BILKENT UNIVERSITY –- SPRING 2016

Last Update: March 9, 2016; 5:34 pm

CS533: INFORMATION RETRIEVAL SYSTEMS

TWITTER-RELATED PAPER RECOMMENDATIONS BY PROJECT GROUPS

February 26, Friday by 5 pm: The project roups must be formed.

March 4 Friday by Class Time: Each group must recommend three papers related to Twitter and your term project. You must provide a brief description of each paper, with your own words, with a few sentences.

Each project group must compile and send recommendations as a group in one email by the group leader. The group leader is the person with the last name with the smallest lexicographic value (a < b, etc.). Group leaders in the email subject line please use "CS533 Twitter Paper Recommendation".

Before sending your recommendation please look at the papers listed in the document "TWITTER-RELATED PAPER RECOMMENDATIONS BY STUDENTS" and determine the category of your suggestion. If it should be under a different category not listed there suggest a brief title of the category. Categories and suggestions within categories will be listed from newest to oldest.

After compiling all suggestions I will select one of the papers suggested by each group and the group will give a joint 8-minute class group presentation. In the presentation you will explain the background work of the study, its motivation, method, and strong and weak points. Each presentation will be followed by a short question and answer period.

March 15, Tuesday, Class Time - Presentations.
Guidelines: **1**. Prepare a power point (or similar) presentation. Send a pdf copy to me after your presentation. **2**. Bring a one-page handout to the classroom for your classmates. **3**. Each group member must contribute to the presentation in the classroom. **4**. The presentation time is 8 minutes, group no. 6 with one member will be given 5 minutes for presentation. **5**. It is expected that you will fill the allocated time with reasonable material. 6. Presentations will be done in the order of group numbers.

No.	Members	Topic	Paper Assigned
1	Didem Demirağ, Anisa Halimi, Nora von Thenen	Emotion Detection	4.2: EMOTEX
2	Yalım Doğan, Mustafa Enes Karaca, Soner Koç	Spam in Popular Hashtags	2.2: HSpam14
3	F. Tuğba Doğan, M. Ali Yeşilyaprak, Simge Yücel	New Event Detection	8.3: Beyond Trending
4	Emre Doğan, Arif Yılmaz	Friend Recommendations	10.2: Graph-Based
5	Celal Öner, Arda Ünal	Event Detection	9.2: Learning
6	Mestan Fırat Çeliktuğ	Misinformation Detection	12.3: Rumor
7	Selim Eren Bekçe, Fouad Amira	New Event Detection	8.2: Erathquake
8	Nazanin Jafari, Naushin Faramarzi	Friend Recommendation	10.3: Followee
9	Volkan Küçük, Can Taylan Sarı	Language Detection Using Topic Modeling	11.1: Latent

Note that some of the papers are not directly related to the group project.

1. USER CLASSIFICATION

1. Marco Pennacchiotti, Ana-Maria Popescu: A Machine Learning Approach to Twitter User Classification. ICWSM 2011 Rec. by Group 2, February 22, 2016.

   This paper is mainly focused on classification of tweets in terms of their political orientation and ethnicity by leveraging observable information about users. Those are user behavior, their network structure(circle of friends) and their linguistic content of the user's Twitter feed. Rec. by Group 2 on Feb. 22, 2016.

2. SPAM DETECTION

1. F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, "Detecting Spammers on Twitter", Electronic messaging, Anti-Abuse and Spam Conference (CEAS), 2010. Rec. by Group 2: Yalım Doğan et al. on Feb. 22, 2016.

   Trending topics in Twitter, popular hashtags, becomes an opportunity for spam tweets containing URL's that leads users to unrelated websites for advertisement. The spammers are also adaptive, as they use URL shorteners to bait users easier because it gets harder to predict the website URL links to. They proposed a system for detection of spammers and spam tweets with a high classification ratio. Human volunteers' vote on high amount of tweets retrieved, in order to be used in classification with Support Vector Machine (SVM) together with defined parameters for both tweet context and users. They have used standard information retrieval metrics of recall, precision, Micro-F1 and Macro-F1.

2. Surendra Sedhai and Aixin Sun. 2015. HSpam14: A Collection of 14 Million Tweets for Hashtag-Oriented Spam Research. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '15). Rec. by Group 2: Yalım Doğan et al., February 22, 2016.

   Hashtags are important channels for creating virtual communities to aggregate information from all users of Twitter. Therefore, these hashtag channels may become target for spamming purposes, especially popular and trending hashtags. For this paper, 14 million tweets are collected and matched some trending hashtags. On this collection of tweets and matched hashtags, systematic annotation of the tweets being spam and ham is conducted. Annotated dataset is named as HSpam14. "annotation process includes four major steps: (i) heuristic-based selection to search for tweets that are more likely to be spam, (ii) near-duplicate cluster based annotation to firstly group similar tweets into clusters and then label the clusters, (iii) reliable ham tweets detection to label tweets that are non-spam, and (iv) Expectation-Maximization (EM)-based label prediction to predict the labels of remaining unlabeled tweets."

3. Wang, A.H. (2010) Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach. Lecture Notes in Computer Science, 6166, 335-342. Rec. by Group 2: Yalım Doğan et al., February 23, 2016.

Throughout this paper Twitter is studied as an example of spam bots detection in online social networking sites. A machine learning approach is proposed to distinguish the spam bots from reals. To facilitate the spam bots detection, three graph-based features, such as the number of friends and the number of followers, are extracted to explore the unique follower and friend relationships among users on Twitter. Three content-based features are also extracted from user's most recent tweets. A real data set is collected from Twitter's public available Tweets using two different methods.

3. TREND PREDICTION

1. Ma, Z., Sun, A. and Cong, G. (2013), On predicting the popularity of newly emerging hashtags in Twitter. J. Am. Soc. Inf. Sci., 64: 1399–1410. Rec. by Group 2: Yalım Doğan et al., February 22, 2016.

   Due to popularity of Twitter and the "viral nature of information dissemination" on Twitter, prediction of which Twitter topics will be popular in the near future will be prioritizable issue. Hastags are used in order to annotate tweets. This article proposes methods to predict the popularity of new hashtags on Twitter by formulating the problem as a classification task. Naïve bayes, k-nearest neighbors, decision trees, support vector machines, and logistic regression are classification models that are used for prediction.

4. EMOTION DETECTION

1. K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, S. M. Harabagiu, "EmpaTweet: Annotating and Detecting Emotions on Twitter", Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012. Rec. by Group 1: Didem Demirağ et al., on Feb. 19, 2016.

   Detecting emotions from tweets has its own challenges as tweets are short. Hence, there are new forms to express the emotions. In this paper, they analyze how emotions are distributed in the data. They used the annotated corpus to train a classifier that can automatically detect the emotions in tweets. They focus on seven emotions: Anger, disgust, fear, joy, love, sadness and surprise. They claim that this analysis should lead to the design of novel supervised and unsupervised emotion detection techniques.

2. M. Hasan, E. Rundensteiner, E. Agu, "EMOTEX: Detecting Emotions in Twitter Messages", ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, pages 27–31. Rec. by Group 1: Didem Demirağ et al. on Feb. 19, 2016.

   In this paper, they propose a new approach that can classify tweets of users in order to define their emotions. Therefore, they use hashtags as labels. Furthermore, they train supervised classifiers to detect multiple classes of emotion on large datasets. They consider different features like emoticons, negations and punctuations and define their utility in emotion detection. After comparing different machine learning algorithms like SVM, KNN, Decision Tree and Naïve Bayes, they show their technique has a 90% accuracy.

5. GENDER DETECTION

1. Z. Miller, B. Dickinson and W. Hu, "Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features," International Journal of Intelligence Science, Vol. 2 No. 4A, 2012, pp. 143-148. Rec. by Group 1: Didem Demirağ et al., on Feb. 19, 2016.

In today's world online social networks (OSNs) have become a popular way of communication. Most of the people are part of at least one social network producing an enormous amount of data. In this study, they identify the gender of the users on Twitter by using Naive Bayes and Perceptron. Stream application of these algorithms were used to handle the volume of tweets. Since tweets are regarded as informal text, they cannot be evaluated by using traditional dictionary methods. That is why n-gram features are used. Due to their large number, informative n-gram features were chosen using multiple selection algorithms. Both algorithms performed really well with a high precision.

2. J. D. Burger, J. Henderson, G. Kim, G. Zarrella, "Discriminating Gender on Twitter" Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11. Rec. by Group 1: Didem Demirağ et al., on Feb. 19, 2016.

   Burger et al. are using various different fields of the user's profile to determine their gender. They investigated statistical models for determining the gender of uncharacterized Twitter users. Not only they analyze the user's name and tweets, they also take into consideration fotos and preferences. This way they claim to out- perform other existing methods, as well as humans on the same task. As a result they define a construction on a large, multilingual dataset labeled with gender.

6. LOCATION DETECTION

1. Heravi, Bahareh, and Ihab Salawdeh. "Tweet Location Detection." Computation+ Journalism (2015). Rec. by Group : Didem Demirağ et al., on Feb. 19, 2016.

   Mapping Twitter data is an interesting way of visualizing Twitter conversations during certain events. Nevertheless, not many Twitter user share their location information which results into a little amount of tweets being location tagged. As stated in the paper these tweets are only about 1% of all the tweets at e.g. an event. In this paper Heravi and Salawdeh present an algorithm that determines a tweet's location. They compare their results to the existing CartoDB maps. By determining a tweet's location with their algorithm, they are able to visualize more tweets than the other methods.

7. USER PROFILING

1. P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh, K. P. Gummadi, "Inferring User Interests in the Twitter Social Network", RecSys'14, Foster City, Silicon Valley, CA, USA , 2014. Recommended by Group 1: Didem Demirağ et al., on Feb. 19, 2016.

   In the traditional methods, the tweets a user posts or receives are used to infer his topics of interest. In this paper, they propose a new method. This new method is based on observations. In Twitter users generally follow some experts on the topics they are interested in to acquire more information about those. So first, they find the expertise of the topics for popular Twitter users and then infer the interests of the users who follow them. According to the conducted experiments this approach is far more superior than some of the previous techniques. By using the above method they built a system Who Likes What, which tries to infer the topics of interests of Twitter users. This system will really be helpful for recommendation services to the user.

8. EVENT DETECTION

1. Li et al. "TEDAS: A Twitter-based Event Detection and Analysis System", 2012 IEEE 28th Conference on Data Engineering. Rec. by Group 7: Selim Erke Bekçe et al.

   This is an intuitive Twitter event detection infrastructure which mainly focuses on the detection crime and disaster related events w.r.t. spacial and temporal locality. They developed keyword

based tweet streaming, keyword rule generator, related tweet classification, location prediction heuristics for this task. Their infrastructure has both offline and online components.

2. Earle et al. "Twitter earthquake detection: earthquake monitoring in a social world", Annals of Geophysics, 54, 6, 2011. Rec. by Group 7: Selim Erke Bekçe et al.

   They ask the question whether it is possible to detect earthquakes by looking at tweets w.r.t. spatial and temporal locality. They developed an online anomaly detection algoritm which signals an event if there is an abnormal increase in targeted tweets. They then compare signaled events to real earthquakes for validation.

3. Becker, H.; Naaman, M.; and Gravano, L. 2011. Beyond trending topics: Real-world event identification on Twitter. Technical Report cucs-012-11, Columbia University. Rec. by Group 3: F. Tuğba Doğan et al.

   In this paper, real-world events and non-event messages are determined by analyzing Twitter messages' contexts. They create clusters from topically similar tweets using incremental clustering algorithm. They identify real-world events and non-event messages by using features of each cluster.

4. H. Abdelhaq, C. Sengstock and M. Gertz, "EvenTweet", Proc. VLDB Endow., vol. 6, no. 12, pp. 1326-1329, 2013. Rec. by Group 3: F. Tuğba Doğan et al.

   In "Even Tweet: Online Localized Event Detection from Twitter" paper they create a system, called Even Tweet, to detect localized events from a stream of tweets in real-time. Their system is different from the other evet detection systems, because they focuses on detecting localized events in real time by adopting a continuous analysis of the most recent tweets within a time-based sliding window. During this process they used a stream of tweets from the 2012 UEFA European Football Championship.

5. Ozdikis O, Senkul P, Oguztuzun H (2012) Semantic expansion of hashtags for enhanced event detection in Twitter. Workshop on online social systems (WOSS). Istanbul, Turkey. Rec. by Group 3: F. Tuğba Doğan et al.

   They present an event detection method in Twitter which focuses clustering of hashtags. They analyze the contexts of hashtags and their co-occurrence statistics with other words, we identify their paradigmatic relationships and similarities. They compare hashtag based methods with contents based methods.

9. COMMUNITY DETECTION

1. J. Yang, J. McAuley and J. Leskovec. Community Detection in Networks with Node Attributes. Rec. by Rec. by Group 5: Celal Öner et al.

   No summary.

2. J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012. Rec. by Group 5: Celal Öner et al.

   No summary.

10. FRIEND RECOMMENDATION

1. Collaborative and Structural Recommendation of Friends using Weblog-based Social Network Analysis (William H. Hsu, Andrew L. King, Martin S. R. Paradesi, Tejaswi Pydimarri, Tim Weninger) Rec. by Group 4: Emre Doğan et al. No summary.

2. A Graph-Based Friend Recommendation System Using Genetic Algorithm (Nitai B. Silva, Ing-Ren Tsang, George D.C. Cavalcanti, and Ing-Jyh Tsang). Rec. by Group 4Ş Emre Doğan et al. No summary.

3. Armentano, Marcelo Gabriel, Daniela Godoy, and Analía A. Amandi. "Followee recommendation based on text analysis of micro-blogging activity." *Information systems* 38.8 (2013): 1116-1127. Rec. by Group 8: Nazanin Jafari et al.

   Nowadays, number of people that are using social media increased finding reliable user to follow as an information seeker user is essential. In this paper they proposed a recommendation system which try to recommend user base on the content of micro-blog to detect user's interest. They want to profile users based on text analysis and another factors which let them to be the best source of following. Their proposed algorithm traverse a social graph to find a best candidate and then rank this candidate according to the inferred interest.

4. Manca, Matteo, Ludovico Boratto, and Salvatore Carta. "Mining User Behavior in a Social Bookmarking System-A Delicious Friend Recommender System." *DATA*. 2014. Rec. by Group 8: Nazanin Jafari et al.

   Social bookmarking systems are a form of social media system that allows to tag bookmarks of interest for a user and to share them. In this paper they want to recommend user base analyzing bookmarks tagged and frequency of tag. They have stated that using both of the measurement enhance the accuracy result and this paper want to infer user's interest from content.

11. LANGUAGE DETECTION

1. D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, January 2003. Rec. by Group 9: Hasan Kocaman et al.

No summary

2. T. Griffiths and M. Steyvers. Finding scientific topics. Proc. National Academy of Science, 2004. Rec. by Group 9: : Hasan Kocaman et al.

No summary

12. MISINFORMATION DETECTION

1. Kwon, Sejeong, et al. "Prominent features of rumor propagation in online social media." *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013. Rec. by Group 6 : M. Fırat Çeliktuğ.

   In the social media, it's very easy to disseminate an idea with respect to previous times.But, it creates the problem of validity of the information. As the valid informations, the fake(or false) infos have at least some general characteristics. In this paper, the characteristics of the rumors and

their propagations are examined in three aspects such as temporal, structural, and linguistic aspects. In terms of temporal characteristics, the newly proposed model is said to be used to see temporal features, furthermore, linguistic&structural differences between a valid and fake online information are said to be identified.

2. Friggeri, Adrien, et al. "Rumor Cascades." *ICWSM*. 2014. Rec. by Group 6 : M. Fırat Çeliktuğ.

   When there is a false information, it could be transferred from one person to the other easily. It's obvious that online social networks provide great opportunity for propagation of almost any type of information. In the other words, it creates the cascade of information sharing. In the paper, rumor propagation in Facebook is examined. The examination is said to be made by the help of referencing from "snopes.com" which is told to be a popular website docmenting memes and urban legends. It could be said that they generally tried to understand the propagation change and rate w.r.t. selected aspect according to given info. One mentioned interesting finding is that rumors change over time which is so similar to ear-to-ear game.

3. Qazvinian, Vahed, et al. "Rumor has it: Identifying misinformation in microblogs." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011. Rec. by Group 6 : M. Fırat Çeliktuğ

   Rumors are said to be either misinformation or disinformation. Besides, it's clearly seen in the society that identifying(or deciphering) the rumors are highly important for public&personal benefit due to problems sourced from wrong perception management. In this paper, detection of a rumor is said to be achieved by looking at effectiveness of three different aspects of features. The context(i.e. respective microblog) is chosen as twitter. The aspects are contentbased, network-based, and microblog-specific. By the title suggests, the focus is said to be on misinformation. But, effect of the given aspects on disinfomers(or disinformation) is said to be investigated.
4. One interesting thing about the paper is the large amount of annotated tweets(i.e. 10000 annotated