Project Proposal, CS 410

Group: chadyuu

1 Team member

Captain: Yutaro Nishiyama (yutaron2@illinois.edu)

2 Topic

The theme chosen for this project is **Intelligent Browsing**. In particular, we will develop a Google Chrome extension to retrieve 10 Q&A pages from Stack Overflow that relates the most to a current web page about text retrieval, text mining, and NLP.

We can deepen our understanding from related Q&A discussions, which shed light on the topic from different perspectives. However, we usually feel tedious to search for related questions. The Google Chrome extension developed in this project requires only one click to realize this and will help us enhance our understanding much easier.

This project utilizes techniques for text retrieval and natural language processing, such as scraping, document vectorization, and content-based filtering, which we have learned in this class.

3 Dataset

We will retrieve Q&As by Stack Exchange API as follows.

1. Use `/search/advanced` API to retrieve the Q&A list that contain words related to this class, such as text mining and natural language processing.
2. Use `/questions/{ids}` API to retrieve each Q&A content.

4 Algorithms & Techniques

1. Create the Q&A database by preprocessing each Q&A content as follows.
    1. Remove stop words.
    2. Lemmatize.
    3. Generate a document vector which stores the frequency of each word.
2. Implement a Google Chrome extension.
    1. Apply content-based filtering with cosine similarity between the current webpage and each Q&A.
    2. Return top 10 related Q&As.

5 How to demonstrate

We will demonstrate the Chrome extension through the video recording where it actually works, in addition to its source code.

6 Programming language

- Python for backend
- Javascript for the Google Chrome extension

7 Predicted workload

We predict the total workload as 20 hours for a team of one member.

- Create the Q&A database: 10 hours
- Implement the Google Chrome extension: 10 hours