Topic Modeling of COVID-19 Tweets

Brandon Le, Yves Wienecke, Angie McGraw, Tu Vu, Keaton Kraiger

1. Project Description

The COVID-19 pandemic is a major world event, which has had an extensive impact on the ebb and flow of society. The pandemic has significantly disrupted the manner in which people communicate and learn. As people turn to social media to voice their thoughts and emotions regarding the new 'normal' of everyday life, websites like Facebook, Twitter, and Instagram reflect the turbulence and confusion during this stressful time. Given the sheer magnitude and disorganized nature of the data that are exposed by these websites, our group is interested in applying machine learning techniques to identify patterns and make sense of the madness. We hope to utilize natural language processing (NLP) to learn about how people are communicating during the pandemic. Moreover, we believe this project will help elucidate *what* people are talking about and how these topics may change by geographic location, time, or both. Our motivation for this project stems primarily from the significant effect that COVID-19 has had on each of us individually. In addition, some of us have a connection to the medical field and see this situation as an opportunity to interweave experience in the fields of medicine and computer science. Although we are isolated from the community and living in our new realities, we are nevertheless interested in understanding the new realities of others. For the final project, our group would like to perform topic modelling on tweets related to COVID-19 by geographic location. Furthermore, we would like to perform this topic modelling utilizing various different models and approaches. We will also perform an analysis of our results.

Topic modeling is a useful form of analysis for large amounts of unlabeled text[1]. Given some number of 'topics', topic modelling algorithms learn to partition a dataset (corpus) into sets that best describe these topics. In our project, we hope to employ topic modelling to extract the salient features of COVID-19 related tweets and discover the salient topics present in the corpus. Our corpus will be broken down into subcorpora defined by geographic location and time. By training models on these subcorpora, we can perform a qualitative analysis of the topical drift and topical differences of tweets. We can also perform a quantitative analysis of the convergence or divergence of these topics. Finally, as a stretch goal, we would like to utilize various topic modelling approaches and compare the differences of the results.

COVID-19 topic modelling is relevant to machine learning in many aspects, and requires careful consideration of the dataset and models that we choose to use. For this project, we are learning about the challenges associated with finding and working with unstructured text and unlabelled data. Notably, tweets may include the usage of various languages, text encodings, slang, and may exhibit non-standard

[1] "Topic Modeling - Mallet." http://mallet.cs.umass.edu/topics.php. Accessed 18 May. 2020.

syntax. We will have to carefully filter and preprocess our corpus to account for these discrepancies. Next, our focus is on saliency, which is defined in our project as the importance or significance of a given keyphrase or topic. Here, we are analyzing topics extracted from COVID-19 tweets from different geographic areas. Topic modeling involves the usage of machine learning for automatic keyphrase extraction and associating these keyphrases with topics. Our project defines keyphrases as notable words or sequences of words that stand out due to a variety of factors. These factors may include word frequency, uniqueness, and grammatical complexity. We hope to use several popular machine learning approaches in order to perform topic modeling, from purely statistical approaches, such as TF-IDF, to approaches involving probability and word embeddings, such as LDA and word2vec. Evaluation of the performance of each model will be done qualitatively and quantitatively, potentially through the usage of several metrics for comparing topic distributions, including perplexity, KLD, Jaccard, and Hellinger measurements.

Our goals with this project are to gain experience working with novel datasets and to employ a few of the preprocessing and machine learning modeling techniques briefly discussed in class. In addition, we seek to share our project and findings to provide insight into the experience of people around the world during the pandemic. We expect to see high significance in topics related to health, employment, misinformation, and motivation. These topics may vary in importance by geographic area in accordance with the unique circumstances of the region.

2. Methods

We will be using the Corona Virus (COVID-19) Tweets Dataset[2], a dataset composed of unique tweet IDs and associated geographic coordinates, which have been filtered through the Twitter API to include only tweets written in English and that contain certain keywords related to COVID-19. To "hydrate" the tweets (turn tweet IDs into a JSON representation of the tweet and its metadata), we will be using DocNow's Twarc[3]. Next, topic modelling will involve the usage of pretrained models and analyzation techniques implemented by the Gensim[4] python library. The Gensim library exposes functions for Latent Dirichlet Allocation (LDA), Term Frequency Inverse Document Frequency (TF-IDF), and Word2Vec, Doc2Vec, and other machine learning algorithms. We aim to use at least LDA for topic modeling, but we hope to use the other algorithms for comparison among the models; if the time constraints and potential complexity of the Gensim model permits. Lastly, we intend on using Matplotlib[5] and pyLDAvis[6] to visualize our results. This project will be implemented using the Python 3 programming language.

---

[2] "Corona Virus (COVID-19) Tweets Dataset | IEEE DataPort." https://ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset. Accessed 18 May. 2020.
[3] "DocNow/twarc: A command line tool (and Python ... - GitHub." https://github.com/DocNow/twarc. Accessed 18 May. 2020.
[4] "gensim: About - Radim Řehůřek." 1 Nov. 2019, https://radimrehurek.com/gensim/about.html. Accessed 18 May. 2020.
[5] "Matplotlib." https://matplotlib.org/. Accessed 18 May. 2020.
[6] "pyLDAvis's documentation! - Read the Docs." https://pyldavis.readthedocs.io/en/latest/. Accessed 18 May. 2020.

3. References

DocNow/twarc. (2020, May 11). Retrieved from https://github.com/DocNow/twarc

Gensim: Topic Modelling for Humans. (n.d.). Retrieved May 18, 2020, from https://radimrehurek.com/gensim/about.html

Lamsal, R. (2020, May 18). Corona Virus (COVID-19) Tweets Dataset. Retrieved May 18, 2020, from https://ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset

Latent Dirichlet allocation. (2020, April 28). Retrieved May 18, 2020, from https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Tf–idf. (2020, May 3). Retrieved May 18, 2020, from https://en.wikipedia.org/wiki/Tf–idf

Topic Modeling. (n.d.). Retrieved May 18, 2020, from http://mallet.cs.umass.edu/topics.php

Visualization with Python. (n.d.). Retrieved May 18, 2020, from https://matplotlib.org/

Welcome to pyLDAvis's documentation! (n.d.). Retrieved May 18, 2020, from https://pyldavis.readthedocs.io/en/latest/

Word2vec. (2020, May 6). Retrieved May 18, 2020, from https://en.wikipedia.org/wiki/Word2vec