UNSTRUCTURED AND DISTRIBUTED DATA MODELING AND ANALYSIS MSBX - 5420

CBC News COVID 19 Articles

Team Mount Albert

Date: 4/11/2020

Prepared By:
Junji Wiener
Fairy Gandhi
Whitney Carter
Meredith Synnott
Aldo Peter

REQUIREMENT PHASE I

GLOSSARY

Kaggle: https://www.kaggle.com/ryanxjhan/cbc-news-coronavirus-articles-march-26
Github: https://github.com/MSBX5420/team-mount-elbert

INTRODUCTION

For our project, we are using the Kaggle dataset on COVID-19 (as suggested by Professor Zhang). Our analysis will use Spark with Python, AWS Classroom, and Google Collab in order to examine the effects of COVID 19 in the news and how the media has covered this new pandemic. We plan on using Latent Semantic Analysis - an unsupervised machine learning model. We are currently brainstorming potential research questions independently and exploring the visualization tools necessary to strengthen our analysis.

FUNCTIONAL AND NON-FUNCTIONAL REQUIREMENTS

To understand how the main focus of CBC news articles has evolved during this COVID -19 affected time period by doing the following:

- Analyzing the trend with the word count of the articles.
- Analyzing the keywords of articles every month using the wordcloud.
- Building an Abstractive News Summarizer with Seq2Seq (Sequence to Sequence Networks) Model (NLP).

PERFORMANCE REQUIREMENTS

Performance Requirements will be set under conditions to detect the words used the most and least in a news article. The Latent Semantic Analysis determines the similarity of meaning of words and set of words to each other. This will help us interpret what the most words and least words mean for the public and their understanding of COVID-19.

TIMELINE

- Requirement Phase is **due April 12th, 2020.**
- Design, Development, Test is **due April 25th, 2020.**
    - We will meet twice a week for the next two weeks to work on the project (data cleanup, building the model, etc.).
- Deployment is **due April 28th, 2020.**
- The presentation is **due April 28th, 2020.**