Team Bingle
Intelligent Browsing – Highlight Search Engine
Progress Report - Nov 15, 2021

Name	Net ID	Captain
Shubha Sundar	Shubhas2	Yes
Mony Chhen	Monyrc2	
Gene Horecka	Geneeh2	
Sushma Ponna	Ponna2	

1. What progress has been made?

- **Frontend Chrome Extension Application**
  - ○ Created a Google Chrome Extension application that fetches dummy data from an API and displays it on the app's frontend UI.
  - ○ Created the functionality to activate the Chrome extension by highlighting some text on a webpage, right-clicking on the page, then selecting "Search on Bingle".
- **Backend Node.js Server**
  - ○ Created an API server that currently runs on localhost that can receive, process, and respond to HTTP requests.
- **Dataset Topic Coverage and Analysis**
  - ○ Implemented data retrieval from dataset downloaded and parsed from wikidump download.
  - ○ Initial Exploratory Analysis of the dataset (corpus metadata) by inverted index on the dataset.
  - ○ Topic coverage graph to aid in test set creation

Rank Algorithm Development:

- ○ Search algorithm using BM25 ranker
  - ☐ Created a test data set out of the wikipedia data
  - ☐ Created queries for testing the model.
  - ☐ Indexed the test dataset and validated the query.
- ○ **IR Algorithm Training**
  - ☐ Pretrained Word Embeddings using Word2Vec and Glove models
  - ☐ Started Ranking Algorithm Development: Created Inverted Index for the 463K pages and tested the end-to-end framework from parsing the query and evaluating it against a dummy query-set

2. What tasks remain?

- **Frontend Chrome Extension App**
  - Finalize the UI design of the Chrome Extension app, which will display the ranked information from our Wikipedia dataset in an easy-to-read way.
  - Connect the Chrome extension frontend to the backend Node.js server.
- **Backend Node.js Server**
  - Incorporate the Python scripts that run all the text retrieval algorithms to the Node.js server. From the server, be able to feed input to the Python scripts and retrieve their output, process it as a JSON object, which can then be sent back to the client.
  - Make sure the communication between the Chrome extension and the server works without any CORS policy/permissions issues.
- **Dataset Topic Coverage and Analysis**
  - Execute on the entire dataset with a larger number of topics and other parameters.
- **Rank Algorithm Development:**
  - Explore different algorithm implementations for ranking. (Eg: BM25, LSA)
  - Create Query test set
  - Create Relevance judgements for the query set for calculating AP/MAP/nDCG
  - Fine tune the parameters to get better search results and performance
  - Error check and result analysis
- **Prepare documentation and presentation**

3. What challenges/issues were faced?
- Ran into an issue while downloading and parsing Wiki dataset from br2 file due to WINDOW OS laptop and not a macbook and laptop.
  - To Resolved this using virtual machine with ubuntu linux OS
- Learning different Algorithms and understanding what the algorithm is doing.
  - Current selected LSA for cluster grouping
  - Deciding on the total number of topics
- Coming up with query set for wide variety of topics is challenging