SCHOOL OF INFORMATICS
Data & Web Science MSc Program

WEB DATA MINING

AUTH Informatics Department
Spring Semester 2019

Faculty: Prof. Athena Vakali (avakali@csd.auth.gr)]
Teaching Assistant: Dimitra Karanatsiou (dimitrakd@csd.auth.gr)

Project

Social networks are an integral part of everyday life. People use social networks to talk about hot news topics, express their views and comment on a fact, rate a show or a movie etc. The above are just a small sample of the activities that social network users can do. As it becomes clear, the volume of data produced every day is particularly important. For example, according to official statistics, 500M tweets are sent daily to the Twitter social network, with a total of 320M active users.

A social network constitutes a social structure composed of individuals (individual entities or groups) represented as nodes and linked to one or more types of interdependence depending on the type of relationship (explicit or indiscriminate). Due to the large amount of data and information exchanged among users of social networks, it is of particular interest to identify important events in an automated manner with a view to their valid recognition and direct information to the wider public. A key feature is the shift of the journalist community towards social networks, especially microblogging services, both for the dissemination of information and for the identification of key and important events taking place around the world and relayed by ordinary users. An event is characterized by topics that are highlighted by the increase of relative activity in microblogging services, and which describe the development of this. Understanding the evolution of an event by analyzing activity in social media first requires the discovery of emerging topic detection, taking into account both the intensity and the significance of activity. In order to analyze activity in relation to an event, information such as reference frequency on topics, user interaction around the event (reference to users mention-, retweet-, etc.), activity intensity in in relation to the geographical location of the users, the type and quality of content generated by the users. A common method of analysis that is applied to conduct a more detailed analysis is sentiment and emotion detection with respect to the certain topic. Another method of analysis is user's location extraction by using text and metadata contained in each tweet-post or by analyzing the features of the social network and the location inference.

The aim of this project is to study and analyze in depth important points of reference in relation to a selected topic, exploiting the activity of Twitter users and present the results through a web application, aiming at the better overview of the topic.

Part I : Part I is common to all teams and includes issues related to:

Topic 1. Emerging topic detection

Topic 2. Attitude, sentiment and emotion detection about a topic (sentiment analysis)

Topic 3. Extraction of user location and spatial analysis (location inference)

This Part is about a literature review of the state-of-the-art approaches and suggested models for the above problems and each team have to study at least 8 articles from scientific journals and conferences of a topic selected from the above list. The articles listed in Starting points section are suggested as starting points, and additional articles are required to be collected. The bibliographic study should be recorded in the form of conference publication, specifically in accordance with the ACM SIG Proceedings [L1] standard, in English or Greek.
For example, if you chose topic 2, you should focus on methods that allow the understanding of users' opinions and feelings applying text analytics techniques. There are two levels of analysis: (i) opinion mining, where the texts are characterized as positive/negative, (ii) analysis of individual emotions (affective analysis), which characterizes texts with specific emotions (i.e. anger, joy, excitement etc.) In sentiment analysis two commonly used approaches are applied (i) machine learning and (ii) lexicon-based technique. In recent years, great emphasis has been put on Deep Learning techniques or tools like Word2Vec. In the theoretical part you should present work related to one selected approach putting more emphasis on SOTA (State Of The Art) implementations.

This "deliverable" should be of 10-12 pages long and include at least 12 references. References should be used throughout the article. In addition, Part I should be accompanied by a presentation of 20 slides showing the most important elements of the literature review. Indicative deliverable structure:

- Introduction
  - Which the topic of your review
  - Which are the main challenges about this topic nowadays
  - Why this topic is interesting
- Fundamental concepts and key elements
  - Basic and useful definitions and key components of the studied topic
- Models
  - Present the SOTA models and approaches used in the studied topic, with comparative tables and images, that will enhance the understanding about the topic
- Conclusions
  - Conclusions and future work directions
- References
  - Presentation of used references according to the given template

Part II

In this part you are required to focus on a topic of your interest and analyze Twitter content about it in different levels. The first level of analysis is (i) opinion mining, where the texts are characterized as positive or negative, (ii) an analysis of affective analysis, which characterizes texts for specific emotions. In this work we will mainly focus on sentiment (positive or negative) but we will also experiment with affective lexicons. To make such an analysis at any of the above levels, two commonly used approaches are machine learning [1, 2, 3, 4] and the lexicon-based technique [5, 6]. The second level of analysis is to study in depth statistical analysis and more extensive exploration of all data. Data and features visualizations can be of great help in making better use and understanding of the studied topic. In this part you are invited to plot the various correlations between the data (at least 6 charts/tag-clouds) and explain their interpretation. It should be such that their interpretation to add knowledge that is not apparent in advance. Then you are asked to study methods used to assess the location of the individual user. There are two main categories of methods: (i) site extraction using the content of each tweet and its metadata [7, 8] and (ii) analyzing the social network and relationships between users/entities using graph analysis techniques [9, 10, 11]. The second category has shown better results, but it is more complex and certainly requires larger computing resources. In specific, Part II must include:

1st part: Data collection from Twitter and Data storage – Build a Social Listener

The Twitter Streaming API [L2] will be used to collect the data from Twitter, which supports the continuous provision of new tweets based on some initial criteria. Each team should initially choose a different subject/event, and then appropriate key terms and/or users in conjunction with the teaching assistant. The collection of data should be done for a sufficient period of time in relation to the event and expected activity of the users. For the collection of Twitter data, it is suggested to use the Twitter4j [L3] Java library or the Tweepy [L20] Python library which greatly facilitate the process. (Caution: for the tweets to be collected you should keep the complete JSON format that returns the Twitter Streaming API). Tweets must be in English language. Data collected from Twitter API have to be stored in a database that is suitable for efficient storage and retrieval. The MongoDB [L6] (open source, NoSQL, document-oriented database) database will be used for this task.

Subject/Event/Topic ideas: teams may choose from a variety of topics of their interest with regard to a social or political event (i.e. #Oscars2019, #NBAAllStar, #Brexit etc.), a social challenge (i.e. sexual harassment with #metoo, climate change with #climatechange, eating disorders with a keywords such as 'anorexia' and 'bulimia' etc.), a business/brand/product (i.e. hashtags from similar product launches from competitive brands , #GalaxyS9 VS #iphonex) or any other topic of interest.

2nd part: Tweets preprocessing and modeling.

This step involves preprocessing tweets to prepare them for further analysis. This process usually includes steps like: i) filtering low quality tweets, ii) tokenization and recognizing hashtags, URLs, mentions, or other tweets (retweets) included in it; iv) the removal of common words (stop words) such as articles, pronouns, etc. (the, at, this, is, was, were etc.), as well as terms that often appear on Twitter (e.g., RT, via), and v) converting the terms into lowercase. A useful procedure for finding topics is also the discovery of name entities in the text through Natural Language Processing techniques (NLP). This process is algorithmically more complex, but it can be easily implemented using easy-to-use libraries (eg Apache OpenNLP [L4], Stanford NLP [L5], NLTK (Python) [L21]).

3rt part: Emerging topics and information evolution in time

To find emerging topics or events, you are encouraged to use the content and metadata of tweets and apply either one of the methods you have studied, or a combination of more methods, or even develop some of your own extension/method based on what you studied. Keep in mind that an emerging topic in relation to an event is mainly characterized by a sharp increase in popularity over previous times. Generally, some method of time segregation is commonly use popularity metrics, coupled with metrics that capture their burstiness, utilizing tweets text.

4th part: Sentiment and emotion information extraction

This part involves the identifications of the expressed feeling through a technique based on mechanical learning. The ranking of tweets will be in these categories: A) overall sentiment a) positive, b) negative and B) specific emotions a) anger, b) joy, c) disgust, d) fear, e) sadness, f) surprise. The steps to be followed are:

1.Features selection: Features selection depends mainly on the available datasets, as we can choose a set of different features, such as:

a. Simple appearance of a term
b. Words calibrated based on their frequency of occurrence
c. Punctuation
d. emoticons

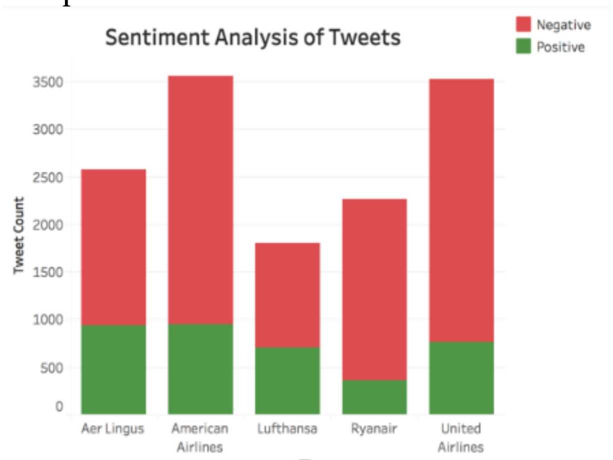
After extracting the features, we have to utilize them in order to train a predictive algorithm (machine learning classification task).

2. Apply different classification algorithms (e.g. Naïve Bayes, Decision Trees, Multinomial Logistic Regression) and choose the one that returns the best results. For the application of the above

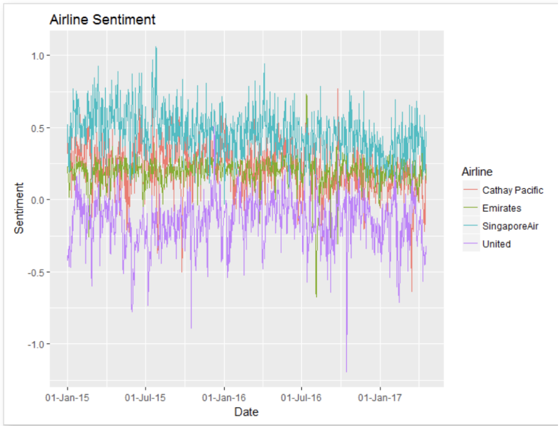
algorithms there is no limitation on the tool to be used (eg Weka [L7], Matlab [L8], Sklearn [L22], Keras [L23]). To train the model, a set of tweets will be given to all groups, which should initially be characterized by the expressed emotions (as described above). Results will then be gathered from all groups to create a data set that includes annotated tweets. Otherwise each team will use one of many available open sets of data that are already annotated or annotate their dataset with the use of affective lexicons (EmoLex (you can find an implementation at [L24]), Wordnet-affect [L25]) or a hybrid technique.

5th part: Statistical analysis and in-depth data exploration

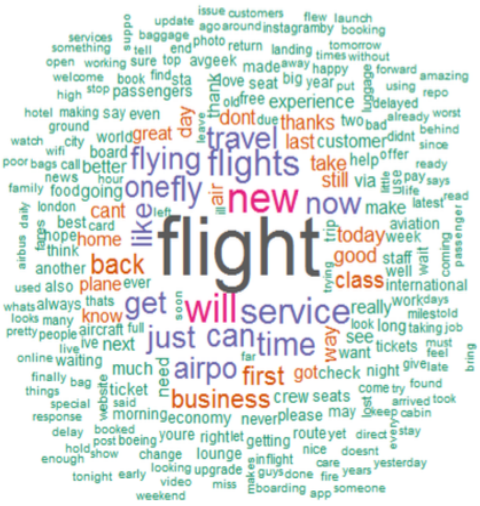
Visualization of data and their features can be of great help in making better use of them. In this part you are invited to plot the various correlations between the data of your collection(s). Should show at least 6 charts/tag-clouds and explain their interpretation. It should be such that their interpretation to add knowledge that is not apparent in advance (except from top tweets, top users etc.). For example:



a) Sentiment analysis of tweets by Airline



b) Sentiment analysis of tweets by airline with time information



c) Word Commonality by airline

6th part: Geo – information extraction

Exporting geographic information is very easy when the user shares his exact coordinates (check-in). Unfortunately (or fortunately) this happens only for ~ 2% of all the tweets that are being wiped out daily. However, this percentage in the total of a large number of tweets is sufficient to draw some relevant conclusions. Specifically, by analyzing tweets that include a geographic location, one can extract a set of words that are indicative of a site (for example, the word howdy is more likely to be used by users with a relationship with Texas). The appearance of such words may be a feature in our attempt to export a user's location. Another feature may be the user's time zone, available for each tweet. By using a logic similar to the 5th part we can achieve a way of determining the user's location, although we will not lead to high-precision conclusions. Alternatively, we can analyze relationships between users. According to the literature (28), relationships between users are indicative of the distance between users. For example, let John follow Jennifer, who in turn follows John. This bi-directional relationship is an indication that users John and Jennifer are well known and in fact suggesting they may be in nearby locations as well. Of course, this can be enhanced when user X refers to user Y, etc. Knowing the user's location X we can also make an estimate of the location of user Y. Using an information dissemination algorithm, we can have geographic information about few users make an estimate for the location of many more. Afterwards, display this information on a map of your choice. You could use the Google Maps library or Open Street Map.

7th part: Creating a web application to display the analysis results.
The results of your analysis should be presented in a web application in an attractive way for the user. The application should present some basic statistics for all the data you analyzed (e.g., production rate of tweets, number of users, etc.), and the important issues you discovered in relation to the time of the event, along with, but not limited to, tweets, possible images included in them, word tag-clouds, hashtag tag-clouds, hashtag co-occurrence network etc. In addition, the application should present the results of the emotional analysis. It is recommended to use libraries that offer interactivity, such as: Google Charts [L9], TimelineJS [L10], d3.js [L11], charts.js [L12]. You can also get ideas from existing apps that show analysis results from Twitter data [L13, L14, L15, L16]. You should use a free web hosting provider for your website [e.g. L17, L18, L19].

Project will be carried out by teams of 2-3 people

Deliverables	Deadline
1.Literature Review: according to the specifications given in Part I. **2.Presentation file**.	**6/5/19**
3. Code of Part II with sufficient commentary. Java or Python will be used to develop the code. 4. **Data Collections** (json files) used for experimentation in Part II. **5.Technical report (~ 10 pages)** which will include: a) a description of the model used to represent the data; b) a description of the implementation of the processing and calculation methods; c) commenting on indicative results; and d) a description of the application developed.	**10/6/19**

References

1. D. Chatzakou, N. Passalis, A. Vakali. MultiSpot: Spotting Sentiments with Semantic Aware Multilevel Cascaded Analysis. Big Data Analytics and Knowledge Discovery (DaWaK), volume 9263, pages 337-350, Springer, 2015.
2. R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. InProceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 151-161.
3. Georgios Paltoglou and Mike Thelwall. 2012. Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media. ACM Trans. Intell. Syst. Technol. 3, 4, Article 66 (September 2012), 19 pages. (sentiment analysis, lexicon-based methodology).
4. B. Pang et al. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proc. Of the 42nd annual meeting on Association for Computational Linguistics, 271, 2004.
5. Yan Dang, Yulei Zhang, and HsinChun Chen. 2010. A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews. IEEE Intelligent Systems 25, 4 (July 2010), 46-53. (sentiment analysis, lexicon-based methodology combined with machine learning).
6. Chatzakou, D.; Koutsonikola, V.; Vakali, A.; Kafetsios, K., "Micro-blogging Content Analysis via Emotionally-Driven Clustering," Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on , vol., no., pp.375,380, 2-5 Sept. 2013.
7. Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews. "Where Is This Tweet From? Inferring Home Locations of Twitter Users." *ICWSM* 12 (2012): 511-514.
8. Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. "You are where you tweet: a content-based approach to geo-locating twitter users." *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010.

9. Jurgens, David, et al. "Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice." *ICWSM*. 2015.
10. Compton, Ryan, David Jurgens, and David Allen. "Geotagging one hundred million twitter accounts with total variation minimization." *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, 2014.
11. Kong, Longbo, Zhi Liu, and Yan Huang. "Spot: Locating social media users based on social network context." *Proceedings of the VLDB Endowment* 7.13 (2014): 1681-1684.

Useful Links

L1. http://www.acm.org/sigs/publications/proceedings-templates

L2. https://dev.twitter.com/docs/streaming-apis

L3. http://twitter4j.org/en/index.html

L4. https://opennlp.apache.org/

L5. http://nlp.stanford.edu/

L6. http://www.mongodb.org/

L7. http://www.cs.waikato.ac.nz/ml/weka/

L8. http://www.mathworks.com/products/matlab/

L9. https://developers.google.com/chart/

L10. https://timeline.knightlab.com/

L11. https://d3js.org/

L12. http://www.chartjs.org/

L13. http://tweettracker.fulton.asu.edu/ (application)

L14.http://twitris.knoesis.org/yolandastorm2013/ (application)

L15. http://topsy.com/ (application)

L16. http://mediafinder.eurecom.fr/ (application)

L17. http://www.hostinger.gr/

L18. http://freehostingnoads.net/

L19. http://freehostingnoads.ga/

L20. http://www.tweepy.org/

L21. http://www.nltk.org/

L22. http://scikit-learn.org/stable/

L23. https://keras.io/

L24. https://github.com/beefoo/text-analysis

L25. https://github.com/DonatoMeoli/WNAffect

Starting points/articles for Part I

For topic 1

1. H. Sayyadi, L. Raschid. “A Graph Analytical Approach for Topic Detection”, ACM Transactions on Internet Technology (TOIT), 2013 (term based, graph model).
2. Unankard, S., Li, X., &amp; Sharaf, M. A. (2015). Emerging event detection in social networks with location sensitivity. World Wide Web, 18(5), 1393-1417.
3. Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. Computational Intelligence, 31(1), 132-164.

For topic 2

4. Ravi, K., Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46.
5. Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89
6. Chatzakou, D., Vakali, A., & Kafetsios, K. (2017). Detecting variation of emotions in online activities. Expert Systems with Applications, 89, 318-332.

For topic 3

7. Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews. “Where Is This Tweet From? Inferring Home Locations of Twitter Users” ICWSM 12 (2012): 511-514.
8. Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. “You are where you tweet: a content-based approach to geo-locating twitter users” Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.
9. Jurgens, David, et al. “Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice” ICWSM. 2015.