EECS 510, Spring 2016

Social Media Mining

Instructor

Professor Alok Choudhary

Technological Institute, L469
Northwestern University
2145 Sheridan Rd. #L469
Evanston, IL 60208

Office Hours: by appointment

TA: Rosanne Liu
Email: rosanne.liu@northwestern.edu
Office: Tech LG65
Office Hours: Tuesday, Thursday 2:00 –4:00pm

Course Description

Over the past decade, social media mining has witnessed a growing interest. There has been a significantly growth in the use of social media (e.g., Facebook, Twitter, YouTube, Google+, MySpace, StumbleUpon, etc.) and in the amount of information generated by users of social media. "Social Media" is producing massive amounts of data, so called "BIG DATA", with Volume, Velocity and Variety (the three "Vs" bigdata challenges) at an unprecedented scale. As a result, many traditional analysis techniques have become insufficient to investigate and predict complex dynamic social phenomena – especially when they require the understanding of many elements interacting in a multidimensional space, or billions of data points.

This course will explore several topics in collecting and analyzing social media, including social networks and analysis, information extraction, link analysis, behavior analytics, and recommendation. We will also address application issues like public opinion, sentiment analysis, privacy, trust, and reputation. This course will prepare students to grapple with the emerging problems from social media, and to learn representative approaches to data collection and innovatively applying multidisciplinary approaches to problem solving.

Prerequisites

Textbooks and Materials

Zafarani, Abbasi, and Liu. Social Media Mining: An Introduction, Cambridge University Press, February, 2014

Easley and Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World, (available in pdf), , Cambridge University Press 2010.

Community Detection and Mining in Social Media, Morgan & Claypool, 2010.

Social Network Analysis - Methods and Applications, Stanley Wasserman and Katherine Faust, Cambridge, 1994.

Course Requirements

The course is composed of readings, assignments and a class project.

- Class participation 10%
- Group Presentation 30%
    - 10% is for paper summaries
- Project 60%
    - 10% is for preliminary report

Approximately, the first two-third of the term will involve lectures on various topics. The rest one-third will involve student teams presenting and discussing papers in their chosen topics.

Schedule

Lecture 1, March 31, 2016 (Thursday, first week)

Topic: Introduction to Course: What is Social Media Mining? New Challenges for Social Media Mining

Lecture 2, April 5, 2016 (Tuesday, second week)

Topic: Graph Basics, Graph Representation, Special Graphs Mining

Lecture 3, April 7, 2016 (Thursday, second week)

Topic: Link prediction, who-follows-who problem in social networks

Lecture 4, April 12, 2016 (Tuesday, third week)

Topic: Data Mining Essentials 1: Data Mining Basics, Data Preprocessing

Lecture 5, April 14, 2016 (Thursday, third week)

Topic: Data Mining Essentials 2: Tutorial Data Mining in Python, IPython Notebook

Lecture 6, April 19, 2016 (Tuesday, fourth week)

Topic: Tools: Weka for General Data Mining

Lecture 7, April 21, 2016 (Thursday, fourth week)

Topic: Data Mining Essentials 3: Introduction to large scale data mining: Map Reduce, Spark, and Storm

Lecture 8, April 26, 2016 (Tuesday, fifth week)

Topic: NLP, Text Mining, document classification with MALLET

Lecture 9, April 28, 2016 (Thursday, fifth week)

Topic: Influence and Applications, Topic Modeling

Lecture 10, May 3, 2016 (Tuesday, sixth week)

Topic: Introduction to Deep Learning, with demos of theano, torch, tensorflow, keras

Lecture 11, May 5, 2016 (Thursday, sixth week)

Topic: Part 1: continue with deep learning demos

Part 2: Social applications of deep learning

Lecture 12, May 10, 2016 (Tuesday, seventh week)

Topic: Social media mining in practice: web scraping and the "Like"

Lecture 13, May 12, 2016 (Thursday, seventh week)

Topic: Big Data Science and Social Media in Business

Lecture 14-19 will be paper presentations

Lecture 14, May 17, 2016 (eighth week)

Lecture 15, May 19, 2016 (eighth week)

Lecture 16, May 24, 2016 (ninth week)

Lecture 17, May 26, 2016 (ninth week)

Lecture 18, May 31, 2016 (tenth week)

Lecture 19, June 2, 2016 (tenth week)

Topic: Student paper presentations and discussion

Logistics

1. Guideline on How to Read a Paper
2. A team of two students will present both papers in one of the topic categories listed below. Each topic should be covered by only one team.
3. Each student (other than the presenters) should submit a two-paragraph summary of the **best** paper **PRIOR TO** the class time in which the topic is being discussed. You will submit these via CANVAS, in the DISCUSSION for the course (each topic has its own thread, which will most likely be pinned prior to the class that discusses it; simply submit your summary as the next post). Note you can read previously submitted summaries, but obviously do not copy other students' work. The paper summaries should cover the following:

    1. A brief summary of the paper and its contributions
    2. Problem solving approach (concepts may or may not have covered in class).
    3. At least one area for improvement in the paper. A suggestion for "next steps" in the same direction that would make interesting future work.
    4. A brief assessment of whether you liked the paper, overall.

Papers

Note: You must be on the Northwestern network to access some of the pdfs.

Topic: Network Measures

1. Freeman, L. 1979 "Centrality in Social Networks: Conceptual Clarification", Soci al Networks 1, No. 3.
2. M. Franceschet 2011 "PageRank: standing on the shoulders of giants" Commun. ACM, Vol. 54, pp. 92 - 101.

Topic: Link Prediction Problem

1. David Liben-Nowell and Jon Kleinberg. 2003. "The link prediction problem for social networks". In Proceedings of the twelfth international conference on Information and knowledge management (CIKM '03). ACM, New York, NY, USA, 556-559.
2. Narang, K.; Lerman, K.; and Kumaraguru, P. "Network Flows and the Link Prediction Problem", In Proceedings of KDD workshop on Social Network Analysis (SNA-KDD), 2013.

Topic: Influence in Social Networks

1. Bakshy, J. M. Hofman, W. A. Mason, D. J. Watts. 2011, "Everyone's an influencer: quantifying influence on Twitter," In Proceedings of Int. Conf. on Web Search and Data Mining (WSDM)

2. Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K.P. 2010, "Measuring User Influence in Twitter: The Million Follower Fallacy, " In Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM).

Topic: Spam in Social Media

1. Grier, C., Thomas, K., Paxson, V., Zhang, M. 2010 "@spam: the underground on 140 characters or less" In Proceedings of the 17th ACM conference on Computer and communications security, pp. 27-37.
2. Xia Hu, Jiliang Tang, Yanchao Zhang, and Huan Liu. "Social Spammer Detection in Microblogging". In Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013).

Topic: Crowdsourcing

1. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M. (2008) "reCAPTCHA: Human-Based Character Recognition via Web Security Measures." Science 321 (September 1, 2008):1465-1468.
2. Yu Cheng, Zhengzhang Chen, Jiang Wang, Ankit Agrawal, and Alok Choudhary. "Bootstrapping Active Name Disambiguation with Crowdsourcing." In the 22nd ACM International Conference on Information and Knowledge Management, October 2013.

Topic: Social Tagging

1. Golder, S. and Huberman, B. 2005. The Structure of Collaborative Tagging Systems. Journal of Information Science, Vol. 32, No. 2.
2. Chi, E. and Mytkowicz, T. 2008. Understanding the efficiency of social tagging systems using information theory, in HyperText'08.

Topic: Modeling Individuals and Collective Behavior

1. Antisocial Behavior in Online Discussion Communities by J. Cheng, C. Danescu-Niculescu-Mizil, J. Leskovec. AAAI International Conference on Weblogs and Social Media (ICWSM), 2015.
2. Wu, F. and Huberman, B. 2007 "Novelty and collective attention" Proceedings of the National Academy of Sciences, Vol. 104, No. 45. (6 November 2007), pp. 17599-17601.
3. Lerman, K., and Hogg, T. 2011 "Using Stochastic Models to Describe and Predict Social Dynamics of Web Users" ACM Transactions on Intelligent Systems and Technology.

Topic: Weblog Analysis

1. Munson, S. and Resnick, P. (2011) The Prevalence of Political Discourse in Non-Political Blogs. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, July 17 – 21, 2011, Barcelona, Spain.
2. Gordon, A. and Swanson, R. (2009) "Identifying Personal Stories in Millions of Weblog Entries", Proceedings of the Third International Conference on Weblogs a

nd Social Media, Data Challenge Workshop, San Jose, CA, May 20, 2009.

Topic: Sentiment Analysis and Opinion Mining

1. Kunpeng Zhang, Yusheng Xie, Yu Cheng, Daniel Honbo, Doug Downey, Ankit Agrawal, Wei-keng Liao, and Alok Choudhary. Sentiment Identification by Incorporating Syntax, Semantics and Context Information. In the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, August 2012.
2. Morency, L., Mihalcea, R., and Doshi, P. (2011) Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. Proceedings of 13th International Conference on Multimodal Interaction ICMI-2011, Nov 14-18, Alicante, Spain.

Topic: Information Cascades and Social Epidemics

1. Information Cartography by D. Shahaf, C. Guestrin, E. Horvitz, J. Leskovec. Communications of the ACM (CACM), 2015.
2. Watts, D. A simple model of global cascades on random networks, in Proceedings of the National Academy of Sciences, Vol. 99, No. 9.
3. Liben-Nowell, D. and Kleinberg, J. 2008 "Tracing information flow on a global scale using Internet chain-letter data", Proceedings of the National Academy of Sciences, Vol. 105, No. 12, pp. 4633-4638.

Topic: Geospatial social networks (Professor Lerman)

1. Cheng, Z., Caverlee, J. and Lee, K. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. 19th ACM International Conference on Information and Knowledge Management (CIKM).
2. Scellato, S., Noulas, A., Lambiotte, R., Mascolo, C. 2011 "Socio-spatial Properties of Online Location-based Social Networks", In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)

Topic: Clustering Social Media data

1. Diana Palsetia, Mostofa Patwary, Ankit Agrawal, and Alok Choudhary. Excavating Social Circles Via User Interests. Journal of Social Network Analysis and Mining, 4(1), 2014.
2. Emilio Ferrara, Mohsen JafariAsbagh, Onur Varol, Vahed Qazvinian, Filippo Menczer, and Alessandro Flammini. "Clustering Memes in Social Media", Proceeding of IEEE/ACM Intl. Conf. on Advances in Social Networks Analysis and Mining ASONAM, *(*2013*)*.

Topic: Natural Disaster or Health care Discovery using Social Media

1. From Tweets to Wellness: Wellness Event Detection from Twitter Streams Mohammad Akbari, Xia Hu, Liqiang Nie, and Tat-Seng Chua AAAI Conference on Artificial Intelligence.
2. Suppawong Tuarob, Conrad S Tucker, Marcel Salathe and Nilam Ram, "An Ensemble Heterogeneous Classification Methodology for Discovering

Health-Related Knowledge in Social Media Messages", (In Press) Journal of Biomedical Informatics (2014).
3. Middleton, S.; Middleton, L.; Modafferi, S., "Real-time Crisis Mapping of Natural Disasters using Social Media", Intelligent Systems, IEEE , vol.PP, no.99, pp.1,1

Topic: Recommender Systems

1. Zhang K, Bhattacharyya S, Ram S. Large Scale Network Analysis for Online Social Brand Advertising. MISQ - Transformational Issues of Big Data and Analytics in Networked Business [Abstract]. 2015.
2. Ido Guy, Michal Jacovi, Adam Perer, Inbal Ronen, and Erel Uziel. 2010. "Same places, same things, same people?: mining user similarity on social media", In Proceedings of the 2010 ACM conference on Computer supported cooperative work (CSCW '10). ACM, New York, NY, USA, 41-50.
3. Liwei Liu, Nikolay Mehandjiev, and Dong-Ling Xu. 2011. "Multi-criteria service recommendation based on user criteria preferences", In Proceedings of the fifth ACM conference on Recommender systems (RecSys '11). ACM, New York, NY, USA, 77-84.

Topic: Topic Modelling

1. Daniel Ramage, Susan Dumais, and Dan Liebling. "Characterizing Microblogs with Topic Models", in Proc. ICWSM 2010, American Association for Artificial Intelligence , May 2010
2. Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. 2013. "Epidemiological modeling of news and rumors on Twitter", In Proceedings of the 7th Workshop on Social Network Mining and Analysis (SNAKDD '13). ACM, New York, NY, USA, , Article 8 , 9 pages.

Topic: Deep Learning and AI

1. Yann LeCun, Yoshua Bengio, & Geoffrey E. Hinton. "Deep learning", Nature, 521, 436–444. 2015
2. Alex Krizhevsky, Ilya Sutskever, & Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks", In Advances in Neural Information Processing Systems 25 (pp. 1097–1105). 2012

Project

A team of up to two students (can be different from the paper presentation group) is expected to complete a term project on social media mining related topics. Project ideas are provided below. You can come up with your own ideas also; be sure to talk to the TA to get approved.

Note: Reports need to be in **pdf** format and turned in via **CANVAS**.

Project proposals are due April 21st at 11:59PM, and expected to be about 1 page in length (single spaced). It should contain title, team member names, project goals and motivation, and related work.

Preliminary Reports are due May 12th at 11:59 PM. Each group must submit a two-page summary of their project progress by discussing 1) Steps you have completed, and any results you have obtained so far 2) The key remaining steps you plan to complete before the end of the quarter 3) Any questions or concerns you have regarding the project.

Final Reports are due June 9th at 11:59PM. The final report is expected to be about 4-10 pages in length (single spaced) and should include your project goals and motivation, along with a concise and clear statement of what results you obtained. Also mention which aspects of future work would be most interesting. Clarity in your report and presentation will contribute significantly to your grade.

Project Ideas

Some Social Media Mining project ideas from Kaggle (note datasets are already provided):

1) Sentiment Analysis on movie reviews
https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews

2) Yelp Recruiting
https://www.kaggle.com/c/yelp-recruiting

3) Influencers in Social Networks
https://www.kaggle.com/c/predict-who-is-more-influential-in-a-social-network

4) Job Recommendation
https://www.kaggle.com/c/job-recommendation

5) Best buy product recommendation
https://www.kaggle.com/c/acm-sf-chapter-hackathon-big

6) Million Song Dataset Challenge
https://www.kaggle.com/c/msdchallenge

7) Personality Prediction based on Twitter stream
https://www.kaggle.com/c/twitter-personality-prediction

8) IJCNN Social Network Challenge
https://www.kaggle.com/c/socialNetwork

9) Consumer Products Contest
https://www.kaggle.com/c/cprod1

10) KDD Cup Challenge 2012
https://www.kddcup2012.org/c/kddcup2012-track1
https://www.kddcup2012.org/c/kddcup2012-track2

More project ideas can be found at:
https://www.kaggle.com/competitions
https://www.kdd.org/kdd-cup

Network Datasets sources:

1) Stanford Large Network Dataset Collection
http://snap.stanford.edu/data/

2) ASU Datasets
http://socialcomputing.asu.edu/pages/datasets

Software for data mining

1. Tableau (http://www.tableau.com)
2. Weka
3. iPython notebook
4. Python libraries: numpy, scipy, keras, tensorflow, scikit-learn, nltk, matplotlib