Project Title

Using Topic Modelling to determine the relationships between Hashtags and Tweets.

Names of the member(s) of the group.

Olashile Adebimpe B00761325

Problem Statement.

In recent years, the tremendous growth in popularity of social media such as Facebook and Twitter has made them a very important source of information. This large amount of data available on these internet platforms has inspired new opportunities for text mining information and analytics which helps understand and solve real-life problems in the world[1] [2].

With the large volumes of dynamic text content generated on these services every second, there is a need for better comprehension to see how relevant are these generated data to their default categorization as well as provide useful information for further insight and analysis.

Our choice of social media platform will be twitter due to its abundance of content from the enormous amount of tweets and our goal is to develop a system which tries to find a correlation between Hashtags on the Twitter dataset and their related tweets.

List of possible approaches with citations to relevant work.

The goal of this research is to develop a system which tries to find a correlation between Hashtags in the Twitter dataset and their related tweets and our approach employs a defined tweet aggregation algorithm based on sentiment analysis on the preprocessed tweet and also the period of time the tweet was posted. We will be using several topic modeling algorithms to analyze the aggregated tweet, in order to extract topics, which are then evaluated by the use topic coherence analysis and some similarity measure algorithm to measure the closeness of our modeled topic to the related Hashtag.

1. Collection of Tweets: In order to access Twitter dataset, we used Tweepy [3], a python library for accessing Twitter for the collection of tweets.

2. Preprocessing: Also called data cleaning, an important step in our approach where we remove stop words, punctuations, mentions, hashtags, URLs and also normalize the corpse.
3. Sentiment Analysis: Sentimental analysis using a high-level library called TextBlob [4]. TextBlob is built on top of NLTK library which uses a labeled movies review dataset to classify tweets into positive, negative and neutral tweets.
4. Tweet Aggregation: Due to the sparse nature of tweets and the length of tweets 140 character, which was recently increased to 280 characters [4], it might not contain enough data to establish satisfactory term co-occurrence, therefore define a pooling technique which aggregate related tweets based on the sentiment analysis of the tweet and the period the tweet was posted.
5. Topic Modelling: After aggregation of tweets with similar pattern and tokenization, we will be trying out different topic modeling algorithm to discover patterns of words-use and comparing how the modeled topic relates to the underlying hashtag. The 3 Topic Models still user review are
    a. Non-Negative Matrix Factorization (NMF) Algorithm in Sklearn [5]
    b. Unsupervised Latent Dirichlet Allocation Algorithm, a generative probabilistic model where documents are represented as random mixtures over latent topics and a topic is characterized by a distribution over words.[6]
    c. Semi-Supervised Guided Topic Model (GuidedLDA) Algorithm, where we will be trying to guide our topic models by providing seed word around which we expect it to converge. [7]. We plan to create seeds words of only Nouns and Noun Phrase using POS tagging phrase.
6. Visualization: Interesting highlights and insights will be presented using Word Cloud visualization in python[8].
7. Evaluation Techniques and Recommendations: Coherence score and similarity measurement are planned for evaluation how close the modeled topics are to hashtag.

Project Plan for the rest of the Term

1. Collection of Tweets: Completed
2. Preprocessing: Completed
3. Tweet Aggregation: Completed
4. Topic Modelling: Work in progress
5. Visualization: Pending
6. Evaluation Techniques and Recommendations: Pending
7. Report writing and Presentation: Pending

List of References.

[1] M. Sokolova *et al.*, “Topic Modelling and Event Identification from Twitter Textual Data,” p. 17, 2016.

[2] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, and R. Van de Walle, “Using topic models for Twitter hashtag recommendation,” *Proc. 22nd Int. Conf. World Wide Web - WWW ’13 Companion*, pp. 593–596, 2013.

[3] J. Roesslein, “Tweepy.” [Online]. Available: http://www.tweepy.org/. [Accessed: 15-Mar-2018].

[4] “Twitter to introduce expanded 280-character tweets for all its users | Technology | The Guardian.” [Online]. Available: https://www.theguardian.com/technology/2017/nov/08/twitter-to-roll-out-280-character-tweets-to-everyone. [Accessed: 15-Mar-2018].

[5] “sklearn.decomposition.NMF — scikit-learn 0.19.1 documentation.” [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html. [Accessed: 15-Mar-2018].

[6] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[7] A. Riddell and T. Hopper, “Welcome to GuidedLDA’s documentation! — GuidedLDA 1.0 documentation.” [Online]. Available: https://guidedlda.readthedocs.io/en/latest/. [Accessed: 15-Mar-2018].

[8] A. Mueller, “word clouds in Python — wordcloud 1.3 documentation.” [Online]. Available: http://amueller.github.io/word_cloud/. [Accessed: 15-Mar-2018].