Team Leader: Kajetan Haas (kahaas@illinois.edu)
CS 410

Project Proposal

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:
   Kajetan Haas – kahaas (Captain/Leader)

2. What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?
   The topic I have chosen is to create a Firefox browser extension that allows for more flexible searching in webpages. It is often difficult to find what you are looking for on a webpage if you are not sure of the exact text you need to find. Using fuzzy/flexible searching you can search for words and phrases that are close to your search phrase, but not exactly the same. They can then be ranked with BM25 and the best results can be displayed to the user. This is related to the Intelligent Browsing theme because it allows for users to be able to search more effectively and uses algorithms to provide more general results for a user's query. This topic is related to the overall class because it is about text retrieval and using search algorithms on text.

3. Briefly describe any datasets, algorithms, or techniques you plan to use:
   Fuzzy Search / Approximate String Matching
   https://en.wikipedia.org/wiki/Approximate_string_matching will be used to search for text in the web page that approximately matches the query. For example it may have a few letters changed, so typos are not an issue.
   BM25 Ranking will be used to rank the results of the closest matching parts of the web page.
   The WordNet dataset https://wordnet.princeton.edu/ or a similar database will be used to find words that are similar to your query word. For example, matching 'bread' with 'loaf'.

4. How will you demonstrate that your approach will work as expected?
   I will use the extension on webpages with non-trivial content and show that it is able to retrieve relevant content in the page from an inexact query. For example searching for 'algorithms' would still yield 'algorithm' or searching for 'bread' might yield 'bun' or 'loaf' since these are related words.

5. Which programming language do you plan to use?
   Since it is a browser extension it will be primarily programmed in JavaScript. It will also likely utilize HTML and CSS so that it can have an appealing interface.

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.

   There is one student on the team so there should be 20 hours of work:

- Setting up basic browser extension – 1 hour
- Indexing all the words on the current webpage – 1 hour
- Use stemming and tokenization to split the webpage and queries into separate and searchable parts – 2 hours
- Implementing an efficient fuzzy search algorithm – 4 hours
- Creating an appealing interface to enter your query and see the results – 4 hours
- Setting up the WordNet database to work in the browser extension – 6 hours
- Rank the results using a heuristic formula and BM25 – 2 hours

Total: 20 hours