ВЪЗМОЖНОСТИ ЗА ИЗВЛИЧАНЕ НА ДАННИ В РЕАЛНО ВРЕМЕ ОТ ПЛАТФОРМАТА TWITTER

Ас. Борис Банков, Икономически университет – Варна, Катедра "Информатика"

***Резюме**: В настоящия доклад е разгледана възможността за извличане на данни в реално време от платформата за социално взаимодействие и микроблогинг Twitter. Отбелязани са различните категории инструменти за работа с Twitter, публично достъпните програмни интерфейси, както и типичните реквизити, съдържащи се в туит съобщенията. За целите на доклада е избрано събитието, свързано с ядрените тестове в Северна Корея от 09.09.2016 и е наблюдавана постоянна емисия от данни в периода 09.09.2016 – 11.09.2016. Изведени са резултати, относно най-често срещаната ключова дума, държави с най-активни потребители, най-използван език на интерфейса на платформата, най-споделяно съобщение и интензивност на публикуване по часове.*

***Ключови думи**: Twitter, неструктурирани данни, извличане на знания, емисия от данни*

1. Въведение

Новите технологични възможности, големите скорости на обмен на данни, разпространението и достъпността на платформите за социално взаимодействие, предразполагат към развитие на алгоритмите, методите и подходите за извличане на познание в реално време от глобалната мрежа. Докато преди десет години се анализираха статични масиви от текстови данни, относно вече състояли се събития, в момента се усъвършенстват услугите за достъп и работа до постоянни потоци от информация, обработена в момента на нейното генериране. Принос към развитието на тези технологии имат високите постижения в усъвършенстването на социалните взаимодействия в Интернет и в частност – социалните мрежи и платформи. Те представляват основния източник на неструктурирани данни под формата на кратки съобщения, изображения, видео и аудио файлове. Освен това системата предоставя възможност за обратна връзка чрез изпращане на отговор (reply), харесване (like) или споделяне (retweet/share). Тази постоянна емисия от информация е

ценен ресурс, спомагащ за определянето на тенденции сред интересите на различни демографски групи, като например политически настроения, отзиви за спортни събития, мнения за продукти и фирми, и др. В настоящия доклад следва да разгледаме възможностите за анализ на подобна постоянна емисия от данни, като източник на информация ще използваме социалната мрежа Twitter, а публикациите, които ще изследваме са относно реакциите на хората след теста на ядрено оръжие в Северна Корея на 09.09.2016.

2. Twitter като постоянен източник на данни в реално време

Извличането на знания от текст представлява приложение на алгоритми за обработка на естествени езици и аналитични методи към текстови данни с цел откриване на ценна информация, разкриване на закономерности и връзки в неструктурираната среда. Интересът към процеса и приложението на извличането на знания нараства експоненциално през последните години поради факта, че се увеличава количеството дигитална текстова информация, генерирана в Интернет под формата на уеб страници, проекти като Google Books и Google Ngram, социални мрежи като Facebook и платформи за микроблогинг като Twitter. Въпреки, че Twitter може да бъде определена като социална мрежа, тя започва да функционира първоначално като среда за публикуване на новини. Съобщенията са могат да съдържат до 140 символа и се наричат туитове (туит ед.ч.). Потоците от данни в Twitter представляват богат източник на информация за всяка възможна предметна област. Тези потоци се използват за бързо определяне на тенденции, измерване на настроения към търговски марки, събиране на препоръки и мнения относно продукти и услуги. С придобиване на широка популярност Twitter става обект на множество приложни разработки, целящи както развитието и ефективността от използването на платформата, така и изследването на непрекъснатия поток от данни.

Според направено изследване на около 100 инструмента за работа с Twitter, основните разработки се разделят в няколко направления, сред които по-важни са:

- Инструменти за маркетингов анализ;
- Инструменти за следене на дискусии/чатове;
- Инструменти за откриване на нова информация и потребители;
- Инструменти за анализ и работа с хаштагове;

- Инструменти за известия и монигоринг[1];

Значителна част от всички разработки осъществяват достъп до Twitter посредством един от следните три интерфейса:

- Twitter`s Search API;
- Twitter`s Streaming API;
- Twitter Firehose.

Чрез Twitter`s Search API се дава достъп до колекция от данни, базирана на туитове, които са настъпили в отминал период от време, като потребителят е първоизточника на заявката. С Twitter`s Streaming API се достъпва постоянна емисия от данни, случващи се в реално време и инициализирани по заявка, изпратена от Twitter, на база на предварително зададени критерии. Twitter Firehose е платено корпоративно решение за извличане на голям обем информация по заявка на Twitter, с гарантирана цялост на данните.

Приблизително по 6000 twitter съобщения биват публикувани всяка секунда в целия свят[2]. Подобен богат поток от информация дава основание за силен интерес към извличане на знания в реално време. Новините за събития и развиването на последващите ги дискусии могат да се следят в момента на тяхното публикуване онлайн. За целите на доклада се спираме на тестовете проведени от Северна Корея на ядреното си оръжие на 9 Септември 2016г. Това може да постигнем, чрез **два програмни сегмента** – единият осъществява постоянна връзка с платформата Twitter, а вторият може да бъде изпълнен по всяко време, за да се получат данни на база на изтеглената до момента информация.

3. Рамки на изследването

За да бъдат използваеми потоците от данни в Twitter във формат подходящ за компютърни изчисления и анализ е нужен достъп до приложния програмен интерфейс на платформата За да тестваме възможностите на платформата избираме да използваме програмния Twitter`s Streaming API поради две причини – желаем да имаме постоянен достъп до актуални данни и може да създаваме комбинации от филтри и критерии за извеждането на разнообразна информация. За обхват на изследването определяме диапазон от 2 дни (09.09.2016 – 11.09.2016), през който да анализираме потоците от данни в платформата Twitter.

---

[1] 91 Free Twitter Tools and Apps to Fit Any Need, www.slideshare.net/Bufferapp/91-free-twitter-tools-and-apps-to-fit-any-need, 30.08.2016

[2] Twitter Usage Statistics, http://www.internetlivestats.com/twitter-statistics/, 30.08.2016

Като първи етап от изследването е необходимо да се установи връзка между Twitter и първия програмен сегмент, необходимо за обработката на данните. За да създадем такава връзка е нужно, посредством уеб услугите на http://apps.twitter.com да генерираме на 4 уникални символни низа:

- API key – ключ на приложението;
- API secret – ключ, разкодиращ API key;
- Access token – потребителски ключ;
- Access token secret – ключ, разкодиращ Access token.

На следващия етап е нужно да определим основните ключови думи, за които Twitter ще следи и при публикуването на туитове, съдържащи една от ключовите думи, то цялата информация за съобщението следва да се запише локално в текстов файл. За целите на изследването избираме North Korea, Северная Корея/Северня Корея/Северной Корее и Nord Korea (от английски, руски и немски).

След като се зададат основните филтри може да се стартира програмният сегмент, който да записва изпратените съвпадения в текстов файл. Информацията, която се изпраща от Twitter, се записва в JSON формат (Javascript Object Notation). JSON обектите са лесно четими от хора и подходящи за обработка от машини. Освен публикуваните съобщения, в които се съдържа една от зададените ключови думи, Twitter предава и информация за всеки отделен туит, като например:

- датата и часът на публикуване;
- езикът на публикуване;
- държавата, в която се намира авторът на туита;
- информация за приложението, от което е пратен туита;
- брой последователите на потребителя;
- брой приятелите, които потребителят следва;
- брой на всички туитове на потребителя;
- и др.

4. Резултати

Крайният файл е с размер 151МБ. В рамките на два дни са извлечени 36710 на брой съобщения, съдържащи една от първоначално зададените ключови думи. За да поставим данните в контекст избираме да допълним филтрите локално, като въведем думите nuclear (ядрен) и nuke(удар). Броят на съобщенията съдържащи nuclear е 12175, а nuke - 1758. Съобщенията съдържащи една от двете думи и ключовата дума North Korea са 12254.

Ключовата дума North Korea се среща 30220 пъти, Северная Корея/Северня Корея/Северной Корее - 1790, а Nord Korea - 14. Топ 3 държавите с най-много съобщения по темата са:

1) САЩ
2) Япония
3) Южна Корея

Туитовете, написани от устройство използващо английски интерфейс са 28023, на японски - 378, а тези, които използват корейски интерфейс - 248. Най-споделяното съобщение с 2344 позовавания е на Хилари Клинтън[3].

На Фиг. 1 е изобразен интензитета на съобщения, които съдържат една от ключовите думи в периода на едно денонощие по Американско централно стандартно време (GMT -5).

Фиг. 1. Брой туитове за едно денонощие

Заключение

Twitter е социална платформа за споделяне на кратки съобщения. Благодарение на приложни интерфейси като Twitter`s Search API, Twitter`s Streaming API и Twitter Firehose става възможно

[3] North Korea's decision to conduct another nuclear test is outrageous and unacceptable, https://twitter.com/HillaryClinton/status/774330575230472192, 11.09.2016

изследването и отразяването на тенденции в потребителското мнение относно новини, отразяващи случващото се в реално време. Информацията, която Twitter предоставя за публикуваните съобщения и техните автори е обемна и подходяща за провеждането на задълбочени анализи и разкриването на скрити зависимости между възрастта, народността, езика, активен период на публикуване и събития от глобален мащаб.

ЛИТЕРАТУРА

1. Врагов, Г., Лечов, Г,. Анализ на прекъсванията в Twitter. http://www.math.bas.bg/infres/IS-publ/IS-2011-Vragov_Lechov-ICAI.pdf , 30.08.2016.
2. Статева, Г., Статистическите изследвания и „големите данни": допълващи се източници или конкуренти, (2015). http://hdl.handle.net/10610/2722, 30.08.2016.
3. Кунева, И., Маркетинг в социалните мрежи: Twitter и практиката на българските фирми, (2010), http://research.bfu.bg:8080/jspui/handle/123456789/166 , 30.08.2016
4. Bonzanini, M., Mastering social media mining with Python, (2016).
5. Chae, B., Insights from hashtag# supplychain and Twitter analytics: Considering Twitter and Twitter data for supply chain practice and research, International Journal of Production Economics 165 (2015): 247-259.
6. Kumar, S., Twitter data analytics, New York: Springer, 2014.
7. Kwak, H., What is Twitter, a social network or a news media?, Proceedings of the 19th international conference on World wide web. ACM, 2010.
8. MacEachren, A. M., Geo-twitter analytics: Applications in crisis management." 25th International Cartographic Conference. 2011.
9. Perera, R., Twitter analytics: Architecture, tools and analysis, MILITARY COMMUNICATIONS CONFERENCE, 2010-MILCOM 2010. IEEE, 2010.
10. Suh, B., Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. Social computing (socialcom), 2010 ieee second international conference on. IEEE, 2010.
11. Twitter Usage Statistics, http://www.internetlivestats.com/twitter-statistics/, 30.08.2016
12. 91 Free Twitter Tools and Apps to Fit Any Need, www.slideshare.net/Bufferapp/91-free-twitter-tools-and-apps-to-fit-any-need, 30.08.2016