**Text Analysis of USF Reddit Discussions**

**Data Mining Project – Final Paper**

Abdul Ahad Ahmed, Andres Herrera Ropero, Samrat Korupolu, Prithvi Telkapally

University of South Florida

ISM 616: Data Mining

Professor Balaji Padmanabhan

November 18, 2020

Contents

[1. Introduction 3](#_Toc56411027)

[2. Dataset Overview & Data Exploration 4](#_Toc56411028)

[3. Process Flow 5](#_Toc56411029)

[4. Predictions with Test Data 7](#_Toc56411030)

[5. Acknowledgment 8](#_Toc56411031)

[6. References 9](#_Toc56411032)

## Introduction

**Text Analytics**

Text analytics is an artificial intelligence (AI) innovation that utilizes natural language processing (NLP) to change the free, unstructured content in records and information. It bases on standardized, organized information appropriate for examination and drives Machine Learning (ML) calculations to discover Patterns and Trends

**Topic Modeling LDA**

Topic Modeling is the process of providing methods for automatically organizing, understanding, searching, and summarizing extensive electronic archives in order to identify topics in a set of documents This can be very useful for search engines, customer services automation, and any social media platform where knowing the documents' topics is essential. There are multiple methods of doing this. However, for the project, we will explain and use **Latent Dirichlet Allocation (LDA).**

**Latent Dirichlet Allocation (LDA)**

Is the process of forming a document by assuming D as a corpus (Doll, 2019), which is a collection of documents (d) as follows:

1. Choose a random topic.
2. Randomly select a topic distribution for the document (d).
3. For each word in the document:
   1. Randomly select a topic in the topic distribution.
   2. Choose a word randomly from the topic that corresponds to the vocabulary distribution.

**Problem Description**

USF Reddit has different existing questions, comments, suggestions, and reviews where they are unanswered, unlabeled, and need to be categorized to fit the next step of the machine learning pipeline.

**Identify Problems**

This is a data mining project where the team learns to find various topics presented on the USF Reddit page. We choose this topic because we feel it’s very important for USF to be able to answer to some questions that students, potential students, faculty and members might have. With time, Reddit is used by many more people where there exists different subreddits like for example, Personal Finance, and people ask their questions and other respond to it.

**Objectives**

The main goal is to identify the number of topics and determine the theme of each topic. We will also test whether our model can predict the topic theme given any new questions.

**Approaches**

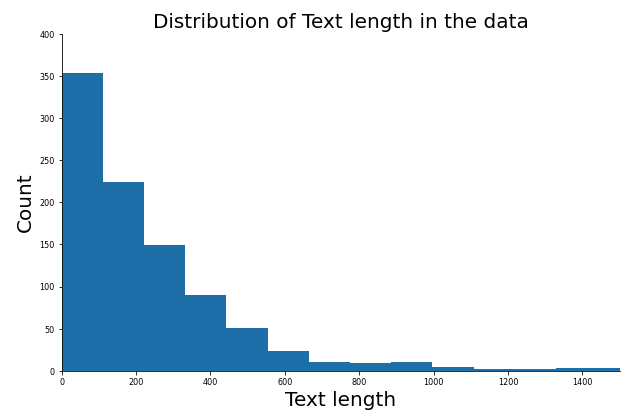
Scrap the data from USF Reddit and then apply an LDA model to the USF Reddit data to determine the number of topics and themes of each topic.

## Dataset Overview & Data Exploration

We used [Universal Reddit Scrapper](https://github.com/JosephLai241/URS) to scrape the USF subreddit. Data was scraped from Aug 2020 to Nov 2020. Scrapping the data gave us Title, upvotes, comments, upvote ratio, text. In our analysis, we are only interested in the Text column. We have around 950 rows for our analysis.

**Text Preprocessing**

Removing the non. ASCII characters using VBA spirit in excel. Also, in this phase, we will convert the text to lowercase, removing stop words, bracket, punctuation and numbers, and lemmatization. We are not going to stem because it often leads to incorrect meanings and spelling errors.



**Word Cloud**

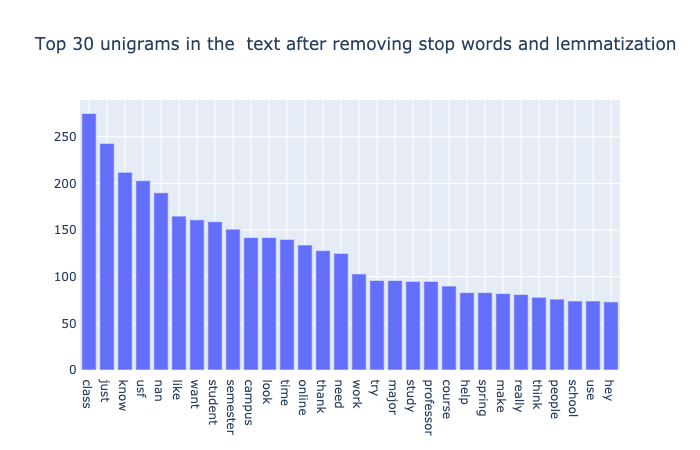
**A** Word Cloud is a visual representation of text data. It displays a list of words, the importance of each being appeared with text dimension or shading. This format is valuable for rapidly seeing the most noticeable terms. The below word cloud is the most prominent terms used in our text data.



## Process Flow

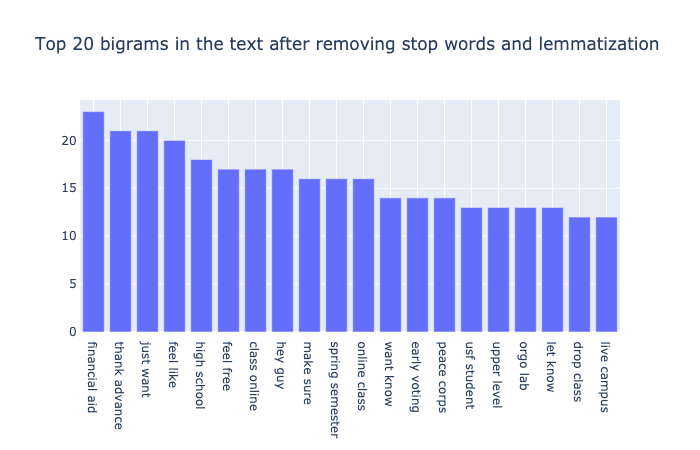
**Using N-Grams**

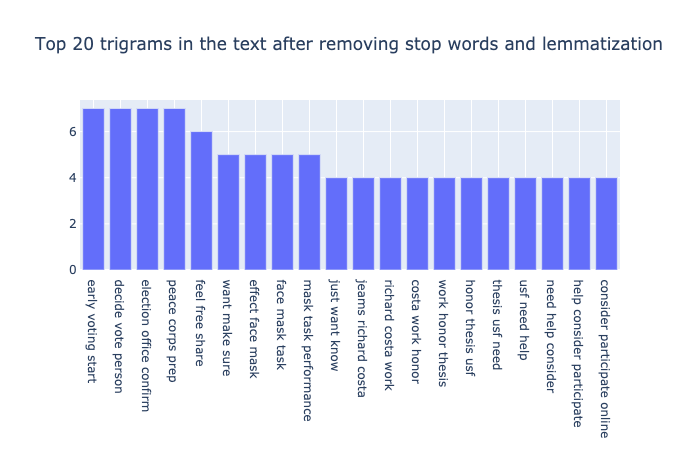
N-grams of texts are extensively used in text mining and natural language processing tasks. They are a set of co-occurring words within a given window.



**Unigrams**

From the above unigrams plot, we cannot conclude the topics discussed, so we are also going to bigrams and trigrams.





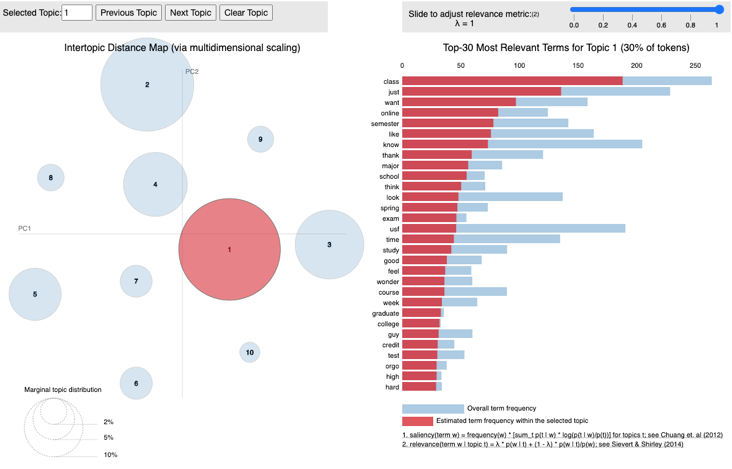
**Bigrams & Trigrams**

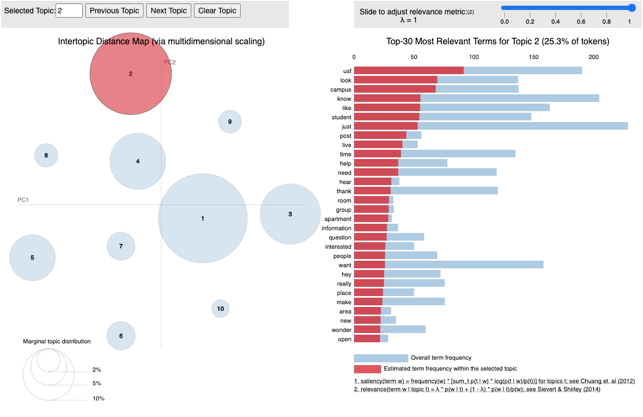
From the above two plots, we can know what people were discussing.

**Observation**:

* Much text was on “election and voting.” (We assume the college elections and presidential elections played a significant role in these discussions)
* Much text was also on the face mask. (We assume these discussions were related to the Coronavirus pandemic)
* Many people were discussing "honor and thesis" (We assume they were talking about course selection)
* The text was also on "online & participation" (We assume these texts were on online classes)

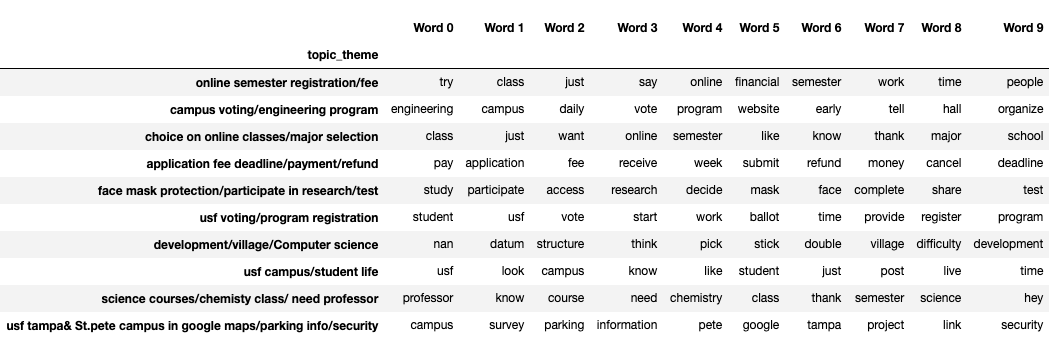
**Applying LDA:**



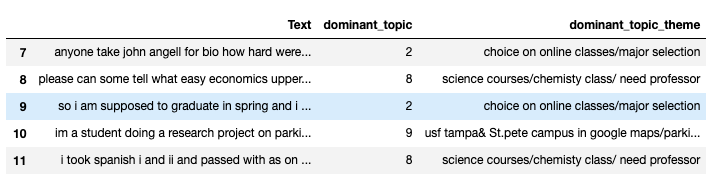


We kept the top 10 most frequent words from each topic that found by LDA. Then search for top words in each topic, manually assigning the theme to each topic. We just used the top 3 or 4 words as the theme.

The following is a partial table on the theme-keywords assignment.



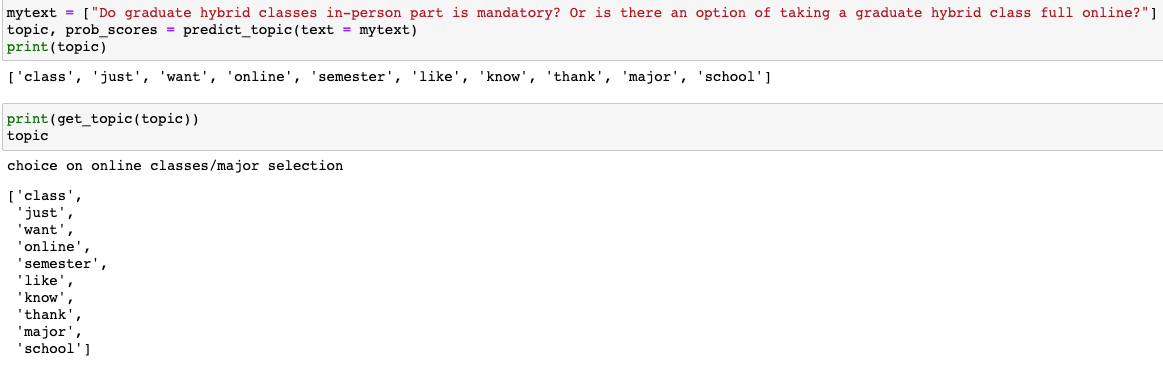
**Observation:** From the above screenshot, we have manually assigned nine topics for the words in the text; this needs some creativity and domain expertise.



For our text data, we are assigning it to the dominant topic and dominant topic theme.

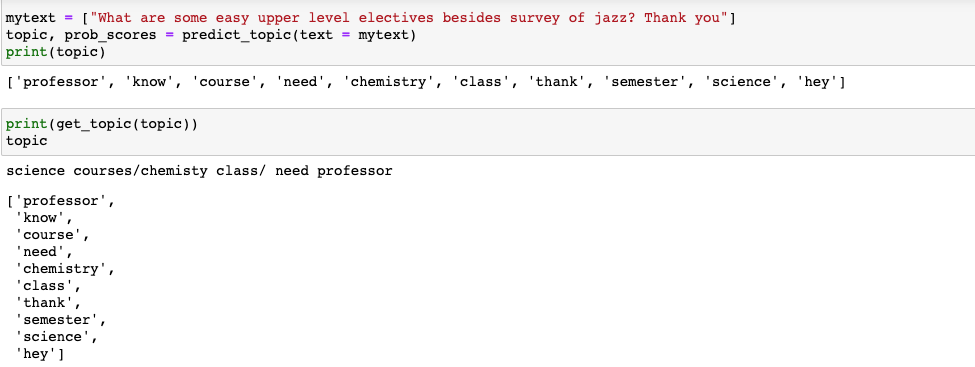
## Predictions with New Data

The first text we picked from the latest USF subreddit to test my model was, “ Do graduate hybrid classes in-person part is mandatory? Or is there an option of taking a graduate hybrid class full online? “



**Observation:** As we see here from the input, we gave our model has classified it as the topic "online classes/major selection."

The second text I picked from the latest USF subreddit to test my model was, "What are some easy upper-level electives besides survey of jazz? Thank you "



**Observation:** As we see here from the input, we gave our model has classified it as the topic “science courses/chemistry class/ need professor

## Acknowledgment

We want to thank Dr. Balaji Padmanabhan for all his guidance and support in every project step. Due to his explanations and teaching during the course, every concept and doing is where we could incorporate it into the project.

## References

Doll, T. (2019, March 11). LDA Topic Modeling. Retrieved November 16, 2020, from <https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd>

Li, S. (2018, June 01). Topic Modeling and Latent Dirichlet Allocation (LDA) in Python. Retrieved November 16, 2020, from https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

JosephLai241. (n.d.). JosephLai241/URS. Retrieved November 16, 2020, from <https://github.com/JosephLai241/URS>

Pascual, F. (2019, November 20). What Is Text Analytics? Retrieved November 16, 2020, from <https://monkeylearn.com/blog/what-is-text-analytics/>

How college students are handling COVID-19

<https://towardsdatascience.com/how-college-students-are-handling-covid-19-3705016205fe>