Theme 1: Intelligent Browsing Requirements

Group: **Intelligent Group**

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.

   Ali Saad (alins2) - Captain
   Jingtian Yao (yao28)

2. What topic have you chosen?

   Intelligent Browsing.

   Why is it a problem?

   The user may not be completely familiar with the content they are trying to search for. Our goal is to improve the overall search experience of a user by inspiring them to refine their search query terms.

   How does it relate to the theme and to the class?

   This relates to the class by taking the idea of what we did in MP2.1 and 2.2 with searching relevant topics that had to do with the course. We based our idea of doing intelligent browsing by adding a chrome extension using BM25 which helps users find the right terms in order to find the article that best fits their needs.

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

   The tool will read pages returned by Google search, then the tool will do following tasks:
   1. Re-rank the documents given a BM25 retrieval function
   2. Provide word-level statistics (frequency, existing in x% of documents etc.), excluding common words
   3. Implement topic mining given the returned documents
      a. If we have the bandwidth we will implement this as it is optional for now

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

   We will demonstrate that our approach worked as expected by inputting a set of query terms and seeing the rankings BM25 function provides for each word as well as showing the percentage of how often each word is used.

5. Which programming language do you plan to use?

   JavaScript + Python

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.

   1. Create web crawler - 5 Hours
   2. Preprocessing of crawled documents - 5 Hours
      a. Tokenization of words
      b. Creation of corpus, terms vs documents matrix etc.
   3. Implementation of word statistics - 10 Hours
   4. Implementation of topic mining - Optional but if we do then 5 Hours
   5. Integrate backend scripts with frontend JS - 10 hours
   6. Documentation - 10 Hours