Advanced Topics in NLP (CSCI 6509)

P1 - Project Statement

Project Title:

Detecting Dalhousie's Official Twitter Account's Main Interests

Mir Erfan Gheibi — B00798315

Last Edited on: April 1, 2020

1 Problem Statement

Nowadays, social media is a part of the daily life of people, and it is no more a luxury. Not only people but also organizations and legal entities are using social media to express their feelings, goals, accomplishments, and their plan for their future. Dalhousie, as a pioneer university in Atlantic Canada, has several twitter accounts to post events, news, alerts, and anything related to the university and students.

The value of detecting the main interests of Dal twitter accounts is in observing the possible relation between some events on the university and its policies and stance, as well as its priorities in the tweets of it.

Natural Language Processing (NLP) offers topic modeling, also known as topic analysis and topic detection, employing statistical and probabilistic toolboxes for extracting abstract topics in a collection of documents. Topic modeling is being used for discovering hidden semantic structures within a text corpus. Therefore, deriving the main interest(s) of a twitter account can be done using topic detection.

2 Possible Approaches and Related works

Not only there are several technology blog posts such as what Jonathan Keller has written on "towards data science"[1] for applying LDA algorithm on different datasets, but also in academia, topic detection is a hot topic and lots of research papers about topic detection, especially using or adopting LDA algorithm. I have chosen some papers on the LDA algorithm and its variations that show the feasibility of this project.

As one of the most recent related works, Sayan Unankard used the LDA method for topic detection in the feedback sections of online courses such as what Coursera provides [1]. The incentive for the research is that due to the high volume of feedback in online courses, it is almost impossible to see an read them all in an appropriate time after receiving them. Additionally, Sayan uses the word cloud visualization to depict the topics.

As another endeavor, Yunseon Choi, and Soohyung Joo used the LDA algorithm to detect topics in online book reviews [2]. The LDA generated some topic terms from online reviews. Then they categorized the topic terms into 11 categories, and finally, they conducted sentiment analysis on the categories above. They concluded that emotion and opinion could be

[1] https://towardsdatascience.com/building-a-topic-modeling-pipeline-with-spacy-and-gensim-c5dc03ffc619

acknowledged as essential hallmarks of book categorization from the readers' points of view.

Bo Huang et al. proposed a method[3] in 2012 to detect topics in microblogs, which twitter is a well-known candidate in this category. They adopted a single-pass clustering technique using the LDA algorithm to detect topics in sparse documents sets like microblogs.

As we go further back in time, there are other valuable papers such as what Loulwah AlSumait et al. published in late 2008 [4]. They proposed online LDA with the idea of incrementally improve the model and named it OLDA for a stream of documents. They claim that OLDA is even better in some cases than LDA in topic detection in unseen documents.

3 Project Plan

After choosing the project subject, I started to read related research in the field of using LDA for topic detection in documents. Then I started to gather the data from Dalhousie's official twitter account. I used the developer account of twitter and the tweepy[2] library for Python programming language. The main Twitter account of Dal has around 20k tweets, but the upper bound of a user's available tweets to retrieve is 3200 tweets according to the official documentation of twitter APIs[3].This limitation pushed me to infuse other official accounts of Dal such as DSU, Dal Science, Dal Libraries, and Dal Alumni to the main account's data to gather an ample amount of data. The gathered data has two fields, the date of the tweet and the full text of the tweet.

For the rest of the current semester, I am going to prepare the gathered data to apply the LDA algorithm on it. Furthermore, according to one of the related works[1], word cloud can be helpful for visualizing topics in the LDA algorithm; hence, I may use that for providing extra insight about Dal's main interests. For the LDA method, I am going to use pre-built libraries such as Gensim, and for other text processing tasks, I want to use SpaCy and/or nltk. After the P1, in the first week, I will apply the LDA algorithm on my dataset after preprocessing it. For the second two weeks, I will focus on building the pipeline of the system as well as the proper visualization.

[2]https://www.tweepy.org/

[3]https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline

References

[1] S. Unankard and W. Nadee, "Topic detection for online course feedback using lda," in *International Symposium on Emerging Technologies for Education*, pp. 133–142, Springer, 2019.

[2] Y. Choi and S. Joo, "Topic detection of online book reviews: Preliminary results," in *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 418–419, June 2019.

[3] B. Huang, Y. Yang, A. Mahmood, and H. Wang, "Microblog topic detection based on lda model and single-pass clustering," in *Rough Sets and Current Trends in Computing* (J. Yao, Y. Yang, R. Słowiński, S. Greco, H. Li, S. Mitra, and L. Polkowski, eds.), (Berlin, Heidelberg), pp. 166–171, Springer Berlin Heidelberg, 2012.

[4] L. AlSumait, D. Barbar, and C. Domeniconi, "On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking," in *2008 Eighth IEEE International Conference on Data Mining*, pp. 3–12, Dec 2008.