Team Bingle
Intelligent Browsing – Highlight Search Engine

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

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

2. What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?

Topic: Intelligent Browsing – Highlight Search Engine

Problem Resolution: This project allows users to quickly access relevant information that they are reading up on without having to open an additional browser tab and search for the word/query. The idea behind this project is to allow users to highlight the word/query that they are interested in searching, and instantly get information on the same page they are looking on with a chrome app extension plugin.

Description: This project consists of two main topics that were discussed in this course, data retrieval and ranking, and user feedback data training. The plan is to download a dataset from Wikipedia and develop algorithms for data retrieval, ranking, and users feedback (users' click) and attach it to the frontend chrome extension app where the user can highlight a word/query to pass into the search engine and the top relevant document will be sent back to the chrome app.

3. Briefly describe any datasets, algorithms, or techniques you plan to use

We will be using a Wikipedia dataset obtained from https://dumps.wikimedia.org/ which contains Wikipedia XML data of all Wikipedia articles in all languages. We will use the Wikipedia Extractor tool (http://medialab.di.unipi.it/wiki/Wikipedia_Extractor) to clean the raw XML data and obtain only the textual content of the Wikipedia pages and its corresponding Wikipedia page URLs based on the instructions obtained from (https://www.lateral.io/resources-blog/the-unknown-perils-of-mining-wikipedia). Due to the massive size of the entire Wikipedia dataset, we will only use the English version of Wikipedia and retain only those pages with at least 20 page views to exclude the mass of robot generated pages (thereby keeping only those pages that are frequently viewed and edited by humans). This will greatly reduce the size of our dataset to a manageable working set and improve the performance of our test models.

We also plan to use the MeTA toolkit Python library (https://github.com/meta-toolkit/metapy) to implement our text retrieval (BM25), ranking, and implicit feedback algorithms. Implicit feedback (user click) information will be used to retrain our model.

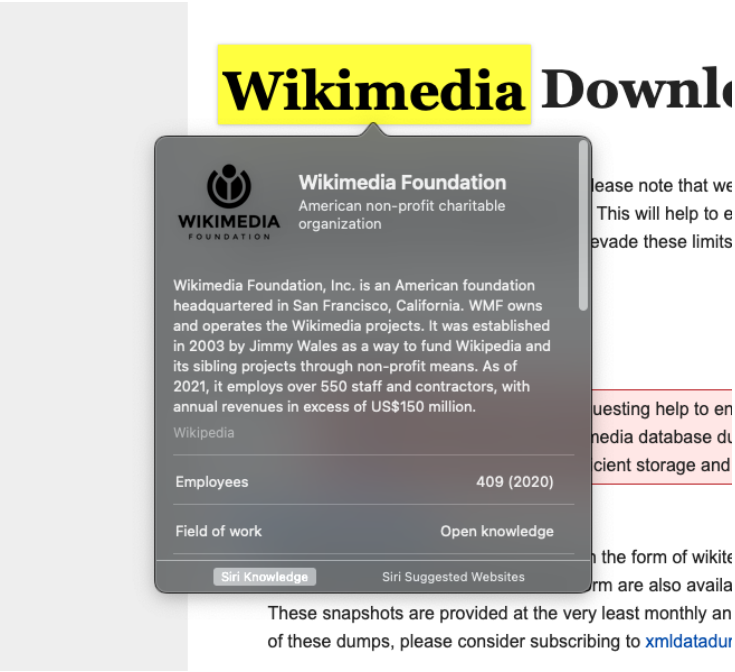
4. How will you demonstrate that your approach will work as expected?

We will show that our approach works by applying the core concepts that we have gone through in this course and by providing an end product that works functionally as described below.

Planned Approach:

1. The user first installs the Highlight Search Engine Google Chrome extension on their Google Chrome browser. Once the Chrome extension application is installed, the user can toggle the search functionality of the chrome extension by highlighting some text on a webpage (which acts as the query), right-clicking on the highlighted text, and then selecting “Search on Bingle”. This will pop up the Chrome extension’s mini UI display containing a short summary of the most relevant Wikipedia page retrieved from our dataset, along with the corresponding URL link to that Wikipedia page. In addition, we will provide an additional next top 4 relevant Wikipedia URL page links within the pop-up so that the user can choose from these other 4 options if the top suggested information is not relevant to the user.

   High-level example of the functionality of the Highlight Search Engine:



2. We will maintain a corpora for this project (freely available datasets from Wikipedia as cited above) to start with and this set can be incrementally updated by students who wish to add to this project in the future. We plan to store the corpora and models as part of our binaries. We will train the retrieval algorithm on the corpora using different text retrieval methods and language models.
3. The application will be built to work locally on localhost.

4. We plan to collect feedback by implementing an Implicit Feedback method and adjusting the weights.
5. Data flow:
    a. The user activates search on some highlighted text (query) on a webpage.
    b. The query is sent as a HTTP GET request to our backend application server, which will process the request and run our retrieval algorithm to provide the top 5 most relevant URLs of Wikipedia pages, as well as a text summary extracted from the most relevant Wikipedia page. This application server will interface with our Wikipedia dataset and model.
    c. Once the top 5 candidates have been selected, the application server will return these results to the application's frontend client. The frontend client will then display these results to the user in the frontend mini-display.
    d. Implicit feedback (user clicks), can be captured by issuing a HTTP POST request to the application server every time the user clicks on a Wikipedia URL link displayed in the mini-display. The application server will then record this information as a relevant link and update the retrieval algorithm accordingly.

5. Which programming language do you plan to use?

    Python, Javascript and SAX parser.

6. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

Tasks		Estimate Hours	Assigned To
Frontend	Chrome App Extension (Creating, Deploying)	15	TBD
	React.js Framework	5	TBD
Backend	Python Programming as Backend (Flask Framework)	5	TBD
	Information Retrieval Algorithm Development	20	TBD
	Implementing Ranking Algorithm	20	TBD
	Implicit Feedback Implementation (user Click)	20	TBD
Data Selection/Storage	Curate and Exploratory Analysis of Data Set	20	TBD
Total Hours		110	

N = 4
Required Hours: 20 * 4 = 80
Estimated Hours = 110 hours