Creating Semantic Spaces Using Document Clustering

By: Tristan Weger, Ryan Rubadue, Yahya Emara

Project Purpose and Goals

Goals:

- Create a semantic space using long document clustering.
- Compare different embedding methods (word, sentence, chunk).
- Display semantic space in large dimensionality along with a simplified 2 dimensional map.
- Create a space that is easy to interpret and understand.

Purpose:

- Provide a continuous flow of new and relevant ideas for brainstorming.

About the Team

Members:

- Tristan Weger (EE) wegerta@mail.uc.edu
- Ryan Rubadue (CS) rubadurs@mail.uc.edu
- Yahya Emara (CE) emarayk@mail.uc.edu

Advisor:

- Professor Ali Minai minaiaa@ucmail.uc.edu

Project Abstract

- The general goal of this project is to organize and analyze text documents embedded into a d semantic space to obtain a domain-specific cognitive map.
- The project will be split into three main sections:
  - Embedding
  - Clustering
  - Dimensionality Reduction

User Stories

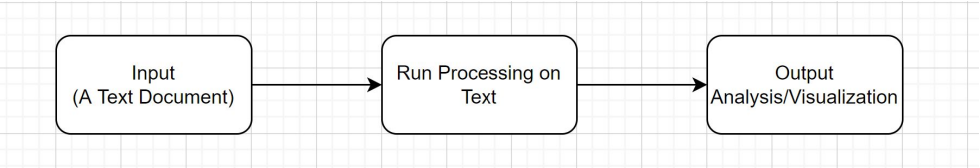
1. As a person interested in politics, I would like to easily visualize frequent words/phrases appearing in congressional speeches and compare that content to previous speeches.

2. As a person brainstorming an idea, I would like to be shown other words/phrases related to the idea to help with my brainstorming and know if I am on the right track.

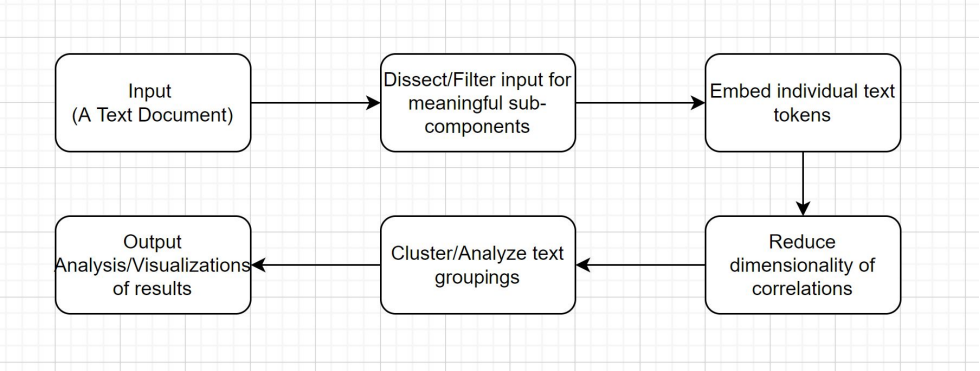
3. As a researcher, I would like to know the best way to break down long documents to attain useful classifications of data clusters.

Design Diagrams

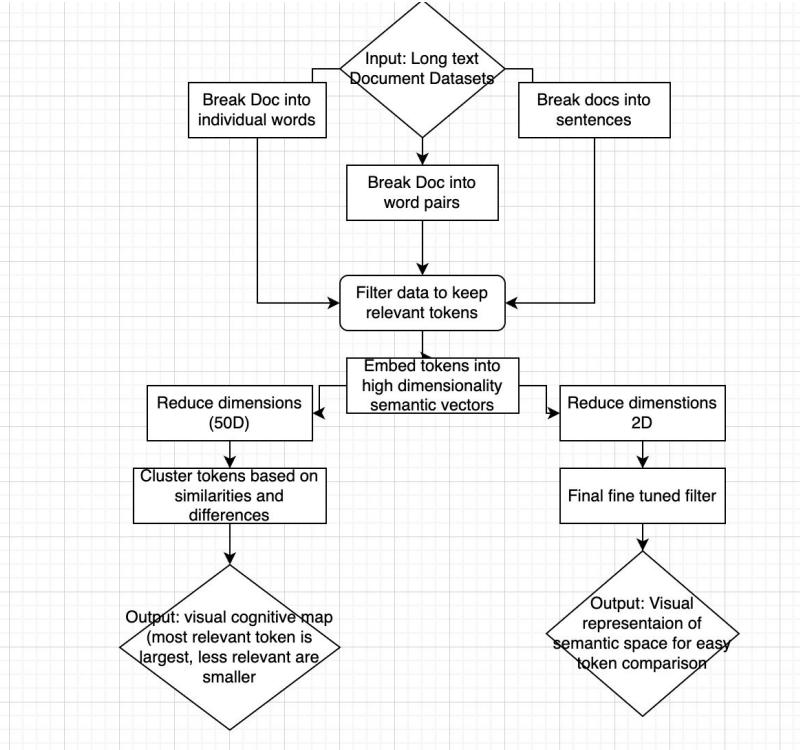
1.



2.



3.



Major Project Constraints

Economic

- Open source Datasets
- Freeware/Shareware
- Free networks for Importing Data

Professional

- Team's lack of prior exposure in project area
- Working with advisor who Specializes in project area

Legal

- Documents used non-copyrighted and open to public
- Ensure system follows ethical code and meets our ethical standards

Project Progress

- Research key concepts and potential tools to perform project functionalities
- Identify high-quality free to use datasets
  - Presidential Speeches and Thinking Fast and Slow
- Parse documents within corpus and properly store information
- Embed ‘tokens’ from the dataset, using individual sentences
- Filter sentences for relevence

Expected Accomplishments For This Term

- Parsing Documents
- Embedding the document using Sentence Embedding methods
- Plot the data points into a vector space
- Comparing pair distances in vector space using cosine similarity

Division of Work

All members are going to collaborate on tasks; however, each task has a designated leader ultimately responsible for proper completion

Tristan - Data clustering, performance analysis

Ryan - Dimension Reduction, 2D Vector Space Visualizations,

Yahya - Document Parsing/Separation, token embedding

Full Task List

1. Tristan - Determine what kind of machine learning method to use (supervised, reinforcement, unsupervised, transfer, etc.)
2. Tristan - Investigate how to partition data into training, validation and test subsets.
3. Tristan - Run the clustering algorithm using the different data subset partitions.
4. Tristan - Select a cross validation method to get a more accurate rating of the performance of each model.
5. Tristan - Interpret the clustering results and make adjustments if needed.
6. Ryan - Determine optimal amount of vector space reduction that can be achieved without losing anything above minimal classification data.
7. Ryan - Decide methodology of dimension reduction and document rationale behind choice.
8. Ryan - Actual implementation of dimension reduction.
9. Ryan - Compare application results when reducing to different numbers of dimensions.
10. Ryan - Document if the optimal number of dimensions is relatively consistent across different document embedding techniques.
11. Ryan - Design quality and informative visualizations for reduced 2D vector space.
12. Ryan - *Potential. Employ different method/tool in order to reduce dimensions. Convey whether results remain consistent or differ.
13. Yahya - Break document into word pairs
14. Yahya - Break document into sentences
15. Yahya - Filter data to keep relevant word tokens
16. Yahya - Embed tokens into high dimensionality semantic vectors
17. Yahya - Compare embedding methods and optimize using the best method