjistit zda se očekávání zákazníků setkávají s tím, co produkt nabízí. Marketingový výzkum je jedinečný, protože informace jsou k dispozici jen pro zadavatele výzkumu. Tyto informace musí být aktuální a tím by měly mít vysokou vypovídací schopnost. Aby informace byly získány v odpovídající kvalitě, je nutné na jejich získání vynaložit relativně vysoké finance a je potřeba mít dostatečně kvalifikované pracovníky, kteří zajistí jejích získání. (7)

Základní předpoklady správně prováděného výzkumu jsou objektivita a systematičnost. Jde o vědeckou metodu, která zahrnuje vědecké postupy ze statistiky, psychologie sociologie a tak dále. Výzkum je zároveň i tvůrčí práce, která má vyhledávat nové příležitosti a nové přístupy. Proto se využívá intuice, která je použita pro tvorbu hypotéz, které jsou následně buď přijaty, nebo vyvráceny na základě zjištěných informací. (7)

Nejvyšší formou marketingového výzkumu je formalizované systematické získávání dat se specifikovanými cíli a rozhodnutími, které má ovlivnit, se specifikovaným rozsahem a metodami a finančními a lidskými zdroji.

K marketingovému výzkumu lze ale přistoupit i méně formálními metodami:

- Soustavné nepřímé sledování bez konkrétního cíle
- Podmíněné sledování jen vymezené oblasti
- Neformální výzkum (omezené a nesystematické vyhledávání informací)

Výstupy této práce mohou sloužit pro všechny výše uvedené typy marketingového výzkumu. Záleží jen na jeho uživatelích, jak systematicky k použití výstupu budou přistupovat.

Pro manažerská rozhodnutí je nutné, aby byly informace dodány v optimální kvalitě, množství a čase. Podmínka včasného dodání informací bývá často nesplněna a informace jsou dodávány se zpožděním v řádu měsíců, protože získání dat je časově náročná práce. Z toho vyplývá, že manažeři nevědí, podle čeho se rozhodovat a nemají přehled o aktivitách firmy a výsledcích podnikatelské aktivity. (7)

Kvalitní informace jsou podkladem pro rozhodování na všech úrovních:

- Strategické – kam zaměřit racionálně marketingové úsilí
- Taktické – kterými aktivitami toto úsilí podpořit
- Kontrola – informační zpětná vazba

Manažeři jsou v posledních letech nuceni se stále více zajímat o to, jaké potřeby má jejích zákazník, aby je mohli plnit a vytvářet dlouhodobý vztah.

Jedním z kanálů marketingového výzkumu je používání internetu pro získání informací. V knize Marketingový výzkum (7) je uvedena možnost použití internetových vyhledávačů a e-shopů jako zdroje pro zachycení aktuální poptávky a jako kanálu, který pomáhá generovat příjmy. Další možností zjištění potřebných podkladových dat je vytvoření online průzkumu, ovšem s upozorněním na problém relevantnosti dat, která nemusí být zaručena. Zde se ukazuje jako možnost monitoring online médií, který přináší další zdroj dat o zákaznickém chování.

Online media zahrnují také sociální media, kde je obsah vytvářen komunitou - obecně uživateli internetu. Sociální sítě, blogy, fóra vytvářejí pro firmy možnost, jak se zviditelnit a přesvědčit svými sděleními o kvalitě svých služeb. Zároveň ale mohou být zdrojem informací, protože na internetu uživatelé sdílejí své osobní zkušenosti a na ty pak navazuje jejich předávání mezi dalšími uživateli.

Průzkum trhu je dílčí součástí marketingového výzkumu. Podle průzkumu (8) ze září 2014, kde respondenti dostali seznam metod výzkumu a měli označit ty, které využívají a plánují využívat, se ukázalo, že analýza dat z online komunit (blogy apod.), analýza dat ze sociálních sítí a analýza textu, jsou jako první čtyři nejrozšířenější metody hned za průzkumem trhu pomocí mobilních telefonů.

2.2. Word Of Mouth marketing

Na poli internetu a původně mimo něj se setkáváme s fenoménem ústně a písemně předávaným hodnocením označovaným jako Word Of Mouth (WOM). Tento termín se používá i marketingové pojetí. Word Of Mouth marketing pracuje s ústním případně písemným předáním doporučení produktů. Spokojený nebo nespokojený zákazník o svém zážitku napíše na internetu a to si mohou přečíst další uživatelé, ať už přímo znají původního zákazníka nebo ne - vzniká tedy neplacená pozitivní ale i negativní reklama předávaná mezi lidmi.

Na počátku procesu WOM je zkušenost zákazníka s konkrétní firmou, produktem nebo službou. Tento zákazník se nazývá iniciátor. Ten tuto zkušenost sdílí se svou rodinou, svými přáteli a známými ale i neznámými uživateli internetu. Může doporučit služby, které vyzkoušel, nebo naopak před nimi varovat další uživatele. Někteří z nich se stávají dalším článkem, který se jmenuje zprostředkovatel, a dále šíří tuto převzatou informaci pokud na podobné téma budou diskutovat.

Cílem Word Of Mouth marketingu je ovlivnit mínění tak, aby informace od iniciátorů a následně zprostředkovatelů, byly spíše pozitivní a to i v případě, že iniciátor chce přestat službu používat. (9)

V internetovém prostoru lze u určitých tržních segmentů nalézt osoby, které mají prostřednictvím svých názorů vliv na široké publikum. Taková osoba je infuencer nebo opinion maker. Jsou to novináři, recenzenti, blogeři, veřejně známé osoby anebo aktivní uživatelé sociálních sítí.

První možností jak chování zákazníků při doporučování upravit je zajistit, aby byl zákazník vždy spokojen. Další možnosti je považována za amorální. Společnost sama o sobě píše na internetových diskusích a propaguje sebe a svoje služby pod účty, které na první pohled nevypadají jako její.

Metody výzkumu WOM: (10)

- Text mining dat na internetu
- Focus groups
- Laboratorní experimenty
- Modelové experimenty
- Retrospektivní průzkumy

Žádná z metod ale nemá stoprocentně věrohodné výsledky. Každá má svá slabá místa. Pokud se budeme zabývat internetovými zdroji, pak East (10) říká, že není těžké nalézt na internetu data, ale může být problém v tom, že pokud jsou z jednoho zdroje / serveru, může být WOM ovlivněna a být spíše negativní nebo pozitivní. Z toho důvodu ve své práci použiji více než jeden zdroj dat.

2.3. Hlas zákazníka (Voice Of Customer)

Tento termín se používá ve dvou oblastech, v ICT a marketingu. V ICT se jedná o vyhodnocení sesbíraných požadavků a očekávání zákazníků, které se týkají služby nebo produktu ve vztahu k softwaru. V marketingu je pojetí obecnější, vždy jde ale o metodu průzkumu, jejímž výstupem je hierarchický seznam potřeb zákazníků. (11)

Dřívější postupy pro získání zpětné vazby byly například průzkumy Focus Group, telefonické dotazování nebo formuláře (dotazníky) zjišťující spokojenost. To jsou časově, finančně a personálně náročné úkoly. Podobné informace lze získat i na internetu na diskusních fórech a v komentářích, či na sociálních sítích. Výhodou těchto zdrojů je fakt, že uživatelé píší své zkušenosti s produktem nebo službou tak, jak je opravdu cítí. Píší přátelům v případě sociálních sítí nebo dalším uživatelům diskusních fór, ne ale výrobci produktu, který výše zmíněné průzkumy pořádá. Proto jsou zpravidla otevřenější a upřímnější. Problém může nastat v případě, že je diskuse zneužita pro marketingovou kampaň. Pak je důležité v diskusi vyhledat jen relevantní neovlivněné informace. (11)

Techniky pro získání hlasu zákazníka mohou odpovědět na několik různých otázek:

- Co je klíčový element zákaznické zkušenosti? Co zákazník na produktu hodnotí?
- Jak dobře dokáže společnost dodávat to, co zákazník očekává?
- Plní firma to, co slíbí v reklamě a další komunikaci?
- Která vlastnost produktu je pro zákazníka nejdůležitější?
- Co se zákazníkovi na produktu líbí a co ne a co by firma měla dělat jinak? (12)

Získat odpovědi na jednotlivé otázky lze více způsoby. Každá metoda má svá specifika a nehodí se pro každou z výše uvedených otázek.

Získat hlas zákazníka lze například těmito způsoby:

- Průzkum trhu a zákazníků pomocí dotazníků, telefonátů apod.
- Sledování chování zákazníků
- Skupinové rozhovory, interview
- Mystery shopping
- Data ze sociálních sítí a diskusí na internetu (12)

Použití dat ze sociálních sítí je vhodné pro všechny typy otázek. Není ale tou nejvhodnější metodou. Jejich hlavní výhodou je, že jde o přímou zpětnou vazbu, mohou být detailní a lze jednoduše sledovat jejich vývoj v čase. Nevýhodou pak je, že uživatelé produktu musí být dostatečně aktivní a komentovat produkt v online prostředí. Obecně jsou více slyšet nespokojení zákazníci než spokojení. A v případě malého množství dat mohou být výsledky rozporuplné. (12)

2.4. Net Promoter Score

NPS je jednoduchou metodou jak zjistit pocity a postoje zákazníků. Zákazník obvykle odpovídá na jednoduchou otázku: Jak pravděpodobné je, že byste doporučil tento produkt/službu/značku svému příteli nebo kolegovi. Odpovídá se bodovým hodnocením, kde 10 je určitě doporučil a 0 určitě nedoporučil. (13)

Z tohoto pohledu pak lze rozlišit tři skupiny zákazníků (bodové hodnocení):

- Promoters Příznivci (9-10)
- Passives Pasivní (7-8)
- Detractors Kritici (0-6)

Příznivci jsou lidé, kteří jsou loajální k firmě, nakupují opakovaně a za více peněz. Mluví o firmě, produktech a službách se svými přáteli a kolegy v dobrém světle. Každá společnost by u těchto osob měla udržovat dobrý vztah k firmě a naučit se jak ekonomicky jejich řady rozšiřovat.

Pasivní lidé jsou sice spokojeni s nákupem a firmou, ale nejsou loajální a firma by podle těchto zákazníků měla upravit své služby a produkty. Tito zákazníci jsou náchylní k nákupu u konkurenční firmy, pokud dostanou slevu, nebo si všimnou konkurenční reklamy. Kritici (pomlouvači) byli s firmou nespokojeni a mohou firmu pomlouvat u svých

kolegů a přátel. Firmy by u těchto osob měly zjistit, co je hlavním problémem, kvůli kterému ohodnotili firmu známkou šest a menší. Pokud neexistuje ekonomicky racionální způsob, jak by firma mohla vylepšit své působení tak, aby ji tito lidé hodnotili lépe, měla by se naučit, jak tento typ zákazníků svou nabídkou neoslovovat vůbec. (13)

Samotná hodnota Net Promoter Score je definována jako procento příznivců mínus procento kritiků. Nabývá tedy hodnot od -100% do 100%. Síla NPS je v jeho jednoduchosti, zpracovává se i vyhodnocuje velmi rychle a lze se respondentů dotazovat kontinuálně. Nevýhoda je, že zužuje jedenáctibodovou škálu na tři skupiny. Z toho důvodu je potřeba provést více měření, aby byl výsledek stejně přesný jako při použití průměru nebo střední hodnoty za použití jedenáctibodové škály. (14)

NPS slouží nejen jako prostředek pro měření zpětné vazby na společnost a její produkty, ale také jako impuls pro vnitřní reflexi a zlepšování procesů ve firmě, trénink zaměstnanců a transformaci fungování k větší zákaznické spokojenosti. Z tohoto důvodu je v (13) tato metoda označována také jako Net Promoter System. Dokládá to také to, že by se NPS mělo týkat všech oddělení ve firmě od financí přes provozní oddělení, marketing, produktové oddělení až po lidské zdroje a oddělení IT.

2.5. Analýza sentimentu

Z předchozího textu vyplývá, že firmy by měly chtít získat přehled o tom, co si o nich lidé myslí. K tomu může napomoci analýza sentimentu, která zkoumá postoj člověka, emoce vztažené k určitému tématu, jeho názor. (15) Pro marketingové účely může být analýza sentimentu zdrojem informací o tom, jak lidé mluví o produktech nebo službách společnosti. Pomocí rozhovorů a osobního kontaktu se zákazníky lze poměrně snadno zjistit jejich názor a hodnocení služby nebo produktu. Tímto způsobem lze zjistit ale jen omezené množství názorů. Díky internetu je k dispozici mnohem více výpovědí uživatelů ve formě recenzí, blogů, příspěvků na sociálních sítích a tak dále. (15)

Zjištění sentimentu z textu je jedním z problémů oboru zpracování přirozeného jazyka (Natural Language Processing - NPL). Výzkum v této oblasti započal ve větším měřítku po roce 2000 také díky nově dostupným internetovým zdrojům a dalším datovým zdrojům, které bylo možné analyzovat. (15)

Cílem analýzy sentimentu je klasifikace na různé úrovni granularity vstupního textu. Lze zkoumat povahu celého textu, vět anebo jednotlivých entit. Klasifikace celého textu ohodnotí celek jednou z předdefinovaných kategorií. Tím je zjištěno, jaké vyznění nese

celý příspěvek – například recenze produktu. Předpokladem pro toto hodnocení je, že celý text popisuje jednu entitu – produkt. Jemnějším způsobem sentiment analýzy je klasifikace věty, kdy kategorie sentimentu je přiřazena pro každou větu textu. Nejjemnějším způsobem lze zjistit, jaký sentiment je přiřazen konkrétní entitě v analyzovaném textu. Kategorie je přiřazena konkrétnímu slovu v textu. Tento přístup tedy kromě výběru vhodné kategorie také říká, co bylo takto ohodnoceno. Příkladem z citovaného zdroje může být věta „I když obsluha není vynikající, mám tuto restauraci rád“. Celkové vyznění je pro restauraci pozitivní, ale méně pozitivní pro obsluhu. (15)

Přístupů k získání emočního zabarvení, sentimentu je více. Liší se použitým teoretickým základem a přesností. Základním způsobem ohodnocení textů je detekce na základě klíčových slov. Úspěšnost této metody závisí na slovníku (databázi) klíčových slov, která jsou předem definována a ohodnocena mírou příslušnosti k emočním třídám – tedy jak moc vyjadřují negativní nebo pozitivní postoj. Problém přesnosti je zde způsoben například nejednoznačností některých slov, šířkou databáze klíčových slov a četností jejich výskytu v hodnoceném textu. (16)

Dalším přístupem je použití různých metod strojového učení a postupů přípravy dat pomocí NLP. Metod a algoritmů NLP, které lze použít je několik, stejně tak existují různé kombinace metod přípravy dat, které mohou zlepšit celkovou úspěšnost hodnocení. Proces je vždy obdobný. Po přípravě dat následuje trénování modelu klasifikace textů, kdy se tvoří báze, která bude následně určovat kategorii neohodnocených příspěvků. (16) Kombinace přístupů je označována jako hybridní přístup. Používají se při tom různé metody předzpracování textu, klíčová slova, lexikální báze, sémantika a strojové učení. (16)

Tato práce nemá za cíl zkoumat, který přístup je pro hodnocení sentimentu u zpracovávaných datových zdrojů nejlepší, cílem je použít jednoduchou metodu i s tím, že nejde o nejlepší možnou variantu. I přesto že sentiment je jedním z poměrně důležitým parametrem, který je v následujících analýzách hodnocen, je tento postup zvolen i proto, že tématem sentiment analýzy se zabývají celé závěrečné práce. Pro inspiraci v možném rozšíření této práce lze tedy použít například přístupy pro kategorizaci dokumentů podle informací z následujících zdrojů.

Analýzou sentimentu českých textů ze sociální sítě Twitter se zabývá práce Michaela Kollera (17). Ten zkoumá nástroje dostupné na internetu, které poskytují funkcionalitu

analýzy sentimentu, a hodnotí je z několika pohledů. Například zda je služba zdarma, jaké datové zdroje lze analyzovat, export a import dat, jazykové možnosti. Shrnuje funkce jednotlivých služeb, ale nehodnotí kvalitu hodnocení sentimentu.

Analýza sentimentu je název bakalářské práce Jiřího Pelíška (18). Porovnává také nástroje pro analýzu sentimentu na sociálních sítích a dále v textu. Uvádí širší teoretický základ než výše uvedená bakalářská práce, kde jsou popsány problémy a postupy spojené s analýzou sentimentu obecně a speciálně s českým jazykem.

Ještě více teoreticky zaměřenou prací s vlastním programovým řešením je diplomová práce Michala Patočky ze Západočeské univerzity (19). Důkladně popisuje teorii předzpracování dat a pak algoritmy strojového učení, které používá dále i při vývoji vlastních programů. Výsledky použití algoritmů pro sentiment analýzu pak mezi sebou porovnává a navrhuje způsob jejich kombinace pro dosažení nejlepších výsledků.

Podobně zaměřená diplomová práce jako předchozí od pana Patočky s názvem Rozpoznávání emocí v česky psaných textech je od Radka Července (16). Také popisuje algoritmy strojového učení a vytváří vlastní řešení pro kategorizaci vstupních textů.

Tyto práce mohou být inspirací pro rozšíření a zpřesnění programu pro určení sentimentu, který je použit v této práci, a jehož popis je v kapitole 4.2.1.

2.6. Závěry kapitoly

Z výše uvedených zdrojů vyplývá potřeba manažerů získávat rychle kvalitní informace o tom, co si zákazníci myslí o službách firem. Jsou zmíněny možnosti získávat část potřebných informací z internetu a to díky snadnému přístupu a rychlé dostupnosti dat. Výše uvedené metody jsou ale založeny spíše na manuálním zpracování dat a ručním procházení fór, diskusí a sociálních sítí. Mým cílem je automatizovat získávání dat a jejich předzpracování pro tyto analýzy. Dále pak automatické hodnocení v definovaných metrikách, které shrnují jednotlivé datové aspekty a v agregované podobě jsou pak porovnatelné mezi sledovanými podniky. Tato hodnocení pomohou zjistit, jaký je hlas zákazníka a jeho Word of Mouth, které předává dál. Nelze samozřejmě podchytit celý přenos zkušeností, který zákazníci kdekoliv na internetu píší. Ovšem alespoň na sledovaných veřejně dostupných zdrojích lze analyzovat témata, která uživatele zajímají. Dále lze tato témata automatizovaně ohodnotit kategorií sentimentu a tím získat rozložení témat s pozitivní nebo negativní zákaznickou zkušeností.

3. Principy stahování a dat z internetu

Tato kapitola popisuje základní principy, které jsou používány pro stahování dat z webových stránek a sociální sítě Facebook. Tyto principy jsou dále použité pro naprogramování aplikací pro stažení dat.

3.1. Webové stránky

Pro stahování dat z webových stránek se používají weboví roboti (web crawlers). Jsou to programy, které automaticky procházejí obsah internetových stránek. Začínají na takzvaných seed URL – tedy počátečních adresách, od kterých procházení začíná. Na těchto stránkách načtou zdrojový HTML kód a vyhledají podle předem definovaných pravidel požadovanou část stránky (může jít o článek, komentář a další části, nebo uloží celou stránku). Zároveň vyhledají všechny odkazy podle pravidel, které jsou definovány, a vyberou například jen adresy, které jsou na určité doméně (například jen adresy začínající www.mesec.cz/) a ty pak dále prochází. Navštívené adresy jsou ukládány v programu robota do databáze nebo do paměti programu a je tedy zaručeno, že každou stránku robot projde jen jednou. Pokud stránka ještě navštívena nebyla, je zařazena do fronty stahování. (20)

V některých případech se obsah stránky pod jednou URL během času změní - například přibyly nové komentáře. Pokud nás takové změny zajímají, je možní si ukládat do databáze například kontrolní součet celé stránky a ten při dalších spuštěních webového robota kontrolovat. Pokud se kontrolní součet liší, nastala na stránce změna a je potřeba ji stáhnout znovu. Stahováním dat pomocí webových crawlerů se zabývá zdroj (21).

Pro své potřeby stahování dat z webových stránek jsem použil jazyk Java a knihovnu crawler4j. Knihovna poskytuje jednoduché rozhraní pro procházení webu. Je vydána pod licencí Apache Licence verze 2 jako open source software. V crawler4j lze jednoduše nastavit pravidla, na které doméně se chceme pohybovat a případně specifikovat pravidla pro procházené URL, které jsou zajímavý svým obsahem. V případě této práce je to například pravidlo, které programu říká, že na doméně www.mesec.cz, jsou to stánky obsahující v URL řetězec „nazory/“. Tyto stránky totiž obsahují komentáře uživatelů.

Po spuštění tedy program prochází všechny URL, které jsou na stránkách. Ale používá dále jen ty, které obsahují řetězec „nazory/“. V dalším kroku jsou identifikovány části HTML kódu stránky s uživatelskými komentáři, které jednoznačně identifikují jméno

komentujícího, text (komentář) a datum zadání komentáře. Tyto části HTML kódu jsou separovány a připraveny pro další zpracování.

3.2. Facebook

Facebook nabízí několik způsobů, kterými lze obsah získat. K datům lze přistupovat přímo z prostředí Graph API Explorer (https://developers.facebook.com/tools/explorer/), které je přístupné všem uživatelům Facebooku po přihlášení a je vhodné pro prvotní seznámení se s daty, které lze z Facebooku získat. Toto prostředí nabízí dvě možnosti pro dotazování a zobrazování dat. (22) První je Graph API, které pomocí menu umožnuje konstruovat dotazy. Druhá možnost je FQL Query, kde lze psát složitější dotazy pomocí FQL (Facebook Query Language), což je dotazovací jazyk podobný databázovému jazyku SQL. Třetí možností je REST API, které je nezávislé na Graph API Exploreru. Dotazování zde probíhá pomocí konstrukce URL a výsledky jsou navráceny jako objekty JSON v prohlížeči nebo do aplikace, která dotaz vyvolala. (4)

Všechna rozhraní pro získávání dat jsou závislá na přihlášení uživatele. V Graph API Exploreru je potřeba vygenerovat Access Token, ten se pak používá pro přihlášení. Při jeho generování zároveň uživatel specifikuje na jaké objekty a jejich data, ke kterým má přístup, se bude umožněno pomocí tokenu dotazovat. Více o kategoriích přístupu je v práci Martiny Hradské (4).

Další možností je vytvoření vlastní facebookové aplikace. Pokud se chceme přihlásit k datovým službám pod aplikací, je nutné vygenerovat App Secret. To ve spojení s App Id, které je přiděleno při založení aplikace, slouží jako přihlašovací údaj. Výhoda tohoto přihlášení je stálý přístup k datům bez nutnosti obnovovat platnost přihlašovacích údajů jako u Access tokenu. Pro potřeby této práce jsem vytvořil aplikaci, která mi poskytuje App Id a App Secret, které zajistí časově neomezený přístup k datům.

Stahování dat z Facebooku jsem vyřešil s pomocí Java knihovny RestFB. Knihovna obsahuje třídy pro práci s facebookovými objekty. Zjednodušeně tedy stačí knihovně nastavit autorizační údaje k FB účtu, pod kterým se data stahují a vybrat si z dostupných tříd objektů. RestFB pak po spuštění provede zalogování do FB a začne tato data stahovat. Mým cílem je získat data z úvodních stránek („zdí“) českých bank. Tyto objekty se jmenují post. Mohou být reprezentovány jen textovým příspěvkem, fotografií s popisem, odkazem a podobně. U každého objektu typu post může být vlákno komentářů. Oba objekty – posty i komentáře – jsou předmětem mého zájmu. V programu je stažení těchto

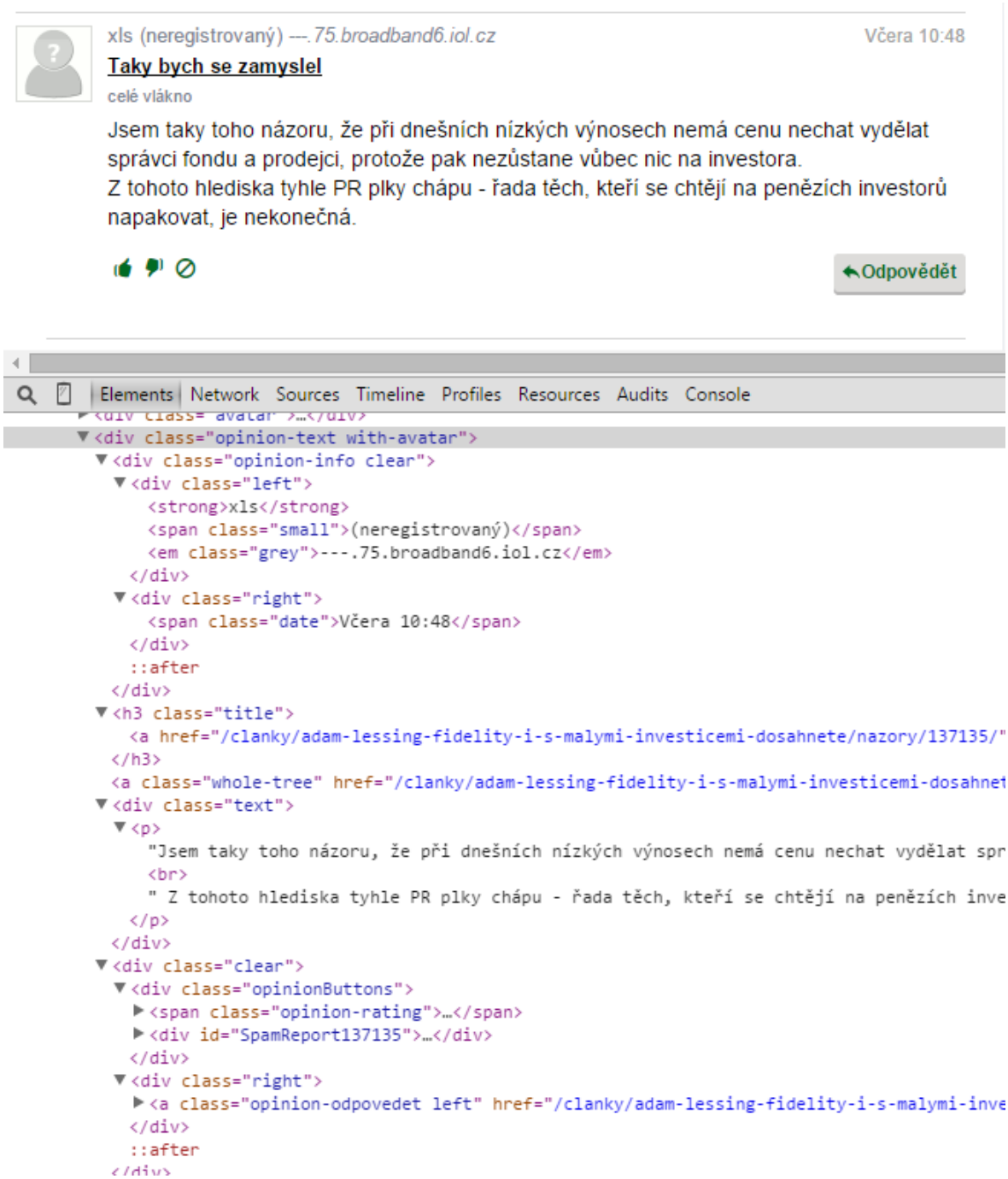
objektů docíleno tak, že se nejprve stahují všechny feed objekty a pro každý se zjišťuje, zda obsahuje komentář. Pokud ano, jsou tyto komentáře staženy. Komentáře jsou na Facebooku ve dvou vrstvách. Na každý komentář mohou uživatelé reagovat subkomentáři, tyto jsou stahovány také.

3.3. Výběr webových datových zdrojů

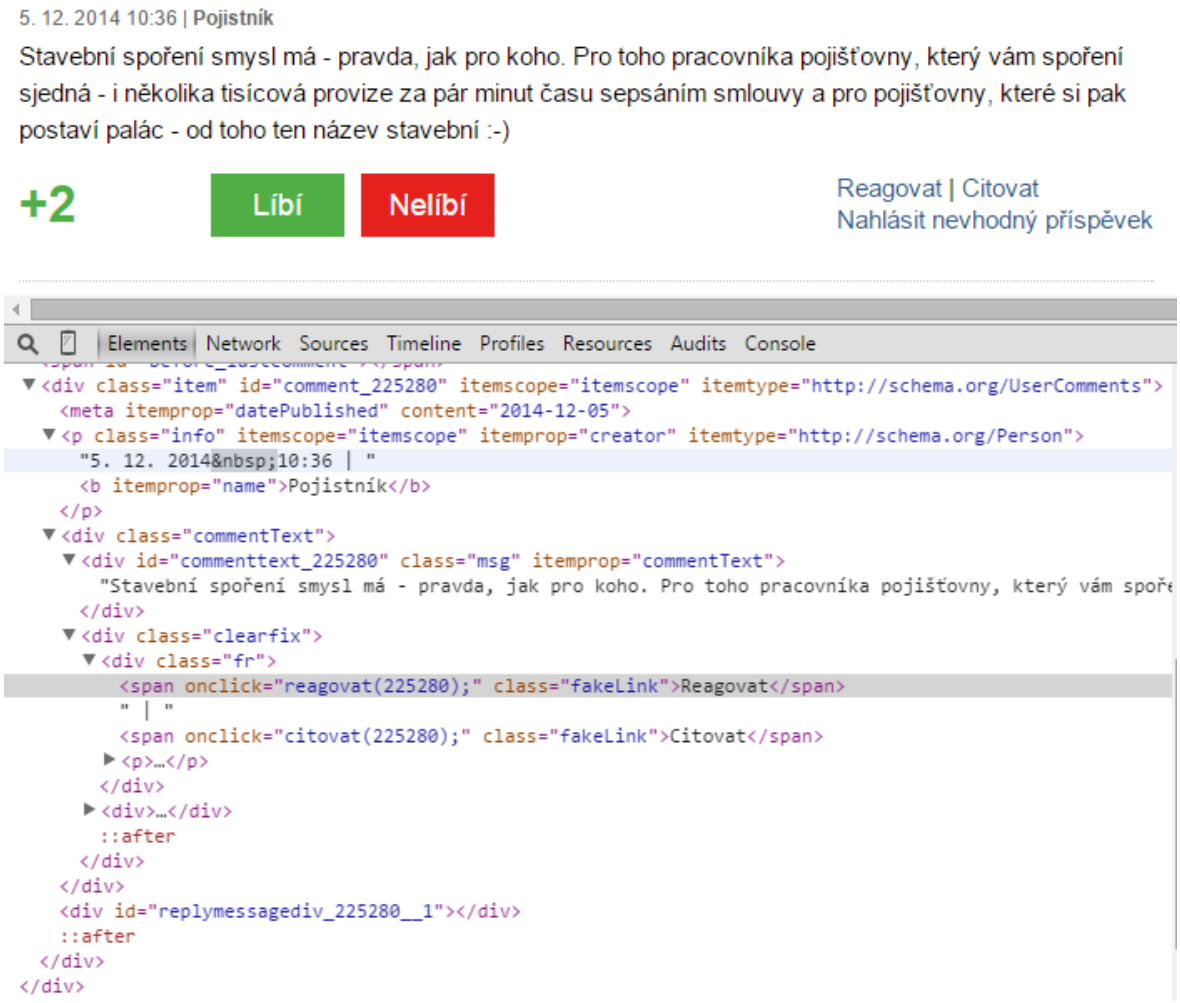
Stránky, ze kterých jsou stahovány webové diskuse, musí splňovat několik předpokladů. Stránky se musí týkat tématu financí a bank, aby se dal předpokládat vysoký podíl diskusí, které se zabývají bankovními službami. Dále diskuse musí mít dobře strukturovaný a otagovaný HTML kód, aby je bylo možné jednoduše identifikovat v celém HTML skriptu stránky.

Z webových zdrojů jsem vybral stránky http://www.mesec.cz/ a http://www.penize.cz/. Nepoužívají pro komentáře Facebookový plugin a lze pro stahování komentářů pod články snadno použít web crawler.

Struktura objektového modelu dokumentu (DOM) HTML stránek obou zdrojů obsahuje u klíčových polí, jejichž obsah je stahován, atributy s jednoznačnými názvy v rámci jednoho příspěvku. Stejně tak příspěvek má vždy identifikovatelnou třídu (atribut class). Lze tedy jednoduše získat tyto objekty například pomocí funkce getElementsByClass z Java knihovny Jsoup. (23) Další možností je použití xPath výrazů pro extrakci požadovaných dat. Výsledek bude stejný jako použití metod z knihovny Jsoup.



Obrázek 1 Ukázka HTML kódu stránky mesec.cz s detailem struktury jednoho příspěvku. (Zdroj: autor)



Obrázek 2 Ukázka HTML kódu stránky penize.cz s detailem struktury jednoho příspěvku. (Zdroj: autor)

Na Obrázek 1 a Obrázek 2 je vidět struktura části HTML kódu, který představuje jeden příspěvek na obou vytipovaných stránkách. Důležité je, že se struktura pro každý příspěvek opakuje, případně se mění jen v detailech.

3.4. Výběr profilů bank na Facebooku

Pro účely analýzy jsem vybral pět bank působících na českém trhu, které poskytují služby univerzálního bankovnictví. Vybral jsem pět s nejvyšší bilanční sumou v roce 2012, které mají český profil na Facebooku:

Název banky	Profil na Facebooku
Československá obchodní banka	nemá český profil na FB
Česká spořitelna	ceskasporitelna
Komerční banka	komercni.banka
UniCredit Bank Czech Republic	UniCreditBankCZ
Raiffeisenbank	RaiffeisenbankCZ
GE Money Bank	gemoney.cz

Tabulka 1 Seznam vybraných profilů bank na Facebooku, které jsou stahovány (Zdroj: autor)

4. Aplikace použité pro stažení a přípravu dat

V této kapitole je uveden popis naprogramovaných či převzatých programů, které sloužily pro stažení a obohacení dat. Stažení probíhalo ze dvou zdrojů, které byly zmíněny výše. Stahovaná data byla přímo ukládána do Elasticsearch a pak dále obohacena pomocí dvou dalších programů, které připojily analýzu sentimentu příspěvků a vyhodnocení jmenných entit v příspěvcích obsažených. Všechny programy byly napsány v jazyku Java.

Funkčnost a návrh aplikací zde nebude detailně představen, protože není hlavním cílem vytvořit tyto programy. Tyto jsou jen prostředkem pro získání dat, která jsou následně hodnocena. Proto také k programům nebude přiložena žádná dokumentace, jen komentáře ve zdrojovém kódu.

4.1. Stažení dat z webových stránek a z Facebooku

Následující podkapitoly popisují programy - konektory - použité pro samotné stažení dat a uložení do Elasticsearch. V obou případech byly použity volně dostupné bezplatné opensource knihovny napsané v jazyku Java.

4.1.1. Webcrawler – stažení dat z webových diskusí peníze.cz a měšec.cz

Jako základ aplikace pro procházení kýžených stránek slouží knihovna crawler4j. Samotné procházení webových stránek je jen rozšířením příkladu použití, který je dostupný na webu knihovny. (24) Rozšíření se týkalo hlavně nastavení kritérií pro procházení obou webových stránek. Bylo nutné nastavit doménu, na které se má crawler pohybovat a dále výčet textových řetězců v URL, které stránky nemají obsahovat, aby bylo procházení efektivnější. Tímto crawler dostal informaci, které části webu nenavštěvovat, protože neobsahují žádné uživatelské komentáře.

Pro obě webové stránky byl vytvořena Java třída s názvy BasicCrawlerMesec a BasicCrawlerPenize, které obsahují výše zmíněná pravidla procházení. Při procházení stránek prohledávají jejich obsah, a pokud narazí na specifickou část HTML kódu – stránka obsahuje komentáře a má tedy HTML strukturu se specifickými znaky jako je div s konkrétními atributy, dojde k rozparsování („zpracování“) této stránky. HTML kód se rozdělí na podsložky, které odpovídají HTML kódu jednoho příspěvku. Tyto části jsou zpracovány v cyklu, kde je v nich vyhledán požadovaný obsah. Tím jsou v tomto případě položky jméno autora komentáře, vlastní komentář, datum uložení komentáře, případně počet hodnocení tohoto komentáře dalšími uživateli. Každý dokument (v tomto případě

souhrn dat jednoho komentáře zmíněný výše) je v Elasticsearch uložen pod jedinečným ID. Protože na webových stránkách nebylo ID příspěvků dostupné, slouží pro tento účel uměle generovaný hash složený z data přidání příspěvku, textu příspěvku a jména autora. Toto ID umožňuje spouštění crawleru stále dokola třeba každý den, protože Elasticsearch pod jedním ID uloží jeden dokument, tedy jeden komentář. Při uložení dokumentu se stejným ID dojde k jeho přepsání. Každý dokument je v rámci cyklu procházení komentářů na jedné stránce ihned ukládán ve formátu JSON do Elasticsearch. To je popsáno níže.

4.1.2. Facebook

Základem připojení k datům na Facebooku je Java knihovna restFB. (25) Ta poskytuje třídy pro připojení k Facebooku více metodami, které byly už popsány výše. Pro přihlášení byly použity přihlašovací údaje k aplikaci, kterou jsem vytvořil pod svým soukromým facebookovým profilem (App Id a App Secret). Tyto dva údaje jsou aktivní neomezeně dlouhou dobu.

Předmětem stahování z Facebooku jsou příspěvky na zdi (typ post), jejich komentáře a podkomentáře (typ comment) a data o facebookové stránce, ze které jsou data stahována. Pro získání jednotlivých typů dat obecně slouží v knihovně restFB metoda fetchConnection. Protože všechny objekty typu post a comment jsou na zdi jednoho profilu (v mém případě jedné z vybraných bank), stačí použít konstrukci:

```
facebookClient.fetchConnection(fbUser + "/feed", Post.class);
```

Tímto získáme pro každého uživatele (fbUser) všechny položky na jeho profilu v první vrstvě. Zanořené komentáře je nutné získávat pro každý prvek zvlášť pomocí dalšího dotazu, který je podobný předchozímu:

```
facebookClient.fetchConnection(elementId + "/comments", Comment.class);
```

Takto jsou extrahovány komentáře komentářů.

Pro každou položku typu post je také nutné zjistit počet liků, tedy kladných hodnocení objektu. To je zajištěno samostatnou funkcí. Facebook sice už v prvním uvedeném příkazu vrátí počet liků postu ale jen do maximálního počtu 25. Pro přesný počet liků je nutné zavolat nový dotaz.

```
facebookClient.fetchConnection(postID + "/likes", Post.Likes.class);
```

Schéma průběhu získávání dat:

1. Získej položku na zdi
   1.1. Zjisti počet liků a ulož položku
   1.2. Zjisti, zda má komentáře a ulož je
      1.2.1. Zjisti, zda má komentář komentáře a ulož je

Data se ukládají opět ve formátu JSON, který v případě Facebooku obsahuje položky:

- Typ (post, comment)
- Id (FB má jedinečné ID, které je převzaté i pro potřeby uložení v Elasticsearch)
- Zpráva (vlastní komentář nebo text v postu na zdi)
- Uživatel, který položku vytvořil
- Stránka, na které je položka (název FB profilu banky)
- Počet liků
- Úroveň zanoření (post – 1, komentář – 0, podkomentář - -1)
- Id rodičovského objektu (jen pro komentáře)

Data získaná popsaným způsobem jsou předávána třídě ESConnect, která obstará jejich uložení. Následuje popis této třídy.

4.1.3. Společný kód uložení dat do Elasticsearch

Ukládání do Elasticsearch bylo v obou aplikacích realizováno podobným způsobem. Obě aplikace obsahují třídu ESConnect. Ta má konstruktor, který vytvoří připojení do Elasticsearch na základě IP adresy a portu, kde je spuštěn. Připojení je pak využíváno po celý běh programu (většinou trvá i několik hodin). Před prvním spuštěním těchto programů Elasticsearch neobsahuje žádná data a metadata, popisující dokumenty, které se budou vkládat. Pro přípravu prostředí – tedy definici metadat – třída obsahuje metody, které umožní vložit mapping polí a základní analyzéry (popis mappingu a analyzérů je v následující kapitole). Zavoláním této metody a předáním patřičných údajů dojde k přípravě prostředí v Elasticsearch. Toto je nutné provést ještě před ukládáním dat, protože následně lze provádět jen některé úpravy mapování. Alternativně lze tato metadata ukládat pomocí curl[1] požadavků. Naprogramování této funkcionality ale zjednodušilo práci

[1] Curl je nástroj příkazové řádky a knihovna pro posílání dat pomocí syntaxe URL (Zdroj: http://curl.haxx.se/)

a ladění, protože bylo možné jednoduše smazat celý index (soubor uložených dat a metadat) a znovu spustit stahování dat, které zajistí i nastavení metadat.

Třída ESconnect poskytuje metodu pro uložení dat do Elasticsearch. Data se předávají ve formátu JSON, které je vstupním parametrem do této metody. Pro uložení se použije připojení z konstruktoru a data jsou uložena do definovaného indexu, typu pod jedinečným ID.

4.2. Obohacení dat

Protože jsou použity dva datové zdroje a k nim dva konektory, zdá se jako vhodnější použít pro obohacení dat programy, které nejsou přímo vloženy programů pro stahování popsaných výše. I když je tato varianta možná, zvolil jsem kvůli pružnosti a lepší organizaci pro obohacení dat spíše vytvoření externího programu, který získá data z Elasticsearch a zpět vrací další pole, která se uloží do původního dokumentu v ES.

4.2.1. Analýza sentimentu

Sentiment analýza je provedena pomocí knihovny OpenNLP. Jde opět o volně dostupnou opensource knihovnu, která slouží pro programování nejrůznějších úloh zpracování přirozeného jazyka. Disponuje širokou škálou použití – například detekce vět, tokenizace, kategorizace dokumentu apod. (26) Vyhodnocení sentimentu příspěvků je provedena pomocí kategorizačního modelu (OpenNLP Document Categorizer), který je založen na principu maximální entropie. Nejdříve je nutné kategorizační model natrénovat. K tomu byla použita data, která vznikla na Západočeské univerzitě jako výstup analýzy sentimentu dat z českých sociálních medií za použití strojového učení s učitelem. (27) Přesněji jsou to data z českých facebookových stránek a filmové recenze z webu CSFD. Tato data byla použita, protože jsou velmi podobná datům, která jsou analyzována v této práci. Nebylo tedy nutné vytvářet vlastní ohodnocený soubor dat pro učení kategorizačního modelu.

Program obohacující data v Elasticsearch o hodnocení sentimentu se skládá ze tří částí:

- Příprava kategorizačního modelu a jeho trénink
- Vyhodnocení sentimentu (kategorizace) podle vytvořeného modelu
- Připojení do Elasticsearch pro získání a zpětné uložení dat

Kategorizační model vznikne načtením trénovacích dat a jejich vyhodnocováním. Vytvořený model se následně uloží do souboru na disk pro potřeby kategorizace. Připojení do Elasticsearch probíhá podobně jako v případě konektorů. Rozdíl je v tom, že je nejdříve

nutné data, která chceme kategorizovat, z Elasticsearch získat. K tomu je použita funkce ScrollSearch, která dávkově stahuje data podle zadaných kritérií. V mém případě je definován index, kde jsou data uložena a název pole, který je potřeba stáhnout (pole message). Stahování probíhá v cyklech. Uvnitř cyklu se provádí hodnocení sentimentu pomocí připraveného kategorizačního modelu. Do Elasticsearch jsou pak zpět pro stejné ID dokumentu vkládána data název kategorie (positive, negative, neutral), váha kategorie, což je hodnota, s jakou pravděpodobností je sentiment kategorizován, a vyjádření kategorie pomocí čísla (-1, 0, 1). Tento cyklus proběhne vždy pro všechna data uložená v daném indexu. Jednoduchou úpravou lze nastavit tak, aby kategorizoval jen ještě nekategorizované dokumenty.

Úspěšnost modelu byla hodnocena na vzorku 100 náhodně vybraných příspěvků a pohybuje se okolo 65%. Použití sentiment analýzy slouží pro demonstraci relativně jednoduché možnosti, jak sentiment analýzu zapojit a obohatit jí data. Úspěšnost modelu lze v zvýšit například vytvořením vlastního datového korpusu s klasifikovanými texty, nad nimiž bude vytrénován model. Případně lze zapojit jiný systém pro hodnocení sentimentu nebo kombinovat více přístupů. Kvalitou sentiment analýzy v českém prostředí se zabývají jiné absolventské práce zmíněné v kapitole 2.5.

4.3. Závěr kapitoly

Byly popsány programy, které jsem vyvinul pro získání a obohacení dat. Jsou zde jen rámcově popsány, protože tato práce se nezabývá vývojem softwaru a úrovní návrhu. Programy jen umožní další postup práce a umožní uložení dat. Další kapitoly se věnují popisu Elasticsearch a analýzou uložených dat.

5. Elasticsearch

Jako uložiště stažených dat a analytický nastroj, byl vybrán software Elasticsearch, který vychází z knihovny Apache Lucene. Tento nástroj jsem použil, protože je volně dostupný a má vlastnosti, které jsou pro tento typ analýzy potřeba, například díky typu uložení dat (bezschematově), schopnosti analyzovat text pomocí například stemmingu, dostupnosti vizualizačního nástroje atd. Elasticsearch je distribuovaný škálovatelný systém pro fulltextové vyhledávání v reálném čase a analytický nástroj. Podporuje strukturované vyhledávání, geolokaci a zaznamenání vztahů mezi daty.

Společně s aplikacemi Logstash a Kibana tvoří takzvaný ELK stack. Souhrnně se používají například pro analýzu logů a další analýzy podle zdroje dat. V ELK stack Elasticsearch slouží pro uložení dat a Logstash slouží jako konektor do různých datových zdrojů. Je možné získávat data například ze souborových systémů, různých logovacích systémů, emailových schránek pomocí IMAP protokolu, z Twitteru, nebo z Elasticsearch. V Logstash lze data dále upravovat a ukládat na další systémy například do Elasticsearch ale i mnoha dalších. (28) Kibana se používá pro zobrazení dat. Její popis je v podkapitole níže.

Data jsou v Elasticsearch uložena v bezschématové databázi. Ukládají se v podobě, v jaké jsou vkládána ve formátu JSON. Komunikace se systémem probíhá pomocí REST API. (29) Data se po předzpracování ukládají do invertovaného indexu. Indexace obsahu Facebooku a webových diskusí je před uložením potřeba nastavit. Toto nastavení je popsáno v následujících odstavcích.

5.1. Zpracování dat

Surová data stažená z webu a z Facebooku je vhodné podrobit dalším úpravám, než se budou posílat k uložení do databáze Elasticsearch. Lze využít několik vrstev předzpracování – ihned při stáhnutí dat, nebo stáhnutá data posílat dále do další na stahování nezávislé komponenty, která bude podle definovaných pravidel data čistit a upravovat. Toto předzpracování může zahrnovat obohacení stažených dat (přidání dalších polí), vyčištění dat (odstranění nechtěných dat, odstranění HTML tagů, převod formátů data, nahrazení prázdných hodnot). Další zpracování už probíhá na straně Elasticsearch, kde lze nastavit, jak se ukládaný text bude dále chovat při vyhledávání. (29)

Data se ukládají do entity zvané index (anglický název index), ta může obsahovat více typů (_type). Každý typ pak může obsahovat jinou soustavu polí (field), což jsou už jednotlivé položky ukládaného dokumentu. Dokumentem se rozumí jedna ukládaná entita, například email (do něhož může patřit odesílatel, příjemce, předmět, datum, zpráva…).

Každé pole má svůj datový typ, které lze nastavit předem před uložením dat anebo se vytvoří automaticky při vkládání dat. Elasticsearch má několik základních datových typů:

- String – text
- Integer / long – celé číslo
- Float / double – desetinné číslo
- Boolean – logický datový typ
- Null

Každé pole pak může obsahovat další vlastnosti jako defaultní hodnotu, zda bude analyzováno (bude pro něj platit nastavení popsané níže) a podobně. Pole může obsahovat také objekt či vnořené objekty. Lze tedy využít například vnoření potomka do dokumentu předka.

5.2. Analýza

Aby bylo možné indexovaná data snadněji prohledávat, tedy aby bylo možné například vyhledávat slova ve všech pádech a podobně je potřeba v Elasticsearch nastavit u každého indexu základní sestavu následujících parametrů.

Analyzér

Souhrnné označení pro soustavu úprav dat a obsažených polí je název Analyzér. Zahrnuje různé možnosti zpracování textu usnadňujících následnou analýzu a vyhledávání. Pro vytvoření invertovaného indexu v Elasticsearch je potřeba text rozdělit na jednotlivá slova a ta ještě upravit. (30)

- **Char filtr**

Tento filtr přidává nebo odebírá, případně nahrazuje znaky nebo řetězce znaků. Je vhodný například pro odstranění HTML tagů. Stále pracuje s celým textem.

- **Tokenizér**

Pro oddělení jednotlivých slov podle definovaných pravidel je zde tokenizér. Slova odděluje například pomocí mezer, čárek, teček, pomlček a podobně. Výstupem jsou již jednotlivá slova.

- **Token filtr**

Upravuje dále jednotlivá slova. Cílem je rozšířit možnosti při následném vyhledávání. Uloží se do indexu tedy původní slovo, ale na dokument se navážou i další varianty vzniklé při pracování tímto filtrem.

Varianty filtru:

- Nalezení kořene slova – například infinitivu, prvního pádu a podobně
- Nalezení synonym slova
- Vytvoření n-gramů
- Odstranění diakritiky

Stop slova

Pro další zlepšení vyhledávacích schopností Elasticsearch je vhodné definovat stop slova. Tato slova nenesou význam, a pokud jsou definována, jsou vyřazena z indexu. Takže nezabírají v indexu místo a při následném vyhledávání například pomocí věty „Až naprší a uschne“, se budou vyhledávat jen slova „naprší“ a „uschne“, protože slova „až“ a spojka „a“ jsou právě stop slova.

5.3. Popis použitého nastavení Elasticsearch

Elasticsearch poskytuje zabudovanou podporu analýzy pro Český jazyk. Tato nastavení jsem ve své práci použil. Jsou nastaveny tyto analyzéry a filtry (30):

- Lowercase

Analyzovaný text je normalizován na malá písmena.

- Czech_stop

Z indexu jsou vyřazena slova, která jsou obsahem stop filtru. Tato slova pak nejsou použita pro vyhledávání.

- Czech_keywords

Tento typ analyzéru je vhodný pro texty, které mají jako celek vystupovat jako jeden token. Jsou to například poštovní směrovací čísla, různá ID a podobně.

- Czech_stemmer

Stemming je přístup k nalezení základu slova, který algoritmicky vyhledává a vypouští předpony, přípony a podobně.

- Czech_syno

Tento token filtr umožnuje expanzi slova na synonyma. Při vyhledávání jsou nalezeny i pak synonymní výrazy k hledanému textu.

Následující Obrázek 3) ukazuje mé nastavení indexu banky.

Obrázek 3 Analyzéry a tokenizéry použité v mém nastavení Elasticsearch, výstup je z aplikace Sense, která instalována jako doplněk do prohlížeče Chrome (Zdroj: autor)

5.3.1. Systém uložení dat

Data jsou uložená ve třech indexech. Toto rozložení jsem zvolil kvůli oddělení dat, podle jejich významu.

Banky

Slouží pro uložení dat z facebookových zdí a webových diskusí.

Bankyusers

Ukládá detailnější informace o uživatelích, kteří jsou autory příspěvků na Facebooku.

Bankypages

Ukládá detail dat o stránkách stahovaných bank, zejména počet liků stránky, popis atd.

Index banky dále dělí data podle toho, kde byla stažena. Pro webové diskuse slouží typ discussion. Facebooková data jsou rozdělena do dvou indexů. Jeden je pro objekty typu post (typ post), a druhý pro komentáře (typ comment). Toto rozložení bylo zvoleno, protože ve všech typech mohou být trochu odlišná pole. Uložení do více typů pak také umožňuje jednoduché filtrování dat. Jeden záznam uložený do indexu je označován jako dokument, ten zahrnuje pole podle mapování anebo podle skutečně uložených dat.

Mapování polí

V indexu Banky jsou uložena následující pole (pro všechny tři typy):

Název pole v Elasticsearch	Význam pole
id	Id objektu
idRaw	Neanalyzované pole s hodnotou Id objektu
message	Zpráva – komentář nebo příspěvek na zdi
messageRaw	Zpráva bez použití analyzérů
userId	Id uživatele
userName	Jméno původce objektu
created	Datum vytvoření příspěvku
postId	Id příspěvku na zdi (i u komentářů)
likes	Počet liků
page	Stránka banky, ze které je objekt stažen
level	Úroveň zanoření objektu (1,0,-1)
source	Zdroj dat (Facebook nebo web)
sentiment	Kategorie sentimentu slovy
sentimentN	Kategorie sentimentu čísle (1,0,-1)
sentimentWeight	Váha sentimentu

Tabulka 2 Seznam polí uložených v indexu Banky (Zdroj: autor)

V rámci indexu Banky obsahuje typ discussion navíc ještě pole url, kde je url stránky na které byl příspěvek na webových stránkách měšec.cz nebo peníze.cz nalezen. V indexu BankyPages, který obsahuje informace i Facebookové stránce, jsou následující sloupce:

Název pole v Elasticsearch	Význam pole
id	Id stránky
about	Úvodní popis o stránce
category	Facebooková kategorie stránky
description	Popis stránky
link	Odkaz na stránku
created	Datum vytvoření příspěvku
likes	Počet liků
page	Stránka banky, ze které je objekt stažen

Tabulka 3 Seznam polí uložených v indexu BankyPages (Zdroj: autor)

Poslední index je BankyUsers, který sdružuje veřejně dostupné informace o uživatelích Facebooku, kteří alespoň jednou přispěli ať už postem nebo komentářem na stránku některé sledované banky. Tento index není v analýzách používán a mapování polí používá defaultní datové typy.

Název pole v Elasticsearch	Význam pole
id	Id uživatel
about	O uživateli
gender	Pohlaví
hometown	Město
birthday	Datum narození
userName	Uživatelovo jméno
page	Stránka, na které přispíval
source	Zdroj
locale	Národní prostředí

Tabulka 4 Seznam polí uložených v indexu BankyUsers (Zdroj: autor)

Datové typy jednotlivých polí odpovídají uloženým datům. Id je typu string protože může obsahovat i nenumerické znaky a v případě Facebooku také většinou obsahuje.

V základním nastavení Elasticsearch ukládá data do pěti shardů. Toto nastavení jsem neměnil. Shard je označení pro jednu instanci Lucene. Je to logické uložiště dat. Pokud je Elasticsearch provozován jako cluster, tedy je rozložen na více počítačů nebo serverů, jsou shardy rozloženy mezi jednotlivé počítače (nody), aby byl optimalizován výpočetní výkon. Shard může mít repliku, což je kopie pro zálohu dat, která je v případě provozu v clusteru uložena na jiném nodu, aby při selhání počítače bylo možno obnovit data z jiného. (30)

5.4. Rozšiřující pluginy a aplikace

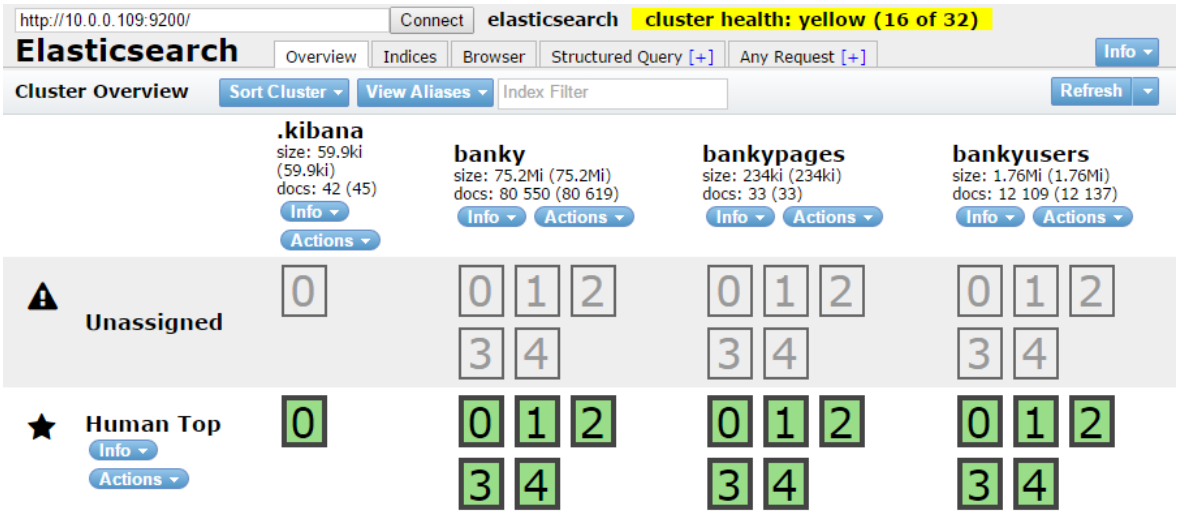
K Elasticsearch jsou použity další pluginy a aplikace. Plugin Head pro zjednodušení práce s indexy (datovým souborem), aplikace Kibana pro zobrazení dat a aplikace carrot2 pro clusterizaci dokumentů.

5.4.1. Head

Plugin Head je webová aplikace, která slouží ke kontrole stavu clusteru Elasticsearch. Zobrazuje topologii uzlů (rozmístění indexů na uzlech clusteru) a jejich aktuální status.

Pomocí Head lze přímo prohlížet uložená data v jednoduchém tabulkovém zobrazení a data filtrovat.

Další funkce je možnost vyhledávání v indexech pomocí JSON dotazů, které lze vytvářet přímo ve formuláři v prohlížeči. Formulář ještě před odesláním To je užitečné pro ladění složitějších dotazů, protože není nutné dotaz editovat a posílat například z příkazové řádky. (31)



Obrázek 4 Obrazovka pluginu Head se třemi indexy se staženými daty a s indexem pro uložení dat aplikace Kibana (Zdroj: autor)

Na obrázku jsou vidět názvy tří indexů, které jsou popsány výše a také pro každý index pět shardů (zelené obdélníky). Každý shard má jednu neaktivní repliku, protože repliku nejde umístit na jiný počítač - moje instalace Elasticsearch je spuštěna jen na jednom PC.

Každý node při spuštění Elasticsearch dostane vlastní název. Lze ho definovat pevně v konfiguraci Elasticsearch, nebo je přiřazen automaticky náhodný název. Na obrázku je aktuální node pojmenován názvem Human Top.

V horní části obrázku jsou záložky, kde je možné využít další funkcionality pluginu.

Přehled indexů – indices, zobrazí jen počet dokumentů uložených v každém indexu a velikost úložiště disku.

Browser umožnuje zobrazovat a filtrovat data. Spíše než pro detailní analýzu, složí pro jednoduchou kontrolu, jaká data index obsahuje.

Pro detailnější a komplikovanější dotazy je možné použít části Structured query a Any request. Obě poskytují možnost skládat složitější dotazy. Structured query uživateli napomáhá s konstrukcí dotazu pomocí výběrů možností z comboboxů. V části Any request

je nutné napsat dotaz pomocí curl požadavku a případně dodat parametry dotazu ve formátu JSON.

5.4.2. Kibana

Vizualizační nástroj Kibana se připojuje do Elasticsearch. Uložená data lze různými způsoby zobrazovat, prohledávat a filtrovat. Vlastností Kibany jsou určitá omezení, která vychází z určení tohoto nástroje. Nejde o klasický vizualizační nástroj Business Intelligence. Omezením je možnost připojení dat jen z Elasticsearch. Možnosti dotazování jsou také omezené jen na query, která poskytuje Elasticsearch, i když je možné vytvářet skriptovaná pole (Scripted Field) nad poli, která obsahuje pro analýzu připojený index v Elasticsearch. Skriptovaná pole mohou vzniknout jen z numerických polí a práce s nimi je omezena jen na několik operací. Tento typ virtuálních polí není v analýze použit.

Klasické Business Intelligence nástroje většinou umožňují oddělení tvorby reportu (v tomto případě dashboardu) a prohlížení pomocí uživatelských oprávnění, která oddělují analytiky, kteří report sestrojí, a uživatele, kteří mají omezenou možnost s reportem pracovat. Uživatelé většinou mají možnost použít filtry a vyhledávání. Nemají ale možnost vytvářet nová zobrazení a upravovat analytiky předdefinovaná zobrazení. Oddělení rolí v Kibaně v základu neexistuje. Každý uživatel může vytvářet a definovat všechny vizualizační prvky a sestavovat si dashboardy podle potřeby. V případě Kibany mohou nezkušení uživatelé upravit vlastnosti zobrazení a dashboardu, která mohou vést ke znehodnocení zobrazení. Data zůstanou nedotčena, ale zobrazení nebude funkční, nebo bude zobrazovat nechtěné nebo nesmyslné výstupy.

Aplikace je určena pro zkušenější uživatele i proto, že nezobrazuje informace hned na první pohled. Je nutné s nástrojem pracovat – pokládat dotazy, filtrovat data a podobně. Síla Kibany je v možnosti rychle měnit pohled na data. Díky typu uložení dat v Elasticsearch (tedy Lucene) jsou požadovaná data velmi rychle vyhledána a zobrazena. To umožňuje uživatelů rychlou analýzu a získání odpovědí na jejich otázky.

Prvním krokem pro zobrazení požadovaných dat je výběr indexu a označení pole z indexu, které obsahuje datum (v mém případě datum vytvoření příspěvku). Tím Kibana dostane informaci, která data má zobrazovat a jak je časově členit. Základní nastavení je zobrazení dat za posledních patnáct minut. Pokud nejsou tak čerstvá data k dispozici, je zobrazeno varování, že nejsou žádná dostupná data.

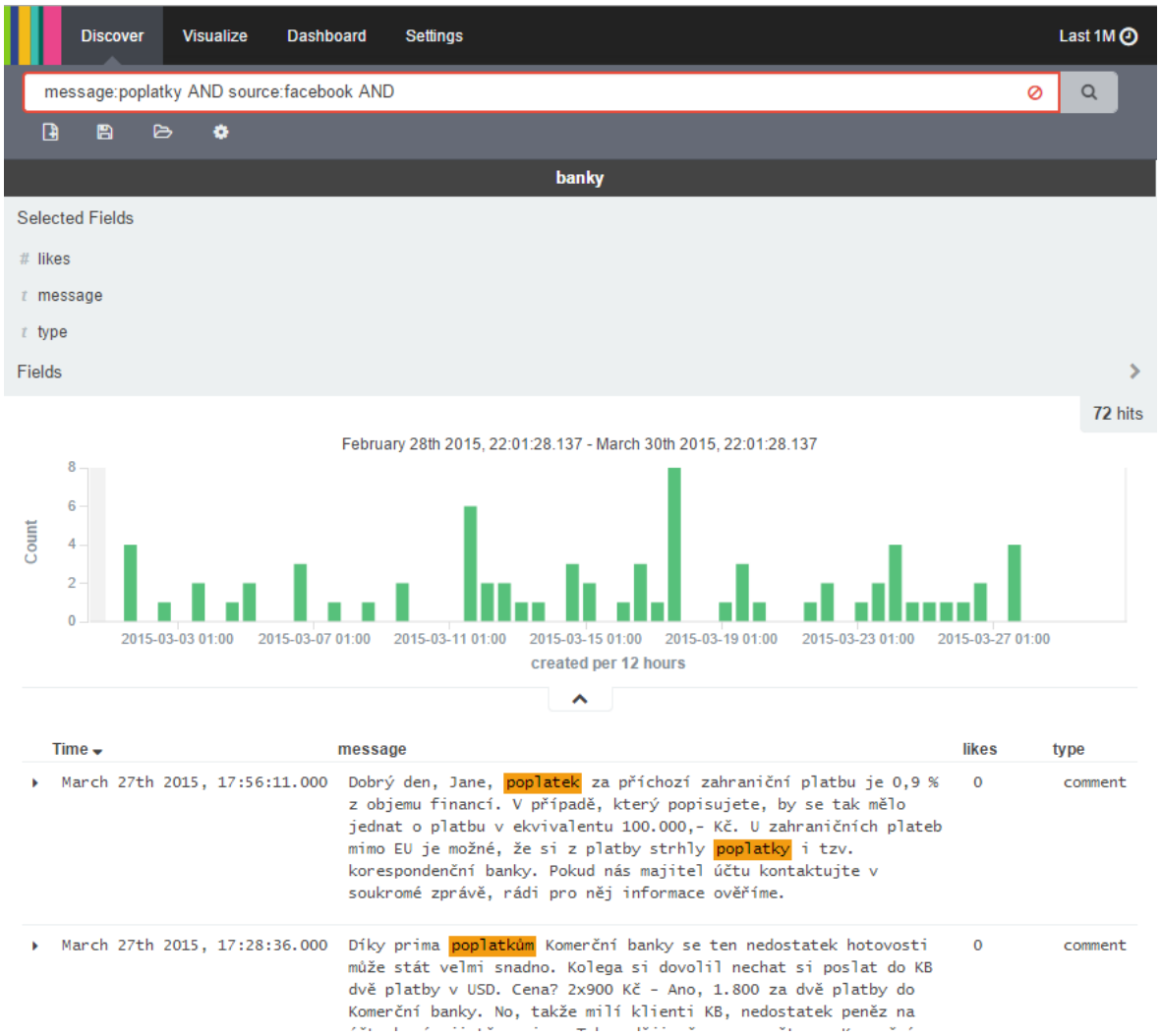
Dále analýza dat probíhá ve třech částech aplikace, kde jejich postupným průchodem lze vytvořit ucelený pohled na zkoumaná data v poslední části Dashboard.

3 základní části jsou:

- Discover
- Visualize
- Dashboard

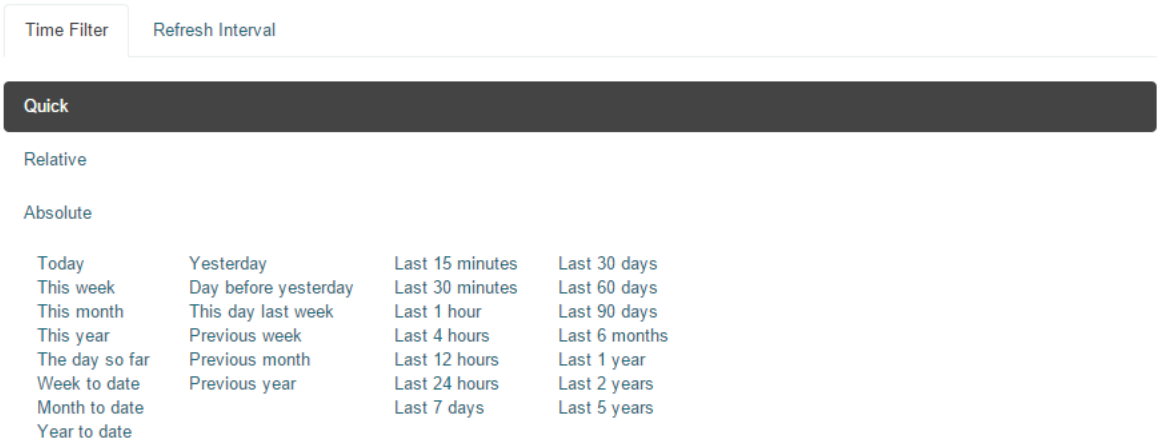
Discover

umožňuje prohlížení a filtrování dat (Obrázek 5). Zobrazuje podle vybraného indexu z Elastichsearch seznam obsažených polí a jejich základní charakteristiku (počet nejčastějších hodnot, typ). Tato pole lze skládat do zobrazení s názvem search. Výsledkem je tabulka s vybranými poli a výpis jejich hodnot. Zároveň je zde v histogramu zobrazen časový vývoj počtu záznamů. Data lze filtrovat podle mnoha kritérií podobně, jako lze skládat query do Elasticsearch. Tuto tabulku i s použitými filtry lze uložit a použít v dalších částech aplikace. Na obrázku 5 je použit vyhledávací výraz, který z dat vybere záznamy obsahující slovo poplatky v různých tvarech (to je vidět na zvýrazněných polích) a zároveň je zdrojem záznamů Facebook. V Pravém horním rohu obrázku je pak časový filtr nastavený na zobrazení dat za poslední měsíc klouzavě.

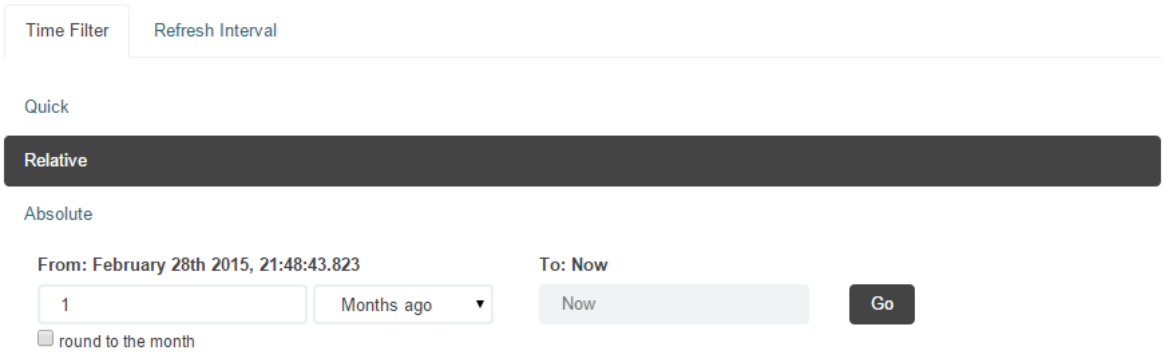


Obrázek 5 Ukázka části Discover aplikace Kibana s použitým vyhledávacím výrazem, zobrazením polí message, likes a type, histogramem a všemi nalezenými hodnotami. (Zdroj: autor)

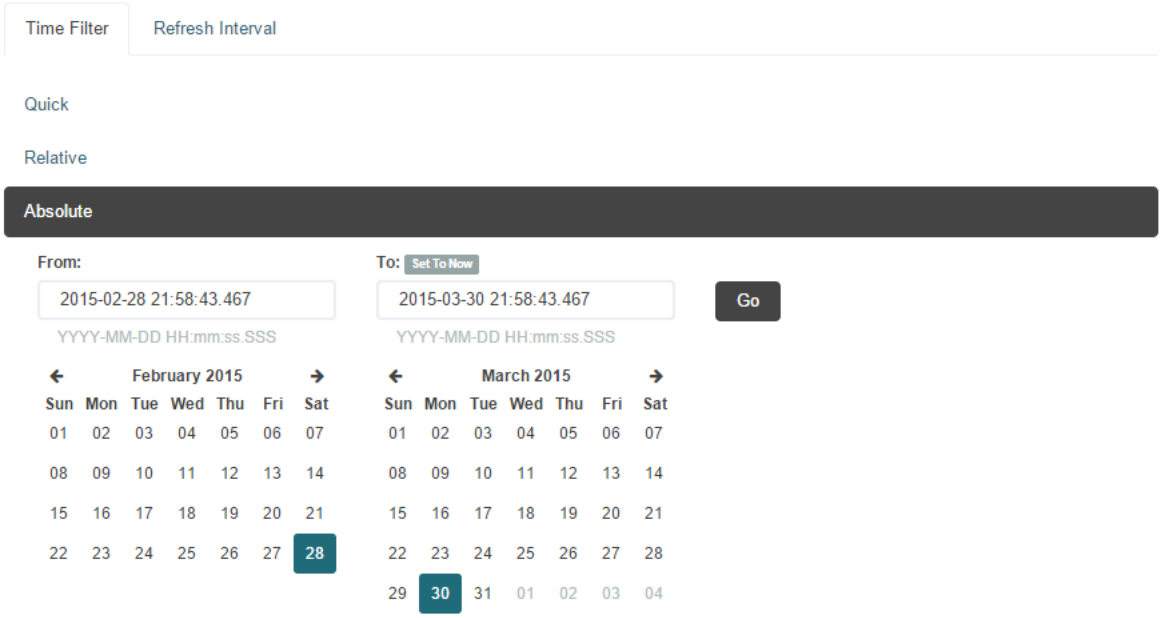
Časové filtrování dat je dostupné ve všech třech částech. Lze filtrovat několika způsoby – předdefinovanými filtry anebo uživatelem přesně nastavenými hodnotami. Tyto typy jsou v následujících obrázcích (Obrázek 6, Obrázek 7, Obrázek 8). Pokud jsou data stahována průběžně, lze nastavit interval obnovení dat, díky kterému se budou všechny objekty ve všech částech Kibany obnovovat a načítat nejnovější data.



Obrázek 6 Předdefinované časové filtry v aplikaci Kibana (Zdroj: autor)



Obrázek 7 Uživatelský relativní časový filtr v aplikaci Kibana (Zdroj: autor)



Obrázek 8 Uživatelský absolutní časový filtr v aplikaci Kibana (Zdroj: autor)

Visualize

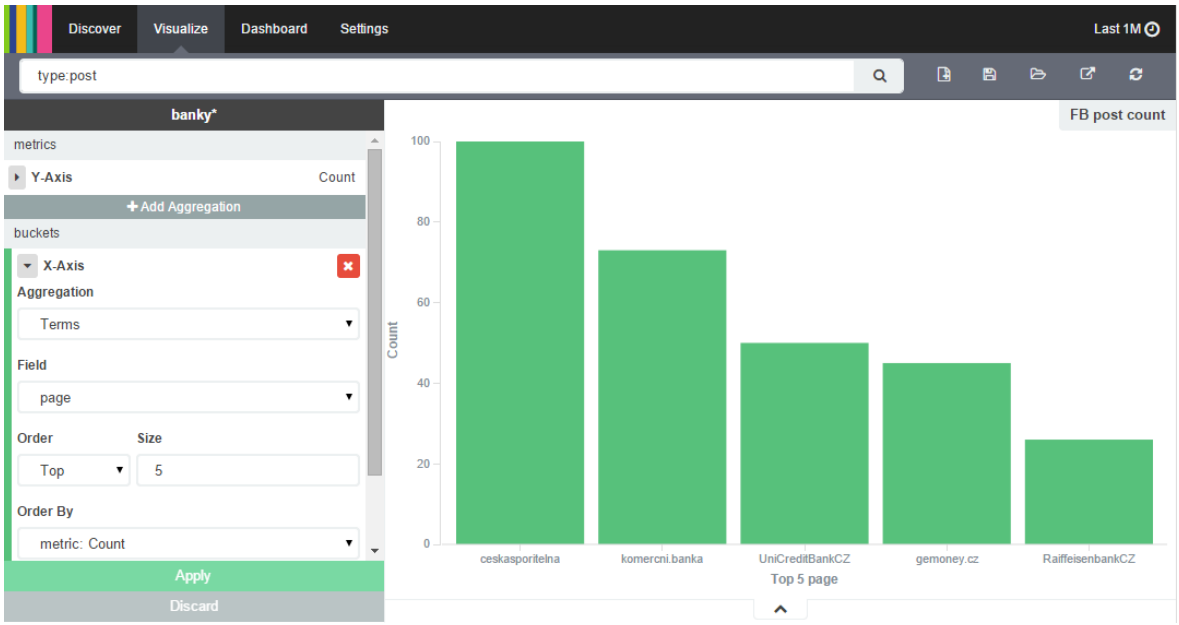
Část Visualize (Obrázek 9) umožňuje vytvářet grafické objekty – různé typy zobrazení dat a agregací. Všechny typy vizualizací lze vytvořit na základě uloženého vyhledávání z předchozího kroku v části Discover anebo vytvořit nové vyhledávání. To může být výhodou, pokud je potřeba vytvořit více typů vizualizací při stejném datovém základu – tedy za použití stejného dotazu vyhledávání. Je také možné upravit už uloženou vizualizaci.

Uložené vizualizace jsou dostupné v části Dashboard.

Kibana poskytuje následující typy vizualizací:

- Area chart - plošný graf
- Datatable – datová tabulka s agregacemi
- Line chart – spojnicový graf
- Markdown widget – komentářový prvek
- Metric – metrika (jedno číslo s různými možnostmi agregace)
- Pie chart – koláčový graf
- Tile map – mapa
- Vertical bar chart – sloupcový graf

Všechny vizualizace musí mít definovanou metriku, kterou budou zobrazovat na základě jedné ze základních možností agregace (počet, maximum, minimum, průměr apod.). Dále se nastavuje takzvaný bucket, což je členění metriky podle různých dalších agregací nebo kategorií – v terminologii převzaté z Business Intelligence by se dalo říci přidání dimenze. Následující obrázek ukazuje uloženou vizualizaci – sloupcový graf, který má na ose y vynesen počet dokumentů a na ose x bucket s agregací podle termu, což je indexovaná hodnota z konkrétního pole. V tomto případě z pole page, které obsahuje název stránky z webu nebo profilu banky. Ve vyhledávacím poli je zároveň omezena datová množina na dokumenty, které jsou typu „post“. Proto jsou vyfiltrované jen stránky z Facebooku a nejsou vidět názvy webových stránek. Výstup je omezen na pět nejčastějších hodnot. Data jsou zobrazena za poslední měsíc, to lze měnit otevřením časového filtru.



Obrázek 9 Ukázka vizualizace v aplikaci Kibana – sloupcový graf zobrazující názvy Facebookových stránek s nejvyšším počtem příspěvků typu post. (Zdroj: autor)

Podobným způsobem se vytvářejí všechny ostatní vizualizace. Specifické jsou jen Markdown widget a Tile map. Markdown widget lze jen naplnit textem, který může sloužit například pro informování uživatele Dashboardu o jeho vlastnostech. Tile map zobrazuje mapu, která je použitelná pokud index obsahuje geolokační data.

Dashboard

Poslední částí aplikace Kibana je Dashboard. Při prvním zobrazení je vidět jen prázdná plocha. Tu je možné vyplnit uloženými vizualizacemi a vyhledáváními podle potřeby uživatele. Oba prvky lze vložit na pracovní plochu dashboardu a pomocí drag and drop přeskupovat, případně upravit velikost jednotlivých vložených objektů tak, aby bylo dosaženo přehledného zobrazení dat.

Pohledů na data, tedy dashboardů, lze vytvořit mnoho. Ukázka dashboardu pro analýzu dat, která je předmětem této práce, je kapitole ÚvodTabulka 6 Seznam názvů produktů bank (Zdroj: autor)

Příprava vyhledávacích výrazů

Výrazy obsahující název banky a názvy jejich produktů jsou spojeny pro každou banku do jednoho vyhledávacího řetězce spojené operátorem OR. Při vyhledávání stačí, aby byl v textu zprávy obsažen jen jeden vytipovaný výraz, aby byl příspěvek nalezen a zobrazen nebo započítán do výstupu dashboardů. Takto připravené vyhledávací výrazy jsou volně kombinovatelné, takže lze zjistit i průniky mezi příspěvky a více vyhledávacími výrazy.

Jednotlivé výrazy lze zadat do vyhledávacího pole i samostatně ale pro využití vyhledávacích schopností Elasticsearch je potřeba klíčová slova uzavřít do složitějšího

výrazu. Klíčová slova jsou hledána jen v poli message, proto je potřeba vyhledávání začít názvem tohoto pole. Pokud tedy budeme chtít vyhledat příspěvky, které obsahují názvy produktů České spořitelny, bude vyhledávací výraz následující:

m*essage:(„Osobní účet ČS II“ „SERVIS 24“ „Kreditní karta Odměna“ iBod)*

Takto budou vyhledány příspěvky, které obsahují alespoň jeden z názvů. Lze také změnit logické OR na AND – pokud požadujeme, aby vyhledané příspěvky povinně obsahovaly více vyhledávaných slov, stačí přidat + (znak plus) před vyhledávaný termín. Víceslovné názvy je potřeba uvádět v uvozovkách.

Návrh dashboard.

5.4.3. Carrot2

Carrot2 je knihovna a sada aplikací, které slouží pro vytváření clustrů (clusters)[2] nad vyhledanými daty. Knihovna Carrot2 není vyhledávač, ale program, který data převede do výsledků podle společných témat tedy clusterů. Jsou k dispozici dva algoritmy pro clusterizaci – Lingo a Suffix tree clustering. Podporuje připojení různých datových zdrojů (Bing, Google, Lucene, Solr) a také API v různých jazycích. Lucene je pro možnost připojení dat z Elasticsearch důležitý, protože jak bylo řečeno, data jsou v ES uložena pomocí Lucene. (32)

Pro použití s Elasticsearch je dostupný jako plugin, který nainstaluje knihovnu a obslužné javascriptové soubory. Zpracování dat probíhá v prohlížeči. Funkcionalita je bohužel oproti plné instalaci v podobě programu (workbench, nadstavba na IDE Eclipse) omezena a je nutné ručně upravovat dotazy ve zdrojovém kódu webové stránky.

Stránka pluginu, kde jsou zároveň zobrazovány výsledky, slouží jako demo aplikace. Po instalaci pluginu jsou k dispozici tři kroky, které předvedou jeho funkcionalitu. Prvním krokem je uložení vzorku dat do Elasticsearch, zadání dotazu, podle kterého budou data clusterizována a výběr algoritmů pro vytvoření clusterů a nakonec spuštění algoritmu, který předvede na demo datech výstup.

Pro použití v této práci byl upraven zdrojový kód stránky – především dotaz do Elasticsearch, který je zdrojem pro data, která jsou cluterizována. Upravená funkcionalita vybírá data z indexu banky, nastavuje časový filtr od 6. 4. 2015 do 13. 4. 2015 a omezuje počet možných výsledků na 10 000[3]. Je nastaveno Elasticsearch query (dotaz do ES) na pole message (zpráva), které přebírá hodnoty ze vstupního pole webové stránky. Příklad výstupu je zobrazen na Obrázek 10.

Carrot2 je použit pro nalezení témat, která souvisí s jednotlivými klíčovými pojmy, definovanými v kapitole 6.2. Kvůli úpravám v nastavení pluginu je nutné vždy zadat vyhledávací výraz. To je z toho důvodu, že byl upraven způsob dotazování do Elasticsearch tak, aby na vyhledávaný výraz byly použity analyzéry a tak byly vyhledány

[2] Clusterizace je algoritmus pro automatickou klasifikaci vzorů v datech. Výsledkem jsou skupiny vstupních dat, která mají něco společného. (33)

[3] Tento počet je vzhledem k počtu dat, která jsou za jeden měsíc uložena naprosto dostačující. Zároveň se počítá s tím, že pro reálné nasazení je nutné systém upravit, jak popsáno na v závěru práce.

všechny formy slova. Chceme-li vytvořit clustery bez zadaného dotazu, lze zadat hvězdu („*“) Pokud tedy zadáme slovo poplatek, budou vyhledány dokumenty, které obsahují toto slovo ve všech pádech.

Obrázek 10 Příklad výstupu z pluginu carrot2 při použití vyhledávacího výrazu „poplatek“ (Zdroj: autor)

6. Metriky a další charakteristiky popisující obraz bank v komentářích na webu a Facebooku

Ze stažených a indexovaných dat je nutné vyvodit metriky a další charakteristiky, které ohodnotí to, jak se která banka jeví v očích komentujících. Tyto charakteristiky, pokud jsou kvantitativního typu, jsou zobrazeny v níže navržených metrikách. Pro lepší pochopení obsahu příspěvků a komentářů jsou navrženy další postupy, které zobrazí témata, o kterých se v příspěvcích nejvíce hovoří.

6.1. Kvantitativní charakteristiky

6.1.1. Metriky z Facebooku

Na Facebooku lze jednoduše zjistit několik základních charakteristik, které ukazují, jak je profil banky populární. Níže navrhnuté metriky jsou základem pro takovou analýzu.

Počet liků profilu banky

Kromě stažení komentářů a jejich kvantitativních charakteristik lze zjistit u každé banky, která má profil na Facebooku, také počet lidí, kteří tuto stánku likují. Mají jí tedy jako oblíbenou stánku a její příspěvky se těmto lidem zobrazují na zdi (pokud nenastaví jinak). Tato metrika zjednodušeně zobrazuje oblíbenost banky na Facebooku. Čím více liků, tím lépe. Tím větší počet uživatelů, kteří si přečtou sdělení banky.

Počet příspěvků (typ post) na zdi banky

Tato metrika ukazuje aktivitu banky a jejích fanoušků. Říká, kolik objektů typu post je na její zdi. Tyto objekty jsou vytvářeny bankou ale i uživateli. Metrika zobrazuje, jak aktivní je banka a další uživatelé.

Počet komentářů (typ comment) na zdi banky

Celkový počet komentářů, které vytvořila banka anebo další uživatelé. Metrika ukazuje kolik komentářů je celkem vytvořeno na dané stránce banky.

6.1.2. Metriky na webu

U komentářů pod články sledovaných webových stánek je oproti Facebooku situace komplikovanější. Nelze jednoduše říci, že se některý článek týká jen jedné banky. Neexistuje počet liků a počet komentářů, který by se dal připočíst jen k jedné bance, respektive k facebookovému profilu banky.

Poměr počtu komentářů obsahujících název banky nebo produktu vůči všem komentářů

Tato metrika hodnotí, jak významné je sledovat danou webovou stránku a její diskuse. Předpokládá se zapojení více webů a pomocí této metriky lze nalézt web, který přispívá k dalším analýzám nejvíce. Naopak pokud některý web obsahuje v porovnání s dalšími velmi nízký poměr, mohl by být z fronty stahování příspěvků odebrán. Tento přístup je navržen pro další rozšíření této práce.

6.1.3. Metriky a charakteristiky společné pro oba zdroje

Počet komentářů obsahujících název banky

Protože nelze jednoduše jednotlivé komentáře na webových diskusích připočíst konkrétně k jedné bance bez analýzy jejich obsahu, je nutné indexovaný text komentáře prohledat a zjistit, zda obsahuje jedno z vytipovaných klíčových slov. V tomto případě jde o název banky a to jak oficiální (Raiffeisenbank), tak například zkratky (RB), nebo neoficiální mezi lidmi běžně používané názvy (raifka, raiffka). Tato metrika, tedy ukazuje počet komentářů, ve kterých byl zmíněn název banky v jakémkoliv z výše zmíněných podob.

Počet komentářů obsahujících název produktu banky

Podobně jako názvy bank je analyzován název produktu. Banky některé své produkty označují specifickými názvy, tak aby byly mezi sebou odlišitelné. Na základě vytipovaného seznamu názvů produktů jsou spočítány komentáře, které takové názvy obsahují.

Seznam témat

Tato charakteristika shrnuje témata, tedy společné znaky komentářů, které se vyskytují nejčastěji. Pokud se například v komentářích velmi často vyskytuje problematika bankovních poplatků. Výstupem v této charakteristice je téma poplatky. Předpokladem je, že v příspěvcích budou velmi různorodá témata, proto u této kategorie bude nutné podívat se na data v různých pohledů. Například podle typu zdroje dat, podle typu komentář/post a podobně.

Témata vyskytující se společně s klíčovými slovy

Při filtrování dat pomocí klíčových výrazů, které jsou definovány níže, můžeme nahlížet na další témata, která se s klíčovým slovem vyskytují. Například pokud by klíčovým slovem

bylo internetové bankovnictví, můžeme například očekávat, že nalezená témata se mohou týkat výpadků služeb.

Témata s pozitivním nebo negativním hodnocením

Pokud jsou indexované příspěvky ohodnoceny pomocí sentiment analýzy, je možné je také přes tuto dimenzi filtrovat. Pak lze zjistit, která témata se vyskytovala v příspěvcích, které byly ohodnoceny pozitivně, neutrálně, nebo negativně. Podobně jako v předchozím případě, lze očekávat, že negativně hodnocené příspěvky se budou týkat například bankovních poplatků, nefunkčních bankomatů a podobně.

Nejaktivnější uživatelé

Ve webových diskusích ale i na Facebooku mohou být uživatelé, kteří svou aktivitou výrazně předčí ostatní. Tito uživatelé mohou být potenciální opinion makeři.

Celkové vyznění příspěvků pro jednotlivé banky

Abychom mohli jednotlivé banky mezi sebou porovnat, je navržena metrika, která se skládá z počtu příspěvků v kategoriích sentimentu jejich poměru ke všem příspěvkům o dané bance. Pro jednu banku jsou vybrány příspěvky z obou datových zdrojů pomocí klíčových slov, které identifikují názvy banky a její produkty (klíčová slova jsou definována v následující kapitole) a dále příspěvky nebyly vytvořeny profilem banky na Facebooku. Výpočet metriky je:

(počet kladných – počet záporných příspěvků)/počet všech příspěvků

6.2. Kvalitativní charakteristiky

Pro řešení otázky o kterých tématech se psalo, je potřeba definovat klíčová slova, pomocích kterých bude možné tuto analýzu vytvořit.

6.2.1. Seznamy klíčových slov pro vyhledávání příspěvků

Klíčová slova slouží pro přesnější vyhledání příspěvků podle uživatelských požadavků. Pokud chce uživatel více informací o České spořitelně, použije jako parametr vyhledávání seznam klíčových slov, která jsou vytipovaná níže. Mohou obsahovat například různé typy a varianty oficiálního názvu banky, „zlidovělé“ a hovorové formy. Obdobně různé formy názvů konkrétních produktů, případně opět jejich hovorovější formy. Tím, že jsou tato klíčová slova použita při vyhledávání, je větší šance, že bude nalezen větší počet příspěvků, které jsou relevantní.

Názvy bank

Každý banka se dá v textu příspěvků vyhledávat pod oficiálním názvem. Lidé ale pro jejich označení používají zkratky, proto jsou tyto názvy pro každou banku vytipovány. Nejspíše existují i další formy zkratek, které uživatelé používají. K nim lze dospět analýzou jednotlivých příspěvků. Tato analýza nebyla provedena.

Název banky	Další označení
Česká spořitelna	spořitelna, spořka, čs
Komerční banka	komečka, kb
UniCredit Bank CR	unicredit, ub
Raiffeisenbank	raifka, raiffka, rb
GE Money Bank	GE, GEmoney

Tabulka 5 Seznam názvů bank a jejich hovorových forem (Zdroj: autor)

Názvy produktů

Obecné názvy produktů jsou připravené podle typů produktů, které sledované banky nabízejí. Názvy konkrétních produktů bank jsou vytipované podle toho, zda obsahují výraz, který je pro daný produkt specifický a je tedy možné pomocí názvu produktu odlišit banky mezi sebou.

Obecné názvy produktů:

- Účet
- Úvěr
- Půjčka
- Hypotéka
- Kreditní karta
- Debetní karta
- Kontokorent
- Internetové bankovnictví, internet banka

Názvy produktů bank:

Název banky	Produkty a služby
Česká spořitelna	Osobní účet ČS II, SERVIS 24, Kreditní karta Odměna, iBod
Komerční banka	A karta, MůjÚčet, Lady Karta, MojeBanka, Konto G2.2
UniCredit Bank CR	U konto, Konto PREMIUM, PRESTO Půjčka
Raiffeisenbank	eKonto Smart/Komplet/Student
GE Money Bank	Genius Free & Flexi, Genius Gratis, Genius, bene+, Expres půjčka

Tabulka 6 Seznam názvů produktů bank (Zdroj: autor)

Příprava vyhledávacích výrazů

Výrazy obsahující název banky a názvy jejich produktů jsou spojeny pro každou banku do jednoho vyhledávacího řetězce spojené operátorem OR. Při vyhledávání stačí, aby byl v textu zprávy obsažen jen jeden vytipovaný výraz, aby byl příspěvek nalezen a zobrazen nebo započítán do výstupu dashboardů. Takto připravené vyhledávací výrazy jsou volně kombinovatelné, takže lze zjistit i průniky mezi příspěvky a více vyhledávacími výrazy.

Jednotlivé výrazy lze zadat do vyhledávacího pole i samostatně ale pro využití vyhledávacích schopností Elasticsearch je potřeba klíčová slova uzavřít do složitějšího výrazu. Klíčová slova jsou hledána jen v poli message, proto je potřeba vyhledávání začít názvem tohoto pole. Pokud tedy budeme chtít vyhledat příspěvky, které obsahují názvy produktů České spořitelny, bude vyhledávací výraz následující:

m*essage:(„Osobní účet ČS II“ „SERVIS 24“ „Kreditní karta Odměna“ iBod)*

Takto budou vyhledány příspěvky, které obsahují alespoň jeden z názvů. Lze také změnit logické OR na AND – pokud požadujeme, aby vyhledané příspěvky povinně obsahovaly více vyhledávaných slov, stačí přidat + (znak plus) před vyhledávaný termín. Víceslovné názvy je potřeba uvádět v uvozovkách.

7. Návrh dashboardů

Zobrazení dat je realizováno pomocí dashboardů v aplikaci Kibana. První dashboard nazvaný Overview ukazuje definované metriky a druhý, Topic analysis ukazuje témata, respektive slova, která se vyskytovala často, či mohou být potenciálně zajímavá. Dashboardy slouží pro analýzu stažených dat a jsou přípravou pro konečnou vizualizaci. Vyhledávání umožňuje zadat dotaz a jsou zobrazena data, která osahují konkrétně tento výraz nebo tvary toho výrazu.

7.1. Zobrazení kvantitativních údajů

Dashboard Overview obsahuje sadu vizualizací, které odpovídají na kvantitativní otázky o uložených datech. Na data lze v tomto dashboardu nahlížet z různých úhlů, lze použít vyhledávání pro zobrazení specifické podmnožiny dat. Tento pohled umožňuje uživateli zobrazit velmi široké množství náhledů na data. Všechny objekty slouží zároveň jako filtry, které umožňují zobrazit data podle zájmu uživatele.

Dashboard obsahuje tyto objekty:

Activity – Vývoj počtu příspěvků v čase

Vývoj počtu příspěvků je v horním nejširším grafu. Je možné filtrovat pomoci označení oblasti v grafu myši. Tím je možné omezit data třeba na období vysoké aktivity uživatelů.

Page – Název stránek

Objekt v levém horním rohu pod grafem aktivity zobrazuje seznam názvů zdrojů dat (název webových stránek a název Facebookových profilů) a k nim počet dokumentů. Vizualizace je typu Data table, která umožňuje filtrování klinutím na název a dále změnu řazení podle počtu dokumentů. Kliknutím na název například ceskasporitelna, dojde k novému načtení dat a zapojí se filtr na data, která pochází ze stránky ceskasporitelna.

Source - Zdroj

Sloupcový graf source umožnuje porovnat počet dokumentů, které pochází z Facebooku a webu. Opět lze data filtrovat kliknutím na tělo sloupce.

Type – Typ

Sloupce v grafu reprezentují počet dokumentů uložených v jednotlivých typech v indexu banky.

Page Likes – Počet liků stránky

Spojnicový graf ukazuje vývoj počtu liků facebookových stránek bank v čase, zároveň lze porovnat popularitu stránek mezi sebou.

FB comment count – Počet komentářů na Facebooku

Počet komentářů na jednotlivých stránkách bank, kde se sčítají komentáře, které vytvořili uživatelé i banka sama.

FB post count – Počet postů na Facebooku

Podobné zobrazení jako předchozí, jen zobrazuje počet postů jak od banky, tak od uživatelů.

Sentiment – Rozložení sentimentu mezi dokumenty

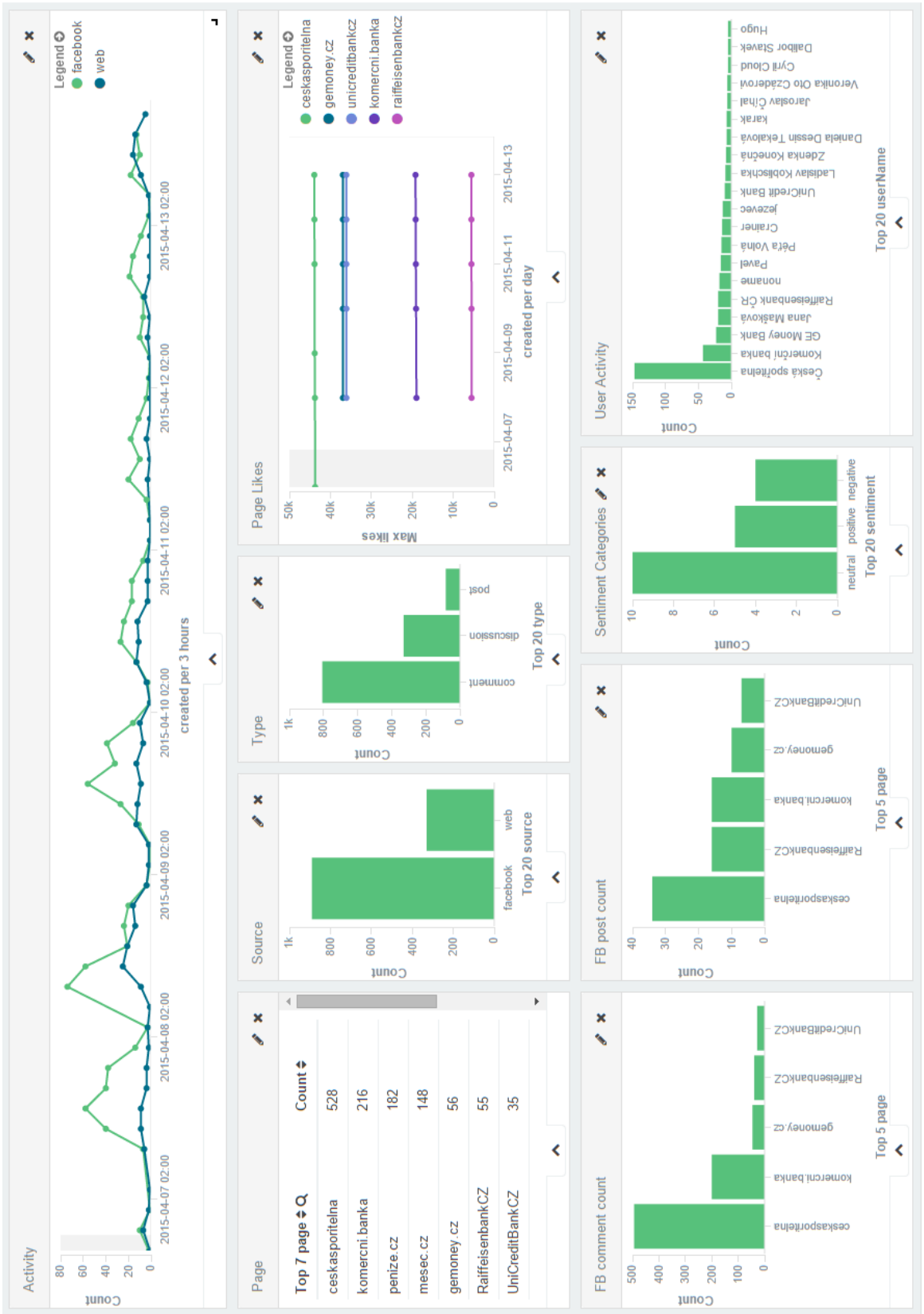
Souhrn hodnocení sentimentu je zřetelné v předposledním sloupcovém grafu.

User Activity – Aktivita uživatelů

Nejaktivnější uživatelé jsou vidět v grafu dole vpravo.

7.1.1. Použití dashboardu

Data, která dashboard zobrazuje, nejsou skoro nijak filtrována – až na časové omezení, které je povinně nastavené. Lze ho samozřejmě libovolně upravovat pomocí popsaných možností. Dashboard umožňuje uživatelům jednoduchým poklikáním na požadované objekty sestavit si zobrazení, které je zajímá. Lze například zjistit, kde se vyskytují nejvíce negativní příspěvky, který zdroj způsobil výkyv po počtu příspěvků a podobně. Další možností je zadat dotaz do vyhledávání a tím například zjistit, zda příspěvky obsahovaly některé z klíčových slov, jak často se vyskytuje název banky a podobně.



Obrázek 11 Dashboard Overview je určený pro zodpovězení kvantitativních otázek (Zdroj: autor)

7.2. Přehled nejčastějších témat

Dashboard s názvem Topic analysis je určen pro získání náhledu na probíraná témata v rámci stažených dat. Vizualizační prvky jsou z části stejné jako v předchozím dashboardu, což umožňuje uživateli vyfiltrovat stejná zobrazení s tím, že získá také informaci o tom, jaká témata se v daném období, a struktuře zdrojů dat probírala. K tomu jaká témata byla předmětem zájmu komentujících lze dojít více způsoby. Napomoci k tomu mohou tabulky nejčastější pojmy a neobvyklé pojmy, a pak ještě četnost příspěvků v časovém průběhu v grafu v horní části.

Obsahuje tyto objekty:

Activity – Vývoj počtu příspěvků v čase

Stejné jako v Overview.

Page – Název stránek

Stejné jako v Overview.

Source - Zdroj

Stejné jako v Overview.

Sentiment – Rozložení sentimentu mezi dokumenty

Stejné jako v Overview.

Frequent Terms – Nejčastější pojmy

Tato tabulka ukazuje nejčastější slova, která obsahují vybrané příspěvky. Použít lze k zjištění témat, kterých se příspěvky týkají. Je zde zobrazeno slovo ve tvaru po provedení stemmovacích pravidel a četnost výskytu.

Top unusual terms – Neobvyklé pojmy

Další tabulka obsahuje pojmy, která jsou ze statistického hlediska nezvyklé. Tyto pojmy se, vyskytují častěji, než by podle statistického modelu podle ostatních dat měly. Tato tabulka tedy může poukázat na nezvyklosti ve vybraných datech.

Obrázek 12 Dashboard Topic analysis – zodpovídá otázky, o čem lidí hovořili (Zdroj: autor)

7.3. Interpretace dat

Předchozí vizualizace jsou nutným dílčím výstupem, díky kterému jsou data analyzována a připravena pro návrh pravidelného přehledu o vývoji WoM na internetu.

Díky dashboardům lze zjistit fakta o chování uživatelů a jejich názoru na služby bank. Pravidelný přehled je v návrhu této práce koncipován jako týdenní, případně na širší časové periodě. V následujícím textu je shrnuta interpretace a popis možností použití vizualizací.

Vycházejme nejdříve z analýzy počtů nebo nápadných výkyvů v počtu příspěvků zobrazených ve dashboardu Overview.[4] Jsou zde zobrazena data za posledních sedm dní (do 13. 4. 2015). Celkově bylo za toto období stažení 888 příspěvků s Facebooku a 330 z webových stránek. Nejvíce příspěvků celkově (typu post i komentář) za toto období bylo přidáno na facebookovém profilu České spořitelny (ČS) a to více než dvakrát tolik, než na profilu druhé Komerční banky. To je možné vysvětlit také rozdíly v počtu liků stránky. Česká spořitelna má 43 920 liků – tedy lidí, kteří sledují její stránky zatím co Komerční banka má 19 209 liků. Pokud ale vztáhneme k sobě jen počet liků a aktivitu uživatelů na FB profilu banky, zjistíme, že předchozí vztah úplně neplatí. Unicredit a GE Money mají na facebookových profilech větší počet liků, ale nemají tak vysoký počet příspěvků. Je zde tedy vidět určitá disproporce v počtech příspěvků a počtu liků stránky u jednotlivých bank. Kteří uživatelé tedy přispívají na stránky České spořitelny na Facebooku? Pomocí vyfiltrování profilu ČS lze zjistit, že nejaktivnějším uživatelem je sama Česká spořitelna, která na svých stránkách vytvořila v daném období 147 příspěvků. ČS je v daném období také nejaktivnější uživatel, následuje Komerční banka (43) a GE Money (23). Podle hodnocení sentimentu jsou příspěvky rozděleny do skupin negativní (527 příspěvků), neutrální (401) a pozitivní (289).[5]

Filtrováním podle tříd sentimentu můžeme získat přehled o tom, kolik příspěvků a od kterých autorů jsou negativní nebo pozitivní. Také je vidět, které stránky jsou pozitivněji nebo negativněji naladěné.

Všechny předchozí filtry také mění rozložení počtu příspěvků v grafu jejich časového vývoje v hodní části dashboardu v grafu Activity. Je proto také možné zaměřit se na časové

[4] Data pro tuto práci byla sbírána průběžně a nelze tedy zaručit, že si budou odpovídat data z předchozích obrázků s daty, která jsou popisována v této části.

[5] Součet počtu příspěvků z datových zdrojů (1218) a součet počtu příspěvků v třídách sentimentu (celkem 1 217) není stejný, protože jeden příspěvek neobsahoval text. Nebylo tedy možné přiřadit třídu sentimentu.

období, které je co se týče počtu příspěvků nápadně odlišné od ostatních. Takové časové období je možné vyfiltrovat výběrem myší v grafu. Tím se data omezí jen na toto období a zle zjistit, díky, kterým stránkách a uživatelům výkyv nastal.

Předchozím postupem je kvantitativně zanalyzováno, jak se data v čase vyvíjela a jaká vyla jejich struktura. Abychom zjistili, jaká témata uživatelé probírali, přepneme na dashboard Topic analysis.

Ten poskytuje podobné možnosti jako předchozí dashboard. Systém práce filtrování a zadávání dotazů je také stejný. Vhodné je také mít připraveny otázky z předchozího průzkumu dat v dashboardu Overview. Pokud nějaký výkyv v datech je, je možné na dashboardu Topic analysis zjistit, o čem uživatelé psali a také si jednotlivé zprávy přečíst. Dva základní prvky, které shrnují slova, která jsou společná pro příspěvky a která jsou potenciálně zajímavá svým netypickým statistickým rozložením, jsou Frequent terms a Unusual Terms. Tato pole zjednodušeně ukazují, která témata uživatelé v daném období probírali nejčastěji. Pro lepší pochopení je vhodné si postupně vyfiltrovat pro uživatele dashboardu zajímavé pojmy z obou zmíněných vizualizací a přečíst si několik příspěvků, které jsou zobrazeny v tabulce Messages. Tím uživatel získá lepší přehled o kontextu témat. Program v tomto případě dokáže automaticky navrhnout, co může být zajímavé, ale jen uživatel dashboardu může data interpretovat a následně výsledek použít.

Posledním nástrojem, kde lze zjistit další informace o datech je plugin carrot2. Díky upravené funkcionalitě (popsané v kapitole 5.4.3), lze jen zadat výraz, který mají mít příspěvky společný – tím omezíme vstupní množinu dat a výsledky se budou týkat jen příspěvků, které obsahují například slovo poplatky (v různých tvarech).

Na základě dotazu poplatek je vidět výsledná vizualizace clusterů na obrázku Obrázek 10. Z obrázku je zřejmé, že některé clustery nejspíše nebudou mít větší význam, protože jsou uvedeny souhrnně pod klíčovými slovy nějaké, době, díky nebo mimochodem. Naopak významné mohou být clustery Vstupní poplatek, Osobní účet a Zrušení poplatku. Clustery samy o sobě většinou nenapoví, co přesně uživatelé ve zdrojových datech probírali. Proto je vhodné vybrat zajímavé clustery a zjistit pomocí vyhledávání v dashboardu Topic analysis zadáním názvu clusterů další informace o příspěvcích. Tomu napomůže buď seznam Frequent Terms a nebo Unusual terms, případně pak přečetní jednotlivých příspěvků.

8. Návrh pravidelného přehledu Word of Mouth z analyzovaných zdrojů

Protože vizualizace aplikaci Kibana není v základním nastavení možné uvolnit veřejně, kvůli možnosti editovat všechna nastavení a vizualizace kterýmkoliv uživatelem, je potřeba navrhnout jiné zobrazení a distribuci toho zobrazení. Existují způsoby jak aplikaci Kibana zabezpečit a oddělit uživatele a tvůrce vizualizací pomocí přihlašovacích údajů. To ale poskytne jen omezenou možnost distribuce. Zároveň vizualizace v aplikaci Kibana není vhodná pro všechny uživatele. Vyžaduje určitou zkušenost s podobnými nástroji, kvůli komplexnosti možností použití.

Hlavním cílem této práce je navrhnout vizualizaci, která předá širšímu okruhu uživatelů z různých zájmových okruhů data a informace, která jsou popsána výše, a který předpokládá distribuci přehledu pro široké spektrum uživatelů, například i ve sdělovacích prostředích, kde uživatelé uvidí, jak se za poslední období vyvinula WoM uživatelů na sledovaných sociálních sítích a webech. Už v tom je jistý rozpor. Potenciál informační hodnoty, který poskytuje navrhnuté prostředí dashboarů v aplikaci Kibana není možné předat v jen statickém zobrazení dat. Z nástinu možností pohledů na data v předchozích odstavcích je vidět, že vždy je nutné hledat souvislosti v konkrétním uspořádání dat. V těch jsou skryté informace, jejichž důležitost různé skupiny uživatelů může být různá. Proto různé skupiny uživatelů bubou s dashboardy pracovat jiným způsobem.

Ve vizualizaci pro distribuci uživatelům proto budu vycházet z definovaných metrik, které vizualizace bude ukazovat. Dalším problém je typ média, který by měl výstup vydávat. Jiný typ vizualizace by měl být pro tištěná média nebo emailovou distribuci. Online media naopak mohou umožnit určitou interaktivitu. Mým záměrem je navrhnou možnost předat informace emailovou formou, případně formou offline media (například tisk).

8.1. Návrh podoby pravidelného WoM přehledu

Návrh respektuje definované metriky v předchozí části práce. Jak bylo řečeno, různé typy uživatelů vyžadují rozdílné typy informací. Aby uživatelé nebyli zahlceni, je navržena stručnější a méně vizuálně komplexní podoba výstupu. Preferuji spíše tabulkové zobrazení dat, protože media, která by měla výstup přenášet, nemusí podporovat složitou grafickou podobu, a jednodušší zobrazení je obvykle přehlednější.

Pokud má být výstup pravidelně dodáván uživatelům, je důležité, aby měl vždy stejnou formu. I z tohoto důvodu se jeví jako jednou z variant zasílání emailové zprávy se screeny dashboardů z aplikace Kibana. Zobrazují veškeré údaje, které jsou předmětem definovaných metrik. Ale forma je poměrně složitá a náročná na prostor. Pokud ale předpokládáme i možnost distribuce v tištěné formě, není výstup z Kibany vhodný, protože nelze jednoduše vizualizaci dat zmenšit, aby obsahovala minimum nevyužitého místa.

8.1.1. Tabulkové zobrazení výstupu WoM

Definované metriky a výstupy v kapitole 6 by měly být zobrazeny ve formě jednoduchých tabulek, aby bylo možné je jednoduše distribuovat kvůli omezením, která jsou dána přenosovými medii. Protože jde o pravidelně podávanou informaci, vždy musí být uvedeno, za jaké období jsou data zobrazena.

Oproti návrhu zobrazení v Kibaně, můžeme data zobrazit v souhrnnější formě. Komplexnost dat a možnosti pohledů na ně jsou velmi široké, proto je potřeba vybrat údaje, které jsou potenciálně důležité pro všechny uživatele. Kvůli spektru uživatelů budeme předpokládat, že nejdůležitější jsou informace, které se týkají základních kvantitativních údajů z Facebooku (které jsou definované v kapitole 6), a pak hlavně témat diskusí z obou zdrojů.

Kvůli omezeným možnostem oproti dashboardům v Kibaně jsou výstupy pro vizualizaci dat méně obsáhlé. Vybrané údaje pro zobrazení jsou mým návrhem. Pokud by měl být výstup zaměřen na konkrétní skupinu uživatelů, je nutné zjistit jejich potřeby a reflektovat je ve výstupu. Zjištění těchto potřeb není předmětem této práce. Existuje k tomu několik důvodů. V úvodu byly nastíněny marketingové pohledy na to, proč je důležité získávat zpětnou vazbu od uživatelů. Specifika této potřeby můžou být různá pro různé uživatele výstupu - rozdílná bude motivace získat informace pro zaměstnance bank na různých odděleních a pro čtenáře novin, kde by mohl být výstup vytisknut.

Komplexnějšímu obsahu lze porozumět jen s pomocí detailnější analýzy dat, která většinou bude muset být podpořena i přečtením příspěvků pro získání povědomí o jejich kontextu. Proto má tento navržený výstup menší potenciál předat informace než dashboardy v Kibaně. Výstupem jsou následující zobrazení.

První tři metriky složené z dat z Facebooku můžeme ukázat v podobě tabulky, která zobrazuje název banky a následně tři sloupce s hodnotami metrik. [6]

Profil banky	Počet liků profilu	Počet postů	Počet komentářů
Česká spořitelna	43 949	29	397
Komerční banka	19 264	14	169
Reiffeisen bank	5 632	21	76
GE Money	36 950	13	48
UniCredit bank	36 071	8	51

Tabulka 7 Návrh zobrazení metrik Facebookových profilů bank (Zdroj: autor)

Následující tabulka zobrazuje údaje týkající se témat, o kterých se hovořilo v komentářích na sledovaných webových stránkách. Cílem je ukázat, jak je významné sledovat komentáře na dané stránce z pohledu četnosti výskytu klíčových slov a témat, která sledujeme.

Klíčové slovo	Počet na stránce mešec.cz	Počet na stránce peníze.cz
Česká spořitelna	2	0
Komerční banka	1	0
Reiffeisen bank	1	1
GE Money	0	1
UniCredit bank	0	0
Obecné názvy produktů	21	12
Konkrétní produkty bank	3	1

Tabulka 8 Návrh přehledu počtů komentářů, diskutovaných v souvislosti s bankami (Zdroj: autor)

Nejzajímavější údaje pro uživatele pravděpodobně budou souhrny témat, která se objevila v datových zdrojích. U facebookových komentářů, lze předpokládat, že se týkají bankovní problematiky a marketingových aktivit banky. U komentářů z webových stánek tento předpoklad neplatí, protože záběr tématiky sledovaných webů je širší. Z toho důvodu data vyfiltrujeme pomocí klíčových slov.

[6] Data jsou zobrazena za období 8.4. – 13.4.2015

Obecná témata z obou zdrojů jsou zobrazena v následující tabulce Tabulka 9. Jsou to slova a fráze, která byla probírána nejčastěji bez ohledu na banku, se kterou byly svázány (případně ani nemusely být) a bez ohledu na zdroj dat.

Název tématu	**Počet výskytů**
Karta (nálepka, platební)	135
Banka	81
Nálepka (bezkontaktní)	64
Účet (zdarma, osobní)	55

Tabulka 9 Nejsilnější témata bez ohledu na sentiment a kontext s bankou (Zdroj: autor)

Témata podle názvu banky (a jejich produktů), jsou zobrazeny v tabulce Tabulka 10. Zároveň je vidět převažující třída sentimentu (nejvíce zastoupené) v názvu tématu pomocí barvy - červený (negativní), černý (neutrální), zelený (pozitivní). Jsou zobrazeny jen nejsilnější témata, proto není u všech bank stejný počet, je to tím, že pro jednotlivé banky není nalezen podle vybraných klíčových slov stejný počet příspěvků. Ve většině případů byla také nejsilnější neutrální kategorie sentimentu.

Název Banky	**Téma**	**Počet výskytů**
Česká spořitelna	nálepka	115
Česká spořitelna	karta	90
Česká spořitelna	ibod	11
Komerční banka	karta	135
Komerční banka	bezkontaktní	41
Komerční banka	nálepka	39
Unicreditbank	konto	12
Unicreditbank	nálepka	6
Reiffeisen bank	eKonto	4
GE Money	GE	24

Tabulka 10 Nejsilnější témata zmiňovaná v kontextu s každou bankou a jejich sentiment (Zdroj: autor)

O kterých tématech se mluví s jakým citovým zabarvením je zobrazeno v následující tabulce témata podle sentimentu – opět pomocí barevného zvýraznění. Data jsou zobrazena bez ohledu na zdroj, bez použití dalších filtrů, kromě časového. Jsou zobrazena témata, která byla podle autora dostatečně silná a vypovídající.

Téma	Počet výskytů
Banka	80
Nálepka	49
Bezkontaktní karty	74
Účet zdarma	42
Nálepka	10

Tabulka 11 Nejsilnější témata pro každou kategorii sentimentu (Zdroj: autor)

Sentiment příspěvků podle názvu banky nehledě na zdroj příspěvků je vidět v tabulce Tabulka 12. Výběr je proveden podle klíčových slov názvů bank a jejich produktů.

Profil banky	Pozitivní	Neutrální	Negativní
Česká spořitelna	114	205	229
Komerční banka	56	140	128
Reiffeisen bank	27	30	36
GE Money	24	24	15
UniCredit bank	20	18	19

Tabulka 12 Počet příspěvků podle kategorií sentimentu pro každou banku (Zdroj: autor)

Nejaktivnější uživatelé z obou zdrojů dat obsahuje tabulka Tabulka 13 Seznam deseti nejaktivnějších uživatelů (kromě profilů bank) v obou datových zdrojích.

Jméno uživatele	zdroj	Počet příspěvků
noname	web	23
Jana Mašková	facebook	20
Pavel	web	16
Péťa Volná	facebook	16
Crainer	web	14
jezevec	web	14
Ladislav Koblischka	facebook	10
Zdenka Konečná	facebook	10
karak	web	9
Daniela Dessin Tekalová	facebook	7

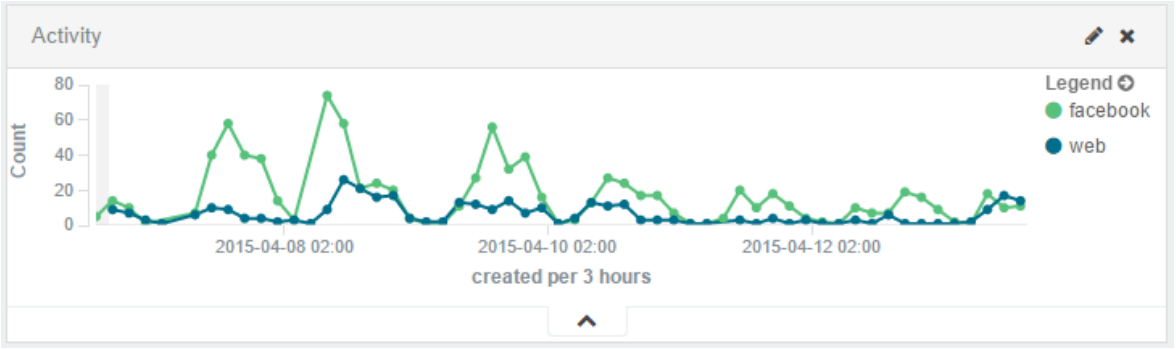
Tabulka 13 Seznam deseti nejaktivnějších uživatelů (kromě profilů bank) v obou datových zdrojích (Zdroj: autor)

Abychom mohli jednoduše jedním číslem porovnat banky mezi sebou je navržen metrika Celkové vyznění příspěvků pro jednotlivé banky. Výpočet je definován v kapitole 6.1.3. Data pro výpočet a výsledné skóre zobrazuje následující tabulka. Nejlépe si stojí GE Money, má ale nejmenší počet příspěvků, relevance je tedy slabší než například u České spořitelny.

Profil banky	Počet příspěvků	Pozitivní	Neutrální	Negativní	Skóre
Česká spořitelna	394	72	145	177	**-0,27**
Komerční banka	279	42	116	121	**-0,28**
Reiffeisen bank	68	13	22	33	**-0,29**
GE Money	37	6	18	13	**-0,19**
UniCredit bank	41	7	15	19	**-0,29**

Tabulka 14 Výpočet Celkového vyznění příspěvků jednotlivých bank (Zdroj: autor)

Posledním přehledem je vývoj počtu příspěvků v čase. Ten je zobrazen pomocí výstupu z aplikace Kibana na obrázku 13.



Obrázek 13 Vývoj počtu příspěvků za dané období

Všechny přechozí tabulky byly vytvořeny ručně pomocí dat z aplikace Kibana. Některé tabulky by bylo možné vytvořit přímo v Kibaně, ne ale všechny. Je to dáno možnostmi aplikace, která není typickou Business Intelligence aplikací, ve kterých jsou možné větší úpravy vizualizací.

8.2. Závěr kapitoly

V této kapitole byla navržena zobrazení, která mohou sloužit jako přehled o příspěvcích na sledovaných zdrojích za poslední období, které v mém příkladu bylo jeden týden. Tento přehled je díky formě tabulek možné distribuovat emailem, případně zobrazit tabulky v na internetu. Jako další možnost předat data a informace uživatelům bylo navrženo zpřístupnění dashboardů v aplikaci Kibana, což si ale vyžádá zavedení bezpečnostních prvků. Důležitá je nutnost přihlášení uživatele před vstupem do aplikace a rozlišení uživatelů a tvůrců reportu rozdílnými oprávněními.

Tyto dva typy zobrazení nejsou zaměnitelné. Kibana poskytuje mnohem širší možnost práce s daty - pomocí filtrování a vyhledávání. Tím je také pro uživatele mnohem cennější, protože ti mají možnost zjistit odpovědi na otázky, které je zajímají. Oproti tomu tabulkové je snazší na pochopení a bylo zvoleno jako vhodnější zobrazení a hlavní výstup této práce zvoleno pro svou jednoduchou distribuci.

9. Závěr

Komentáře vytvořené uživateli na sociálních sítích a internetových diskusích jsou doplňkem pro získání zpětné vazby a dalších informací o jejich zájmech a chováních při marketingovém výzkumu. Tato práce se zabývá technologiemi potřebnými pro získání a analýzu těchto dat a návrhem zobrazení, které mohou použít zkušenější uživatelé v aplikaci Kibana a nebo souhrnnější průřezové informace v tabulkovém zobrazení. Následující odstavce rekapitulují dílčí cíle této práce a uvádí, jakým způsobem byly vyřešeny. Hlavní cíl je zhodnocen v na konci této kapitoly.

Východiskem této práce byl popis marketingového výzkumu a jeho dílčích součástí, které se zabývají získáním informací od uživatelů služby nebo produktu. Jedním z konceptů je Word of Mouth marketing, který podobně jako v běžné komunikaci mezi lidmi je součástí i internetové komunikace, kde fungují podobné principy. Další metody, které slouží nejen jako součást marketingového výzkumu ale i jako samostatné metody v širším rámci, jsou popsány Voice of Customer, Net Promoter Score a analýza sentimentu. Tím je definován rámec zájmu práce a důvody, které vedou k potřebě získat informace od zákazníků. Tyto základy práce jsou popsány v kapitole 2 Marketingový význam analýzy.

Jakým způsobem jsou data z internetu získána, je popsáno v kapitole 3 Principy stahování a dat z internetu. Nejdříve je nastíněn princip stažení komentářů pod články z internetových stránek a následně z facebookových profilů. Dále jsou popsány důvody výběru stránek a FB profilů, které jsou v této práci analyzovány. Jsou to facebookové profily bank Česká spořitelna, Komerční banka, Reiffeisenbank, Unicredit bank a GE Money Bank a komentáře článků peníze.cz a měšec.cz

Použitá technologie pro stažení, uložení a obohacení dat je v specifikovaná v kapitole 4. Všechny aplikace byly vytvořeny s pomocí knihoven v jazyce Java. Zde je uveden základní popis funkcí aplikací pro stažení dat z vybraných webových stánek a z Facebooku. Detailní návrh aplikací není obsažen, protože samotné aplikace nejsou hlavním cílem této práce. Proto na „čistotu“ vytvořených aplikací nebyl kladen takový důraz. Obohacení dat o je realizováno programem pro klasifikaci sentimentu příspěvků. To je rozděleno na trénování klasifikačního modelu a zpracování dat z uložiště pro přidání kategorie sentimentu. Všechna data jsou ukládána do Elasticsearch, tím je determinována také použitá technologie pro stažení dat a úpravy.

Data jsou uložena v systému Elasticsearch, který staví na knihovně Lucene od Apache, jehož specifika jsou uvedena v kapitole **Chyba! Nenalezen zdroj odkazů.**. Součástí je popis použitého nastavení prostředí v tomto systému a jeho základní možnosti. Následně jsou popsány pluginy a aplikace, které jsou použity v rámci Elasticsearch pro jednodušší prací a pro získání vizuálních výstupů datových analýz.

Jaká data jsou zobrazována, popisuji v kapitole 6. Výstupy, které jsou předmětem této práce, jsou navrženy a definovány právě zde. Facebookové a webové metriky a charakteristiky jsou rozděleny a kvantitativní a kvalitativní. Je uvedeno, jaký způsobem může být počítána metrika, která ukazuje, jak jedním číslem ohodnotit výslednou pozici banky podle sentimentu komentářů, které o bance vytvořili uživatelé.

Podoba vizualizací pomocí aplikace Kibana je navržena v textu kapitoly 7. Tato část je nejvíce analytická. Ukazuje možnosti rozboru stažených dat z různých pohledů. I když Kibana není klasický BI nástroj, jsou možnosti široké. Díky možnosti přímého čtení, dat jejich filtrování a téměř okamžitému přehledu všech charakteristik v navržených dashboardech, jde o nejlepší možnost, kterou lze implementovat s minimálním náklady. Součástí kapitoly je analýza dat na období jednoho týdne a jejich interpretace. Tento výstup je dílčím cílem, který vede k vytvoření zjednodušených vizualizací pro distribuci emailem nebo obecně webovými prostředky.

Hlavní cíl je naplněn v kapitole 8. S pomocí vizualizací v Kibaně, byla vytvořena tabulková zobrazení navržených metrik, která dávají základní přehled o chování uživatelů ve sledovaných datových zdrojích. Základem je přehled kvantitativních charakteristik, ale důraz je kladen spíše na kvalitativní v podobě seznamů témat, která byla probírána v příspěvcích, v různých pohledech. Tato analýza byla provedena ručním zpracováním dat v aplikaci Kibana.

Vyhodnocení těchto výstupů a shrnutí zkušeností z analýzy dat ukazuje, že je nutné data vyhodnotit v širším kontextu. Tabulková zobrazení jsou základním vodítkem pro získání vhledu do témat a číselných charakteristik analyzovaných dat. Je dobře dostupný pro velmi širokou skupinu uživatelů. Avšak tento výstup může být pro řadu uživatelů dostačující. Pro marketing firem ale dostatečný není. Důležitější jsou podle autora kontextuální souvislosti dat, o nichž lze získat přehled jen důkladnější analýzou. Kvůli rozmanitosti dat a možností průřezů je nejpodstatnější zobrazení v aplikaci Kibana.

Výstupy v Kibaně jsou určeny možnostmi této aplikace. Protože je Elasticsearch možné používat a dotazovat přes Java API. Je také možné vytvořit vizualizace naprogramované přímo na míru požadavkům konkrétním uživatelům. Rozšířením této práce proto může být získání přehledu o požadavcích na informace konkrétních uživatelů a vytvoření zobrazení dat přímo pro jejich potřeby. Zobrazení tedy může být formou úpravy dashboardů v aplikaci Kibana, nebo vytvořením vlastní vizualizace za pomoci různých vizualizačních frameworků a knihoven dostupných na internetu. Tento výstup ale musí být vytvořen na míru požadavkům konkrétních uživatelů. Musí také být zvážen přínos oproti nákladům na implementaci podobného řešení, které by vytvořeno pomocí dalších vizualizací mimo aplikaci Kibana.

Jiné rozšíření této práce může zhodnotit použití datových zdrojů a vybrat širší spektrum webových stránek nebo facebookových profilů a případně dat z jiných sociálních sítí pro hodnocení širšího portfolia bankovních produktů a bank obecně. Nebo se lze zaměřit na jinou oblast podnikání, například na telekomunikace, nebo politické strany.

Dílčím rozšířením této práce je také zkvalitnění modelu sentiment analýzy, který byl popsán a pro analýzu příspěvků používán.

Výsledky této práce jsou také použitelné pro zaměstnance, kteří uvažují o použití systémů pro monitoring medií a zvažují možnosti nasazení různých technologií a od různých dodavatelů. V této práci je předveden postup pro analýzu dat z vybraných zdrojů s minimálními náklady na programové vybavení a i na hardwarovou podporu. Celý systém běží i na počítači (resp. notebooku) s omezenými výpočetními možnostmi (AMD 2 Core, 4GB RAM, Win 8)

Získání příspěvků z internetu – webových diskusí a sociálních sítí – je zcela běžně dostupnou možností, jak obohatit marketingový výzkum o údaje od velkého množství uživatelů. Za nejdůležitější pokládám možnost získávat tato data průběžně. Většina standardních zdrojů dat pro marketingový výzkum je dočasná. Získávání dat z internetu je průběžný a nekončící proces, který má potenciál ovlivnit rozhodování manažerů a uživatelů.

9.1. Shrnutí

Díky analýze dat bylo ukázáno, že data na internetu opravdu mají potenciál být doplňkem pro marketingový výzkum. V aplikaci Kibana byly předvedeny široké možnosti práce

s těmito daty a pro přehlednější, čitelnější výstup byl navržen zjednodušený tabulkový výstup dat, co je hlavním výstupem této práce.

Tato práce je potenciálním návodem pro firmy, které chtějí podobná data získávat a analyzovat. Cele řešení je postavené na opensource aplikacích, které mohou firmy pro vybudování vlastního řešení, použít.

Rozšířením může být zvýšení přesnosti sentiment analýzy, zapracování dalších datových zdrojů – dalších webových diskusí, nebo facebookových profilů. Případně lze toto řešení použít na získání přehledu o zcela jiné doméně zájmu – například telekomunikace.

10. Bibliografie

1. **Gleich, Viktor.** *Marketing na sociálních sítích – metriky a měření.* Praha : Vysoká škola ekonomická v Praze, 2014.

2. **Kotásková, Ljuba.** *Budování značky na sociálních sítích (Facebook).* Zlín : Univerzita Tomáše Bati ve Zlíně, 2014.

3. **Linhart, Ondřej.** *Využití dat ze sociálních sítí pro BI.* Praha : Vysoká škola ekonomická v Praze, 2015.

4. **Hradská, Martina.** *Získávání nestrukturovaných dat ze sociální sítě Facebook.* Brno : Masarykova univerzita, 2013.

5. **Šverák, Martin.** *Analýza nestrukturovaného obsahu z veřejně dostupných sociálních médií za pomocí nástroje Watson společnosti IBM* . Praha : Vysoká škola ekonomická v Praze, 2014.

6. **Kotler, Philip, a další.** *Moderní marketing.* Praha : Grada, 2007. 978-80-247-1545-2.

7. **Kozel, Roman, Mynářová, Lenka a Svobodová, Hana Svobodová.** *Moderní metody a techniky marketingového výzkumu.* Praha : Grada Publishing a.s., 2011. 802473527X, 9788024735276.

8. **Poynter, Ray a Henning, Jeffrey.** The Top 20 Emerging Methods In Market Research. *GreenBook Charting the future of market research.* [Online] 29. září 2014. [Citace: 1. duben 2015.] http://www.greenbookblog.org/2014/09/29/the-top-20-emerging-methods-in-market-research-a-grit-sneak-peek/.

9. **Zikmund, Martin.** Word of mouth – moderní strašák každého businessu. *BusinessVize.* [Online] 6. květen 2010. [Citace: 1. duben 2015.]

http://www.businessvize.cz/zakaznici/word-of-mouth-moderni-strasak-kazdeho-businessu.

10. **East, Robert.** Researching Word of Mouth. *Australasian Marketing Journal.* 2007, 15.

11. **Nečas, Miroslav a Marc, Michal.** Monitoring hlasu zákazníka. *Časopis Systémová integrace.* 2011, 2-příloha.

12. **Brandt, Randall.** Hearing Aids How well are you capturing the vioce of customer. [Online] 2012. [Citace: 1. duben 2015.]

http://www.maritzresearch.com/~/media/Files/MaritzResearch/Social-Media/Quality-Progress-Voice-of-the-customer.pdf.

13. **Reichheld, Frederick F.** *The Ultimate Question 2.0: How Net Promoter Companies* . místo neznámé : Harvard Business Press, 2011. 1422173356, 9781422173350.

14. **Sauro, Jeff.** 10 Things To Know About Net Promoter Scores And The User Experience. *MeasuringU.* [Online] [Citace: 1. Duben 2012.]

http://www.measuringu.com/blog/nps-ux.php.

15. **Bing, Liu.** *Sentiment Analysis and Opinion Mining.* Toronto : Morgan & Claypool Publishers, 2012. 9781608458851.

16. **Červenec, Radek.** *Rozpoznání emocí v česky psaných textech.* Brno : Fakulta elektroniky a komunikačních technologií, Vysoké učení technické v Brně, 2011.

17. **Koller, Michael.** *Analýza sentimentu v českém prostředí sítě Twitter.* Praha : Vysoká škola ekonomická v Praze, 2014.

18. **Pelíšek, Jiří.** Analýza sentimentu. Praha : Vysoká škola ekonomická v Praze, 2014.

19. **Patočka, Michal.** *Metody strojového ucení pro analýzu sentimentu.* Plzeň : Západočeská univerzita v Plzni, 2013.

20. **Vural, A. Gural, Cambazoglu, Berkant Barla a Karagoz, Pinar.** Sentiment-Focused Web Crawling. *ACM Transactions on the Web.* 8, 2014, 4.

21. **Moravec, Petr.** *Monitoring internetu a jeho přínosy pro podnikání nástroji firmy SAS Institute.* Praha : Vysoká škola ekonomická v Praze, 2013.

22. **Russell, Matthew A.** *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More.* místo neznámé : O'Reilly Media, Inc., 2013. 1449368220, 9781449368227.

23. **Hedley, Jonathan.** jsoup: Java HTML Parser. *jsoup.* [Online] [Citace: 1. duben 2015.] http://jsoup.org/.

24. **yasserg@github.** crawler4j Open Source Web Crawler for Java. *crawler4j.* [Online] [Citace: 1. duben 2015.] https://code.google.com/p/crawler4j/.

25. **Allen, Mark.** RestFB - A Lightweight Java Facebook Graph API and Old REST API Client. *restfb.* [Online] [Citace: 1. duben 2015.] http://restfb.com/.

26. **The Apache Software Foundation.** Apache OpenNLP Developer Documentation. *Apache OpenNLP.* [Online] [Citace: 1. duben 2015.]

http://opennlp.apache.org/documentation/1.5.3/manual/opennlp.html.

27. **Habernal, Ivan, Ptáček, Tomáš a Steinberger, Josef.** Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. Atlanta, Georgia : Association for Computational Linguistics, 2013, stránky 65-74.

28. **Podnikový autor.** logstash. *logstash.* [Online] [Citace: 1. duben 2015.] http://logstash.net/docs/1.4.2/.

29. **Elasticsearch.** the definitive guide. *elasticsearch.* [Online] [Citace: 1. duben 2015.] http://www.elasticsearch.org/guide/.

30. **Vlček, Lukáš.** Elasticsearch: Vyhledáváme hezky česky. *zdroják.cz.* [Online] 1. červenec 2013. [Citace: 1. duben 2015.] http://www.zdrojak.cz/clanky/elasticsearch-vyhledavame-cesky/.

31. **@mobz.** elasticsearch-head. *elasticsearch-head.* [Online] [Citace: 1. duben 2015.] http://mobz.github.io/elasticsearch-head/.

32. **Stanisław Osiński, Dawid Weiss.** Carrot2. *User and Developer Manual for version 3.9.0-SNAPSHOT.* [Online] [Citace: 1. duben 2015.]

http://download.carrot2.org/head/manual/index.html.

33. **Jain, A. K., Murty, M. N. a Flynn, P. J.** Data clustering: a review. *ACM Computing Surveys* . 1999, 31.

10.1. Seznam obrázků

10.2. Seznam tabulek

10.3. Přílohy

Uložené na CD

- Zdrojové kódy aplikace pro stahování dat z webových stránek.

  Webcrawler.zip

- Zdrojové kódy aplikace pro stahování dat z Facebooku.

  Facebook.zip

- Zdrojové kódy aplikace pro trénování modelu a kategorizaci sentiment analýzy.

  Sentiment.zip