Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

Analýza dat v sociálních sítích

Plzeň, 2016 Zuzana Mikolášová

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracovala samostatně a výhradně s použitím citovaných pramenů.

.....

V Plzni dne 12.5.2016 Zuzana Mikolášová

Abstract

Submitted master's thesis deals with social network Twitter and with analysis of tweets which are about discussed event. The main goal was to create a system for getting tweets about the event and for their ranking which is based on informativness of their content. Another goal was to rank hashtags, nouns, authors and URL which is extracted from the tweets and to present the results to a user. Based on the requirements the system was created for ranking tweets that are about current events or events that are couple years old.

Abstrakt

Předložená diplomová práce se zabývá sociální sítí Twitter a analýzou příspěvků týkajících se diskutované události. Hlavním cílem bylo vytvořit systém použitelný pro stahování příspěvků, které se týkají zadané události, a jejich ohodnocení na základě informativnosti jejich obsahu. Dílčím úkolem bylo stejným způsobem ohodnotit příslušné hashtagy, podstatná jména, autory a URL extrahované z příspěvků a výsledky prezentovat uživateli. Na základě těchto požadavků vznikl systém, který umožňuje získávat a analyzovat příspěvky o událostech jak aktuálních, tak i několik let nediskutovaných.

Obsah

1 Úvod

V posledních letech se nebývalým tempem rozrostlo množství uživatelů, kteří využívají sociální sítě. Uživatelům slouží pro sdílení svých názorů s ostatními, pro interakci s přáteli či pro přehled o různých probíhajících či plánovaných událostech.

Sociální síť, která je nyní na vrcholu v žebříčku mikroblogujících systémů, je Twitter. Uživatelé přispívají na síť krátkými zprávami, kterými se vyjadřují k aktuálním tématům či pomocí nich sdílejí s ostatními zajímavosti ze života. Příspěvky jsou opatřeny tzv. hashtagy, které slouží jako klíčová slova a pomáhají s přiřazením příspěvku k diskutovanému tématu.

Trendem na této sociální síti je mimo jiné i komentování právě probíhajících událostí, ať už jde o tragické události či o sportovní utkání. Díky velkému množství uživatelů je pak velmi pravděpodobné, že informace o průběhu události se na Twitteru objeví mnohem dříve než v oficiálním zpravodajství. Mezi těmito informacemi se ale vyskytuje i velké množství zpráv, kterými uživatelé vyjadřují své pocity ohledně dané události. V rámci příspěvků k danému tématu se zároveň vyskytuje velké množství spamu, který zneužívá klíčová slova pro diskutované téma. Je tedy poměrně obtížné v tomto neustále přibývajícím množství zpráv najít přínosnou informaci, která se týká průběhu události.

Cílem této práce je vyfiltrovat příspěvky, které obsahují informace o probíhajících či již ukončených událostech. Zpracovávané příspěvky jsou zaměřeny na tragické události, ať už se jedná o teroristické útoky nebo o zemětřesení. Výsledkem práce jsou příspěvky ohodnocené dle míry informativnosti, které se týkají dané události. Součástí výstupu jsou také ohodnocené hashtagy, autoři příspěvků, extrahovaná slova a URL.

Diplomová práce přímo navazuje na projekt KIV/OPSWI, který se zabýval výběrem prostředků pro zpracování textu. Základem práce je článek From Chirps to Whistles – Discovering Event-specific Informative Content from Twitter, jehož autory jsou Debanjan Mahata, John R. Talburt a Vivek Kumar Singh [1].

2 Twitter

Aplikace pracuje se sociální sítí Twitter [2]. Tato sociální síť slouží pro spojení se světem pomocí zpráv, které jsou omezeny na délku 140 znaků. Zprávy jsou určeny primárně pro stručná a výstižná sdělení. Součástí zprávy mohou být mimo textu i fotografie, video či odkaz na další zdroj. Tyto zprávy jsou zveřejňovány na profilu přispěvatele a jsou zobrazeny na hlavní stránce všech odběratelů příslušného uživatele. Příspěvky jsou zároveň dostupné i ve vyhledávání.

Twitter není pouze sociální síť pro party kamarádů, či pro udržování rodinných vztahů. Počet uživatelů a dostupnost zpráv je výhodná z hlediska proudu informací. Pokud se ve světě děje nějaká událost, je díky této sociální síti šance, že se zpráva o události dostane k uživatelům dříve než přes běžná média, jakými jsou zprávy v televizi nebo na internetu. Je to mimo jiné dáno přítomností uživatelů Twitteru na místě dění, přičemž právě 140 znaků je velmi rychlou cestou, jak dát vědět světu nejaktuálnější informace z místa dění. V extrémních případech lze tedy např. narazit na zprávy od lidí, kteří se někde ukrývají před teroristy. Než média takovéto zprávy zpracují, většinou se objeví na předních příčkách ve vyhledávání, které se týká dané události, pokud se událost ještě nevyskytuje mezi světově nejdiskutovanějšími tématy (tzv. trendy). Trendy slouží k přehledu nejdiskutovanějších témat na Twitteru v rámci celého světa nebo vybrané lokality. Díky trendům lze zachytit aktuální dění ve světě.

2.1 Twitter API

Twitter API [3] je knihovna funkcí, která umožňuje číst a zapisovat data Twitteru. Pro možnost využití této knihovny je nutné zaregistrovat vyvíjenou aplikaci na stránkách Twitteru. Poté jsou uživateli poskytnuty přístupové klíče. Veřejná Twitter API se skládá ze dvou částí, kterými jsou REST API a Streaming API.

REST API poskytuje přístup k datům na Twitteru. Umožňuje např. vkládání nových příspěvků, vyhledávání na základě dotazu, získání příspěvků daného uživatele nebo získání seznamu jeho odběratelů. Součástí REST API je Twitter Search API. Ta umožňuje vyhledávání příspěvků na základě dotazu. Vyhledávání probíhá mezi nejpopulárnějšími příspěvky, které jsou staré nejvýše 7 dní. Search API je zaměřena na relevanci, proto některé příspěvky nebo uživatelé nejsou do výsledků zahrnuty.

Streaming API umožňuje čtení příspěvků v reálném čase. Twitter poskytuje několik datových toků. Jedná se o veřejný stream, který je vhodný pro odchytávání příspěvků týkajících se aktuálního tématu. Dalším je uživatelský stream, obsahující příspěvky účtů, které přihlášený uživatel odebírá. Posledním streamem je tzv. „site stream“, který je určen webovým stránkám, přes které se na Twitter přihlašuje větší počet uživatelů. Tento přístup neumožňuje využití klíčových slov (narozdíl od veřejného streamu). Pro tento přístup je nutné se přihlásit.

Komunikace API s Twitterem funguje na základě HTTP dotazů. Výsledky jsou ve formátu JSON. API nabízí dva druhy přístupu: uživatelský a tzv. „application-only“. Pro oba typy je používáno ověření pravosti přes OAuth protokol. Při využití „application-only“ přístupu není k dispozici kontext připojeného uživatele. Z tohoto důvodu je zablokována možnost zasílání statusů. Také není přístupné vyhledávání uživatelů či připojení ke streamu.

2.2 Twitter Rate Limit

Z důvodu omezení zneužití či zasílání spamu na sociální síť jsou zavedena některá omezení. Ta jsou vázána na přístupové klíče. Omezení jsou rozdělena dle typů přístupu na omezení v rámci aplikace a na jednoho uživatele. Limity jsou definovány jako maximální možný počet dotazů za 15 minut. Veškerá omezení se dají rozdělit do dvou základních skupin. Nejvyšší možný počet dotazů během 15 minut je většinou 15 nebo 180. Přesná omezení jsou k dispozici v dokumentaci API [4].

Běžné vyhledávání příspěvků či informací o konkrétních uživatelích, je omezeno na 180 dotazů. Pokud jsou tedy hledány tweety na základě klíčových slov, výsledkem bude max. 180 příspěvků. Pro méně běžná vyhledávání, kterými jsou získání např. seznamu odběratelů (seznam konkrétních účtů) či nejdiskutovanějších témat, je počet dotazů omezen na 15.

Pro Twitter Streaming API, která slouží k odchytávání nových příspěvků, je limitován počet připojení. Při překročení limitu je připojení na několik minut zablokováno. Konkrétní časové údaje v tomto případě nejsou zveřejněny.

V rámci Twitter API nyní nelze stáhnout příspěvky starší než několik dnů. Toto omezení je v programu kvůli vyhledávání ošetřeno externí knihovnou GetOldTweets.

3 Data

Práce je zaměřena na hodnocení tweetů, které se týkají tragických událostí. Jedná se o události přírodního charakteru (např. zemětřesení), ale i o události, které způsobil člověk (teroristický útok). V rámci tohoto tématu jsou na internetu [5] k dispozici anotovaná data a slovník klíčových slov, se kterými se bude nadále pracovat. Jedná se o celkem 26 různých událostí, které se udály v letech 2012-2013. V těchto událostech jsou zahrnuty např. výbuchy v Bostonu nebo tajfun Pablo. Stažené tweety jsou uloženy v jednotlivých složkách, které obsahují dále popsané soubory.

V první řadě složka obsahuje soubor .json, který obsahuje klíčové informace o celé kolekci tweetů. Lze zde najít počet stažených tweetů, den začátku události, lokalitu, klíčová slova a kategorii (např. událost byla zaviněna člověkem).

Dále je obsažen první soubor .csv, který obsahuje čas zveřejnění tweetu a jeho id. Další soubor .csv obsahuje

- **ID tweetu** – pro přiřazení tweetu k času zveřejnění,
- **text tweetu** – celé znění tweetu,
- **zdroj** – zdroj tweetu, použity jsou následující možnosti:
    - **Business (podniky)** – tweety zveřejněny různými podniky (např. vyjádření soustrasti oficiálním profilem NHL).
    - **Eyewitness (očitý svědek)** – informace v tweetu pochází od očitého svědka.
    - **Government (vládní organizace) –** zdrojem informace je vládní organizace (např. policejní oddělení).
    - **Media (média)** – zdrojem jsou profesionální média (např. CNN, BBC,…).
    - **NGOs (nevládní organizace)** – zdrojem textu tweetu je nevládní organizace (např. Červený kříž).
    - **Outsiders (ostatní) –** ostatní uživatelé, reagující na katastrofu, která se jich bezprostředně netýká.
    - **Not applicable (neaplikovatelné) –** používá se u tweetů, které jsou psány v jiném jazyce, než je angličtina.
- **typ tweetu –** typ tweetu je dán obsahem, který zpráva nese. Mohou to být doporučení pro okolí, vyjádření emocí, apod.
    - **Affected Individuals (uživatelé, jichž se událost týká)** – zprávy o osobách, jež jsou událostí postiženy (např. oběti, zranění).
    - **Caution and advice (upozornění a rady) –** Jedná se např. o informace k případné evakuaci, informaci o podezřelých, apod.

  - **Donations and volunteering (dárcovství a dobrovolnictví) –** tweet nese informaci o krizových kontech, potřebě dárců krve či o možnosti pomoci přímo na místě (shromažďování zásob po katastrofě).
  - **Other Useful Information (další praktické informace)** – tweet obsahuje např. videa, fotky a jiné zprávy (např. výbuchy v Bostonu – tweet o složení bomb)
  - **Infrastructure and utilities (infrastruktua a užitek)** – tweet nese informace o dění v okolí, např. o evakuacích, o místě exploze či uzavírkách
  - **Sympathy and support (soucit a podpora) –** tyto zprávy obsahují vyjádření emocí, týkající se dané události. V tomto případě se tedy jedná hlavně o vztek či smutek.
  - **Not applicable (neaplikovatelné) –** používá se u tweetů, které jsou psány v jiném jazyce, než je angličtina.
- **míra informativnosti –** ukazuje, zda tweet obsahuje informace
  - **Related and informative (informativní) –** tweet se týká dané události a obsahuje relevantní informaci (např. počet obětí, lokalitu události,…)
  - **Related – but not informative (týká se události, ale neinformativní) –** tweet se týká dané události, ale neobsahuje žádnou přínosnou informaci. V tomto případě se jedná o vyjádření emocí, jež se týkají dané události, např. vyjádření soustrasti.
  - **Not related (netýká se události, neinformativní) –** tweet se dané události vůbec netýká. Jedná se o spamy nebo o tweety uživatelů, které mají za cíl je zviditelnit.
  - **Not applicable (neaplikovatelné) –** používá se u tweetů, které jsou psány v jiném jazyce, než je angličtina.
- **počet lidí, kteří tweet označili jako oblíbený,**
- **počet retweetů,**
- **informace o zdroji tweetu – zdali je účet ověřen,**
- **počet odběratelů uživatele,**
- **jméno uživatele, který tweet napsal.**

V další práci jsou využívány všechny informace o příspěvku kromě typu příspěvku a informace o zdroji tweetu.

4 Zpracování textu

Před detailním výpočtem ohodnocení příspěvku dle míry informativnosti je potřeba stanovit inicializační hodnotu pro každý příspěvek. Pro tuto práci jsou k dispozici vhodná anotovaná data, která slouží ke klasifikaci textu [6][7] . Díky anotovaným datům, pomocí nichž je klasifikátor trénován, lze následně hodnotit i příspěvky stažené uživatelem. V první části je pro seznámení s problematikou uvedeno jedno ze základních dělení metod učení. Druhá část této kapitoly se zabývá třemi uvažovanými způsoby klasifikace. Zvolenou metodou je logistická regrese.

4.1 Metody učení

Dle typu dostupných vstupních dat je rozlišováno několik různých metod učení. Rozdílem mezi těmito metodami jsou právě různá vstupní data.

4.1.1 Učení s učitelem

Učení s učitelem zahrnuje metody, které mají k dispozici data, u nichž je označena příslušnost k třídě. Těmto datům se říká trénovací data. Metody tohoto typu učení jsou nazývány klasifikace. Klasifikátor je natrénován pomocí trénovacích (anotovaných) dat. Samotná klasifikace potom u vstupních testovacích dat stanovuje pravděpodobnost příslušnosti dokumentu k dané třídě.

Vzhledem k dostupnosti množiny plně anotovaných dat, je v rámci této práce použita klasifikace.

4.1.2 Učení bez učitele

Metody učení bez učitele jsou využívány v případech, kdy není k dispozici žádná trénovací množina, jsou k dispozici pouze vstupní data. Základem je předpoklad, že dokumenty patřící do jedné třídy jsou si podobnější než dokumenty z různých tříd. Cílem je tedy najít strukturu ve vstupních datech a vzájemné podobnosti, pomocí kterých jsou dokumenty roztříděny do jednotlivých shluků. Metoda, která se používá pro seskupení vstupních dat do tříd, se nazývá shlukování (clustering).

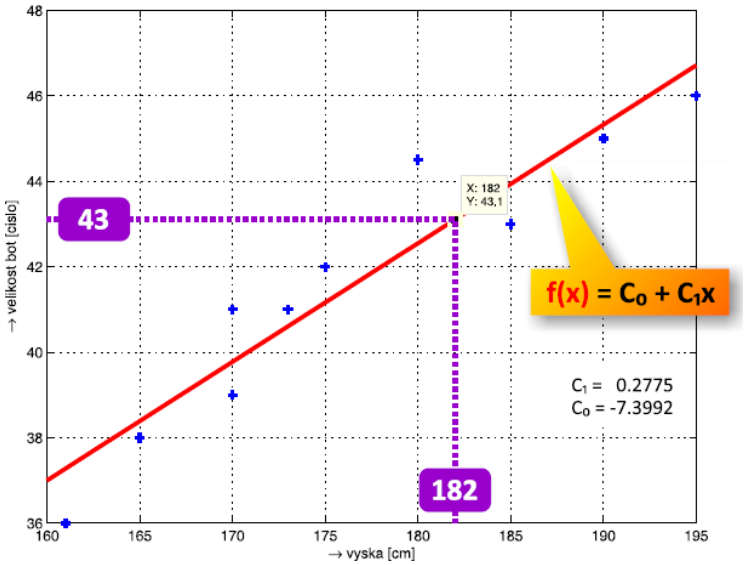
4.1.3 Kombinovaná metoda

Kombinovaná metoda [7] využívá jak anotovaná data, tak data neanotonovaná. Obě tyto množiny jsou využity k natrénování klasifikátoru. Tato metoda je využívána, pakliže je k dispozici malé množství anotovaných dat a velké množství dat neanotovaných. Nejprve jsou k tréninku využita anotovaná data. Následuje klasifikace neanotonovaných dat. Data s nejvyšší pravděpodobností příslušnosti k dané kategorii jsou využita k dalšímu přetrénování klasifikátoru. Metoda se opakuje, dokud není počet neanotonovaných dat minimální. Nevýhodou je časová náročnost metody.

4.2 Vybrané metody klasifikace

4.2.1 Lineární regrese

Lineární regrese [6] je metoda, která se snaží vyjádřit vztah mezi několika proměnnými. Část proměnných je považována za závisloé, zatímco ostatní proměnné jsou považovány za popisné. Cílem lineární regrese je proložit známá data se správnými výsledky lineární funkcí a to tak, aby bylo možné pro libovolnou hodnotu vstupu předpovědět výsledek. Jako příklad lze uvést hledání vztahu mezi váhou a výškou osoby za pomoci lineární regrese.



Obrázek 4.1: Ukázka lineární regrese (převzato z [6])

V rámci lineární regrese jsou možné dva případy, a to:

- Lineární regrese s jednou závislou proměnnou – např. již zmíněný vztah výšky a váhy
- Lineární regrese s více závislými proměnnými – např. vztah mezi váhou, výškou a věkem

Tato metoda je nejjednodušší metodou v rámci učení s učitelem. Pro reálné příklady je prakticky nepoužitelná, velmi dobře však slouží pro pochopení techniky regrese. Poskytuje snadný popis vztahu mezi vstupními a výstupními daty. Pokud je malé množství trénovacích dat, nebo jsou data znehodnocena, je lineární regrese efektivnější než komplikované nelineární metody. Metodu lze také využít pro transformaci vstupních dat.

4.2.2 Lineární regrese s jednou závislou proměnnou

Hypotéza této lineární regrese je reprezentována následujícím vzorcem:

$$h_\Theta(x) = \Theta_0 + \Theta_1 x \tag{4.1}$$

- $h_\Theta(x)$ – hypotéza
- Θ_0 a Θ_1 – volitelné parametry lineární regrese
- x – vstup (proměnná, vektor příznaků)

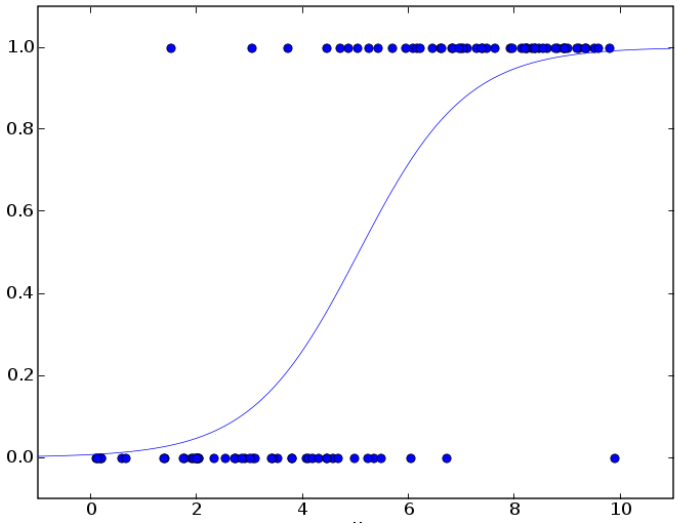
Model závisí na parametrech Θ_0 a Θ_1. Cílem je určit tyto parametry tak, aby hodnoty předpovídané hypotézou pro $x^{(i)}$ z trénovací množiny byly co nejblíže hodnotám $y^{(i)}$. Hledá se tedy minimum cenové funkce $J(\Theta_0, \Theta_1)$.

$$\arg\min_{\Theta_0,\Theta_1}\left(\frac{1}{2m}\sum_{i=1}^{m}\left(h_\Theta\left(x^{(i)}\right) - y^{(i)}\right)^2\right) \tag{4.2}$$

- m – velikost trénovací množiny
- y – výstup (obvykle skalární predikovaná hodnota)

4.2.3 Logistická regrese

Logistická regrese [6] je zobecněný lineární model. Využívá se hlavně v situacích, kdy lineární regrese nestačí. Řadí se mezi pravděpodobnostní klasifikační modely.



Obrázek 4.2: Ukázka logistické regrese (převzato z [8])

Základem modelu je tzv. predikovaná proměnná $y = h_\Theta(x)$. Hypotéza tohoto modelu předpovídá z daného vstupu x pravděpodobnost, že $y = 1$. Logistická regrese se dělí na tři typy dle možných hodnot predikované proměnné:

- **Binární či binomická logistická regrese** – predikovaná proměnná nabývá diskrétních hodnot 0 nebo 1. Jedná se např. o rozpoznávání spamů v e-mailu (je spam/není spam).
- **Ordinální logistická regrese** – proměnná nabývá minimálně tří hodnot, přičemž hodnoty jsou hierarchicky uspořádány. Příkladem je míra souhlasu (ne/nevím/ano).
- **Multinomiální logistická regrese –** predikovaná proměnná nabývá diskrétních hodnot 0 až K. Jako příklad lze uvést lékařskou diagnostiku (chřipka/spalničky/neštovice).

Logistická regrese se používá při modelování pravděpodobnosti výskytu jevu v závislosti na hodnotě spojité proměnné. Výstup má tedy následující vlastnost:

$$y = h_\Theta(x) \in \langle 0; 1 \rangle \tag{4.3}$$

- x – vstup (proměnná, vektor příznaků)
- y – výstup (obvykle skalární predikovaná hodnota)
- $h_\Theta(x)$ – hypotéza

Jelikož je požadována vlastnost $0 \leq h_\Theta(x) \leq 1$, tj. $h_\Theta(x) = P(\Theta^T x)$, má hypotéza tvar

$$h_\Theta(x) = \frac{1}{1 + e^{-\Theta^T x}} \tag{4.4}$$

- Θ^T – transponovaný vektor parametrů, které určují tvar hypotézy

Pravděpodobnost, že výstup $y = 1$ za podmínky x parametrizovaná Θ lze vyjádřit následovně:

$$h_\Theta(x) = P(y = 1|x; \Theta) \tag{4.5}$$

Z toho plyne

$$\begin{gathered} P(y = 0|x; \Theta) + P(y = 1|x; \Theta) = 1 \\ P(y = 0|x; \Theta) = 1 - P(y = 1|x; \Theta) \end{gathered} \tag{4.6}$$

kde Θ je parametr, který je potřeba nastavit tak, aby logistická regrese klasifikovala neznámé vzorky s nejmenší chybou. Je tedy zavedena tzv. funkce Cost, která má následující tvar:

$$\begin{aligned} J(\Theta) &= \frac{1}{m}\sum^{m} Cost\left(h_\Theta\left(x^{(i)}\right), y^{(i)}\right) \\ &= -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} log\left(h_\Theta\left(x^{(i)}\right)\right)\right. \\ &\left. + \left(1 - y^{(i)}\right) log\left(1 - h_\Theta\left(x^{(i)}\right)\right)\right] \end{aligned} \tag{4.7}$$

- m – velikost trénovací množiny

Na tomto modelu strojového učení je založeno mnoho dalších modelů, neboť je velmi populární. Název „logistická regrese“ se používá pouze z historických důvodů, jedná se totiž o klasifikační úlohu.

Pro další práci byla zvolena tato metoda klasifikace, neboť je v porovnání s SVM jednodušší, co se týče multinomiální kategorizace. Zároveň je tato metoda klasifikace použita i ve zdrojovém článku [1]. V případě většího počtu tříd, do kterých je možno dokumenty klasifikovat, je problém převeden na několik binárních problémů.

Pro realizaci tohoto algoritmu je využita knihovna Weka (viz část 4.5.3).

4.2.4 Support vector machine

SVM [6] je metoda strojového učení, jejímž cílem je nalezení optimální rozdělující nadroviny, která separuje jednotlivá trénovací data v prostoru příznaků. To znamená, že jednotlivé body leží na opačných stranách nadroviny a minimum vzdáleností bodů od této nadroviny je co největší. Po obou stranách nadroviny by tedy měl vzniknout

široký pás bez bodů. K popisu nadroviny stačí několik nejbližších bodů – podpůrných vektorů. Metoda je binární, nadrovina v prostoru příznaků představuje lineární funkci.

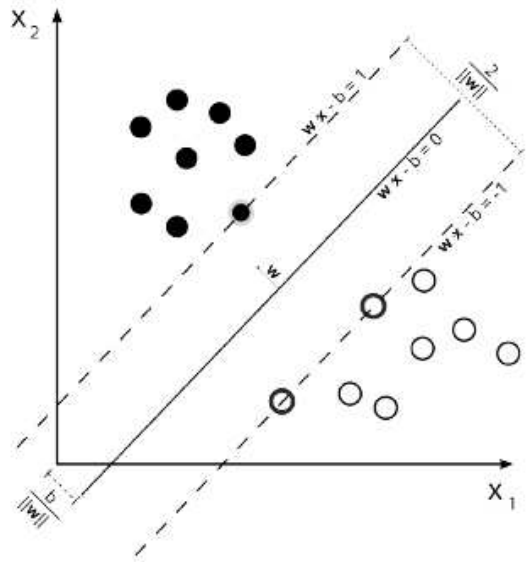
Od binární logistické regrese je tato metoda odlišná. Zatímco logistická regrese vyjadřuje pravděpodobnost výskytu daného jevu, SVM vyjadřuje přímo příslušnost vzorku k třídě 0 nebo 1. Metoda SVM často dosahuje lepších výsledků než např. regrese. Tato metoda je ale komplikovanější, co se týče realizace.

Parametry Θ_i jsou získány optimalizací cenové funkce, která má v tomto případě následující tvar:

$$\min_{\Theta} C \sum_{i=1}^{m} \left[y^{(i)} Cost_1\left(\Theta^T x^{(i)}\right) + \left(1 - y^{(i)}\right) Cost_0\left(\Theta^T x^{(i)}\right) \right] + \frac{1}{2} \sum_{j=1}^{n} \Theta_j^2 \tag{4.8}$$

- $C > 0$ – regularizační faktor, zabraňuje přetrénování
- Θ_i – parametry, určující tvar hypotézy, Θ je vektor
- $y^{(i)} Cost_1\left(\Theta^T x^{(i)}\right)$ – příslušnost klasifikovaného vzorku k třídě 1
- $\left(1 - y^{(i)}\right) Cost_0\left(\Theta^T x^{(i)}\right)$ – příslušnost klasifikovaného vzorku k třídě 0

Hypotéza pak nabývá hodnot $h_\Theta(x) = 1$, pokud je $\Theta^T x \geq 0$ a 0, pokud je tomu jinak.



Obrázek 4.3: Ukázka SVM (převzato z [6])

4.3 Statistiky

Pro vyhodnocení klasifikace existují hodnoty, které ukazují úspěšnost přiřazení hledané proměnné. Jsou jimi přesnost, úplnost a f míra [9][10]. Pro definování přesnosti a úplnosti je nutné zavést následující pojmy:

- true positive (tp) – dokumenty správně označeny jako relevantní,
- true negative (tn) – dokumenty správně označeny jako irelevantní,
- false positive (fp) – dokumenty špatně označeny jako relevantní,
- false negative (fn) – dokumenty špatně označeny jako irelevantní.

4.3.1 Přesnost

Přesnost (precision) představuje procento dokumentů, které jsou správně označeny jako relevantní vzhledem k danému dotazu a to ve vztahu ke všem dokumentům, které jsou označeny jako relevantní. Vztah tedy vypadá následovně:

$$prec = \frac{tp}{tp + fp} \tag{4.9}$$

4.3.2 Úplnost

Úplnost (recall) označuje procento příspěvků, které byly správně označeny jako dokumenty relevantní k dotazu a to ve vztahu ke všem dokumentům, které měly být označeny jako relevantní. Vztah je tedy následující:

$$rec = \frac{tp}{tp + fn} \tag{4.10}$$

4.3.3 F míra

F míra (f measure) představuje harmonický vážený průměr mezi přesností a úplností. Existuje několik podob výpočtu této hodnoty. Nejběžněji používaný vztah vypadá následovně:

$$F = 2 \cdot \frac{prec \cdot rec}{prec + rec} \tag{4.11}$$

4.5 Dostupné knihovny

Pro sestavení klasifikátoru existuje nepřeberné množství knihoven, které jsou více či méně používané či přehledné. Na knihovnu bylo kladeno několik požadavků:

- podporuje programovací jazyk Java,
- je možnost realizace logistické regrese,
- přehledná dokumentace,
- jednoduchá implementace.

Knihovny byly vybrány na základě doporučení na různých diskuzích. Průběžně byly odzkoušeny tři následující knihovny, z nichž se jedna ukázala jako efektivní a jednoduchá na použití, a proto nebyly testovány další knihovny.

4.5.1 LingPipe

LingPipe [11] je nástroj pro zpracování textu. Využívá metody počítačové lingvistiky. Používá se např. pro rozpoznávání jmen nebo při autokorekci. Tato knihovna je psána v programovacím jazyku Java a pro akademické účely je zdarma. Na stránkách jsou k dispozici návody pro realizaci funkcí a obsáhlá dokumentace.

Knihovna obsahuje návod na realizaci klasifikace s přiloženými ukázkovými soubory. Přesto nebyla pro další využití vybrána pro komplikace s nastavením atributů.

4.5.2 Mallet

Mallet [12] je knihovna, která byla v rámci metod pro předzpracování textu zkoumána v KIV/OPSWI. Co se týče klasifikace, je k dispozici metoda maximální entropie.

Před samotnou klasifikací je potřeba data převést do formátu, který vyhovuje klasifikátoru Malletu. Pro tvorbu existuje dokumentace a návody, ze subjektivního pohledu jsou ale nepřehledné. Proto tato knihovna nebyla využita.

4.5.3 Weka

Weka [13] je projektem novozélandské univerzity. Tato knihovna byla v rámci metod pro předzpracování textu zkoumána v předmětu KIV/OPSWI. Co se týče klasifikace, je k dispozici hned několik metod, mezi ně patří např. lineární regrese, logistická regrese, či SVM.

Weka pracuje s daty, která musí být v požadovaném formátu. Formát souboru pro Weka klasifikátor je označen koncovkou .arff. Soubor je rozdělen na hlavičku a tělo.

Hlavička obsahuje názvy atributů a v případě nominálního typu výčet hodnot, kterých může atribut nabývat. Pro výčet hodnot jsou používány složené závorky. Pokud je jako atribut využito libovolné reálné číslo, je označen jako numeric. Na posledním místě se obvykle uvádí atribut, který se klasifikátor učí. V případě učení daného atributu jsou metody rozděleny na dvě skupiny. Jedna část metod vyžaduje po daném atributu, aby byl nominální. To znamená, že atribut je diskrétní.

V těle souboru jsou uvedeny konkrétní hodnoty ve sledu, který je shodný s pořadím uváděných atributů v hlavičce.

Pro tento projekt byla vybrána tato knihovna. Důvodem byla rychlá orientace ve tvorbě souboru pro klasifikátor a v návodech pro realizaci klasifikátoru.

5 Postup při hodnocení tweetů

5.1 Zpracování tweetů

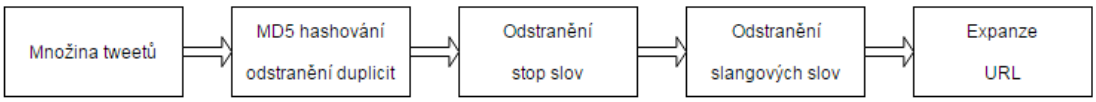
Po klasifikaci je dalším krokem zpracování tweetů [1]. Jedná se o odstranění duplicitních tweetů, odstranění stop slov, odstranění slangových slov a nakonec o expanzi URL. Odstraněním všech slov je vylepšen výsledek následné extrakce klíčových slov z textu pomocí analýzy slovních druhů v textu (POS-tagging).

Prvním krokem při přípravě příspěvků je odstranění tweetů, jejichž text je naprosto shodný. Tento krok je realizován z důvodu urychlení programu, eliminace spamu a odstranění duplicitních položek při počítání skóre. Duplicity by totiž mohly ovlivnit výsledná čísla. Jejich odstranění je prováděno realizací MD5 hashování a následným porovnáním řetězců mezi sebou. Porovnání textu po hashování je rychlejší, neboť takové řetězce jsou kratší než celé příspěvky.

Dalším krokem při přípravě příspěvků pro další výpočty je odstranění stop slov. Tato slova jsou pro jakékoliv hodnocení zbytečná, neboť nenesou žádnou informaci. Jedná se např. o tzv. funkční slova, v angličtině tedy např. „on“ nebo „the“.

Dalšími slovy, která je třeba z textu odstranit, jsou slova slangová. K tomu dopomáhá slovník slangových slov, který vydala FBI. Tento slovník se zaměřuje přímo na prostředí sociální sítě Twitter.

Posledním krokem je expanze URL adresy. Twitter, z důvodu úspory znaků, totiž transformuje URL adresy na kratší verze, za kterými se skrývají adresy plnohodnotné. Tento krok probíhá zároveň s extrakcí URL z textu příspěvku.



Obrázek 5.1: Postup přípravy příspěvků

5.2 TwitterEventInfoGraph

Obecně lze říci, že informativnost tweetů je definována počtem slov, výskytem URL a faktem, je-li příspěvek napsán uživatelem s vysokým počtem odběratelů. Celá událost E_i, které se tweety týkají, je tedy popsaná následujícími položkami:

- Množinou hashtagů $\left(H_{E_i} = \{h_1, h_2, \dots, h_p\}\right)$, které jsou používány pro anotaci tweetů $\left(\in M_{E_i}\right)$ a týkají se události.
- Množinou textových jednotek $\left(W_{E_i} = \{w_1, w_2, \dots, w_r\}\right)$, které jsou používány pro sdílení textové informace ve tweetech $\left(\in M_{E_i}\right)$ o události. Jedná se o podstatná jména, která byla extrahována z textu příspěvku.
- Množinou uživatelů $\left(U_{E_i} = \{u_1, u_2, \dots, u_s\}\right)$, kteří posílají příspěvky $\left(\in M_{E_i}\right)$ o události.
- Množinou URL $\left(L_{E_i} = \{l_1, l_2, \dots, l_t\}\right)$. Ty odkazují na externí zdroje, které se týkají události. Jsou sdíleny uživateli $\left(\in U_{E_i}\right)$ v příspěvcích $\left(\in M_{E_i}\right)$, které posílají.

Tweet je tedy považován za informativní a týkající se dané události, pokud obsahuje:

- informativní hashtagy, které se týkají události (vysoká frekvence výskytu),
- informativní textové jednotky, které se týkají události (vysoká frekvence výskytu v příspěvcích),
- napsal ho jeden z uživatelů, který je potenciálním autorem informativních příspěvků (vyšší šance informativního příspěvku u uživatelů s vysokým počtem odběratelů),
- informativní URL, které se týkají události (vysoká frekvence výskytu, objevují se spolu s důležitými hashtagy, slovy).

Vztah mezi těmito tzv. informačními jednotkami, které se týkají dané události a tweety, týkající se události E_i, tvoří řetězec vzájemného posílení. Graf $G = (V, D)$ je pojmenovaný jako TwitterEventInfoGraph. Ten reprezentuje vztah, kde $V = M_{E_i} \cup H_{E_i} \cup W_{E_i} \cup U_{E_i} \cup L_{E_i}$ je množina vrcholů a D je množina orientovaných hran mezi různými vrcholy. Když se k sobě navzájem vztahují dva vrcholy různého typu, jsou mezi nimi dvě hrany, každá v opačném směru. Hrany mezi vrcholy ze stejné množiny neexistují. Každá hrana má danou váhu, která ukazuje stupeň asociace mezi dvěma vrcholy.

Hodnoty těchto hran jsou dány následujícími vztahy:

$P(h_i\|w_j) = \frac{n_T h_i \& w_j \text{ se objeví společně}}{n_T w_j \text{se objevuje}}$	$P(w_i\|h_j) = \frac{n_T w_i \& h_j \text{ se objeví společně}}{n_T h_j \text{ se objevuje}}$
$P(h_i\|l_j) = \frac{n_T h_i \& l_j \text{ se objeví společně}}{n_T l_j \text{ se objevuje}}$	$P(l_i\|h_j) = \frac{n_T l_i \& h_j \text{ se objeví společně}}{n_T h_j \text{ se objevuje}}$
$P(h_i\|u_j) = \frac{n_T h_i \& u_j \text{ se objeví společně}}{n_T u_j \text{ se objevuje}}$	$P(u_i\|h_j) = \frac{n_T u_i \& h_j \text{ se objeví společně}}{n_T h_j \text{ se objevuje}}$
$P(w_i\|l_j) = \frac{n_T w_i \& l_j \text{ se objeví společně}}{n_T l_j \text{ se objevuje}}$	$P(l_i\|w_j) = \frac{n_T l_i \& w_j \text{ se objeví společně}}{n_T w_j \text{ se objevuje}}$
$P(w_i\|u_j) = \frac{n_T w_i \& u_j \text{ se objeví společně}}{n_T u_j \text{ se objevuje}}$	$P(u_i\|w_j) = \frac{n_T u_i \& w_j \text{ se objeví společně}}{n_T w_j \text{ se objevuje}}$
$P(u_i\|l_j) = \frac{n_T u_i \& l_j \text{ se objeví společně}}{n_T l_j \text{ se objevuje}}$	$P(l_i\|u_j) = \frac{n_T l_i \& u_j \text{ se objeví společně}}{n_T u_j \text{ se objevuje}}$
$P(h_i\|m_j) = P(m_i\|h_j) = P(w_i\|m_j) = P(m_i\|w_j) = P(u_i\|m_j) = P(m_i\|u_j) = P(l_i\|m_j) = P(m_i\|l_j) = 1.0$	

Tabulka 5.1: Výpočet ohodnocení hran mezi vrcholy různých typů

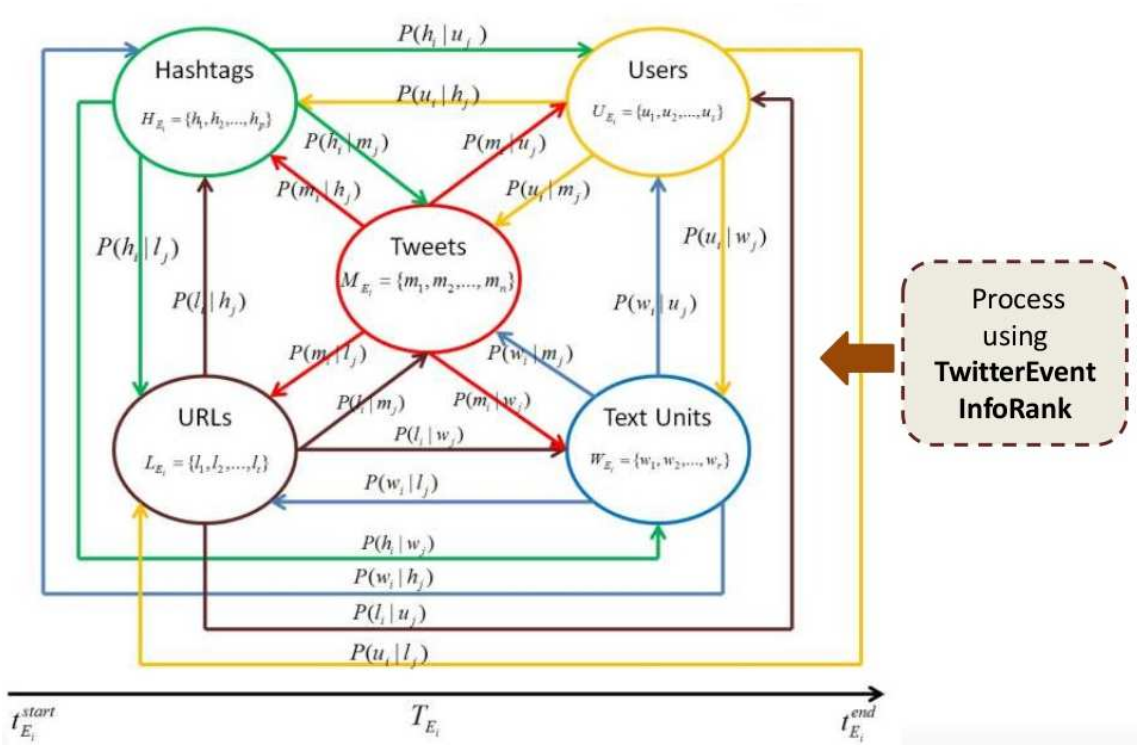
Např. tedy $P(h_i|w_j)$ je pravděpodobnost výskytu hashtagu h_i, pokud se vyskytuje textová jednotka w_j ve streamu tweetů M_{E_i}, které se týkají události E_i. Ty jsou nasbírány v časovém období T_{E_i} a n_T je počet tweetů.

Hodnoty vrcholů grafu jsou dány následujícími výpočty:

$Score(h_i) = \frac{freq(h_i)}{max\{freq(h_1), freq(h_2), \dots, freq(h_p)\}}$	$Score(w_i) = \frac{freq(w_i)}{max\{freq(w_1), freq(w_2), \dots, freq(w_r)\}}$
$Score(u_i) = \frac{followers(u_i)}{max\{followers(u_1), \dots, followers(u_r)\}}$	$Score(l_i) = \frac{freq(l_i)}{max\{freq(l_1), freq(l_2), \dots, freq(l_r)\}}$

Tabulka 5.2: Výpočet počátečního ohodnocení hashtagů, slov, uživatelů, URL

- $freq(h_i)$ – frekvence výskytu hashtagu $(\in H_{E_i})$ na pozici i ve streamu tweetů M_{E_i}
- $freq(w_i)$ – frekvence výskytu textové jednotky $(\in W_{E_i})$ na pozici i
- $freq(l_i)$ – frekvence výskytu URL $(\in L_{E_i})$ na pozici i
- $followers(u_i)$ – počet odběratelů uživatele $u_i \in (U_{E_i})$



Obrázek 5.2: Graf představující zkoumanou událost (převzato z [1])

5.3 TwitterEventInfoRank

Po stanovení hodnot všech vrcholů a hran grafu je potřeba ohodnotit příspěvky podle informativnosti. Vztah mezi jednotlivými vrcholy je dán maticí vztahů. Například $A_{E_i}^{MH}$ označuje matici vztahu $M_{E_i} - H_{E_i}$ pro událost E_i, kde $(i,j)^{th}$ prvek je váha hrany, která označuje stupeň asociace mezi tweetem $(\in M_{E_i})$ na pozici i a hashtagem $(\in H_{E_i})$ na pozici j.

Podobně $A_{E_i}^{WH}$ označuje matici vztahu $W_{E_i} - H_{E_i}$. Tj. ukazuje vztah mezi množinou textových jednotek W_{E_i} a množinou hashtagů H_{E_i} pro událost E_i. Nechť $R_{E_i}^{M}$ označuje ohodnocení množiny příspěvků $(\in M_E)$. Ohodnocení množin vrcholů dalších typů je značena odpovídajícím způsobem. Vztah pro ohodnocení vrcholů různých typů je formulován následovně:

$$
\begin{aligned}
R_{E_i}^{M(k+1)} &= A_{E_i}^{MM(k)} R_{E_i}^{M(k)} + A_{E_i}^{MH(k)} R_{E_i}^{H(k)} + A_{E_i}^{MW(k)} R_{E_i}^{W(k)} \\
&\quad + A_{E_i}^{MU(k)} R_{E_i}^{U(k)} + A_{E_i}^{ML(k)} R_{E_i}^{L(k)} \\
R_{E_i}^{H(k+1)} &= A_{E_i}^{HM(k)} R_{E_i}^{M(k)} + A_{E_i}^{HH(k)} R_{E_i}^{H(k)} + A_{E_i}^{HW(k)} R_{E_i}^{W(k)} \\
&\quad + A_{E_i}^{HU(k)} R_{E_i}^{U(k)} + A_{E_i}^{HL(k)} R_{E_i}^{L(k)} \\
R_{E_i}^{W(k+1)} &= A_{E_i}^{WM(k)} R_{E_i}^{M(k)} + A_{E_i}^{WH(k)} R_{E_i}^{H(k)} + A_{E_i}^{WW(k)} R_{E_i}^{W(k)} \\
&\quad + A_{E_i}^{WU(k)} R_{E_i}^{U(k)} + A_{E_i}^{WL(k)} R_{E_i}^{L(k)} \\
R_{E_i}^{U(k+1)} &= A_{E_i}^{UM(k)} R_{E_i}^{M(k)} + A_{E_i}^{UH(k)} R_{E_i}^{H(k)} + A_{E_i}^{UW(k)} R_{E_i}^{W(k)} \\
&\quad + A_{E_i}^{UU(k)} R_{E_i}^{U(k)} + A_{E_i}^{UL(k)} R_{E_i}^{L(k)} \\
R_{E_i}^{L(k+1)} &= A_{E_i}^{LM(k)} R_{E_i}^{M(k)} + A_{E_i}^{LH(k)} R_{E_i}^{H(k)} + A_{E_i}^{LW(k)} R_{E_i}^{W(k)} \\
&\quad + A_{E_i}^{LU(k)} R_{E_i}^{U(k)} + A_{E_i}^{LL(k)} R_{E_i}^{L(k)}
\end{aligned}
\tag{5.1}
$$

Tento vztah může být reprezentován ve formě blokové matice ΔE_i:

$$
\Delta_{E_i} = \begin{pmatrix}
A_{E_i}^{MM} & A_{E_i}^{MH} & A_{E_i}^{MW} & A_{E_i}^{MU} & A_{E_i}^{ML} \\
A_{E_i}^{HM} & A_{E_i}^{HH} & A_{E_i}^{HW} & A_{E_i}^{HU} & A_{E_i}^{HL} \\
A_{E_i}^{WM} & A_{E_i}^{WH} & A_{E_i}^{WW} & A_{E_i}^{WU} & A_{E_i}^{WL} \\
A_{E_i}^{UM} & A_{E_i}^{UH} & A_{E_i}^{UW} & A_{E_i}^{UU} & A_{E_i}^{UL} \\
A_{E_i}^{LM} & A_{E_i}^{LH} & A_{E_i}^{LW} & A_{E_i}^{LU} & A_{E_i}^{LL}
\end{pmatrix}
\tag{5.2}
$$

Matice, které reprezentují hrany mezi vrcholy ze stejné množiny, jsou nulové, neboť hrany mezi nimi nejsou uvažovány.

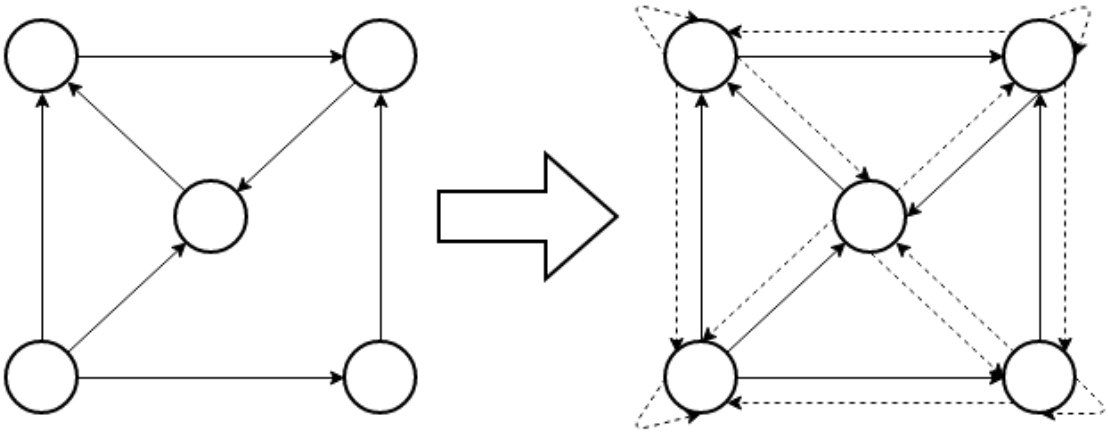
Hodnocení R_{E_i} může být vypočítáno jako dominantní vlastní vektor matice Δ_{E_i}:

$$
\Delta_{E_i} \cdot R_{E_i} = \lambda . R_{E_i} \tag{5.3}
$$

Aby byla garantována unikátní hodnota R_{E_i}, Δ_{E_i} musí být stochastická a nerozložitelná. Aby byla Δ_{E_i} stochastická, je potřeba vydělit každý prvek sloupce matice sumou hodnot všech prvků ve sloupci. Takto ohodnocené hrany představují pravděpodobnost přechodu z jednoho vrcholu na jiný. Poté je matice přeznačena na $\widehat{\Delta}_{E_i}$. Následně je třeba docílit nerozložitelnosti. Graf je potřeba upravit na silně souvislý (viz Obrázek 5.3: Přechod na silně souvislý graf). Tím se zabrání teoretické existenci vrcholů, ze kterých nevede žádná hrana. Tyto vrcholy obsahují pravděpodobnostní vektor p. Takže je matice $\widehat{\Delta}_{E_i}$transformována následujícím způsobem:

$$\overline{\Delta}_{E_i} = \alpha\hat{\Delta}_{E_i} + (1-\alpha)E$$
$$E = p \times [1]_{1\times k} \tag{5.4}$$

Platí, že $0 \leq \alpha \leq 1$a k je rozměr matice $\hat{\Delta}_{E_i}$. Předpokládá se rovnoměrná distribuce, a proto lze nastavit $p = [1/k]_{k\times 1}$. Nyní je $\overline{\Delta}_{E_i}$stochastická a nerozložitelná. Pro matici s těmito vlastnostmi platí, že existuje unikátní vlastní vektor této matice. Jeho prvky jsou nezáporné a jejich součet dává dohromady 1.



Obrázek 5.3: Přechod na silně souvislý graf

Dalším krokem je inicializace vektorů $\left(R_{E_i}^{M(0)}, R_{E_i}^{H(0)}, R_{E_i}^{W(0)}, R_{E_i}^{U(0)}, R_{E_i}^{L(0)}\right)$. Při inicializaci je použito skóre, které je počítáno pro množiny hashtagů, textových jednotek, uživatelů, URL a počáteční skóre příspěvků. Všechny hodnoty leží mezi 0 a 1. Pro tweety je využito ohodnocení logistickou regresí a je přidělena inicializační hodnota, která leží mezi 0 a 1.

Poté je sestaven inicializační vektor

$$R_{E_i}^0 = \left(R_{E_i}^{M(0)}\ R_{E_i}^{H(0)}\ R_{E_i}^{W(0)}\ R_{E_i}^{U(0)}\ R_{E_i}^{L(0)}\right) \tag{5.5}$$

a normalizován tak, že $\left\|R_{E_i}^0\right\|_1 = 1$. Následně je použita power metoda.

5.3.1 Power metoda

Tato metoda je využívána při výpočtu PageRank [14]. Vlastnosti tohoto přístupu jsou následující:

- konverguje k výslednému vektoru, který je unikátní,
- rychlost konvergence je nezávislá na rozměru matice,

- je nutné si pamatovat pouze jeden vektor,
- počítá se s řídkou maticí, ta má malé množství nenulových prvků,
- jednoduchá metoda na realizaci.

Jsou nastaveny následující parametry:

- množina konvergence tolerance: ε = 1e-08,
- $\alpha = 0.85$ – tato hodnota je dána statistikou PageRanku. Bylo zjištěno, že uživatel je ochoten kliknout na 7 různých odkazů, než se přesune jinam nebo zkusí jiný dotaz.

Power metoda probíhá následujícím způsobem:

Inicializuje se počáteční sloupcový vektor $R_{E_i}^0$ (viz výše). Tento vektor je normalizován tak, aby součet v každém sloupci dával dohromady 1, tj.

$$\left\|R_{E_i}^0\right\| = \sum_{i=1}^{n} R_i^0 \tag{5.6}$$

Následně je touto normou vydělen každý prvek vektoru, tj.

$$R_{E_i}^0 = \frac{R_{E_i}^0}{\left\|R_{E_i}^0\right\|} \tag{5.7}$$

Následuje cyklus, ve kterém se opakují následující kroky:

- $R_{E_i}^k \leftarrow \bar{\Delta}_{E_i} R_{E_i}^{k-1}$
- $k \leftarrow k + 1$

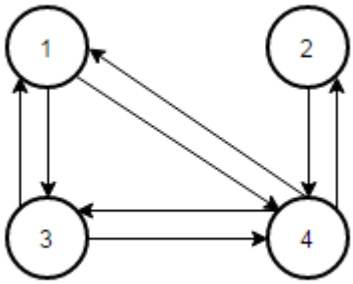
Protože matice $\bar{\Delta}_{E_i}$je stochastická a nerozložitelná a vektor $R_{E_i}^0$ byl normalizován, není potřeba vektor $R_{E_i}^k$ normalizovat. Součet jeho prvků by měl být jedna. Tyto kroky jsou opakovány, dokud neplatí, že $\left\|R_{E_i}^k - R_{E_i}^{k-1}\right\|_1 < \varepsilon$ nebo $k \geq 100$. Platí, že:

$$\left\|R_{E_i}^k - R_{E_i}^{k-1}\right\|_1 = \sum_{i=1}^{n} \left|r_i^k - r_i^{k-1}\right| \tag{5.8}$$

Tímto postupem jsou získány konečné vektory pro každou množinu vrcholů $\left(R_{E_i}^M,\ R_{E_i}^H, R_{E_i}^W, R_{E_i}^U, R_{E_i}^L\right)$. Nakonec jsou tedy získány podmnožiny $\widehat{M}_{E_i}, \widehat{H}_{E_i}, \widehat{W}_{E_i}, \widehat{U}_{E_i}, \widehat{L}_{E_i}$, které obsahují výsledné hodnocení tweetů, hashtagů, textových jednotek, uživatelů a URL. Dle tohoto hodnocení jsou tyto podmnožiny sestupně seřazeny. Čím vyšší je výsledná hodnota, tím vyšší je míra informativnosti, kterou daný vrchol má.

5.4 Příklad

Pro lepší představu aplikace popsaného postupu lze uvést zjednodušený příklad. Následující graf představuje zjednodušený příklad pro výpočet PageRank. V úvahu budou brány pouze ty vrcholy, se kterými bude pokračovat výpočet.



Obrázek 5.4: Ukázkový graf před ohodnocením

Vrcholy představují příspěvky a jednotlivé části, které jsou součástí jednoho nebo obou tweetů. Pro přehled budou uvažovány smyšlené příspěvky v češtině. Jsou uvažovány tedy následující vrcholy:

1. Podezřelí z explozí v Bruselu byl vzat do vazby. #letištěBrusel #exploze
2. Modleme se za pozůstalé v Bruselu. #letištěBrusel
3. #exploze
4. #letištěBrusel

Nejprve jsou příspěvky ohodnoceny klasifikátorem. Jsou tedy nastaveny počáteční hodnoty příspěvků. Pro první příspěvek je hodnota stanovena např. na 0.85 a druhý příspěvek je ohodnocen 0.55. Hodnota vrcholů 3 a 4 je vypočtena dle frekvence výskytu (viz Tabulka 5.2: Výpočet počátečního ohodnocení hashtagů, slov, uživatelů, URL), tedy:

$$freq\ 3 = \frac{1}{2} = 0.5$$

$$freq\ 4 = \frac{2}{2} = 1$$

Dalším krokem je výpočet hran mezi jednotlivými vrcholy. Hrany mezi vrcholy stejného typu nejsou uvažovány, proto nebude počítána hrana mezi uzly 1 a 2. Hrany představují vzájemné odkazy mezi sebou a pravděpodobnost společného výskytu. Výpočet je společný výskyt k celkovému výskytu (viz Tabulka 5.1: Výpočet ohodnocení hran mezi vrcholy různých typů), např.:

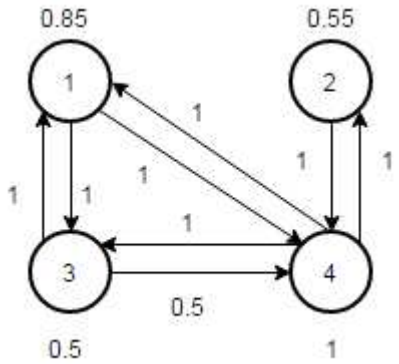
$$P(3|4) = \frac{1(frekvence\ společného\ výskytu\ vrcholů\ 3\ a\ 4)}{2(celková\ frekvence\ výskytu\ vrcholu\ 4)} = 0.5$$

Hrany mají následující hodnoty:

$P(1\|3) = 1$	$P(3\|1) = 1$
$P(1\|4) = 1$	$P(4\|1) = 1$
$P(2\|4) = 1$	$P(4\|2) = 1$
$P(3\|4) = \frac{1}{2} = 0.5$	$P(4\|3) = \frac{1}{1} = 1$

Tabulka 5.3: Praktický příklad výpočtu ohodnocení hran mezi vrcholy

Hodnoty těchto hran představují prvky matic, které představují vztahy mezi podmnožinami události. Pro jednoduchost je předpokládáno, že v tomto případě představují kompletní matici pouze hrany uvedené v tabulce. Graf je tedy celkově ohodnocen následovně:



Obrázek 5.5: Graf po počátečním ohodnocení

Následuje sestavení matice:

$$\Delta_{E_i} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0.5 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

V tomto případě je počet nulových a nenulových prvků shodný. V obecném případě je matice velmi řídká. Dalším krokem je zajištění stochastické a nerozložitelné matice. Matice bude stochastická, pokud sečteme všechny prvky ve sloupci a každý prvek v něm vydělíme touto sumou. Součet v každém sloupci by tedy měl nakonec být 1. Matice získá následující podobu:

$$\bar{\Delta}_{E_i} = \begin{pmatrix} 0 & 0 & 0.5 & 0.4 \\ 0 & 0 & 0 & 0.4 \\ 0.5 & 0 & 0 & 0.2 \\ 0.5 & 1 & 0.5 & 0 \end{pmatrix}$$

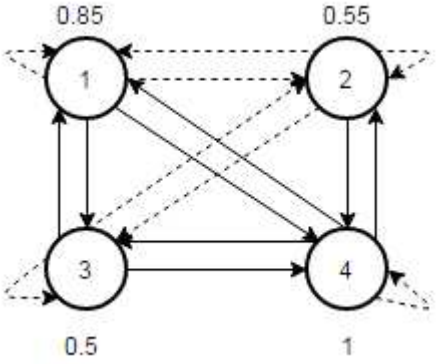
Nakonec je potřeba zajistit, že matice bude nerozložitelná. Toho je docíleno silnou souvislostí matice. K tomu je potřeba zavést parametr $\alpha = 0.85$. Matice je upravena podle následujícího vzorce:

$$\hat{\Delta}_{E_i} = \alpha \bar{\Delta}_{E_i} + (1 - \alpha)E$$

kde E je matice, stejného rozměru jako matice $\bar{\Delta}_{E_i}$ a všechny její prvky jsou rovny $\frac{1}{rozměr\ \bar{\Delta}_{E_i}}$. Každý prvek $\bar{\Delta}_{E_i}$ bude upraven dle rovnice. Matice $\hat{\Delta}_{E_i}$ vypadá následovně:

$$\hat{\Delta}_{E_i} = \begin{pmatrix} 0.0375 & 0.0375 & 0.4625 & 0.3775 \\ 0.0375 & 0.0375 & 0.0375 & 0.3775 \\ 0.4625 & 0.0375 & 0.0375 & 0.2075 \\ 0.4625 & 0.8875 & 0.4625 & 0.0375 \end{pmatrix}$$

Graf vypadá po této úpravě takto:



Obrázek 5.6: Graf po úpravě na silně souvislý

přičemž ohodnocení hran odpovídá výše uvedené matici $\hat{\Delta}_{E_i}$.

Pro aplikaci power metody je kromě matice potřeba inicializovat počáteční vektor. Ten je sestaven z hodnot vrcholů, které jsou dány klasifikátorem a ohodnocením podle frekvence výskytu. Počáteční vektor vypadá následovně:

$$R_{E_i}^0 = \{0.85, 0.55, 0.5, 1\}$$

Po inicializaci je potřeba vektor normalizovat tak, aby součet všech prvků byl 1 (viz rovnice (5.6)). Toho je docíleno jedničkovou normalizací:

$$\left\|R_{E_i}^0\right\| = 0.85 + 0.55 + 0.5 + 1 = 2.9$$

Výslednou sumou je vydělen každý prvek vektoru (dle rovnice (5.7)):

$$R_{E_i}^0 = \{0.2931, 0.1897, 0.1724, 0.3448\}$$

Protože je počáteční vektor normalizovaný a matice je stochastická a nerozložitelná, není nutné normalizovat mezivýsledky. Posledním krokem je násobení matice s transponovaným vektorem. Výsledkem je opět sloupcový vektor, který je znovu násoben maticí. Tento postup je opakován, dokud není splněna jedna ze zastavovacích podmínek. V tomto případě jsou zastavovací podmínky dvě:

1. Počet iterací přesáhne 100.
2. Rozdíl mezi posledními dvěma vektory je menší než 1e-08.

Jako příklad lze uvést první tři iterace:

$$R_{E_i}^1 = \begin{pmatrix} 0.0375 & 0.0375 & 0.4625 & 0.3775 \\ 0.0375 & 0.0375 & 0.0375 & 0.3775 \\ 0.4625 & 0.0375 & 0.0375 & 0.2075 \\ 0.4625 & 0.8875 & 0.4625 & 0.0375 \end{pmatrix} \begin{pmatrix} 0.2931 \\ 0.1897 \\ 0.1724 \\ 0.3448 \end{pmatrix} = \begin{pmatrix} 0.2280020 \\ 0.1547320 \\ 0.2206835 \\ 0.3965825 \end{pmatrix}$$

Nyní je potřeba vypočítat rozdíl mezi vektory (rovnice (5.8)), kvůli případnému splnění jedné ze zastavovacích podmínek. Výsledek tohoto výpočtu je 0.200132.

$$R_{E_i}^2 = \begin{pmatrix} 0.0375 & 0.0375 & 0.4625 & 0.3775 \\ 0.0375 & 0.0375 & 0.0375 & 0.3775 \\ 0.4625 & 0.0375 & 0.0375 & 0.2075 \\ 0.4625 & 0.8875 & 0.4625 & 0.0375 \end{pmatrix} \begin{pmatrix} 0.2280020 \\ 0.1547320 \\ 0.2206835 \\ 0.3965825 \end{pmatrix} = \begin{pmatrix} 0.266128538 \\ 0.172338050 \\ 0.201819875 \\ 0.359713538 \end{pmatrix}$$

V případě této iterace je rozdíl mezi vektory 0.111465175.

$$R_{E_i}^3 = \begin{pmatrix} 0.0375 & 0.0375 & 0.4625 & 0.3775 \\ 0.0375 & 0.0375 & 0.0375 & 0.3775 \\ 0.4625 & 0.0375 & 0.0375 & 0.2075 \\ 0.4625 & 0.8875 & 0.4625 & 0.0375 \end{pmatrix} \begin{pmatrix} 0.266128538 \\ 0.172338050 \\ 0.201819875 \\ 0.359713538 \end{pmatrix} = \begin{pmatrix} 0.245576050 \\ 0.159802603 \\ 0.211755930 \\ 0.382865418 \end{pmatrix}$$

Po poslední iteraci je výsledek rozdílu mezi vektory 0.06617587.

Z ukázky je patrné, že rozdíl mezi vektory se postupně zmenšuje. To znamená, že vektor postupně konverguje k výslednému ohodnocení.

6 Realizace programu

Programová realizace řešení byla rozdělena na několik částí. První a zároveň nejdůležitější částí je soubor tříd pro realizaci výpočtu PageRank a pro tvorbu blokové matice, která se k výpočtu využívá. Další částí jsou různé možnosti získání dat pro vyhodnocení, které jsou realizovány prostřednictvím externích knihoven.

Po získání příspěvků je jejich text zpracován pro klasifikátor. Pro získání některých vlastností příspěvku (analýza výskytu slovních druhů, počet slov) je používána knihovna Stanford CoreNLP [15]. Po zpracování jsou příspěvky ohodnoceny klasifikátorem podle míry informativnosti. Ten je realizován pomocí knihovny Weka (viz část 4.5.3).

Poslední částí byla možnost ukládání dat do databáze, která je realizována pomocí SQLite, a tvorba GUI, jehož cílem je jednoduché ovládání programu.

6.1 Použitý jazyk a software

Mnoho knihoven, které se zabývají strojovým učením a zpracováním textu, je založeno na programovacím jazyku Java. Proto je tato práce realizována v programovacím jazyku Java jdk 1.8. Program byl realizován a testován na operačním systému Windows 7 Ultimate.

V rámci této práce je používáno několik knihoven. Některé byly vybrány v rámci předmětu KIV/OPSWI. Jedná se o následující knihovny:

- *Stanford CoreNLP* – jedná se o knihovnu, která je používána při zpracování textu. Jedná se o počítání slov a identifikaci slovních druhů při klasifikaci a o extrakci podstatných jmen z textu,
- *Twitter4J [16]* – pomocí této knihovny jsou stahovány příspěvky ze sociální sítě Twitter a jsou získávány informace o uživatelském účtu (počet odběratelů a ověření účtu).
- *Weka* – tato knihovna slouží pro realizaci klasifikace příspěvků a pro detekci stop slov v tweetu.
- *Apache commons io [17]* – slouží pro převedení obsahu souboru do pole řetězců.
- *GetOldTweets [18]* – knihovna, která slouží pro získání tweetů, které jsou starší než 7 dní.

- *jlangdetect [19]* – knihovna pro detekci anglického jazyka. Knihovna byla vybrána na základě doporučení.
- *sqlite-jdbc-3.8.11.2* – slouží pro propojení a práci s databází.

6.2 Struktura programu

Program je rozdělen do několika balíků dle funkcí jednotlivých tříd. Diagram reprezentující vazby mezi třídami byl pro přehlednost rozdělen do několika menších diagramů. Všechny tyto diagramy jsou k dispozici v Příloze C.

- **classify**

Balík obsahuje třídy obsluhující klasifikaci příspěvků. Třída *ParseNewData* analyzuje vlastnosti příspěvku pro následující klasifikaci. Obsahuje metody pro zjištění počtu slov, stop slov, citově zabarvených slov a slangových slov. Další metody zjišťují přítomnost URL, počet hashtagů a počet zmíněných uživatelů. Pomocí POS taggeru knihovny Stanford CoreNLP je analyzován počet jednotlivých slovních druhů a určena míra formality textu (viz rovnice (6.1)). Pro klasifikátor je uložen počet retweetů a počet lidí, kteří příspěvek označily jako oblíbený. Výsledek zpracování je uložen do souboru ve formátu pro zpracování knihovnou Weka (viz část 6.5.1).

Druhou třídou je *ClassifyData*, která na daná data aplikuje již připravený klasifikační model. Každému příspěvku pak dle specifických vlastností (viz část 6.5) přiřadí pravděpodobnost příslušnosti k informativním příspěvkům. Platí, že čím je pravděpodobnost vyšší, tím by měl být příspěvek informativnější.

- **comparators**

Comparators je soubor tříd, které obsahují metody pro řazení za různých podmínek. Jako příklad lze uvést třídu *ScoreComparator*, pomocí které jsou řazeny vrcholy podle výsledného skóre pro jejich reprezentaci v tabulce příspěvků. Mezi další patří řazení blokové matie podle sloupců a řazení příspěvků podle výsledného skóre.

- **data**

Balík obsahující objekty použité v programu. Obsahuje třídu *Coordinates*, pomocí které je uchována informace o umístění nenulového prvku v matici. Dalším objektem je *Matrices*, která uchovává dvě matice z důvodu rychlejšího výpočtu ohodnocení hran. Třetím a nejobsáhlejším objektem je *TweetObject*, který uchovává informace o příspěvku včetně řetězce po hashování a původního ohodnocení informativnosti.

Uchovává také informace o autorovi, o počtu retweetů a seznam vrcholů, se kterými je příspěvek svázán. Poslední třídou je třída *Vertex*, která je společná pro všechny množiny kromě příspěvků. Obsahuje informaci o frekvenci výskytu daného řetězce a aktuálním ohodnocení vrcholu. Data souvisí se všemi balíky, diagram by byl nepřehledný, a proto jsou tyto třídy obsaženy v diagramech ostatních balíků.

- **database**

Jak již název napovídá, balík obsahuje třídy pro obsluhu databáze. Databáze slouží pro ukládání informací o zpracovaných událostí. Součástí jsou třídy pro ukládání databáze a pro získávání dat z databáze. Jednou z obsažených tříd je *SaveToDatabase*, která realizuje ukládání dat a vztahů mezi nimi do databáze. Druhou třídou je *GetPreparedData*. Ta slouží k získání dat z již uložené databáze včetně získání jednotlivých vazeb mezi příspěvky a dalšími množinami, které poté lze zobrazit v tabulce. Podrobný popis struktury databáze je v části 6.10.

- **layout**

Balík obsahuje veškeré třídy pro práci s GUI aplikace, včetně hlavní třídy celého programu, kterou je *MainWindow*. Třída obsluhuje hlavní okno, včetně práce s tabulkami, umožnění řazení podle skóre a volání metod pro naplnění tabulek daty. Obsahuje také renderer a editor pro možnost vytvoření tlačítka v tabulce.

- **newTweets**

Tato část programu obsluhuje stahování nových příspěvků z internetu. Ústřední částí je abstraktní třída *GetNewTweets*, která obsahuje získání přístupu k sociální síti Twitter. Zároveň udržuje všechny množiny týkající se zkoumané události a volá veškeré metody pro výpočet výsledného ohodnocení, včetně počáteční klasifikace. Od této třídy dědí celkem třídy dvě.

První z nich je *ListenTweets*, která zajišťuje stahování aktuálních příspěvků pomocí listeneru knihovny Twitter4J. Získané tweety jsou pomocí hashování mezi sebou navzájem porovnány, čímž jsou odstraněny duplicity. Dále jsou zanedbány příspěvky, které nejsou napsány v angličtině. Ostatní příspěvky jsou uloženy do seznamu, který uchovává objekty *TweetObject*. Stahování je ukončeno na pokyn uživatele.

V případě druhé třídy, kterou je *NewTweetsRank* jsou získávány příspěvky týkající se již ukončených událostí. Z důvodu omezení Twitteru, který dovoluje stahovat příspěvky staré pouze několik dní, je využívána externí knihovna GetOldTweets. Pomocí ní získané příspěvky jsou opět roztříděny (viz výše) a výsledek je uložen do seznamu příspěvků.

- **processData**

Soubor všech tříd, které různým způsobem pracují s daným textem. Nejvýznamnější třídou je *ProcessText*, která se stará o získání všech částí příspěvku a výpočet počátečního ohodnocení hashtagů, slov, uživatelů a URL. Obsahuje metody pro extrakci hashtagů, URL a podstatných jmen. Ta jsou získávána pomocí externí knihovny, která je volána ve třídě *RemoveWords*. Tato třída mimo extrakce slov realizuje odstranění zmínek o uživatelích a odstranění stop slov. Dalšími třídami jsou *HashString*, která realizuje MD5 hashování pro rychlejší porovnávání příspěvků mezi sebou, a *ProcessInput*, kde je kontrolována správnost vstupů od uživatele. Poslední obsaženou třídou je *PrintNewData*, kde jsou výsledná data připravována pro odeslání do tabulky.

- **threads**

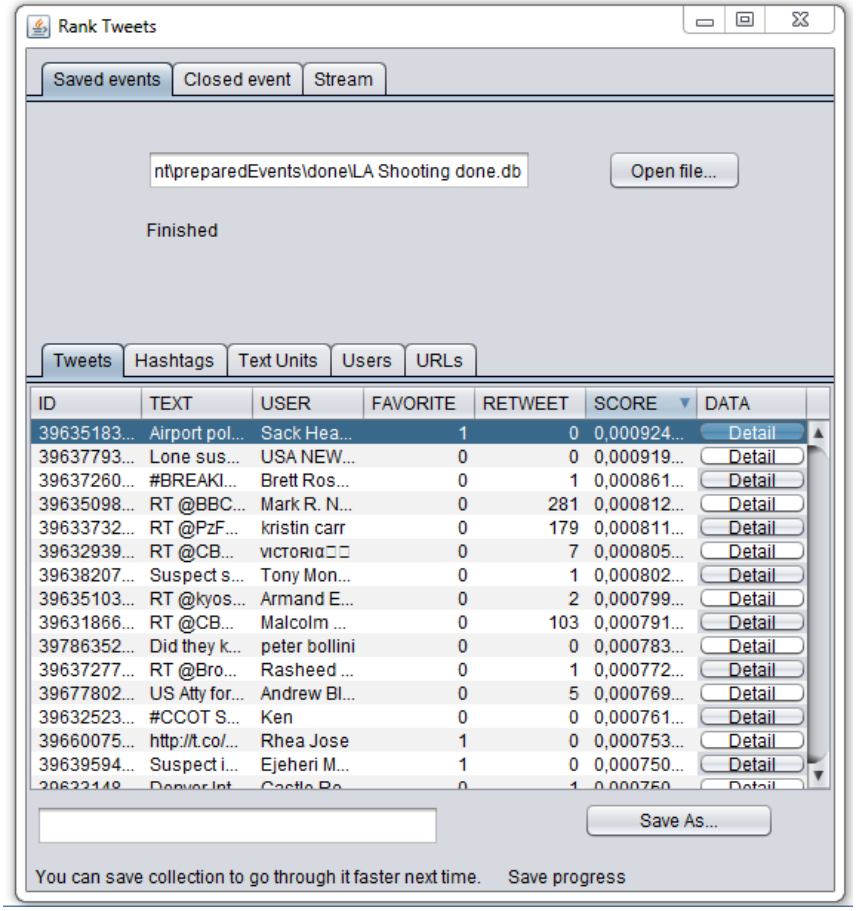
Balík threads obsahuje dvě třídy inicializující vlákna pro zápis do tabulek databáze. Jedna třída realizuje zápis do tabulek s daty, zatímco druhá zapisuje do tabulek vztahů. Před zápisem do tabulek reprezentujících jednotlivé vazby se čeká na ukončení vláken zapisujících do tabulek s daty.

- **utils**

Utils realizuje veškeré výpočty spojené se získáním výsledného ohodnocení příspěvků a dalších množin. Jedná se o třídu *ProcessMatrix* sloužící pro výpočet veškerých matic. Metody realizují výpočet ohodnocení hran mezi jednotlivými vrcholy a zajišťují, že výsledná matice je stochastická a nerozložitelná. Druhou třídou je *ComputeRank*, ve které je realizována power metoda pro výpočet výsledného ohodnocení vrcholu. Je zde zajištěna také počáteční inicializace vektoru a jeho znormování.

6.3 GUI aplikace

Okno aplikace je rozděleno na dvě základní části, kterými jsou získání dat a prohlížení dat (viz obrázek níže). Horní část aplikace umožňuje získat data třemi způsoby: prohlížení uložených příspěvků, získání příspěvků, týkajících se událostí z minulosti a stahování příspěvků, které se týkají aktuálního dění. Každá část má různý počet nutných vstupů. Spodní část aplikace je věnována reprezentaci dat. Tabulky odpovídají množinám, které se týkají dané události. Jedná se tedy o tabulku příspěvků, hashtagů, textových jednotek, uživatelů a URL.



Obrázek 6.1: GUI aplikace

6.4 Získání dat

Uživatel má možnost získat data třemi možnými způsoby: prohlížet data již dříve ohodnocená a uložená, získání příspěvků, které se týkají již ukončené události a odchytávání příspěvků, týkajících se právě probíhající události.

6.4.1 Připravené události

Uživatel má možnost zvolit si některou z již ohodnocených událostí. Jedná se o 26 událostí z let 2012-2013, které jsou připraveny v adresáři *resources/preparedEvents*. Tento adresář je nastaven jako výchozí pro průzkumník souborů, pomocí něhož si uživatel zvolí požadovanou ohodnocenou událost. V tomto případě jsou uživateli zobrazena data, která jsou seřazena dle výsledného hodnocení. Výsledné skóre je uloženo v databázi, opakované zobrazení dat je tedy mnohem rychlejší. Výběr databáze je realizován pomocí průzkumníku. Při zpracování souboru je kontrolován správný formát souboru. Při zjištění problému je uživatel upozorněn pomocí informačního okna.

Data získaná z databáze jsou načtena do příslušných tabulek pojmenovaných podle množin (tweets, hashtags, words, users, URL). Tabulka s příspěvky obsahuje jeden neviditelný sloupec, kde jsou uloženy hashtagy, podstatná jména a URL, které příspěvek obsahuje, jejich frekvenci a skóre. Tyto informace jsou předány do dialogu, který informuje o všech vrcholech dalších typů, které jsou obsaženy v daném příspěvku. Obdobný sloupec obsahuje tabulka uživatelů. V něm jsou uloženy příspěvky, které uživatel napsal.

6.4.2 Ukončené události

Druhou možností je vyhledávání nových příspěvků na základě dotazu a časového rozmezí. Uživatel má možnost zvolit si maximální počet příspěvků, který není garantován. Příspěvky totiž mohou být duplicitní nebo mohou být psány v jiném jazyce. S takovými příspěvky se nadále nepracuje. Před samotnou klasifikací a ohodnocením je potřeba provést kontrolu vstupů od uživatele a možnosti realizace ohodnocení příspěvků. Probíhají následující testy:

- Zadaný počet příspěvků je nezáporné číslo, které je větší nebo rovno 100 (již lze vidět nějaké výsledky, ale ještě není vyčerpán limit pro počet dotazů).
- Maximální délka zadaného tématu je 150 znaků. Minimální počet jsou 3 znaky. 3 znaky už v angličtině dají dohromady jedno slovo, proto je stanovena tato minimální délka.
- Výsledný počet tweetů, které jsou získány na základě dotazu, je větší než 0.

Vyhledání těchto příspěvků probíhá pomocí externí knihovny GetOldTweets, která je založena na čtení json souboru, který je získán po zadání dotazu.

6.4.3 Právě probíhající událost

Třetí a poslední možností je stahování příspěvků, které se týkají právě probíhající události. V tomto případě je kontrolována minimální a maximální délka dotazu (viz výše). Po zadání možnosti stahování aktuálních tweetů se otevře nové okno, kde jsou zobrazeny právě získané příspěvky. Odchytávání nových příspěvků na dané téma je ukončeno až na přání uživatele.

Stahování aktuálních příspěvků probíhá pomocí knihovny Twitter4J. Součástí implementace je listener, který odchytává nově přidané příspěvky.

6.5 Klasifikace příspěvků

Vůbec prvním krokem v tvorbě programu byla klasifikace informativnosti jednotlivých příspěvků na základě několika vlastností tweetu. Pro klasifikaci byla vybrána logistická regrese realizována pomocí knihovny Weka.

Klasifikátor je trénován pomocí kolekce anotovaných dat, která obsahuje tweety o různých událostech z let 2012, 2013. Celkem se jedná o 21 396 příspěvků. Tato data ale byla pro potřeby klasifikace neúplná a bylo potřeba je manuálně doplnit. Jedná se o následující informace:

- počet uživatelů, kteří označili příspěvek jako oblíbený,
- počet retweetů,
- informace o tom, je-li účet ověřen,
- počet odběratelů účtu, ze kterého byl příspěvek odeslán,
- jméno uživatele.

Získání těchto detailů o jednotlivých příspěvcích bylo realizováno pomocí knihovny Twitter4J a ID tweetu, které je dostupné v kolekci anotovaných souborů. Příspěvek je na Twitteru vyhledán dle ID a detaily jsou přidány do csv souboru. Tyto soubory jsou pro možnost prohlédnutí k dispozici v adresáři *resources/csv*.

Cílem klasifikátoru je určit míru informativnosti daného příspěvku. Ta je stanovena pomocí několika vlastností textu. V následujícím výčtu zkoumaných vlastností příspěvku je u každého atributu uvedena jeho váha. Weka stanovila váhy na základě pravděpodobnosti příslušnosti příspěvku k neinformativním. Zkoumané vlastnosti příspěvků jsou následující:

- **Přítomnost odkazu (0.9504)**

  - Vlastnost nabývá hodnot {yes,no}. Text tweetu je porovnáván s řetězcem "http(s)://". Pakliže je tento řetězec v textu obsažen, lze předpokládat, že se v tweetu vyskytuje odkaz.
- **Počet slov (-0.013)**
  - Počet slov je zjišťován pomocí knihovny Stanford CoreNLP a to pomocí tokenizace a anotace textu, kdy jsou vyhledána jednotlivá slova.
- **Počet stop slov (0.1822)**
  - Tato vlastnost je opět realizována pomocí knihovny Weka, kdy je pro každé slovo volána metoda *isStopword(word)*.
- **Počet feeling slov (-0.0915)**
  - Feeling slova jsou taková slova, která v sobě nesou nějakou emoci. Jako příklad lze uvést slovo „devastated". Počet je zkoumán porovnáváním jednotlivých slov se slovy v obsáhlém slovníku [20].
- **Počet slangových slov (0.0393)**
  - Slangová slova jsou počítána stejným způsobem jako feeling slova. Jednotlivá slova v příspěvku jsou porovnávána se slovníkem slangových slov [20].
- **Počet hashtagů (0.3156)**
  - Hashtagy jsou specifické znakem „#" na začátku sekvence. Proto jsou vyhledávány pomocí regulárního výrazu, který má následující tvar: „(#\\w+)\\b".
- **Počet zmíněných uživatelů (-0.0838)**
  - V příspěvku lze zmínit uživatele přidáním znaku „@" před jeho jméno. Princip je tedy shodný s vyhledáváním hashtagů, pouze je upraven regulární výraz: „(@\\w+)\\b".
- **Délka tweetu (-0.005)**
  - Jedná se o celkový počet znaků. Ten lze zjistit zavoláním délky celého příspěvku.
- **Počet unikátních znaků (-0.0699)**
  - Pro zjištění počtu unikátních znaků je zavedeno pole všech znaků, které se mohou v příspěvku objevit. Poté se prochází celý příspěvek a v poli se označují ty znaky, které byly použity.
- **Počet speciálních znaků (0.0615)**
  - Speciálních znaků je poměrně velké množství, a proto je text porovnáván s regulárním výrazem následujícího tvaru: „[^a-z0-9]". Jedná se např. o znaky „$%^&*()", atd.
- **Počet uživatelů, kteří označili příspěvek jako oblíbený (0.018)**
  - Tato informace je součástí kolekce dat. Získá se parsováním posledního úseku.
- **Počet retweetů (0)**
  - Tato informace je součástí kolekce dat. Získá se parsováním posledního úseku. Váha 0 je pravděpodobně dána nahodilostí retweetů. Uživatelé sdílí jak vyjádření emocí, tak informace o události.

- **Typ účtu (0.8388)**
  - Jedná se o to, jestli je účet ověřen nebo ne. Ověřené účty mají tendenci posílat tweety, které v sobě nesou nějakou informaci. Tato informace je součástí kolekce dat. Získá se parsováním posledního úseku.

Následuje počítání slov, které přísluší k různým slovním druhům. Tato informace je získávána pomocí POS taggeru knihovny Stanford CoreNLP.

- **Počet podstatných jmen (-0.0438)**
  - Za podstatná jména jsou považována všechna slova, která jsou označena zkratkou slovního druhu, která začíná na N (nouns).
- **Počet přídavných jmen (-0.0486)**
  - Jako přídavná jména jsou označována ta slova, jejichž označení začíná písmenem J.
- **Počet sloves (-0.3079)**
  - Slovesa jsou taková slova, jejichž označení začíná na V (verbs). Další kategorií jsou modální slovesa, která jsou využívána při stavbě věty v angličtině. Ta jsou označována zkratkou MD.
- **Počet příslovcí (0.0518)**
  - Za příslovce jsou považována všechna slova, jejichž konečné označení začíná písmeny RB.
- **Počet zájmen (0.531)**
  - Zájmena jsou označována zkratkou, která začíná symboly PR nebo WP.
- **Počet citoslovcí (0.4055)**
  - Citoslovce jsou označována zkratkou UH.
- **Počet členů (0.2107)**
  - V angličtině existují celkem tři členy: „a“, „an“ a „the“.
- **Počet předložek (-0.367)**
  - Předložky jsou označovány zkratkou IN.
- **Formalita textu (0.0091)**
  - Formalita textu je zjišťována dle výskytu jednotlivých slovních druhů. Tato hodnota je počítána dle následujícího vzorce:

$$Formalita = (podst.jména + příd.jména + předložky + členy - zajména - slovesa - příslovce - citoslovce + 100)/2 \quad (6.1)$$

Po natrénování je klasifikátor schopen přiřadit příspěvku hodnotu odpovídající pravděpodobnosti, s jakou příspěvek patří mezi informativní tweety týkající se vybrané události.

6.5.1 Vstup a výstup klasifikátoru

Vstupem klasifikátoru je soubor s koncovkou .arff. Jeho obsahem jsou hlavička a data. V hlavičce jsou definovány všechny atributy, představující vlastnosti příspěvků (např. je-li obsažena URL, počet slov, počet retweetů) a typ dané vlastnosti. Je-li vlastnost nominal, je součástí popisu atributu výčet hodnot. V případě URL vypadá popis atributu následovně:

```
@attribute url {yes,no}
```

Druhou částí souboru jsou data. Každá řádka odpovídá jednomu příspěvku, hodnoty jednotlivých atributů jsou od sebe odděleny čárkou. Poslední atribut, jehož hodnota je výsledkem klasifikace, je pro vstupní soubor nahrazen hodnotou 0. Řádka reprezentující příspěvek má ve vstupním souboru podobu

```
yes,6,0,0,51,4,0,1,0,0,0,0,0,1,1,1,61,33,8,0,5,false,0
```

přičemž jednotlivé hodnoty v tomto pořadí představují přítomnost URL, počet slov, stop slov, citově zabarvených slov, míra formality textu, počet podstatných jmen, přídavných jmen, sloves, příslovcí, citoslovcí, počet členů, předložek, počet slangových slov, hashtagů, zmíněných uživatelů, počet znaků, unikátních znaků a speciálních znaků, počet lidí, kteří příspěvek označili jako oblíbený, počet retweetů, informace o ověřeném účtu a budoucí informativnost.

Výstup klasifikátoru se od vstupu liší pouze v poslední hodnotě, kde se místo anotované hodnoty či automaticky přiřazené hodnoty 0 objeví hodnota pravděpodobnosti, se kterou je daný příspěvek informativní.

6.6 Zpracování příspěvku

Před samotným výpočtem skóre jednotlivých příspěvků je potřeba jednotlivé příspěvky připravit [1]. Prvním krokem je detekce anglického jazyka. To je realizováno pomocí externí knihovny jlangdetect. Pokud je příspěvek napsán anglicky, je dalším krokem odstranění duplicitních tweetů. To je realizováno pomocí MD5 hashování. Takto vzniklé řetězce jsou kratší než samotné tweety. Porovnání těchto řetězců urychlí vzájemné porovnání. Platí totiž, že pokud jsou příspěvky shodné, budou shodné i po hashování.

Dalším krokem je odstranění stop slov a speciálních znaků. Odstranění stop slov je realizováno pomocí knihoven Stanford CoreNLP a Weka, kde dojde nejprve k rozpoznání jednotlivých slov a následně určení, zdali jsou daná slova stop slova nebo

ne. Pokud ano, nebudou do výsledného řetězce uložena. Odstranění speciálních znaků probíhá pomocí porovnání řetězce s regulárním výrazem.

Třetím krokem je odstranění slangových slov z příspěvku. K tomuto kroku je k dispozici slovník slangových slov na Twitteru, který dala dohromady FBI. Pokud je zkoumané slovo nalezeno ve slovníku, není uloženo do výsledného řetězce.

Posledním krokem je expanze URL adresy. To je zapotřebí kvůli unikátnosti jednotlivých adres. Je toho docíleno pomocí HttpURLConnection. Pro zrychlení tohoto přístupu je zakázáno automatické přesměrování (takové, jaké dělá prohlížeč) a umístění zkrácené adresy je extrahováno z HTTP hlavičky. Tento přístup je většinou potřeba provést dvakrát. Za zkrácením pomocí Twitteru se obvykle skrývá ještě jedno zkrácení pomocí jiné služby.

6.7 Získání vrcholů

Před samotným sestrojením grafu je potřeba získat vrcholy všech podmnožin týkající se dané události. Následující procesy jsou realizovány pro všechny příspěvky. Při nalezení každého nového vrcholu je příslušná množina testována na přítomnost daného vrcholu. Pokud uzel již existuje, zvýší se frekvence o 1 a daný příspěvek je přidán do seznamu výskytu. Pokud neexistuje, je vytvořen nový uzel, kterému je nastavena frekvence výskytu 1.

První množinou jsou URL, pro jejichž získání z příspěvku je použit regulární výraz, který má následující tvar:

\\(?\\b(https?://|www[.])[-A-Za-z0-9+&@#/%?=~_()|!:,.;][-A-Za-z0-9+&@#/%=~_()|]*

Následuje přes HttpUrlConnection přístup dvojí expanze adresy (viz výše).

Další množinou vrcholů, se kterou se bude nadále pracovat, jsou hashtagy. Ty jsou z příspěvku extrahovány pomocí regulárního výrazu, který vypadá následovně:

(#\\w+)\\b

Další množinou jsou autoři jednotlivých příspěvků. Uživatel, který daný příspěvek napsal, je získán spolu s tweetem pomocí externí knihovny Twitter4J.

Poslední částí příspěvku, kterou je potřeba získat, jsou textové jednotky. Za slova, která nesou informaci, jsou považována všechna podstatná jména, která se v příspěvku vyskytují. Ta jsou extrahována pomocí externí knihovny Stanford CoreNLP. Využívá se konkrétně POS tagger, který identifikuje slovní druh každého slova. Podstatná jména jsou přidána do seznamu textových jednotek, týkajících se hodnocené události.

6.8 Realizace grafu

Realizace samotného grafu [1] je rozdělena do několika částí. V průběhu zpracování tweetů jsou ukládány důležité části tweetu, které tvoří jednotlivé vrcholy grafu. Jedná se o hashtagy, uživatele, url a textové jednotky. Všechny tyto položky jsou ukládány do vlastních HashMap, které v sobě nesou informaci o řetězci, který danou položku tvoří a o frekvenci, která je zvednuta o jedna pokaždé, když se daná položka u tweetu vyskytuje. Zároveň jsou průběžně ukládány identifikátory příspěvků, ve kterých se položka objevila.

6.8.1 Ohodnocení vrcholů

Po průchodu všemi příspěvky proběhne výpočet skóre pro každou položku ve všech množinách týkajících se dané události. Ty jsou uloženy v jednotlivých HashMapách. Výpočet má následující tvar:

$$položka_i = \frac{frekvence\ položky}{maximální\ frekvence\ v\ HashMap} \tag{6.2}$$

Z tohoto vztahu je patrné, že před samotným výpočtem je nutné projít celou „HashMap“ a najít nejvyšší frekvenci výskytu. Poté následuje druhý průchod, kdy je vypočítáno skóre každé položky dle daného vztahu.

Počáteční ohodnocení příspěvků je provedeno pomocí klasifikátoru, který každému příspěvku přiřadí hodnotu, která představuje pravděpodobnost, s jakou je daný příspěvek informativní. Pro další výpočet se uvažují všechny ohodnocené příspěvky. Pro přehled je ale stanoveno, že pokud je přiřazená hodnota menší než 0.3 je příspěvek považován za neinformativní. Příspěvky ohodnoceny číslem větším než 0.7 jsou uvažovány jako informativní.

6.8.2 Ohodnocení hran

Při ohodnocení hran se vždy prochází dvě množiny. Porovnávají se seznamy výskytů jednotlivých prvků v tweetech mezi sebou. Pokud se dva prvky těchto množin vyskytly v jednom příspěvku, hodnota hrany ukazuje pravděpodobnost, s jakou se vyskytují dva vrcholy spojené touto hranou (např. pravděpodobnost výskytu slova „sun“, pakliže se v příspěvku vyskytl hashtag „morning“ a naopak). Ohodnocení hrany tedy náleží do intervalu $\langle 0; 1 \rangle$. Výsledkem je matice o velikosti $seznam_1 \times seznam_2$.

Výpočet ohodnocení hran je proveden pro všechny následující seznamy navzájem (hashtagy, textové jednotky, uživatelé, url). Při ohodnocení hran mezi příspěvky a vrcholy dalších typů mají hrany hodnotu 1.0, pokud se daný vrchol (uživatel/textová jednotka/hashtag/url) vyskytuje v příspěvku. V případě dvou totožných seznamů se hrany mezi nimi neuvažují. Tyto pomyslné matice jsou tedy nulové. Výsledné matice jsou velmi řídké, většina prvků má hodnotu 0. Proto jsou uloženy do seznamů, které obsahují souřadnice výskytu nenulového prvku a hodnotu tohoto prvku.

Takto vznikne celkem 25 matic (z toho 5 nulových), které jsou základem pro blokovou matici. Bloková matice je čtvercová a řídká a je využívána při výpočtu ohodnocení. Bloková matice je vytvořena spojením všech seznamů do jednoho. Takto vytvořený seznam prvků je pro první krok seřazen po sloupcích.

6.9 Výpočet ohodnocení

Po sestavení blokové matice a počátečním ohodnocení vrcholů grafu je realizována power metoda [1][22]. Před samotnou realizací cyklu je potřeba blokovou matici upravit tak, aby byla stochastická a nerozložitelná (viz část 5.3). Díky této úpravě nebude nutné v rámci metody normalizovat vektor po každém násobení. Pro docílení stochastické matice jsou potřeba dva průchody maticí. Během prvního průchodu jsou vypočteny hodnoty jednotlivých hran a jsou připočteny k sumám pro jednotlivé sloupce. Během druhého průchodu je každý prvek vydělen příslušnou sumou.

Nerozložitelnost je vyřešena při dělení prvků danou sumou, kdy je na každý prvek aplikován vzorec (viz rovnice (5.4)). Tento výpočet je aplikován pro nenulové hodnoty. Nulových hodnot je velké množství, proto nejsou uchovány. Při násobení matice a vektoru je to vyřešeno vynásobením prvku vektoru hodnotou $(1 - \alpha)(\frac{1}{k})$.

- α – 0.85
- k – rozměr matice

Následuje násobení matice s vektorem. To je realizováno průchodem všech možných prvků matice. Pokud v seznamu nenulových prvků existuje prvek s danými souřadnicemi, vynásobí se s příslušným prvkem vektoru. Pokud neexistuje, je příslušný prvek vektoru vynásoben hodnotou, která je uvedena výše.

Jakmile je splněna některá ze zastavovacích podmínek (většinou rozdíl vektorů je menší než 1e-08), je získán vektor ohodnocení jednotlivých vrcholů grafu (příspěvků, hashtagů, textových jednotek, uživatelů a url). Jako test správnosti výsledku lze sečíst hodnoty všech prvků výsledného vektoru. Tento součet by měl dát dohromady přibližně 1. Po dokončení této metody jsou připravena data pro možnost uložení do databáze.

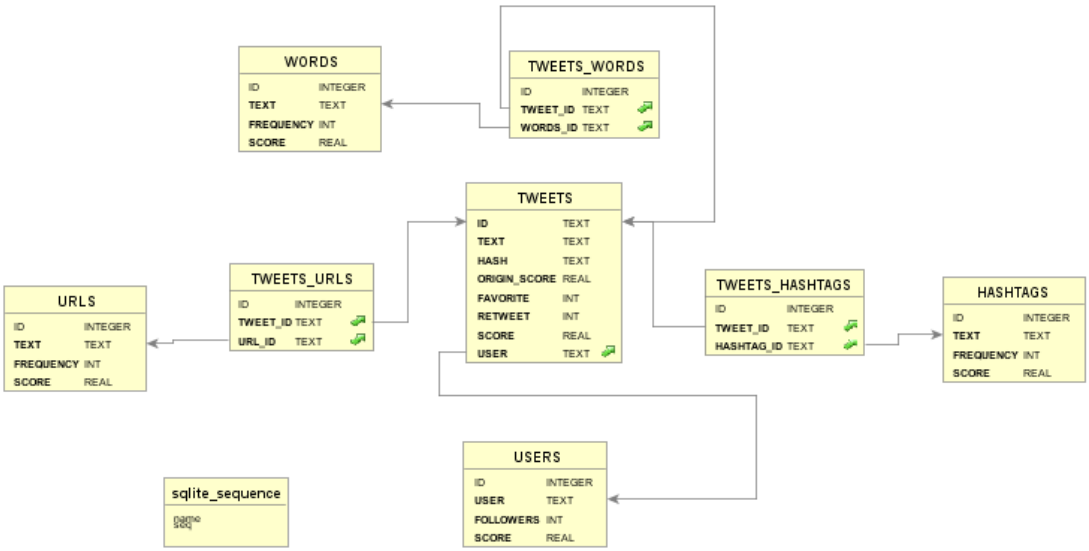
6.10 Zpracování dat

Uživateli je umožněno výsledek uložit. Pro uložení slouží průzkumník, kde si uživatel vybere složku a pojmenuje databázi. Protože je tlačítko pro uložení viditelné i při prohlížení uložených databází, je jeho funkčnost závislá na vybraném typu prohlížení tweetů a na počtu řádků v tabulce příspěvků (musí být větší než 0). Data jsou ukládána v SQLite databázi. Pro každou zpracovanou událost je vytvořena nová databáze. Ta obsahuje několik tabulek s daty:

- *Tweets* – tabulka, ve které jsou uloženy základní informace o tweetu. Jsou uloženy informace, jakými jsou ID příspěvku, text, počet lidí, kteří označili příspěvek jako oblíbený, počet retweetů, hodnocení logistickou regresí, řetězec po hashování a výsledné skóre tweetu. Autor příspěvku je použit jako cizí klíč a odkazuje na uživatele v tabulce *Users*.
- *Hashtags* – tabulka obsahuje text hashtagu, počet výskytů v rámci dané množiny tweetů a výsledné skóre, které ukazuje výslednou významnost.
- *Texts* – tabulka obsahuje všechna podstatná jména, která se v množině příspěvků vyskytla. Jsou uloženy informace o počtu výskytů a výsledné skóre.
- *Users* – tabulka obsahuje jméno uživatele, počet odběratelů na Twitteru a výsledné ohodnocení významnosti uživatele.
- *Urls* – tabulka obsahuje plné znění všech webových adres, frekvenci jejich výskytu a výsledné skóre.

Součástí databáze jsou i relační tabulky, které realizují vztahy mezi příspěvky a dalšími množinami, které definují danou událost:

- *Tweets_Hashtags* – ukazuje vztahy mezi příspěvky a hashtagy. Každá řádka obsahuje ID tweetu a hashtag, který daný tweet obsahuje. Protože každý příspěvek může obsahovat vícero hashtagů a jeden hashtag může být použit ve větším množství příspěvků, jedná se o m:n vazbu.
- *Tweets_Urls* – ukazuje vztah mezi příspěvky a URL. Každý záznam obsahuje ID tweetu a url, kterou daný tweet obsahuje. Protože každý příspěvek může obsahovat více URL a každá URL může mít větší frekvenci výskytu v příspěvcích, jedná se opět o m:n vazbu.
- *Tweets_Words* – obsahuje ID příspěvků a jednotlivá podstatná jména, která se v nich vyskytují. Protože každý příspěvek může obsahovat více podstatných jmen a jedno podstatné jméno se může vyskytovat ve vícero příspěvcích, jedná se o m:n vazbu.



Obrázek 6.2: Struktura databáze

V diagramu (Obrázek 6.2: Struktura databáze) se vyskytuje ještě jedna tabulka, kterou je *sqlite_sequence*. Ta zajišťuje automatické zvyšování hodnot primárních klíčů. Pro rychlejší vkládání do tabulek byla jako primární klíče vybrána po sobě jdoucí čísla.

Centrální tabulkou celé databáze je tabulka s tweety. S tou jsou propojeny všechny ostatní tabulky pomocí odkazu na jednotlivé tweety, kde se dané textové jednotky, hashtagy či URL vyskytují. Také jsou s ní propojeny všechny tabulky představující jednotlivé vztahy mezi příspěvky a ostatními tabulkami. Tabulka uživatelů je s tabulkou příspěvků propojena pomocí cizího klíče v tabulce Tweets.

V připravených databázích se vyskytuje 26 zpracovaných událostí z let 2012 a 2013, které sloužily jako trénovací soubory pro klasifikátor.

7 Vyhodnocení

7.1 Testovací data

Testování proběhlo na 26 souborech, které obsahují příspěvky o různých událostech tragického charakteru. Těchto 26 událostí obsahuje anotované příspěvky s počáteční hodnotou informativnosti nastavenou na 0 nebo 1. Výsledkem celého algoritmu jsou množiny tweetů, hashtagů, textových jednotek, uživatelů a webových adres, které jsou seřazeny dle významnosti v příspěvcích.

Jako ukázku výsledků realizovaného algoritmu lze uvést střelbu na letišti v Los Angeles, která se odehrála 1. listopadu 2013 a vyžádala si jednoho mrtvého a několik zraněných. Tato událost je součástí anotovaných dat. Obsahuje 715 příspěvků z období 1. - 14. listopadu 2013.

V tabulce je u výsledných příspěvků v závorce uvedena hodnota, která byla přiřazena anotátorem. Hodnota 1 znamená, že příspěvek se nevztahuje k události. 2 představuje příspěvek, který se k události vztahuje, ale neobsahuje žádnou užitečnou informaci (vyjádření emocí). Číslem 3 jsou označeny příspěvky, které se týkají události a jsou považovány za informativní.

Střelba na letišti v Los Angeles TOP 5	
hashtagy	#lax, #laxshooting, #tsa, #breaking, #news
textové jednotky	airport, shooting, shooter, tsa, suspect
URL	http://www.cnn.com/2013/11/01/us/lax-gunfire/index.html http://www.facebook.com/ads4LAX http://abc7.com/live http://www.booklinker.net/error.php?e=notfound&type=book https://twitter.com/palomacaraball/status/397854627279802368/photo/1
příspěvky	Airport police engaged and neutralize lone shooter after suspect shot through TSA checkpoint. #BREAKING #LAX #LAXShooting suspect injured (3) Lone suspect in Los Angeles #LAX airport shooting named to US media as Paul Ciancia, 23, by law enforcement officials http://t.co/0LZfFKpYqN (3) #BREAKING: LOS ANGELES (AP) -- Law enforcement officials identify LA airport shooting suspect as 23-year-old Paul Ciancia. #wsbtv (3) RT @BBCBreaking: Man opened fire with assault rifle in security screening area at #LAX Terminal 3, Los Angeles police say http://t.co/swrfb… (3) RT @PzFeed: SHOOTING AT LAX AIRPORT: -2 SUSPECTS -1 SHOT, 1 IN CUSTODY -AT LEAST 2 PEOPLE SHOT -8-10 SHOTS FIRED -SUSPECT HAD A RIFLE -FAA … (3)

Tabulka 7.1: Ukázka nejvýše hodnocených vrcholů podle algoritmu TwitterEventInfoRank

Jak je z tabulky patrné, klíčová slova a hashtagy se týkají typu události (střelba) a lokality. To ukazují i další zpracované události. Další přední příčky zaujímají slova typu „oběť“, „podezřelý“, případně konkrétní jména spojená s událostí. Hashtag #tsa je

spojen s obětí události, která pracovala pro Transportation Security Agency. V případě URL jsou na horních příčkách některé případy spamu. To může být dáno vyšším výskytem URL nebo zneužitím vysokého množství klíčových slov, které příspěvek vyhodnotí jako důležitý. Tyto konkrétní URL jsou navíc svázány s nejvýše hodnoceným vrcholem celého grafu, kterým je hashtag #lax.

Ve výsledcích je také vidět platnost předpokládané vlastnosti informativních tweetů, kterou je délka příspěvku. V následující tabulce, která představuje nejníže hodnocené vrcholy množin, jsou příspěvky většinou kratší než ty, které jsou považovány za informativní. Za příspěvky je opět v závorce uvedena počáteční informativnost stanovená anotátorem.

Střelba na letišti v Los Angeles, nejníže hodnocené	
hashtagy	#info, #worldaffairs, #stayinformed, #nbcn, #sp4zee
textové jednotky	rant, exnsa, hayden, group, responsibility
URL	http://www.radionews.us/2013/11/04/lax-shooting-latest-on-suspect-victims-and-warning-that-may-have-come-too-late/ https://twitter.com/safety/unsafe_link_warning?unsafe_link=http%3A%2F%2Fadf.ly%2FYeBx0 http://trapier.org https://www.yahoo.com/news/gunman-told-police-acted-alone-lax-shooting-174959371.html?ref=gs https://twitter.com/yasirkhanatk/status/397077649568190464/photo/1
příspěvky	@GameHounds is LAX the airport you worked at?? (1) A shooter at LAX … (2) LA airport reopens after shooting http://t.co/Jkpsa3MuWD (3) I'm at Los Angeles International Airport (LAX) - @lax_official w/ @christophechoo http://t.co/f47yS1KQdX (1) RT @nbcnightlynews: LATEST: All shooting victims at LAX were TSA employees; 1 killed and as many as 3 others wounded - @PeteWilliamsNBC (3)

Tabulka 7.2: Ukázka nejníže hodnocených vrcholů podle algoritmu PageRank

První tři hashtagy ve výčtu jsou spojeny s jedním konkrétním příspěvkem, který tyto hashtagy obsahuje. Obecně se na posledních příčkách objevují slova či příspěvky spojená s krátkým vyjádřením soucitu či s konspiračními teoriemi.

7.2 Klasifikátor

Klasifikátor byl realizován pomocí externí knihovny Weka, která mimo jiné umožňuje vyhodnotit výsledky klasifikačního modelu po natrénování. Nejprve je sestaven

klasifikační model podle anotovaných dat a následně je na stejných datech otestována přesnost pomocí pokusu o přiřazení správné hodnoty.

Pro vyhodnocení úspěšnosti klasifikátoru byla použita tzv. křížová validace. Křížová validace rozdělí trénovací soubor na několik množin (v tomto případě 10). Všechny množiny kromě jedné jsou použity pro trénink, přičemž poslední je použita pro vyhodnocení. Stejný postup je aplikován postupně pro všechny množiny. V tomto případě bude tedy každá množina jednou testovací a devětkrát tréninková. Tak vznikne celkem deset klasifikačních modelů, jejichž průměrný výkon je potom prezentován jako celková přesnost, úplnost a f míra. Soubor použitý k natrénování obsahuje údaje o 21 396 příspěvcích, které se týkají 26 různých událostí z let 2012 a 2013.

Klasifikátor zaznamenal 16 025 správně ohodnocených příspěvků. To je 74.8972 %. Neúspěšně ohodnocených příspěvků bylo celkem 5 371, což je 25.1028 %.

Následuje tabulka statistik v rámci příslušnosti jednotlivých příspěvků do tříd.

	Přesnost	Úplnost	F míra
Informativní	0.769	0.858	0.811
Ostatní	0.7	0.563	0.624
Vážený průměr	0.743	0.748	0.741

Tabulka 7.3: Vyhodnocení logistické regrese

7.3 Porovnání algoritmů

Pro správné vyhodnocení úspěšnosti algoritmu slouží tzv. baseline. Baseline jsou další algoritmy vyhodnocení příspěvků, se kterými je zkoumaný postup hodnocení srovnáván. Jako baseline slouží v tomto případě hodnocení pomocí logistické regrese a pořadí dle počtu retweetů. Logistická regrese udává počáteční ohodnocení příspěvků na základě specifických vlastností textu (viz část 6.5). Jako příklad bude sloužit již zmíněná střelba na letišti v Los Angeles. Pro demonstraci je součástí příloh ukázka 10 nejvýše hodnocených příspěvků dle algoritmu TwitterEventInfoRank, logistické regrese a 10 příspěvků s nejvyšším počtem retweetů.

7.3.1 Normalizovaný srážkový kumulativní přírůstek

Srážkový kumulativní přírůstek [1][23] je oblíbeným způsobem měření pro vyhodnocení výsledků vyhledávání na internetu. Je založen na dvou předpokladech:

1. Vysoce relevantní dokumenty jsou užitečnější než okrajově relevantní dokumenty.
2. Čím nižší je pozice ohodnoceného dokumentu, tím je méně užitečný a je nižší šance, že se k němu uživatel vůbec dostane.

Prvním krokem je získání základního hodnocení (1-3) 100 příspěvků, které jsou na nejvyšších příčkách. Toto hodnocení je získáno z anotovaného souboru. Hodnota 1 reprezentuje příspěvky, které nejsou spojeny s událostí. Hodnota 2 je přiřazena k příspěvkům, které se události týkají, ale nenesou žádnou informaci. Číslo 3 je přiděleno příspěvkům, které jsou informativní a týkají se zkoumané události. Dalším krokem je výpočet normalizovaného srážkového kumulativního přírůstku (normalized discounted cumulative gain – nDCG). Výpočet je realizován jako poměr srážkového přírůstku (DCG) a ideálního srážkového přírůstku (IDCG):

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{7.1}$$

kde p je počet příspěvků. IDCG je počítáno dle stejného vztahu jako DCG (viz níže). Pracuje ale s předpokladem, že je ideální zobrazovat nejvíce relevantní dokumenty na prvním místě. Dojde tedy k přeřazení dokumentů od nejrelevantnějších až po ty nejméně relevantní (či irelevantní). Výpočet pak probíhá dle stejného vzorce jako DCG, tj.:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{7.2}$$

kde rel_i je ohodnocení příspěvku (1-3) na pozici i. Další hodnotou, díky které lze porovnávat mezi sebou dva výsledky různých principů, je přesnost. Ta je počítána pomocí následujícího vztahu:

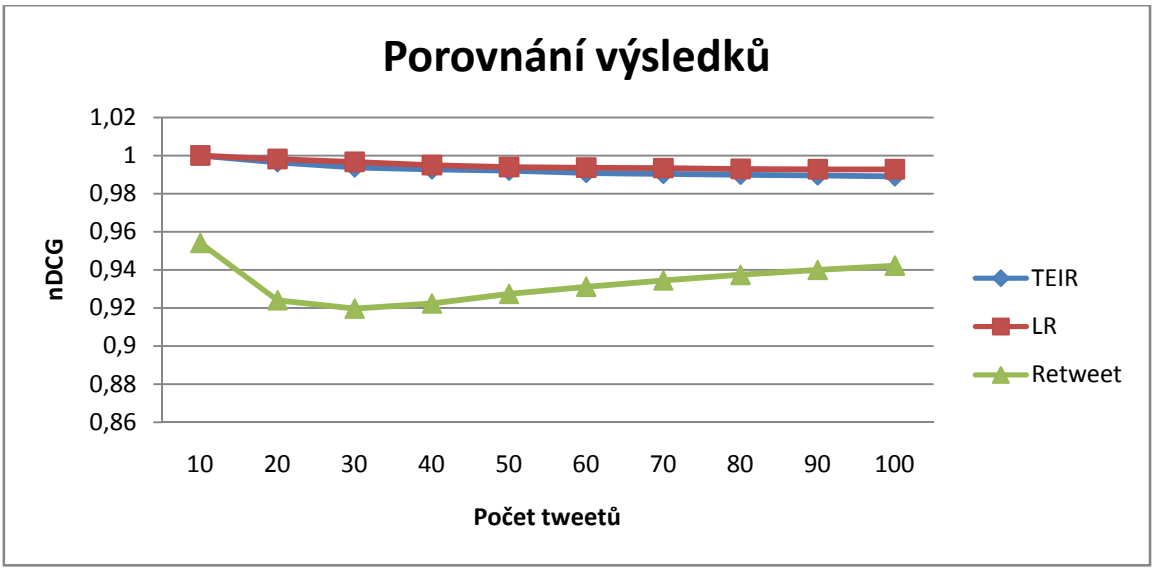
$$přesnost = \frac{počet\ relevantních\ příspěvků\ v\ n}{n} \tag{7.3}$$

kde n je celkový počet příspěvků. Za relevantní příspěvky se považují takové, které jsou s ohodnocením 2 nebo 3.

		Porovnávané způsoby řazení příspěvků					
		TwitterEventInfoRank		Logistická regrese		Řazení podle retweetů	
		NDCG	Přesnost	NDCG	Přesnost	NDCG	Přesnost
Počet příspěvků	10	1.0	100	1.0	100	0.954	100
	20	0.996	95	0.998	100	0.924	100
	30	0.994	96.7	0.997	96.7	0.920	100
	40	0.993	97.5	0.995	97.5	0.922	100
	50	0.992	98	0.994	98	0.927	100
	60	0.991	96.7	0.994	98.3	0.931	98.3
	70	0.990	97.1	0.993	97.1	0.934	98.5
	80	0.990	96.3	0.993	97.5	0.937	97.5
	90	0.989	96.7	0.993	97.8	0.940	97.8
	100	0.989	96	0.993	98	0.942	98

Tabulka 7.4: Porovnání algoritmů pro hodnocení informativnosti příspěvků

V případě této události dosahují logistická regrese a TwitterEventInfoRank podobných výsledků, ačkoliv přesnost regrese dosahuje podle křížové validace přibližně 75 %. Tento výsledek je pravděpodobně následkem velikosti testovaných dat. Toto tvrzení lze podložit pouhým pohledem na nejlépe hodnocené příspěvky událostí obsahujících do 500 příspěvků a událostí, které jsou popsány přibližně 1000 příspěvky. Větší počet příspěvků je více vypovídající než menší počet příspěvků rozličného obsahu. Roli hraje i složení dat. Pokud bude větší poměr příspěvků neinformativních, budou na předních místech díky podobnému obsahu. Tím pádem navyšují váhu vrcholům, které nemají s událostí mnoho společného.



Obrázek 7.1: Porovnání výsledků algoritmů

7.3.2 Spearmanova korelace

Pro další porovnání výsledků hodnocení příspěvků je možno použít tzv. Spearmanovu korelaci [24]. Ta se používá pro měření síly asociace mezi dvěma množinami hodnot. Základem jsou dvě množiny hodnot, přičemž každá hodnota je ohodnocena tak, že nejvyšší hodnotě je přiřazena 1, druhé nejvyšší 2, atd. Pokud se v jedné množině vyskytují dvě stejné hodnoty, jsou ohodnoceny průměrem z hodnot, které by jim byly přiřazeny, kdyby byly odlišné. Pokud by se to tedy týkalo např. dvou nejvyšších hodnot, každé z nich by byl přiřazen průměr, tedy 1.5 a třetí nejvyšší hodnotě je přiřazena již standardně 3. Pro případ bez průměrovaných hodnot je pro výpočet výsledného koeficientu používán vzorec

$$\rho = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} \tag{7.4}$$

kde d_i je rozdíl mezi dvěma výslednými ohodnoceními a n je počet zkoumaných případů. Pokud se v jedné z množin vyskytují dvě shodné hodnoty a je nutné použít průměr ohodnocení, je pro výpočet výsledného koeficientu použit následující vztah:

$$\rho = \frac{\sum_i (x_i - \bar{x})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{7.5}$$

kde i je pár výsledného ohodnocení.

Výsledkem Spearmanovy korelace je tzv. Spearmanův korelační koeficient ρ, pro který platí $\rho \in \langle -1; 1 \rangle$. Pokud se výsledná hodnota pohybuje kolem 0, znamená to, že mezi hodnotami není žádný vztah a pořadí hodnot je náhodně zpřeházené.

Pro tento případ je prvních 10 příspěvků seřazeno dle ohodnocení algoritmem TwitterEventInfoRank. Následuje přiřazení hodnot dle skóre. Díky řazení příspěvků to bude v případě tohoto algoritmu seřazená posloupnost hodnot od 1 do 10.

TwitterEventInfoRank	rank (TEIR)	Logistická regrese	rank (LR)	d	d^2
0.00092480674080316	1	0.827513	5	-4	16
0.000918893111490629	2	0.935029	3	-1	1
0.000861159643233992	3	0.868626	4	-1	1
0.000812445073116697	4	0.947218	2	2	4
0.00081161983809185	5	0.81992	6	-1	1
0.000805518348344131	6	0.809004	7	-1	1
0.000803096011627159	7	0.545487	9	-2	4
0.000799862900275942	8	0.984608	1	7	49
0.000791437874143881	9	0.712338	8	1	1
0.000783509341684967	10	0.205461	10	0	0

Tabulka 7.5: Určení ohodnocení dle pořadí hodnot pro Spearmanovu korelaci

$$\sum_i d_i^2 = 16 + 1 + 1 + 4 + 1 + 1 + 4 + 49 + 1 + 0 = 78$$

Výsledná hodnota je dosazena do vzorce:

$$\rho = 1 - \frac{6 \times 78}{10(10^2 - 1)} = 1 - \frac{468}{990} = 1 - 0.47273 = 0.52727$$

Mezi množinami je tedy silný a kladný vztah. Čím vyšší bude výsledné ohodnocení algoritmem TwitterEventInfoRank, tím vyšší bude ohodnocení logistickou regresí a naopak.

8 Experimenty

Výsledný graf popisující danou událost obsahuje vrcholy pěti typů: příspěvky, hashtagy, podstatná jména, URL a autory příspěvků. Nosnou částí práce jsou příspěvky, které ve svém plném znění nesou informace o dění. Je tedy otázkou, jaký dopad mají na řazení příspěvků vrcholy dalších typů, jakými jsou např. uživatelé. Váhu každého typu lze určit pomocí úvahy. Vzhledem k vysoké frekvenci a tudíž i hodnocení slov a hashtagů v množině příspěvků, jejichž počet je v řádu stovek, je pravděpodobné, že tyto dvě množiny budou mít na hodnocení příspěvků největší dopad. Naopak nejmenší dopad by měli mít autoři příspěvků a URL. Jejich frekvence je poměrně nízká a tudíž se jejich hodnocení nejspíše odvíjí od hodnocení vrcholů, se kterými se v příspěvcích vyskytují.

Teorii lze ověřit pokusem. Množina dat popisující střelbu v Los Angeles je znovu ohodnocena, tentokrát jsou však vynechány vždy vrcholy jednoho typu. Sledována bude změna pořadí v prvních 5 ze 715 příspěvků, které jsou ohodnoceny jako informativní. Ve výsledcích jsou barevně odlišeny příspěvky, které se vyskytly mezi 5 nejvýše ohodnocenými příspěvky podle algoritmu TEIR. U každého příspěvku je v závorce na konci uvedena hodnota přidělená anotátorem.

Při odstranění celé množiny uživatelů z výpočtu výsledného ohodnocení jsou na prvních pěti místech následující příspěvky:

1. Did they know the lax shooter who knew james comey who saw ferris pass out at 31 brothers & a half brother last night ben obenshein mark ? (3)
2. RT @Green_Footballs: The LAX shooter was a libertarian of the Ron Paul/Alex Jones variety (End the Fed, gold standard, NWO, etc.) http://t.… (3)
3. http://t.co/VAhHTIc8cZ #fiverr #StoryOfMyLife #Halloween #music #tcot #androidgames #sats #LAX #Boston #BostonStrong (1)
4. RT @ezralevant: Los Angeles TV stations showing video feed of possible terrorist shooting at LAX airport. Toronto TV showing Rob Ford drivi… (3)
5. Denver Int'l Airport @DENAirport DIA operations are normal. Passengers from DEN to LAX should check flight status w/airlines (3)

Po odstranění množiny URL z grafu popisujícího střelbu z Los Angeles jsou na prvních pěti místech tweety:

1. #libros LA LUZ EN LA TORMENTA http://t.co/i86E8eMqJC http://t.co/GrsqiEfVuC #lax #sf #queleer #vancouver #bizitalk #quebec #montreal #lee (1)
2. Airport police engaged and neutralize lone shooter after suspect shot through TSA checkpoint. #BREAKING #LAX #LAXShooting suspect injured (3)
3. Lone suspect in Los Angeles #LAX airport shooting named to US media as Paul Ciancia, 23, by law enforcement officials http://t.co/0LZfFKpYqN (3)
4. #BREAKING: LOS ANGELES (AP) -- Law enforcement officials identify LA airport shooting suspect as 23-year-old Paul Ciancia. #wsbtv (3)
5. Did they know the lax shooter who knew james comey who saw ferris pass out at 31 brothers & a half brother last night ben obenshein mark ? (3)

Po zanedbání hashtagů z množiny vrcholů byly nejvýše ohodnoceny následující příspěvky:

1. RT @OutFrontCNN: #OutFront tonight: Obamacare cost cancer patient's health care? LAX shooter's anti-govt msg; Dolphins hazing? @ErinBurnett… (3)
2. RT @disturbthoughts: Eps 28 / Brandon A. Cottrell, Mike Klass & Jim Walters / #LAXShooting, Ongoing #NSA Updates, World Economics /#iTunes … (3)
3. RT @PzFeed: SHOOTING AT LAX AIRPORT: -2 SUSPECTS -1 SHOT, 1 IN CUSTODY -AT LEAST 2 PEOPLE SHOT -8-10 SHOTS FIRED -SUSPECT HAD A RIFLE -FAA … (3)
4. RT @kyoshino: LAX shooter pulled assault rifle out of bag at TSA screening area, shot at TSA screeners, 'opened fire' in terminal, LAX poli… (3)
5. Did they know the lax shooter who knew james comey who saw ferris pass out at 31 brothers & a half brother last night ben obenshein mark ? (3)

Nakonec po odstranění množiny podstatných jmen jsou jako nejinformativnější příspěvky označeny následující:

1. http://t.co/VAhHTIc8cZ #fiverr #StoryOfMyLife #Halloween #music #tcot #androidgames #sats #LAX #Boston #BostonStrong (1)
2. #BFL #BUR #FAT #LGB #LAX #MRY #OAK #ONT【Petition】Stop Korean Propaganda in the US Nickname・Email・Country That's it ! http://t.co/R2BoB6xNl6 … (1)
3. #libros LA LUZ EN LA TORMENTA http://t.co/i86E8eMqJC http://t.co/GrsqiEfVuC #lax #sf #queleer #vancouver #bizitalk #quebec #montreal #lee (1)
4. Check out one of Hottest new producers in the Game! https://t.co/WD0tCSiu83 #MoneyTeam #music #PeopleChoice #EMAzing #drake #LilWayne #LAX (1)
5. RT @VicehoodNews: LAX Shooting Victim: Human or Dummy? http://t.co/SXxeAQjv2Y #LAXShooting #Dummy #Conspiracy #TeamWakeEmUp #Wheelchair #Fa… (2)

Je více než patrné, že největší význam na výsledné ohodnocení příspěvků má množina podstatných jmen. Po jejich odstranění totiž převezmou roli nejvýznamnější množiny hashtagy a nejlépe hodnocené příspěvky jsou tedy zpravidla takové, které obsahují velké množství hashtagů. Cílem takových příspěvků ale není informovat o události, ale dostat se mezi největší možné množství typů vyhledávání.

Naopak nejmenší dopad na výsledky má odebrání množiny URL. Výsledné příspěvky se od původních liší ve třech případech z uvedených pěti.

9 Závěr

Cílem této práce bylo vytvořit aplikaci pro získání informací o právě probíhajících či ukončených událostech různého charakteru. Program umožňuje získat přehled o klíčových URL, zajímavých uživatelích, klíčových slovech či přehled o hashtagách, které danou událost provází.

Práce byla rozdělena na několik základních podproblémů, mezi které patřily problematika stahování příspěvků, jejich klasifikace, příprava matice pro výpočet ohodnocení a nakonec samotná realizace ohodnocení příspěvků a dalších množin (hashtagů, URL, uživatelů a podstatných jmen) dle míry informativnosti.

Pro realizaci stahování příspěvků byly vybrány dva přístupy, které se navzájem doplňují a umožňují stahování jak aktuálních tweetů, tak zpráv, které jsou několik let staré.

Při řešení klasifikace byl nejproblematičtější výběr rychlé pochopitelné a přehledné knihovny. Volba nakonec padla na knihovnu Weka, která v rámci klasifikátorů obsahuje velmi přehledné návody a umožňuje práci přes GUI. Klasifikátor byl testován zhruba na 20 000 příspěvcích s výsledky přesnosti – 74 %, úplnosti – 74.9 % a f míry – 74.2 %.

Co se týče výsledků realizovaného algoritmu, pro malý objem dat (v řádu stovek) jsou výsledky logistické regrese v řádu desetin procenta lepší. Se zvětšujícím se objemem dat se ale výsledky algoritmu TwitterEventInfoRank v porovnání s logistickou regresí zlepšují. To je patrné jak ve zdrojovém článku [1], ve kterém je zpracováno velké množství dat, tak z ukázkových databází, kde je celkové zlepšení výsledků patrné již při rozdílu v řádu stovek příspěvků. TwitterEventInfoRank je tedy vhodný hlavně pro stahování velkého objemu dat nebo pro komplexní ohodnocení dalších složek události (používané hashtagy či slova) a jejich analýzu. Logistická regrese je vhodná pro rychlé ohodnocení pouhých příspěvků bez dalších souvislostí.

Protože je práce stavěna na příspěvky pouze v angličtině, je možné rozšíření, kterým je podpora dalších jazyků. Další možností je průzkum využití placených funkcí knihovny Twitter API a jejich přínosu do této práce.

Přehled zkratek

API Application Programming Interface je rozhraní pro programování aplikací. Jde o soubor procedur, funkcí či tříd nějaké knihovny (ale i jádra operačního systému), které může programátor používající knihovnu využít.

DCG Discounted Cumulative Gain je srážkový kumulativní přírůstek a je to jedna ze složek pro výpočet nDCG.

HTTP Hypertext Transfer Protocol je komunikační protokol pro přenos dat na internetu.

IDCG Ideal Discounted Cumulative Gain je ideální srážkový kumulativní přírůstek a je to jedna ze složek pro výpočet nDCG.

JDBC Java Database Connectivity je rozhraní pro přístup k databázi.

JDK Java development kit je prostředí pro vývoj Java aplikací.

JSON JavaScript Object Notation je způsob zápisu dat.

LR Logistická Regrese je používaný způsob klasifikace.

MD5 Message-digest algoritmus je hašovací funkcí.

nDCG Normalized Discounted Cumulative Gain je normalizovaný srážkový kumulativní přírůstek a používá se pro vyhodnocení výsledků algoritmu.

NLP Natural Langure Processing je soubor metod pro zpracování textu, např. lemmatizace, POS-tagging, apod.

POS Part Of Speech je slovní druh.

R, RT Retweet je počet sdílení příspěvku na sociální síti Twitter.

REST Representational State Transfer je architektura rozhraní webových služeb.

SVM Support Vector Machine je algoritmus klasifikace textu.

TEIR TwitterEventInfoRank je realizovaný algoritmus.

URL Uniform Resource Locator se používá pro přesnou identifikaci dokumentů na internetu.

Zdroje

[1] MAHATA, Debanjan, John R. TALBURT a Vivek Kumar SINGH. *From Chirps to Whistles: Discovering Event-specific Informative Content from Twitter*. 7th Annual ACM Web Science Conference. ACM, 2015, At Oxford, UK, 2015. Conference paper. University of Arkansas at Little Rock.

[2] *Twitter Help Center: New user FAQs* [online]. USA: Twitter, 2016 [cit. 2016-05-08]. Dostupné z: https://support.twitter.com/articles/13920?lang=en

[3] *Twitter Developers: Documentation* [online]. USA: Twitter, 2016 [cit. 2016-05-09]. Dostupné z: https://dev.twitter.com/overview/documentation

[4] *Twitter Developers: Rate Limits: Chart* [online]. USA: Twitter, 2016 [cit. 2016-05-08]. Dostupné z: https://dev.twitter.com/rest/public/rate-limits

[5] *CrisisLex: Crisis Lexicons* [online]. USA: AAAI Press, 2014 [cit. 2016-05-08]. Dostupné z: http://crisislex.org/crisis-lexicon.html

[6] EKŠTEIN, Kamil. *Teorie kognitivních systémů*. Plzeň, 2012. Soubor přednášek. Západočeská univerzita v Plzni.

[7] CHAPELLE, Olivier, Bernhard SCHÖLKOPF a Alexander ZIEN. *Semi-supervised learning*. Cambridge, Mass.: MIT Press, c2006. Adaptive computation and machine learning.

[8] Distributed logistic regression. *MathWorks* [online]. USA: MathWorks, 2012 [cit. 2016-05-09]. Dostupné z: http://www.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/39653/versions/1/screenshot.png

[9] *Stanford NLP: Precision, Recall and F measure* [online]. USA: Stanford University, 2012 [cit. 2016-05-08]. Dostupné z: https://www.youtube.com/watch?v=2akd6uwtowc

[10] MANNING, Christopher D., Prabhakar RAGHAVAN a Hinrich SCHÜTZE.*Introduction to information retrieval*. Cambridge: Cambridge University Press, 2008. ISBN 978-0-521-86571-5.

[11] *LingPipe* [online]. USA: Alias-i, 2013 [cit. 2016-05-08]. Dostupné z: http://alias-i.com/lingpipe/index.html

[12] *MALLET* [online]. USA: University of Pennsylvania, 2002 [cit. 2016-05-08]. Dostupné z: http://mallet.cs.umass.edu/

[13] *Weka 3 - Data Mining with Open Source Machine Learning Software in Java* [online]. Nový Zéland: University of Waikato, 2013 [cit. 2016-05-08]. Dostupné z: http://www.cs.waikato.ac.nz/ml/weka/

[14] *The Linear Algebra Aspects of PageRank* [online]. USA: Carolina State University, 2007 [cit. 2016-05-08]. Dostupné z: http://www4.ncsu.edu/~ipsen/ps/slides_dagstuhl07071.pdf

[15] *The Stanford NLP (Natural Language Processing) Group* [online]. USA: The Stanford Natural Language Processing Group, 2010, 2016 [cit. 2016-05-08]. Dostupné z: http://nlp.stanford.edu/software/corenlp.shtml

[16] Twitter4J - A Java library for the Twitter API. *Twitter4J* [online]. Japan: Twitter4J, 2007, 2016 [cit. 2016-05-09]. Dostupné z: http://twitter4j.org/en/index.html

[17] *Commons IO: Commons IO Overview* [online]. USA: The Apache Software Foundation, 2016 [cit. 2016-05-09]. Dostupné z: https://commons.apache.org/proper/commons-io/

[18] *Get Old Tweets* [online]. USA: Jefferson Henrique, 2016 [cit. 2016-05-09]. Dostupné z: https://github.com/Jefferson-Henrique/GetOldTweets-java

[19] *A language detection library* [online]. France: Cédric Champeau, 2014 [cit. 2016-05-09]. Dostupné z: https://github.com/melix/jlangdetect

[20] *We Feel Fine and Searching the Emotional Web* [online]. Hong Kong, China, Feb: WSDM'11 proceedings of the 4th International Conference on Web Search, 2011 [cit. 2016-05-12]. Dostupné z: http://wefeelfine.org/

[21] *Twitter Shorthand* [online]. USA: FBI, 2014 [cit. 2016-05-09]. Dostupné z: https://www.documentcloud.org/documents/1199460-responsive-documents.html#document/p1

[22] *The Mathematics of Web Search* [online]. Cornell University, 2009 [cit. 2016-05-09]. Dostupné z: http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/index.html

[23] *Introduction to Information Retrieval: Evaluation* [online]. USA: Stanford University, 2013 [cit. 2016-05-09]. Dostupné z: http://web.stanford.edu/class/cs276/handouts/EvaluationNew-handout-6-per.pdf

[24] *Sperman's Rank-Order Correlation: A guide to when to use it, what it does and what the assumptions are*[online]. UK: lærd statistics, 2013 [cit. 2016-05-09]. Dostupné z: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Příloha A – Uživatelská příručka

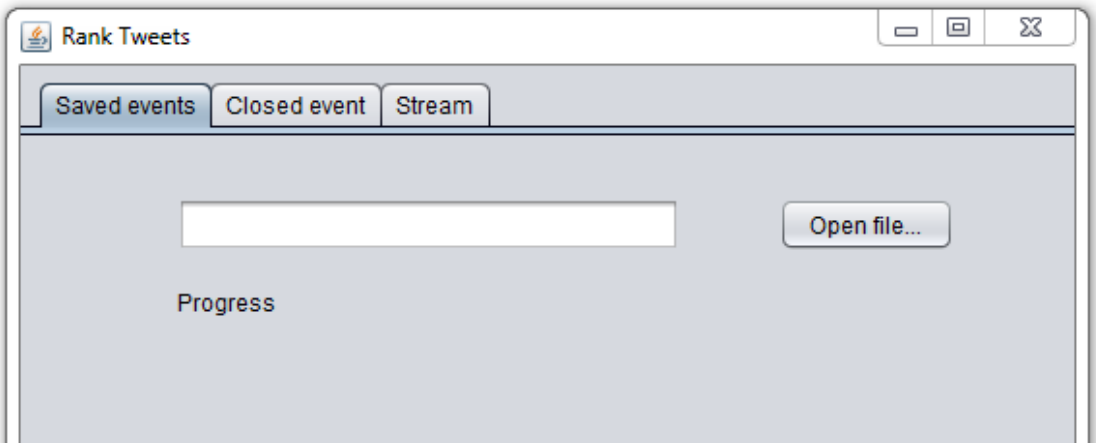
Pro zahájení běhu programu je nutné spustit aplikaci ze složky „app“ nebo z příkazové řádky. Pro tuto metodu spuštění je nutno se přesunout pomocí příkazu „cd“ do složky „app“. Poté stačí pro spuštění programu zadat příkaz

java –jar RankingTweets.jar

Po spuštění je možno vybrat si z několika možností běhu programu, jejichž funkce je popsána dále.

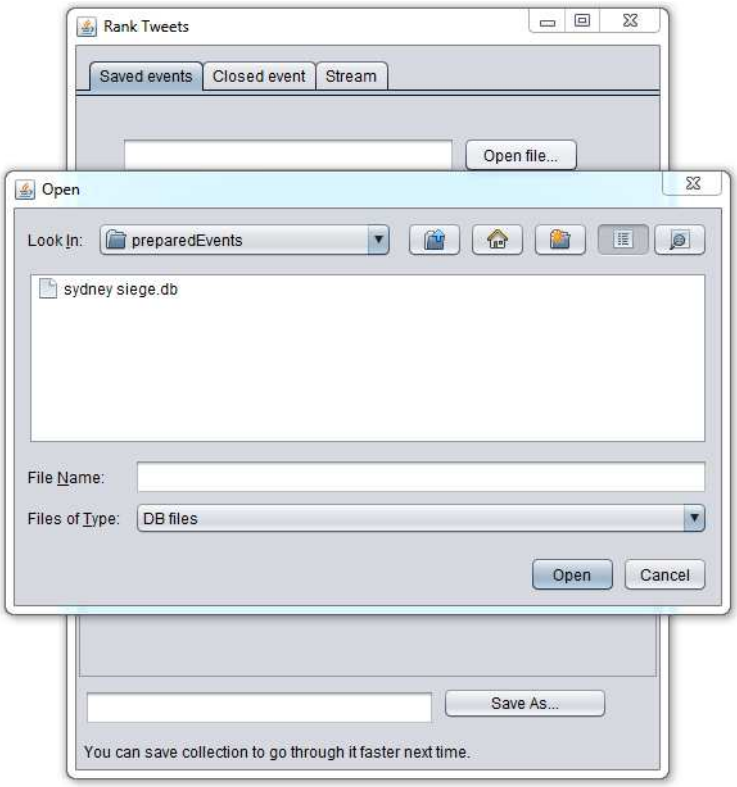
Prohlížení uložených událostí

Pro prohlížení uložených událostí je potřeba být přepnutý na záložce „Saved events“. Pro zvolení načtení databáze je nutné kliknout na tlačítko „Open file…“.



Obrázek 0.1: Část programu pro výběr připravené databáze

Po stisknutí tlačítka „Open file“ se otevře průzkumník souborů ve výchozím adresáři, kde jsou již připraveny některé ohodnocené události. Pro načtení požadované databáze lze buď dvakrát kliknout na vybraný soubor nebo na něj kliknout jednou a potvrdit tlačítkem „Open“. Popisek pod textovým polem ukazuje průběh načítání tabulek do programu.



Obrázek 0.2: Ukázka průzkumníku pro výběr databáze

Po načtení všech tabulek je možné data libovolně prohlížet a řadit dle libovolného sloupce. V případě příspěvků a uživatelů je přítomen sloupec „Data“, který pro každý záznam obsahuje informace o obsažených klíčových údajích v případě příspěvku. V případě uživatele zobrazí tato volba všechny příspěvky, které uživatel na dané téma napsal.

Stažení tweetů o uzavřené události

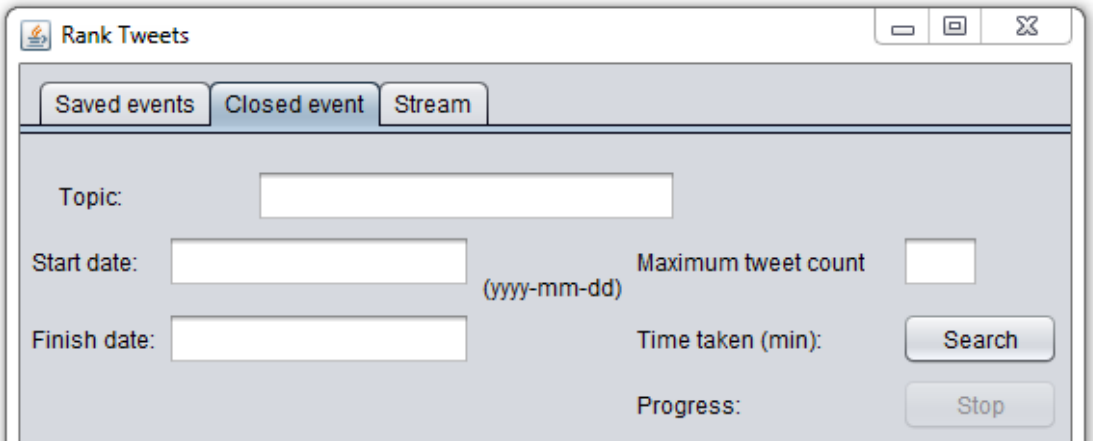
Druhou možností získání příspěvků je možnost stahování příspěvků o událostech, které již nejsou diskutované. Pro tuto možnost je nutné mít zvolenou záložku s nápisem „Closed event“. Po její aktivaci se objeví několik textových polí, přičemž všechna jsou povinná. Jejich obsah je následující:

- „Search event“ – pole pro popis vyhledávané událost (např. Sydney siege). Délka dotazu nesmí být větší než 150 znaků kvůli prevenci zahlcení.
- Start date – počáteční datum
- Finish date – koncové datum, nesmí být shodné s počátečním
- Maximum tweet count – maximální požadovaný počet příspěvků. Přesný počet není garantován, neboť příspěvky jsou v průběhu hodnocení filtrovány.

Minimální zadaný počet je 100. Maximum je 100 001 kvůli omezení zadávání nesmyslných hodnot.

Řetězec „Time taken“ ukazuje dobu trvání stahování s ohledem na omezení Twitteru. Během 15 minut lze získat nejvýše 180 příspěvků. Uvedený čas nebere v úvahu předchozí vyhledávání.

Nápis „Progress“ ukazuje počet zpracovaných příspěvků. Po zpracování vypisuje, kolikátá iterace výpočtu ohodnocení právě probíhá. Po překročení limitu počtu příspěvků se objeví výpočet, ukazující čas ve chvíli překročení limitu s přidanými 15 minutami pro přibližný odhad pokračování běhu programu. Čekání lze přerušit stiskem tlačítka „Stop“ a je možno ohodnotit příspěvky stažené před překročením limitu.



Obrázek 0.3: Část programu pro stažení příspěvků o ukončené události

Pro zahájení vyhledávání je po vyplnění nutné stisknout tlačítko „Search“. Po vyhodnocení se objeví výsledky v jednotlivých tabulkách ve spodní části programu.

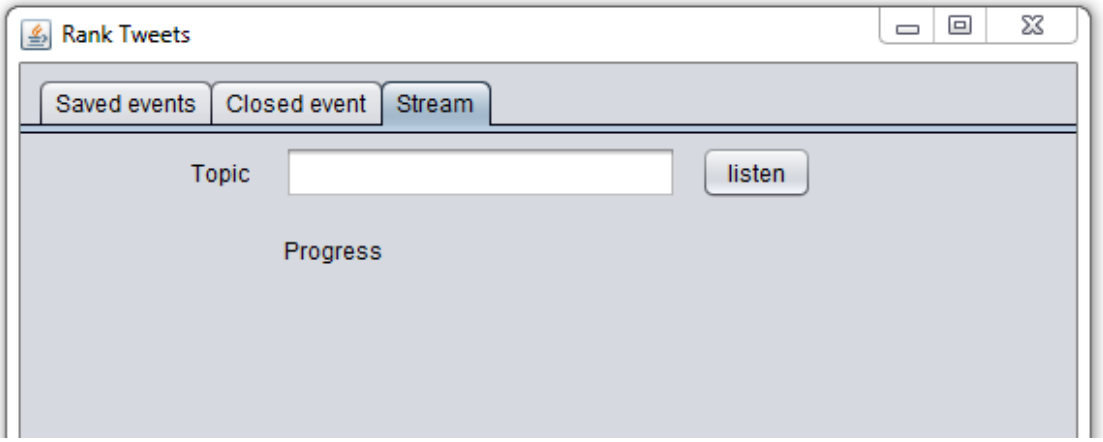
Jako inspirace pro vyhledávání slouží následující dotazy:

- Sydney siege, 2014-12-14, 2014-12-15
- Bataclan, 2015-11-13, 2015-11-14
- Bruselles attack, 2016-03-22, 2016-03-23

Získání aktuálních příspěvků

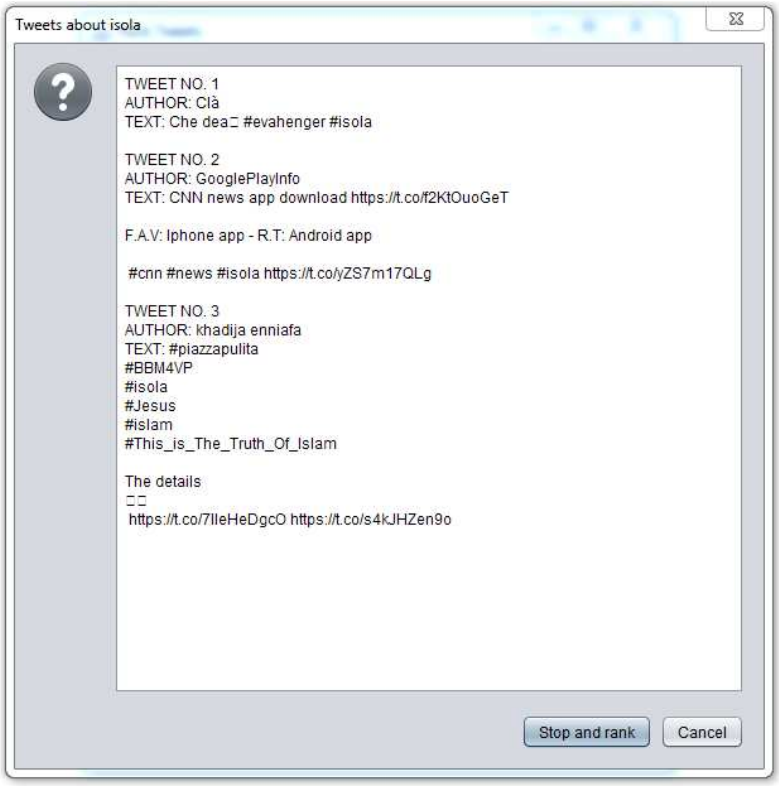
Pro získání aktuálních příspěvků slouží záložka „Stream“. Pro zahájení vyhledávání stačí zadat diskutované téma a stisknout tlačítko „Listen“. Okno obsahuje popisek

ukazující průběh programu. Po dokončení stahování aktuálních příspěvků popisek ukazuje, kolikátá iterace výpočtu právě probíhá.



Obrázek 0.4: Část programu pro stahování aktuálních příspěvků

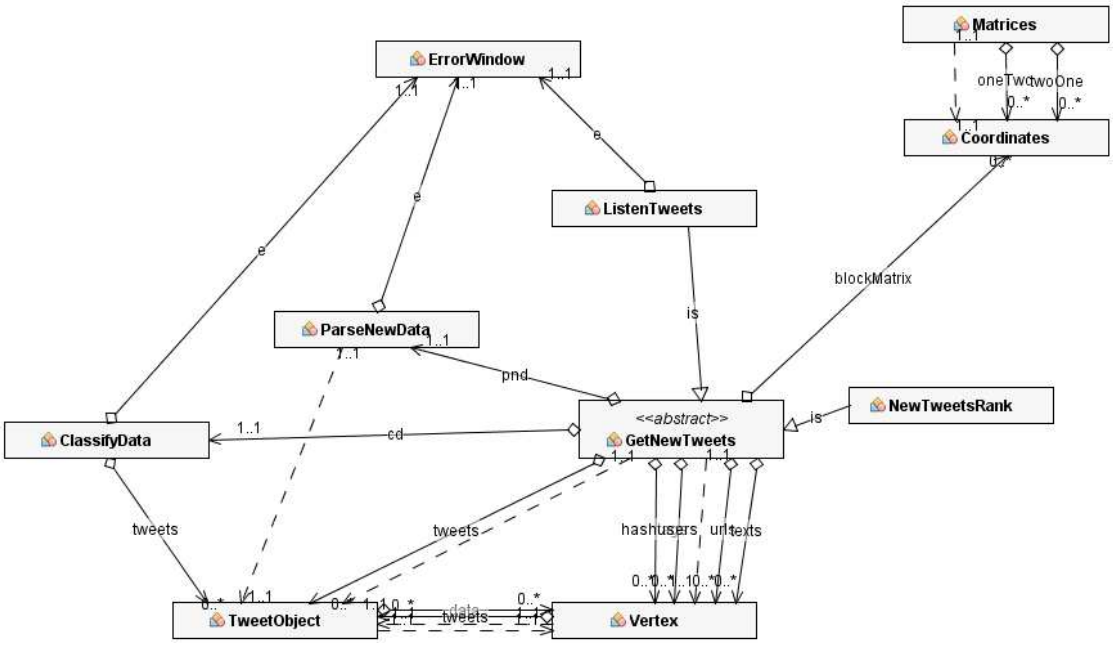
Po stisknutí tlačítka pro zahájení stahování příspěvku se objeví nové okno s přehledem stahovaných příspěvků. Pro ukončení stahování a zahájení vyhodnocení příspěvků dle informativnosti stačí kliknout na tlačítko „Stop and rank“. Pokud si uživatel nepřeje dané příspěvky hodnotit, je pro tuto možnost určeno tlačítko „Cancel“.



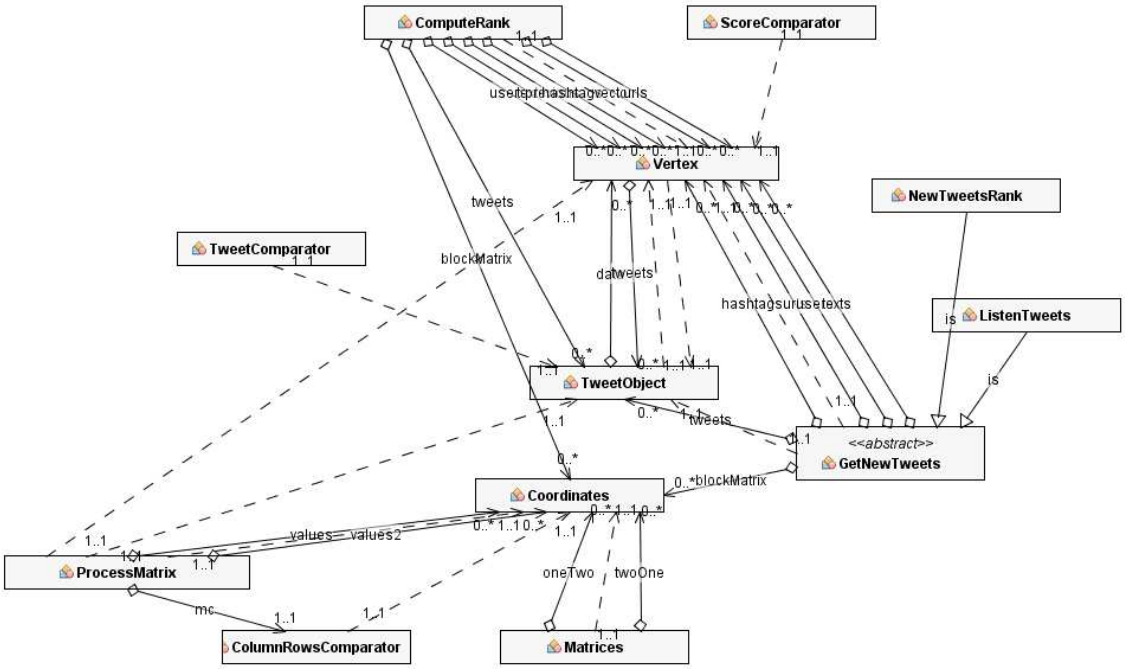
Obrázek 0.5: Okno pro prohlížení stažených příspěvků

V případě stahování příspěvků je po vyhodnocení uživateli umožněno ohodnocené množiny uložit do databáze pro pozdější prohlížení. Stačí kliknout na tlačítko „Save As…“ v dolní části programu. Otevře se průzkumník souborů (viz Obrázek 0.2: Ukázka průzkumníku pro výběr databáze) s rozdílem, že místo tlačítka „Open“ je tlačítko „Save“. Uživatel musí zadat jméno pro databázi a uložit ji nebo stisknout tlačítko „Cancel“. Z důvodu konzistence souboru s uloženými daty nelze ukládání souboru přerušit. Při tvorbě databáze lze sledovat průběh v pravém spodním rohu programu.

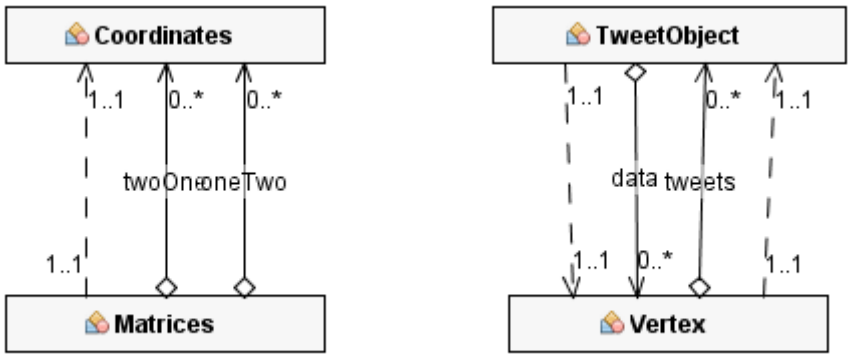
Příloha B – Diagramy balíků



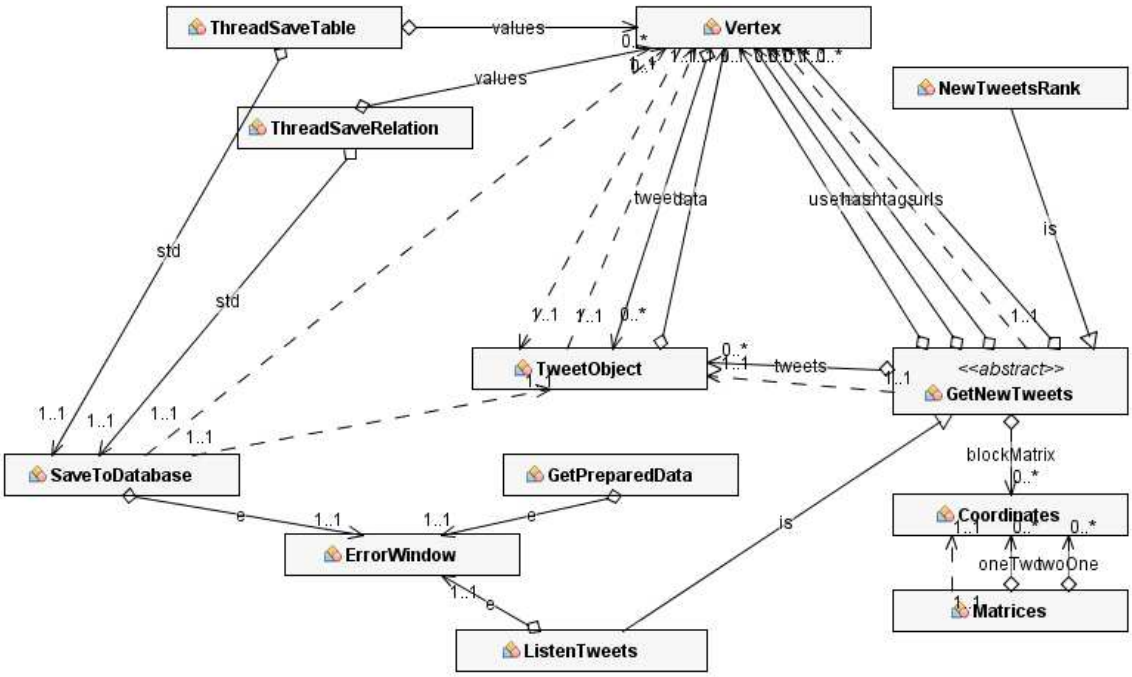
Obrázek 0.6: Diagram pro balík classify



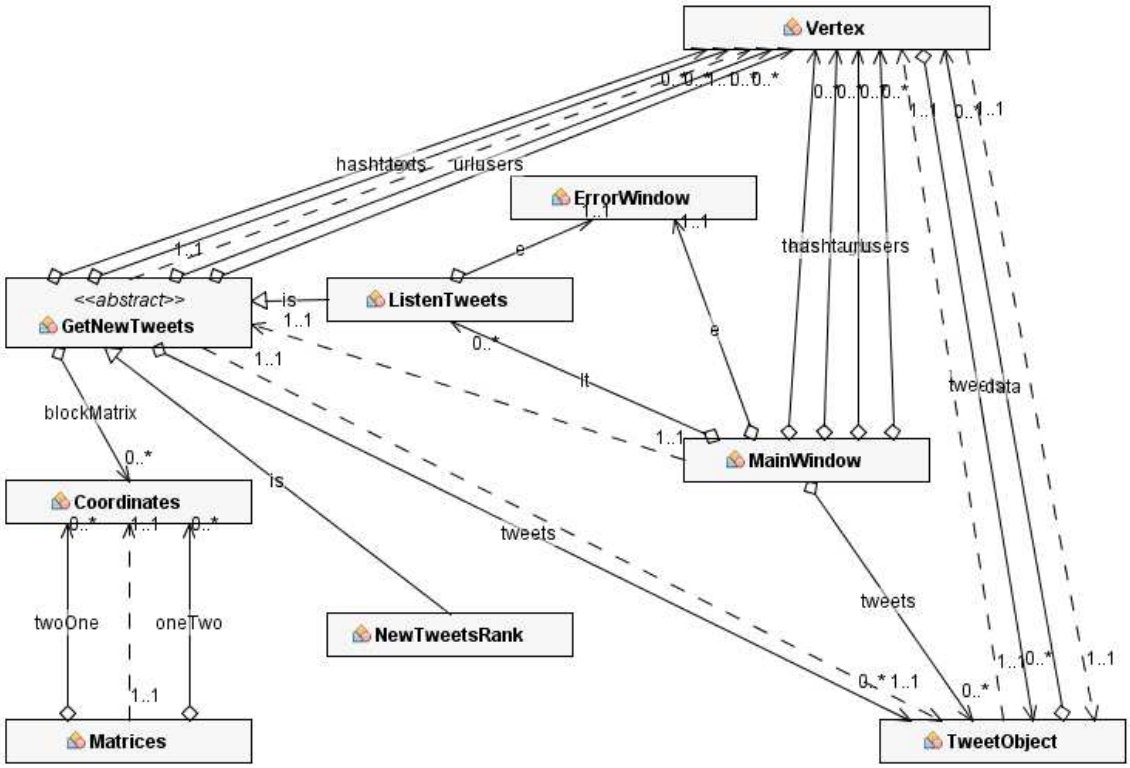
Obrázek 0.7: Diagram pro comparators



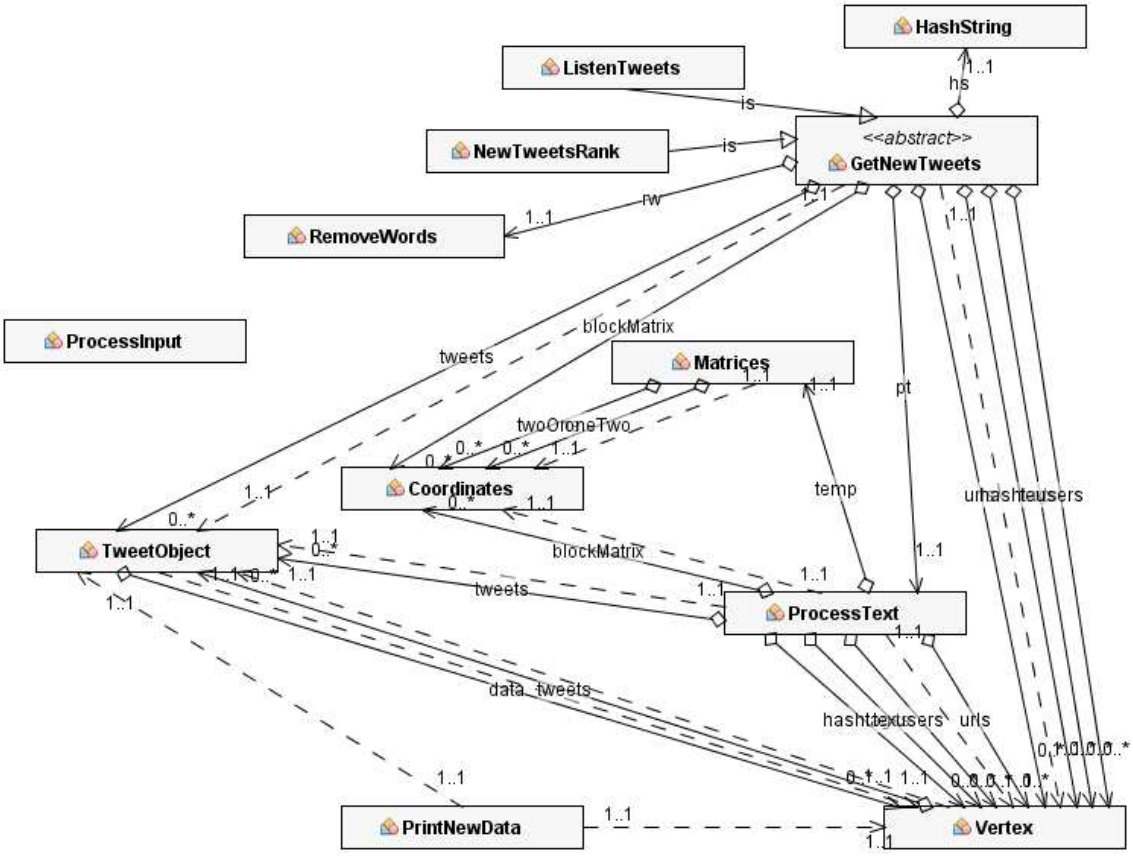
Obrázek 0. 8: Diagram pro data



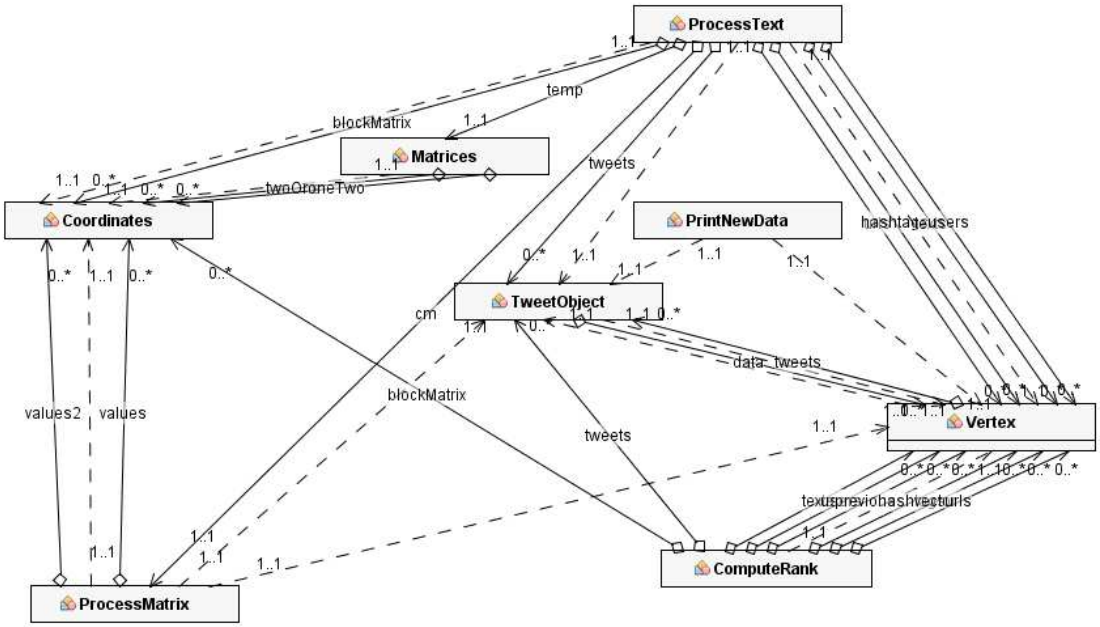
Obrázek 0.9: Diagram pro database



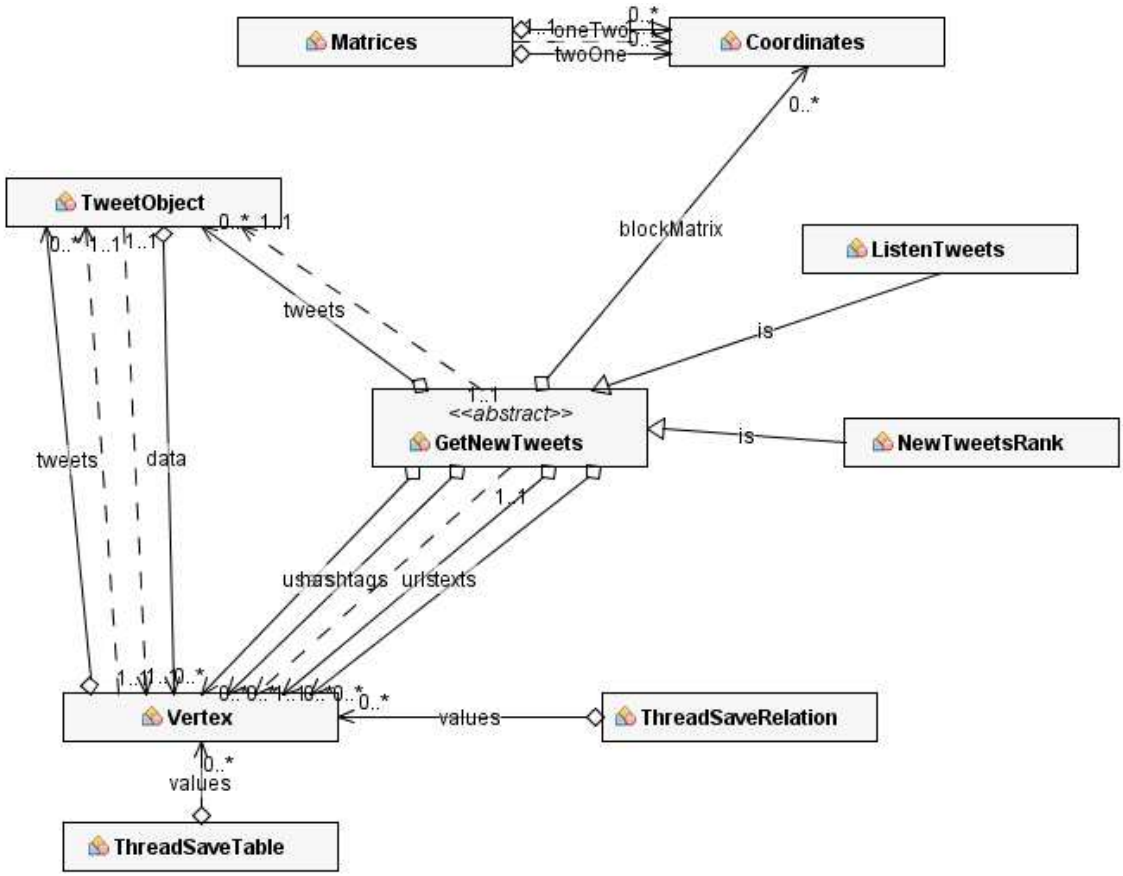
Obrázek 0.10: Diagram pro layout



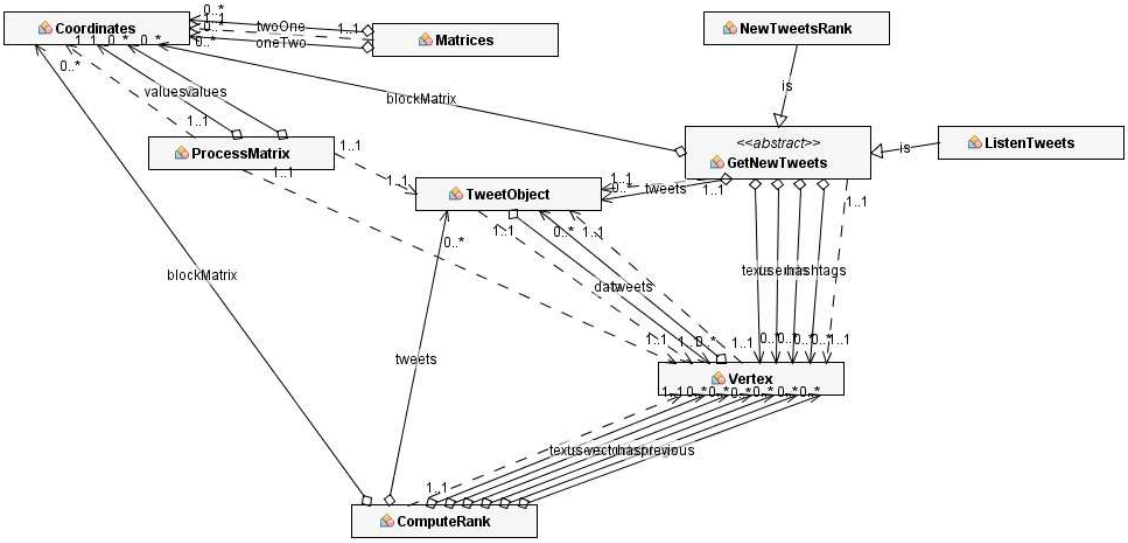
Obrázek 0.11: Diagram pro newTweets



Obrázek 0.12: Diagram pro processData



Obrázek 0.13: Diagram pro threads



Obrázek 0.14 Diagram pro utils

Příloha C – Ukázka výstupu

Jako ukázku výstupu různého hodnocení lze uvést 10 nejvýše ohodnocených příspěvků algoritmem TwitterEventInfoRank, logistickou regresí a 10 příspěvků s nejvyšším počtem retweetů. Ve výsledcích jsou v případě logistické regrese a počtu retweetů barevně odlišeny příspěvky, které se vyskytly v prvních 10 příspěvcích po ohodnocení pomocí algoritmu TwitterEventInfoRank. Celkový počet hodnocených příspěvků byl 715 a příspěvky se týkaly střelby na letišti v Los Angeles. Na konci každého příspěvku je uvedena hodnota přidělená anotátorem. Jedná se o hodnoty 1, 2 nebo 3.

TwitterEventInforRank:

1. Airport police engaged and neutralize lone shooter after suspect shot through TSA checkpoint. #BREAKING #LAX #LAXShooting suspect injured (3)
2. Lone suspect in Los Angeles #LAX airport shooting named to US media as Paul Ciancia, 23, by law enforcement officials http://t.co/0LZfFKpYqN (3)
3. #BREAKING: LOS ANGELES (AP) -- Law enforcement officials identify LA airport shooting suspect as 23-year-old Paul Ciancia. #wsbtv (3)
4. RT @BBCBreaking: Man opened fire with assault rifle in security screening area at #LAX Terminal 3, Los Angeles police say http://t.co/swrfb… (3)
5. RT @PzFeed: SHOOTING AT LAX AIRPORT: -2 SUSPECTS -1 SHOT, 1 IN CUSTODY -AT LEAST 2 PEOPLE SHOT -8-10 SHOTS FIRED -SUSPECT HAD A RIFLE -FAA … (3)
6. RT @CBCAlerts: Initial reports say gunman targeted #TSA agents at Los Angeles Airport. Conflicting reports re: suspect in custody. #LAX (3)
7. Suspect shot and in police custody. Witnesses say it was pandemonium. Chaos at #LAX. A gunman walked in and a TSA agent Killed. #Unreal (3)
8. RT @kyoshino: LAX shooter pulled assault rifle out of bag at TSA screening area, shot at TSA screeners, 'opened fire' in terminal, LAX poli… (3)

9. RT @CBCAlerts: Reports of gunfire at the Los Angeles airport. Police say 'major incident' underway. No word on victims. #LAX (3)
10. Did they know the lax shooter who knew james comey who saw ferris pass out at 31 brothers & a half brother last night ben obenshein mark ? (3)

Prvních 10 příspěvků podle logistické regrese:

1. RT @kyoshino: LAX shooter pulled assault rifle out of bag at TSA screening area, shot at TSA screeners, 'opened fire' in terminal, LAX poli… (3)
2. RT @nbcnightlynews: BREAKING: Shooting reported at Los Angeles Int'l Airport Terminal 3; WATCH LIVE coverage via @NBCLA http://t.co/x1nQ14V… (3)
3. RT @FoxNews: UPDATE: Suspect in LAX shooting wrote note telling of intent to kill TSA employees, report says http://t.co/oGXN1kC2q8 (3)
4. RT @Mediaite: WATCH LIVE: Terminal 3 Evacuated After Shooting Incident Reported at LAX Airport http://t.co/nCkHRaYmZs via @mediaite (3)
5. Suspect in LAX shooting apparently had suicidal thoughts before attack - The man suspected of killing a... http://t.co/xxIbmJIQ1K (3)
6. Shots fired at LAX, airport evacuated: ABC7 in Los Angeles reports that a TSA employee was shot at a security... http://t.co/QbF0KOtN7s (3)
7. RT @CarlosWPLG: Photo: bloodied clothes, gun inside terminal. Shooting began before checkpoint and continued through? #LAXShooting http://t… (3)
8. NBC News: At least 3 hurt, unknown if shooter is among injured. At least 1 in custody. President Obama briefed on #LAXShooting. @NBCLA (3)
9. LAX Shooting: Officials Confirm 'Incident' At Airport, Witnesses Hear Gunshots http://t.co/dzwGwoNtOa via @HuffPostCrime (3)
10. Frightening if genuine. Note: slightly graphic. RT @TJD19083: LAX Terminal 3 Shooter Scene http://t.co/mw90DgRbJ1 (3)

Pořadí dle počtu retweetů:

1. RT @AlfredoFlores: My prayers are with the family of the TSA agent who was killed today at LAX and the other victims injured in this shooti… (3)
2. RT @CodySimpson: stuck in the airport terminal because of the shootings. praying for the victims (2)
3. RT @LAX_Official: Airport officials confirm police incident began at 9:30 a.m. @ Terminal 3 at LAX. More info to come. (3)
4. RT @cnnbrk: Photo obtained by CNN appears to show a weapon on the floor at #LAX . http://t.co/xuWPhFfeje http://t.co/KipB68PPJq (3)
5. RT @alexmorgan13: Praying for those involved in the LAX shooting. Scary thinking about how many times I've gone through that airport. (2)
6. RT @cnnbrk: Suspect at LAX was shot and is in police custody, source says. http://t.co/6iPMwBpriM (3)
7. RT @jricole: So shooting down people at an airport is *not* terrorism, but carrying a flashdrive with evidence of gov't wrongdoing *is* ter… (2)
8. RT @adamlambert: Tragic news about LAX. Thoughts and prayers w victims. (2)
9. RT @wilw: I am shocked — shocked — that LAX shooter used an AR-15. It’s almost like that gun only exists to kill as many people as quickly … (3)
10. RT @dinahjane97: OMG just heard about the shooting at the LAX AIRPORT #prayingNoonegothurt (2)

Z výsledků je patrné, že počet retweetů je vysoký u příspěvků, které především vyjadřují sympatie. V případě logistické regrese se na nejvyšších příčkách umístily informativní příspěvky rozličných druhů. Data ukazují, že lidé mají tendence sdílet příspěvky, které vyjadřují jejich emoce nebo emoce známé osobnosti (RT @adamlambert).