

CS 6604 Middle Term Report

Computational Linguistics PJ

-Explore Correlation between
Newsires and Twitter

Client: Mohamed Magdy Farag

by Tianyu Geng, Wei Huang, Ji Wang, and Xuan Zhang

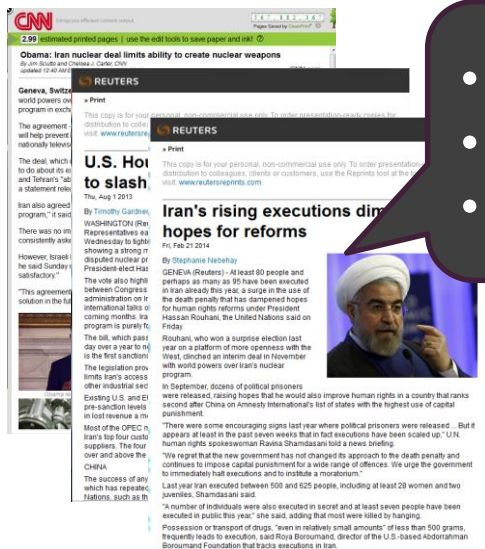
March, 6th, 2014
Blacksburg, VA

Table of Contents

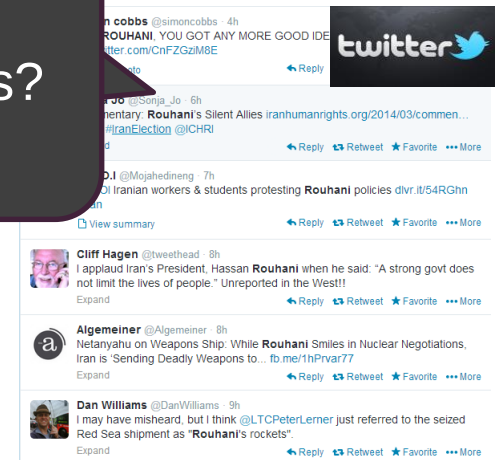
- *Introduction*
- *Solution*
- *Progress*

Problem to Solve

Motivation: Much news, much tweets, little connection...



- Key points in news?
- Relation between news & tweets?
- Major attitudes of audience?



Mainstream News

Tweets

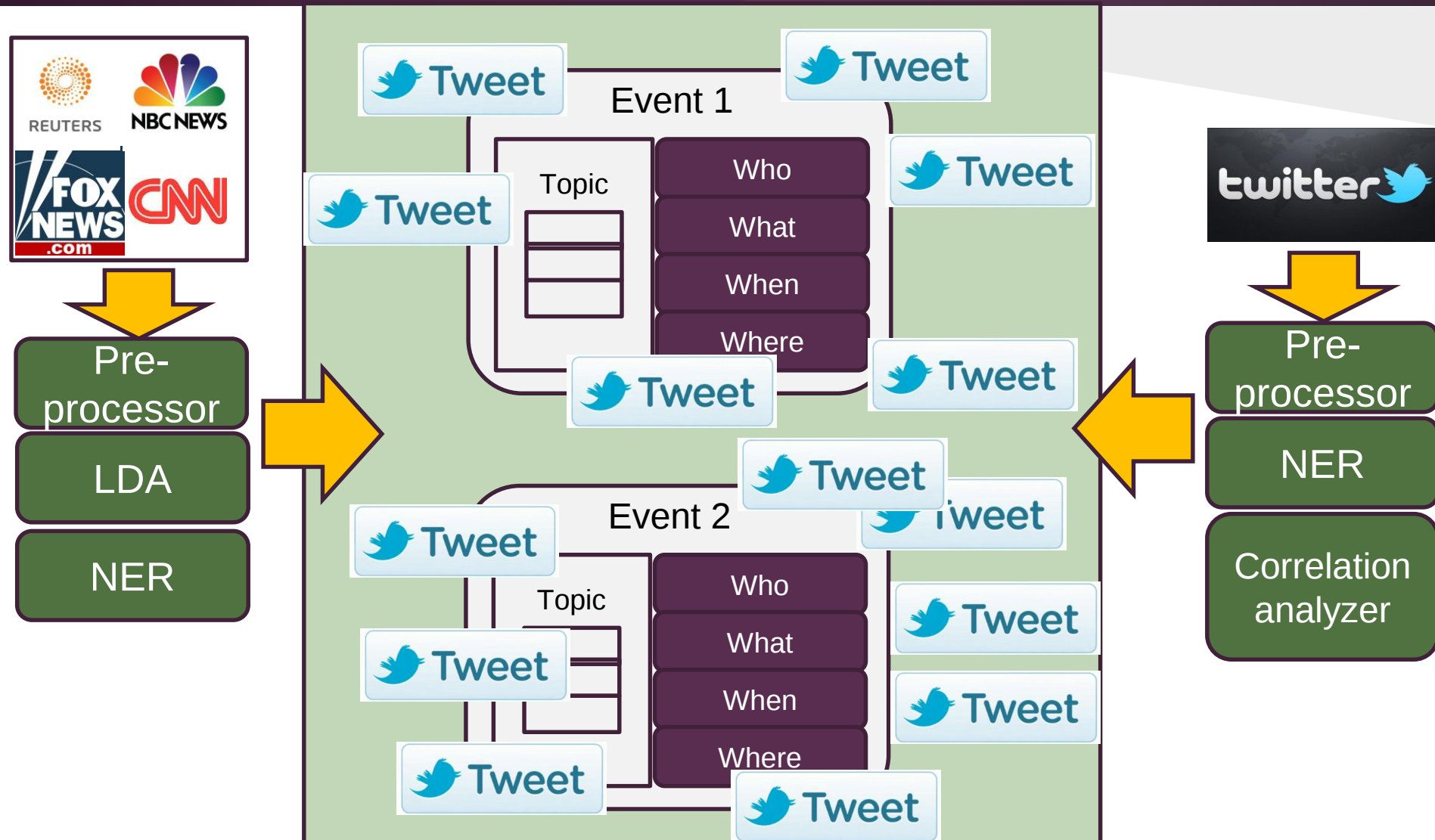
Objects:

1. Summarize info in news and tweets
2. Explore correlation between news & tweets
3. Mine opinions in tweets

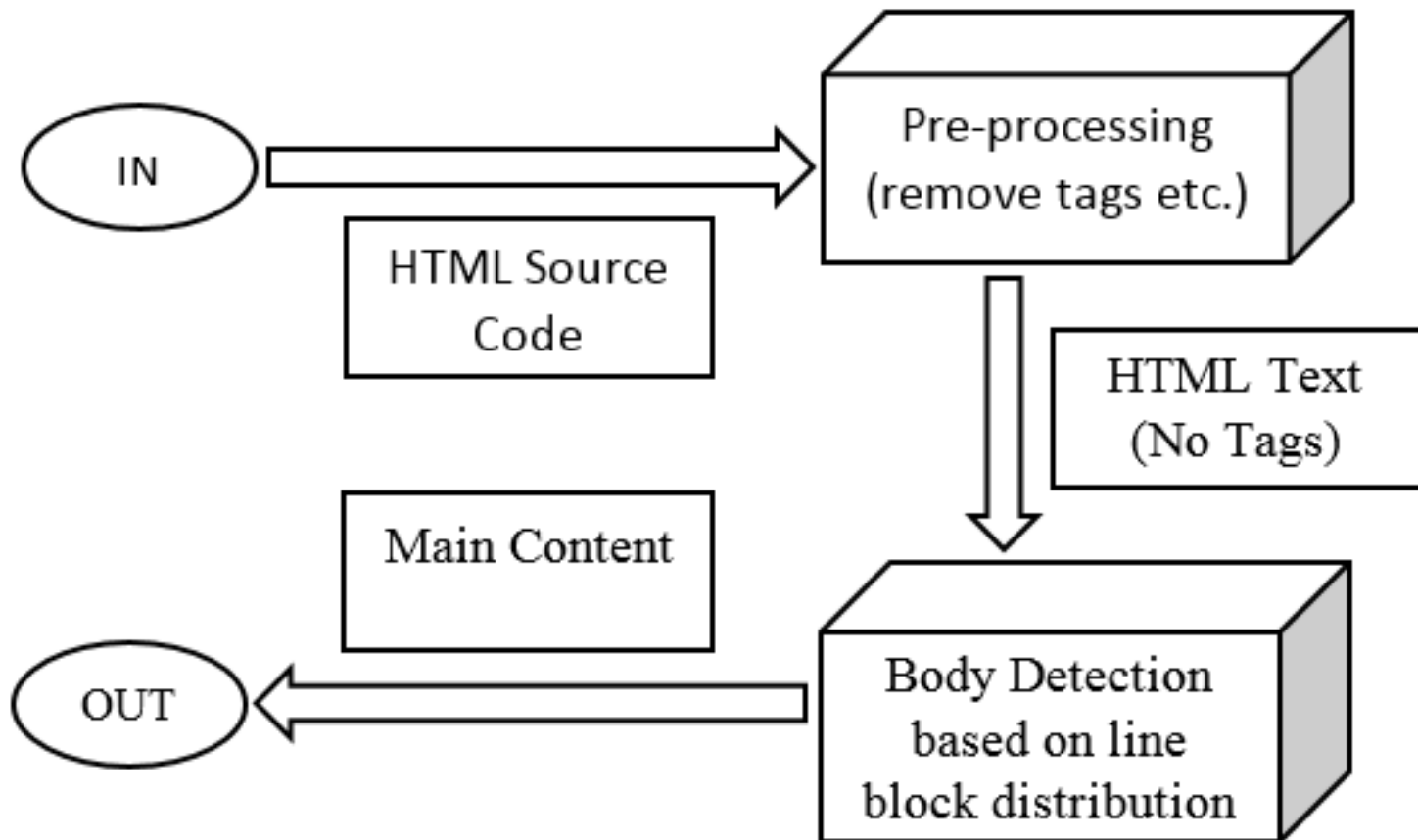
Solution Overview

1. Fetch text from news & tweets respectively
2. Preprocess texts: stemming, stop-word...
3. Extract events from news
Event: [Topic, Named entities(who, what, where, when)]
4. Map tweets to events (correlation model)
5. Mine major opinions around events

Solution: Link Tweets to Events



Progress: Text Extraction



Progress: Tweets Analysis

We ran LDA on a sample of the #Iran collection from IDEAL

- 50,000 tweets
- Feb 13, 2013 23:58:30 ~ Feb 15, 2013 00:00:03
- 4 topics

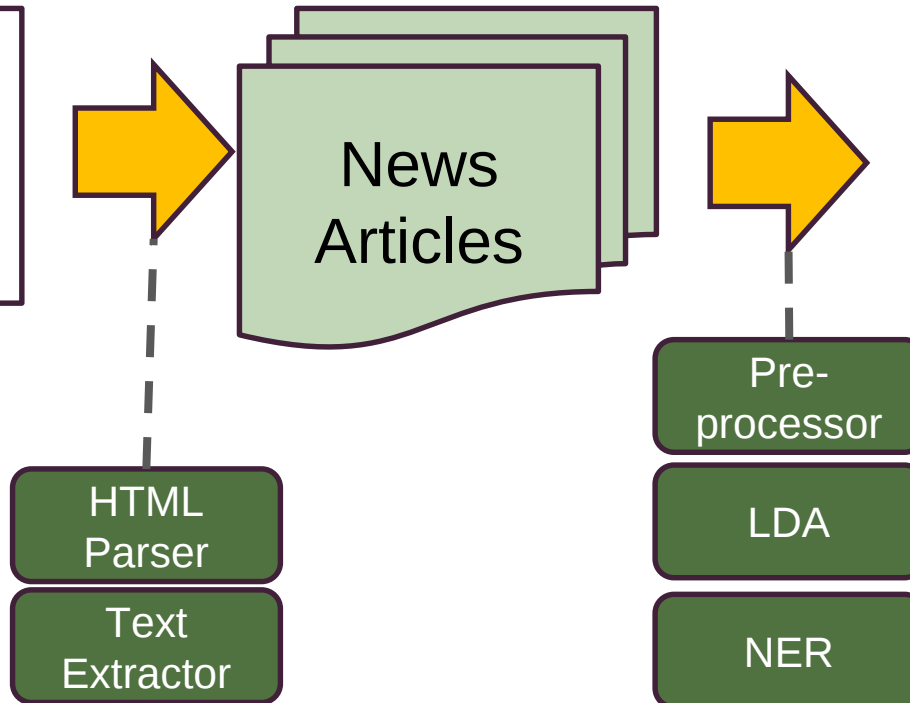
Topic 0		Topic 1		Topic 2		Topic 3	
<i>P</i>	<i>Keyword</i>	<i>P</i>	<i>Keyword</i>	<i>P</i>	<i>Keyword</i>	<i>P</i>	<i>Keyword</i>
0.028	nuclear	0.027	iran	0.018	iran	0.054	camp
0.018	weapons	0.019	time	0.017	via	0.053	liberty
0.015	iranian	0.015	now	0.012	2help	0.020	never
0.011	stop	0.013	opposition	0.011	killed	0.015	martin
0.010	menlu	0.012	u	0.010	sanctions	0.014	kobler
0.008	iran	0.012	feb	0.009	syria	0.013	mr
0.008	executions	0.011	seeking	0.009	guards	0.012	adequate

Progress: News Analysis

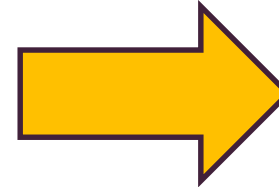
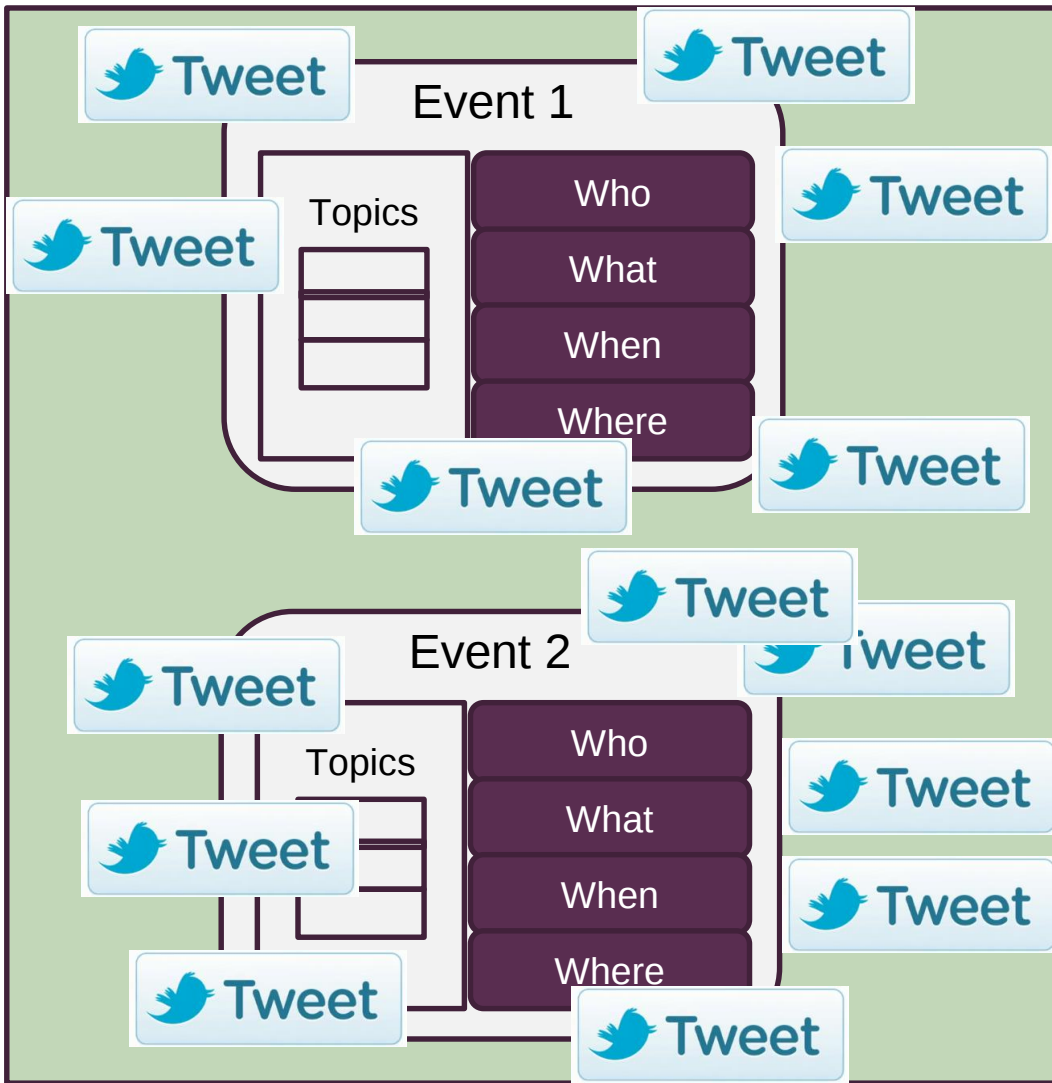
Dataset:

- 1) 2762 news about “Iran Election”.
--Only news titles used for topic modeling
- 2) News articles from CTRnet PJ

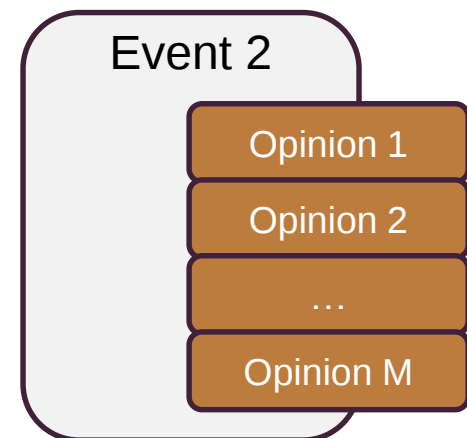
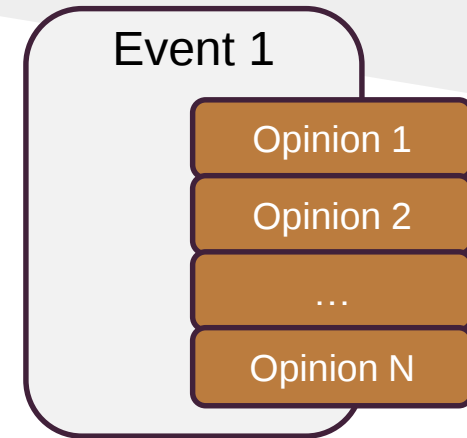
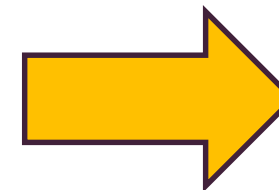
Tools: GibbsLDA, Stanford NLP



Future Work: Opinion Mining



Opinion Mining



Appendix

Literature Review

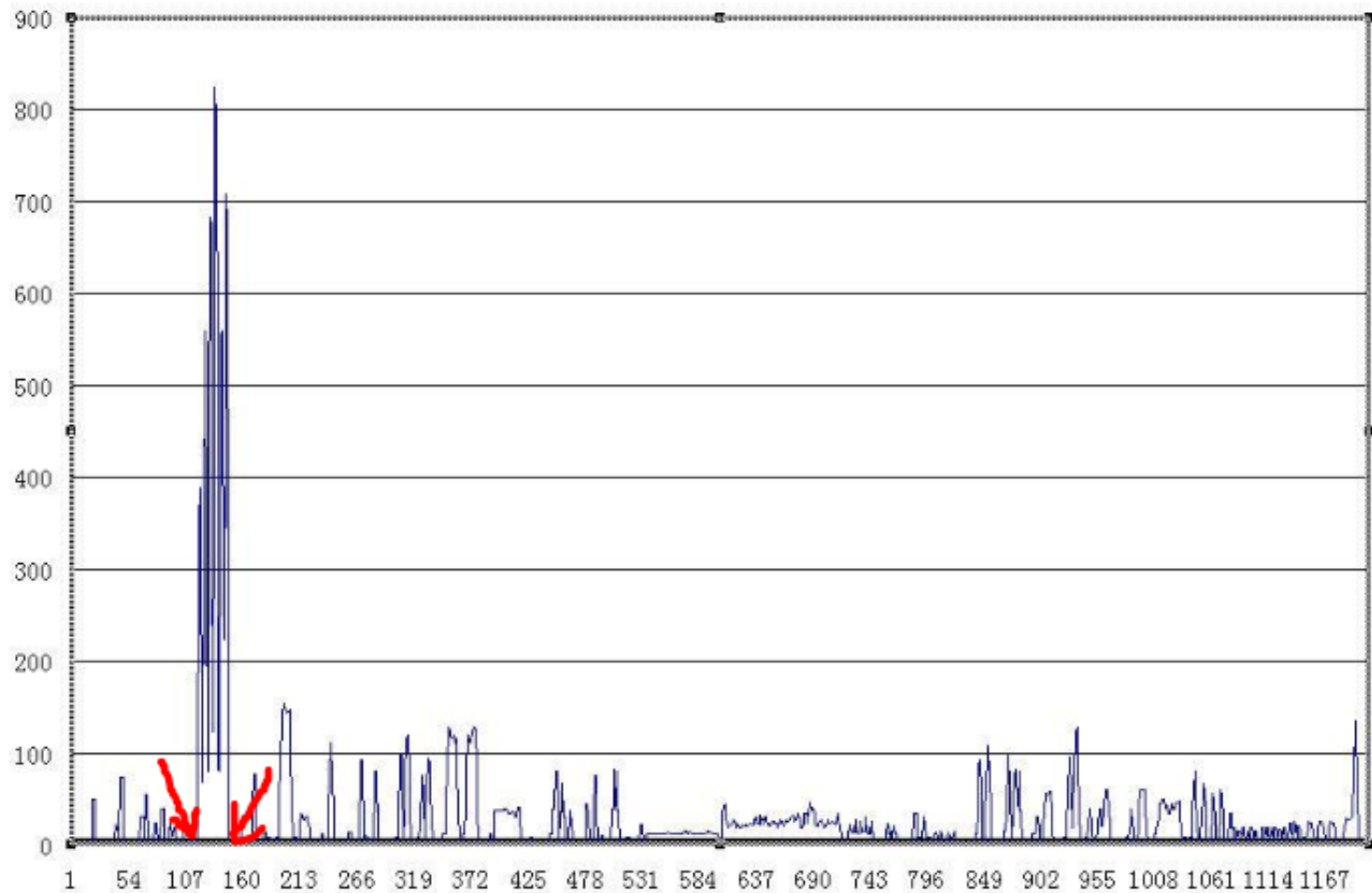
1. Analysis between Tweets and News Articles

- Fact: News providers report events earlier, but Twitter contains more details
- Algorithm: LDA, cosine similarity, sentiment analysis

1. Summary based on Templates

- Systems: SUMMARIST, Artequakt, etc.
- Topic signature is used for selecting summarizing sentences
- Using Apple Pie Parser, GATE and WordNet for knowledge extraction

The Characteristics of HTML



Reference

- [1] Petrovic S, Osborne M, McCreadie R, et al. Can twitter replace newswire for breaking news[C]//Seventh International AAAI Conference on Weblogs and Social Media. 2013.
- [2] Balahur A, Tanev H. Detecting Event-Related Links and Sentiments from Social Media Texts[J]. ACL 2013, 2013: 25.
- [3] Lobzhanidze A, Zeng W, Gentry P, et al. Mainstream media vs. social media for trending topic prediction-an experimental study[C]//Consumer Communications and Networking Conference (CCNC), 2013 IEEE. IEEE, 2013: 729-732.
- [4] Introne J E, Drescher M. Analyzing the flow of knowledge in computer mediated teams[C]//Proceedings of the 2013 conference on Computer supported cooperative work. ACM, 2013: 341-356.
- [5] Hovy E, Lin C Y. Automated text summarization and the SUMMARIST system[C]//Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998. Association for Computational Linguistics, 1998: 197-214.

Reference

- [6] Lin C Y, Hovy E. The automated acquisition of topic signatures for text summarization[C]//Proceedings of the 18th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2000: 495-501.
- [7] Luhn H P. The automatic creation of literature abstracts[J]. IBM Journal of research and development, 1958, 2(2): 159-165.
- [8] Alani H, Kim S, Millard D E, et al. Automatic ontology-based knowledge extraction from web documents[J]. Intelligent Systems, IEEE, 2003, 18(1): 14-21.
- [9] El Hamali S, Nouali O, Nouali-Taboudjemat N. Knowledge extraction by Internet monitoring to enhance crisis management[R]. CERIST, 2011.