CS410 Group Project Proposal

Team Members

1. (Captain) Daocheng Wang (dw30)
2. Ziyue Wang (wang567)

Background & Motivation

The topic that our team has chosen is Intelligent Browsing. When users perform searches in a search engine, there are occasions when the users don't have a clear keyword in mind about what exactly to search. For example, they might want to search for a specific type of cheese (e.g. Camembert) but not remember the exact name of the cheese. Then they might search for the keyword "cheese" and browse information from there.

The goal of our team is to accelerate the information retrieval process for the users when non-exact keywords are used. We would like to design, implement, and deliver a software (Chrome extension or tampermonkey script) such that when a user searches on a website, the software will automatically inspect / retrieve the search results and give a user a list of relevant keyword suggestions for more refined search. This is similar to Google's 'people also ask' feature, but we would like to extend the feature to other applications' search engines such as Reddit or Twitter.

This is related to the course as it's a text retrieval problem, and we are planning to utilize various statistical language models during the implementation process. Regarding the technology involved, besides the language models, we are also planning to use the query likelihood retrieval function. In addition, we would use Javascript / Python as primary programming languages, and we would leverage either web scraping or Twitter / Reddit APIs for data collection. Furthermore, to ensure the relevance (effectiveness) of the suggested keywords recommended by the software, we are also planning to implement an implicit feedback system such that the users would be able to judge the results and then update the algorithm accordingly.

To demonstrate that our project will work as expected, the team members will be picking out a variety of queries and noting the suggestions by the extension. We will then validate that the suggestions are in line with the content from top search results.

Requirement

1. User enters a query in a supported search engine (e.g. Twitter / Reddit - to be determined during implementation)
2. The system would recommend several keywords for additional searches
3. User can optionally click on a suggested to perform a new search with that particular keyword

Assumptions and Non-Goals

1. Users may search for a phrase, but only individual keywords would be recommended to the users. Recommendation of phrases is out scope for this project.
2. The team will only work on integrating the extension with one platform in order to maximize the remaining bandwidth available for the text retrieval parts. Options such as Twitter and Reddit will be assessed during the project implementation phase.
3. The UI / UX will be functional yet basic as web design is not an emphasis of this course.

High Level Design

Text Access

We are planning to use either web scraping or Twitter / Reddit APIs to pull the user entered query and the collections of the documents from the search results.

Text Retrieval

Upon retrieving the collection of search results, the software would be able to build the document LM (language model) and compare it with e.g. collection LM to assign scores to the words from the result based on the differences between the 2 LMs. In this way, the software can rank the words and choose the top words to recommend to users for related searches.

UI/UX

Once the software generates a list of top ranked related keywords, the user would be able to see those words near the search bar and click those to start a new search with the new keywords.

Proof of Correctness

From another perspective, there are some similarities between this problem and topic mining problem - both analyze a list of documents and attempt to extract ‘common ideas’ from the list of docs. To validate the effectiveness of the software, we could potentially analyze the paradigmatic relations between the keywords to be recommended and the keywords that the user has searched via EOWC.

Other proof of correctness options considered

The initial thought for validating the correctness and effectiveness of the software is to implement an implicit feedback system such that the users would be able to judge the results and then update the algorithm accordingly. Upon user clicks the new keywords, the feedback system would be able to capture that and potentially update a feedback LM accordingly. However, there are a number of concerns with this approach:

1. Need to host a database to incrementally build a feedback LM.
2. An extra layer of database network call would increase the latency significantly.
3. Cold start problem: user queries can be very different, leading to slow build-up of the feedback LM

We also thought about using pseudo feedback, but ultimately decided against the idea due to lack of confidence in the feedback reliability.

Scoping

Task No.	Task	Dependency	Effort (dev hours)
1	External API Investigation	-	5
2	Learn about Chrome extension development	-	3
3	Implement business logic for keywords recommendation	1, 2	13
4	UI/UX	1, 2	5
5	Validation system	3, 4	13
6	Testing (Initialization)	5	5

References

(10/19 Update)

API Support

- Reddit:
    - Search API (Free) https://www.reddit.com/dev/api/#GET_search
    - Guides:
        - https://towardsdatascience.com/how-to-use-the-reddit-api-in-python-5e05ddfd1e5c
- Twitter:
    - Free search APIs:
        - Tweets within the past 30 days https://developer.twitter.com/en/pricing/search-30day
            - Limit: 250 requests per month