INFORMATION RETRIEVAL PROJECTS

AUG-NOV 2018

Department of CSE,
IIT Hyderabad

1. TWITTER GEOLOCATION PREDICTION

Given a tweet, predict the geolocation from which the tweet was generated.

References:

A. https://arxiv.org/pdf/1805.04612.pdf,

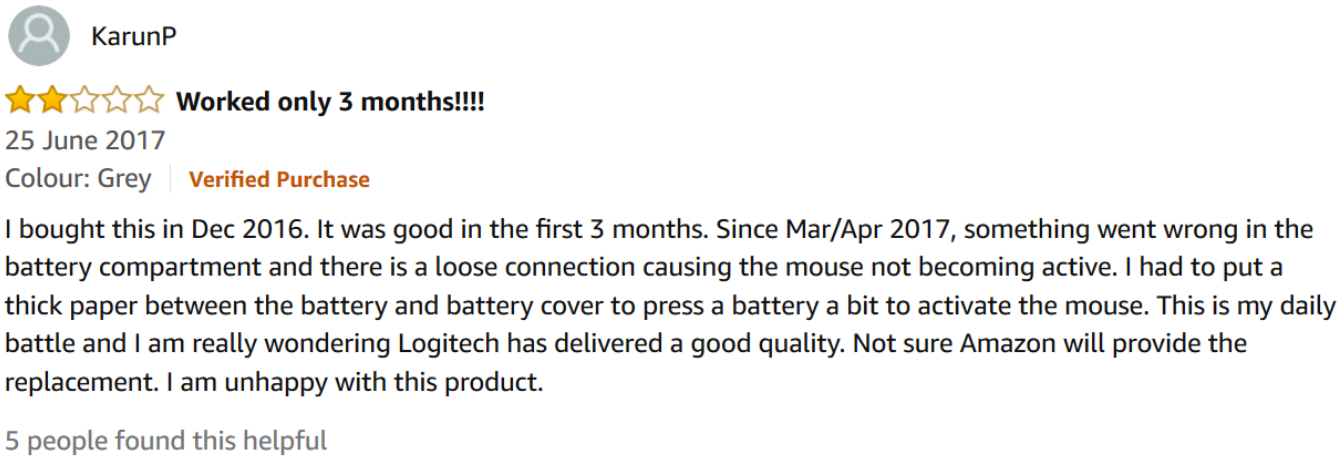
B. http://www.aclweb.org/anthology/D12-1137

C. https://cs.stanford.edu/~jurgens/docs/jurgens-et-al_icwsm-2015.pdf

Data: http://www.cs.cmu.edu/~ark/GeoText/

2. PREDICTING PRODUCT REVIEW HELPFULNESS

Reviews are an extremely important component of any ECommerce system. The purpose of this project is to find, given an input review (and few other associated details), the helpfulness of the review.



KarunP

★★☆☆☆ **Worked only 3 months!!!!**

25 June 2017

Colour: Grey | **Verified Purchase**

I bought this in Dec 2016. It was good in the first 3 months. Since Mar/Apr 2017, something went wrong in the battery compartment and there is a loose connection causing the mouse not becoming active. I had to put a thick paper between the battery and battery cover to press a battery a bit to activate the mouse. This is my daily battle and I am really wondering Logitech has delivered a good quality. Not sure Amazon will provide the replacement. I am unhappy with this product.

5 people found this helpful

3. PREDICTING RATINGS FROM REVIEWS FOR A PRODUCT

There are many systems that allow users to enter reviews as well as ratings. In this problem, we will look at the reviews that also have associated ratings with it. Given a review text, you have to find out the rating that the reviewer gave to the product for which the review was written by him/her.

References:

http://www.cs.ust.hk/~qyang/Docs/2011/IJCAI11-305.pdf

http://www.anthology.aclweb.org/W/W10/W10-1205.pdf

Dataset: http://jmcauley.ucsd.edu/data/amazon/

4. PREDICTING NUMBER OF REVIEWS FOR A PRODUCT

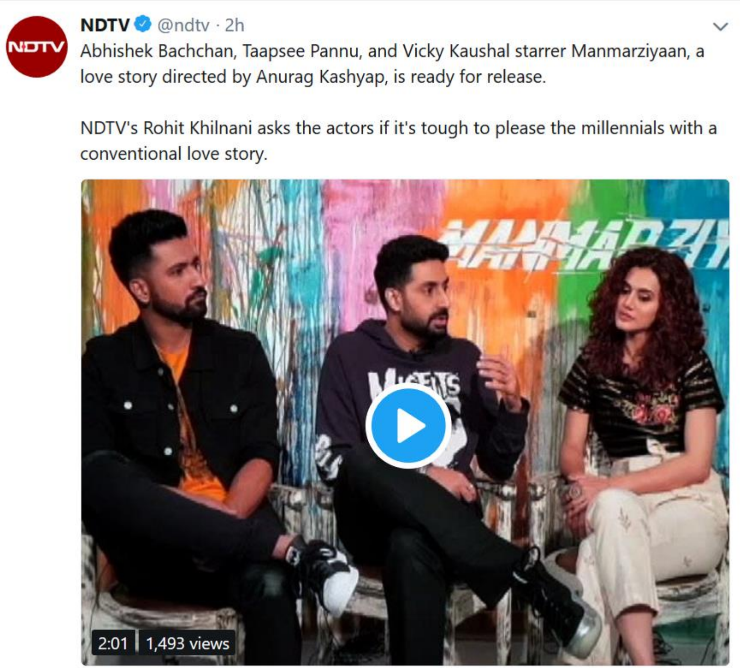
Given a product, predict the number of reviews that the product will have in a future time window. For example, for a product p and at time t, predict the number of reviews it will have between time (t+T1) and (t+T2), with both T1 and T2 as positive integers.

References:

Dataset: http://jmcauley.ucsd.edu/data/amazon/

5. Matching YouTube Videos Against Tweets

Consider a tweet that has a YouTube video link. Consider the scenario where the the video link is masked, and you have to guess the link.



NDTV @ndtv · 2h

Abhishek Bachchan, Taapsee Pannu, and Vicky Kaushal starrer Manmarziyaan, a love story directed by Anurag Kashyap, is ready for release.

NDTV's Rohit Khilnani asks the actors if it's tough to please the millennials with a conventional love story.

Reference:

Data: Available with us, will be shared. The participants may also need to collect some additional data for this task.

6. HIERARCHICAL GRAPH CLUSTERING USING REPRESENTATION LEARNING

In this work, our goal is to come up with a representation of weighted graphs as low dimensional embeddings which preserves the hierarchical relationships among the nodes in the input graph.

Efforts: 80% implementation, 20% research.

Dataset: http://manikvarma.org/downloads/XC/XMLRepository.html

References:

a. https://arxiv.org/abs/1706.07845
b. https://github.com/GTmac/HARP

7. EXTREME CLASSIFICATION

Goal: Extreme Multi-label Classification (XMC) refers to supervised multi-label classification problems where we need to find the most relevant labels for each data point from very large (possibly millions) set of labels. We will try to come up with an low dimensional embeddings of the features and labels using Autoencoders.

Dataset: http://manikvarma.org/downloads/XC/XMLRepository.html

References:

a. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12423
b. http://www.kdd.org/kdd2016/papers/files/rfp0191-wangAemb.pdf
c. https://github.com/suanrong/SDNE

8. DETECTING DUPLICATE QUESTION IN CQA SYSTEM

In community question answering system, many people ask same question in different ways, objective of the project is to find duplicate questions. Moreover system can be scalable to prevent user from adding new duplicate questions.

Dataset: Quora Question Pairs , Stack Overflow

References: Paper 1 , Paper 2

9. FACT CHECKING IN CQA SYSTEM

CQA system are very popular and effective nowadays, but all the information is not always true/fact. Objective is to check the fact/information mention in the answer using multiple things link answer content/ author profile/ other answers/ external sources.

Dataset: CQA-QL-2016-fact: A Dataset for Fact Checking in cQA

Reference: https://arxiv.org/pdf/1803.03178.pdf, AAAI, 2018

10. FINDING ACTIVE EXPERT USERS IN CQA SYSTEM

In every CQA system, people are generally active in certain period. Finding the **active** users who answers most of the question **correctly** and route the newly added questions to those users. (Defining expertness and activeness can be tricky)

Dataset: Any CQA system like Stack Exchange

11. PAPER VENUE PREDICTION

Consider the problem of matching the topics of a scientific paper with those of possible publication venues for that paper. While every researcher knows the few top-level venues for his specific fields of interest, a venue recommendation system may be a significant aid when starting to explore a new research field. The aim is to predict the publication venue based on only title and abstract, differently from previous works which require full-text and reference list: Hence, Future developed system can be used even in the early stages of the authoring process and greatly simplifies the building and maintenance of the knowledge base necessary for generating meaningful recommendations.

References: A. https://ieeexplore.ieee.org/abstract/document/6984588/
B. https://arxiv.org/pdf/1612.05817.pdf

Database: AMiner Computer Science, Database and Information Systems (DBIS)

12. FINDING SIMILAR USERS IN MULTIPLE EVENTS FROM CRYPTOCURRENCY NETWORK

Given the wide array of uses for cryptocurrencies - as a money transmission mechanism, as a store of value, as a way to avoid international remittance fees, as a way to facilitate currency trading, and even as a way to facilitate gambling and trade in illegal merchandise - the importance of finding similar users is of vital importance. The problem can be started with **community detection** and could exploit other information related to cryptocurrencies.

Dataset: Bitcoin OTC trust weighted signed network

13. PATTERN ANALYSIS IN PRICES OF CRYPTOCURRENCY

New cryptocurrencies has received much attention by the media and investors alike due to the assets' innovative features, potential capability as transactional tools, and tremendous price fluctuations. Thus, analyzing evolutionary dynamics of the cryptocurrency market is a topic of current interest and can provide useful insight about the market share of cryptocurrencies. It has been shown that social media data such as Twitter can be used to track investor sentiment, and price changes in the Bitcoin market and other predominant cryptocurrencies. Hence, using Twitter sentiment to analyze price fluctuations of nascent alternative cryptocurrencies could provide valuable insight, and eventually lead to a viable arbitrage opportunity in other emerging alternative cryptocurrencies.

Dataset: Kaggle Bitcoin Historical Dataset

References: Paper

14. INLINE MATHEMATICAL EXPRESSION DETECTION IN SCIENTIFIC DOCUMENTS

One of the issues in extracting natural language sentences from PDF documents is the identification of non-textual elements in a sentence. In-line mathematical expressions include complex mathematical structures, such as “∑” or “∫ ”, in addition to symbols or variables that accompany their explicit natural language definitions, such as “where w is a sequence of words” or “the probability distribution p(W|c).” Identifying this notation is useful not only for reducing the errors of sentence parsing, but also enabling further scientific text mining because mathematical expressions often convey key concepts in scientific information dissemination.

Reference: https://dl.acm.org/citation.cfm?id=3121041

Dataset: We have the dataset with us. It will be shared with the students separately.

15. IRONY DETECTION IN ENGLISH TWEETS

The frequent use of irony on social media has important implications for natural language processing tasks, which struggle to maintain high performance when applied to ironic text. Although different definitions of irony co-exist, it is often identified as a trope or figurative language use whose actual meaning differs from what is literally enunciated.

P1: The first problem is a two-class (or binary) classification where the system has to predict whether a tweet is ironic or not. The following sentences present examples of an ironic and non-ironic tweet, respectively.

- I just love when you test my patience!! #not
- Had no sleep and have got school now #not happy

P2: The second problem is a multiclass classification where the system has to predict one out of four labels describing i) verbal irony realized through a polarity contrast, ii) verbal irony without such a polarity contrast (i.e., other verbal irony), iii) descriptions of situational irony, iv) non-irony

References:

A. http://aclweb.org/anthology/C16-1257
B. https://pdfs.semanticscholar.org/5b12/a7f95cbf7aac388155b8edb5cc2380c1b19f.pdf
C. https://arxiv.org/abs/1602.03426
D. https://dl.acm.org/citation.cfm?id=2745994

Dataset: Link

16. Retrieving Tweets Related to News Articles

Nowadays, social media users react in real-time to local and global events. Therefore, social media can be used to measure the impact of particular topics or events and to analyze public opinion. To this end, identifying and ranking social media posts, such as tweets, associated with a news article is an important information retrieval task.

Ref:

A. http://research.signalmedia.co/publications/signal1m-tweet-retrieval.pdf
B. https://github.com/lucene4ir/lucene4ir
C. https://github.com/castorini/Anserini

Dataset: Link

17. Mining Adverse Drug Reaction Mentions in Social Media

Social media is becoming increasingly popular as a platform for sharing personal health-related information. This information can be utilized for public health monitoring tasks, particularly for pharmacovigilance, via the use of Natural Language Processing (NLP) techniques. However, the language in social media is highly informal, and user-expressed medical concepts are often non-technical, descriptive, and challenging to extract. There has been limited progress in addressing these challenges, and thus far, advanced machine learning-based NLP techniques have been underutilized. Our objective is to design a machine learning-based approach to extract mentions of adverse drug reactions (ADRs) from highly informal text in social media.

Ref:

A. https://www.ncbi.nlm.nih.gov/pubmed/25755127
B. https://www.sciencedirect.com/science/article/pii/S1532046415000362
C. http://diego.asu.edu/psb2016/acceptedpapers/DLIR.pdf
D. https://arxiv.org/pdf/1802.05121.pdf

Dataset: Link

18. Detecting Personal Medication Intake in Twitter

Social media sites (e.g., Twitter) have been used for surveillance of drug safety at the population level, but studies that focus on the effects of medications on specific sets of individuals have had to rely on other sources of data. Mining social media data for this information would require the ability to distinguish indications of personal medication intake in this media.

Ref:

A. http://aclweb.org/anthology/W17-2316

Dataset: Link

19. Aggression Detection in Social Media

As the interaction over the web has increased, incidents of aggression and related events like trolling, cyberbullying, flaming, hate speech, etc. too have increased manifold across the globe. While most of these behaviour like bullying or hate speech have predated the Internet, the reach and extent of the Internet has given these an unprecedented power and influence to affect the lives of billions of people. So it is of utmost significance and importance that some preventive measures be taken to provide safeguard to the people using the web such that the web remains a viable medium of communication and connection, in general.

Goal: To beat the best F1-Score 0.6425 (Baseline)

Ref:

A. https://arxiv.org/abs/1803.09402
B. https://web.science.mq.edu.au/~smalmasi/trac1/pdf/W18-4411.pdf

Dataset: Link

20. Fake News Detection

Fake news, defined by the New York Times as “a made-up story with an intention to deceive”, often for a secondary gain, is arguably one of the most serious challenges facing the news industry today. In a December Pew Research poll, 64% of US adults said that “made-up news” has caused a “great deal of confusion” about the facts of current events.

Ref:

A. http://www.aclweb.org/anthology/P17-2067
B. https://arxiv.org/pdf/1707.03264.pdf
C. https://onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2015.145052010082
D. https://onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2015.145052010083
E. https://arxiv.org/pdf/1803.05355.pdf (FEVER)
F. https://arxiv.org/pdf/1804.08559.pdf

Dataset: Link

21. SEMANTIC TEXTUAL SIMILARITY

Semantic Textual Similarity (STS) measures the degree of equivalence in the underlying semantics of paired snippets of text. While making such an assessment is trivial for humans, constructing algorithms and computational models that mimic human level performance represents a difficult and deep natural language understanding (NLU) problem.

Ref:

A. https://aclweb.org/anthology/S/S16/S16-1081.pdf
B. https://dl.acm.org/citation.cfm?id=2806475

Dataset: Link Look for english sentences.

22. CLICKBAIT IDENTIFICATION

Clickbait refers to a certain kind of web content advertisement that is designed to entice its readers into clicking an accompanying link. Typically, it is spread on social media in the form of short teaser messages that may read like the following examples:

- A Man Falls Down And Cries For Help Twice. The Second Time, My Jaw Drops
- 9 Out Of 10 Americans Are Completely Wrong About This Mind-Blowing Fact
- Here's What Actually Reduces Gun Violence

When reading such and similar messages, many get the distinct impression that something is odd about them; something unnamed is referred to, some emotional reaction is promised, some lack of knowledge is ascribed, some authority is claimed. Content publishers of all kinds discovered clickbait as an effective tool to draw attention to their websites.

Data: https://webis.de/data/webis-clickbait-17.html#download

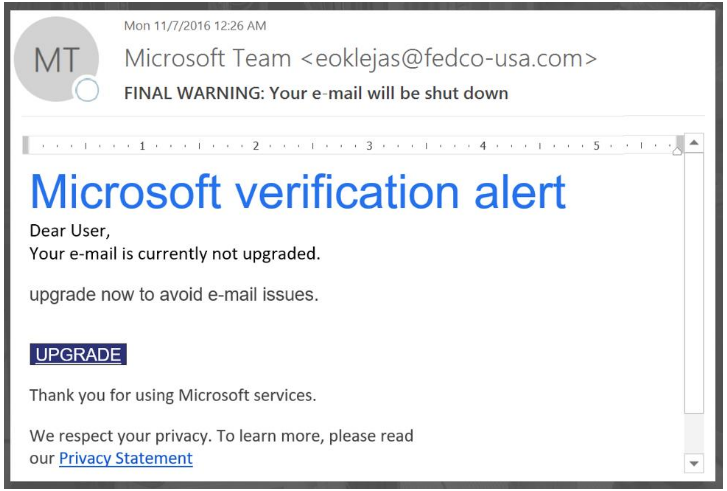
References:
https://arxiv.org/pdf/1710.01507.pdf
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7752207
https://arxiv.org/pdf/1710.02861.pdf
https://arxiv.org/pdf/1710.06699.pdf
https://people.mpi-sws.org/~babaei/FairUMAP2018.pdf

23. DETECTING PHISHING EMAILS

Phishing is the fraudulent attempt to obtain sensitive, often for malicious reasons, by disguising as a trustworthy entity in an electronic communication.

References: https://www.hindawi.com/journals/jam/2014/425731/

Data: https://spamassassin.apache.org/old/publiccorpus/



Mon 11/7/2016 12:26 AM

Microsoft Team <eoklejas@fedco-usa.com>

FINAL WARNING: Your e-mail will be shut down

Microsoft verification alert

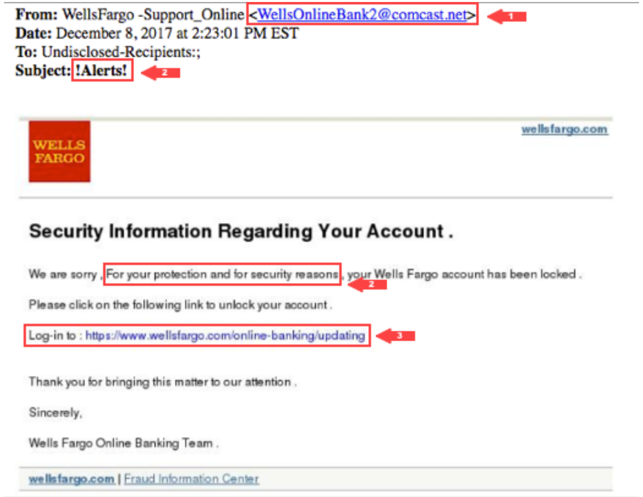
Dear User,
Your e-mail is currently not upgraded.

upgrade now to avoid e-mail issues.

UPGRADE

Thank you for using Microsoft services.

We respect your privacy. To learn more, please read our Privacy Statement



From: WellsFargo -Support_Online <WellsOnlineBank2@comcast.net>
Date: December 8, 2017 at 2:23:01 PM EST
To: Undisclosed-Recipients:;
Subject: !Alerts!

WELLS FARGO

wellsfargo.com

Security Information Regarding Your Account .

We are sorry , For your protection and for security reasons , your Wells Fargo account has been locked .

Please click on the following link to unlock your account .

Log-in to : https://www.wellsfargo.com/online-banking/updating

Thank you for bringing this matter to our attention .

Sincerely,

Wells Fargo Online Banking Team .

wellsfargo.com | Fraud Information Center

24. FINDING SIMILARITY BETWEEN WORDS

Given a pair of words, find similarity between them.

25. FINDING SIMILARITY BETWEEN WORDS

Understanding the various senses of a token in a domain specific corpus and match those senses to predefined global senses and creating a knowledge graph in the process.

Example: Tokens with similar syntax has multiple interpretations according to the context they occur in.

1. Ron went to the *bank* to withdraw some money to sustain his company.
2. Alex likes the cool breeze at the river *bank*, helps him keep the company on track.
3. Tim is someone you can *bank* upon.

Data: To be provided

References:

http://aclweb.org/anthology/W/W16/W16-1620.pdf

https://arxiv.org/pdf/1511.06388.pdf

https://pdfs.semanticscholar.org/2f43/cf76760dd6d945801369f282fe9b38cbad58.pdf

26. VECTOR REPRESENTATION OF WORD SENSES

Understanding the various senses of a token in a domain specific corpus and match those senses to predefined global senses and creating a knowledge graph in the process.

Example: Tokens with similar syntax has multiple interpretations according to the context they occur in.

1. Ron went to the *bank* to withdraw some money to sustain his company.
2. Alex likes the cool breeze at the river *bank*, helps him keep the company on track.
3. Tim is someone you can *bank* upon.

Data: To be provided

References:

http://aclweb.org/anthology/W/W16/W16-1620.pdf

https://arxiv.org/pdf/1511.06388.pdf

https://pdfs.semanticscholar.org/2f43/cf76760dd6d945801369f282fe9b38cbad58.pdf

27. IDENTIFYING IMPORTANT PHRASES

To detect phrases in domain specific corpus using Supervised learning and/or Deep learning methods.

Data: To be generated/provided

References: https://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

28. AUTOMATED TEXT SUMMARIZATION

Given a single text document (can be multiple documents too) generate a summary of the text.

Data:

References:

30. DETECTING TRAFFIC-RELATED EVENTS FROM ONLINE MEDIA

In this project, the purpose will be to

1. Develop a mechanism to identify posts/articles (tweets/blogs/news articles) that report traffic related incidents in an automated manner. Example of traffic-related incidents could be accident, road-blockade, infrastructure damage resulting in blockage of road, rallies etc.
2. Develop clustering/classification mechanism for the traffic related posts. For classification, the classes could be the event types mentioned in the above task.

29. QUERY SEGMENTATION

Query segmentation is the task of breaking (Segment) the query into multiple adjacent phrases so that each segment can refer to something meaningful, and helps in specifying an important aspect of the query.

For example, one possible segmentation for the query “reinforcement learning recent papers” could “[reinforcement learning][recent papers]”.

Data: https://zenodo.org/record/1137746
https://www.uni-weimar.de/en/media/chairs/computer-science-department/webis/data/corpus-webis-qsec-10/#webis-download
References:
https://dl.acm.org/citation.cfm?id=2348401
https://www.cse.iitb.ac.in/~soumen/doc/www2013/QirWoo/HagenPSB2011QuerySegmentRevisit.pdf
https://www.uni-weimar.de/medien/webis/publications/papers/stein_2012q.pdf
http://difabbrizio.com/papers/sigir-ecom-2017-qs.pdf

31. SYSTEM DESIGN: SEARCH ENGINE FOR SENATE DOCUMENTS

The Senate is the academic decision making body of IIT Hyderabad. The IITH Senate often (roughly twice a year) holds meetings where several academic policy decisions are made. The discussions and the decisions made in the senate meetings are documented. Those documents are called senate minutes. In this project, the goal is to index the senate minutes, and have a search functionality over this indexed collection. For this project, Several decisions need to be made that may cater to aspects like granularity of document, use of metadata for the documents/mini-documents, ranking strategy, GUI choice etc. Apache Solr, which provides an open source framework for developing search engines can be used for this purpose.

32. SYSTEM DESIGN: INFORMATION PORTAL FOR FOLK ART IN INDIA

The history of folk art in India is very rich. There are so many traditional art forms that have passed the test of time and draw a lot of attention from people in India and around the globe. Examples of few such art forms include: Madhubani, Warli, Kalamkari, Tanjore, Patachitra etc. Information about these art forms is scattered in the web. As part of this project, we are aiming to create a (searchable) portal that contains various bits of information about such art forms. The portal, for example, may include the following features:

- Searching in a collection of documents on Indian traditional art
- Searching in a collection of images
- Notification when an article on these topics is published in any online newspaper/magazine
- Notification when any exhibition displaying arts of such type is organized