miguelf4, rgyanm2, mcroos2, ajfitch3, ak85; Team Captain: miguelf4

Team WXYZ

Project Proposal

We have chosen "Theme 1: Intelligent Browsing. Specifically, we want to create a chrome extension that indexes the current page to allow users to search over the page using common retrieval functions. This is very similar to the first example, but since our team consists of 5 members that means we will have the opportunity to be more ambitious with the project. The problem that we're trying to solve is that the built-in search for chrome (ctrl+f) only does exact-word match filtering, which is oftentimes not very useful. Being able to search a single page with an intuitive "search query"-based type of search can help a lot.

Since we only plan to be indexing and running queries in a single webpage (as opposed to a collection of web pages), it means that we have to decide what exactly we define as a "document". We plan to have this be a configurable option based on what users are looking at. For example we could treat a document as a single sentence, a paragraph, or in the case of a PDF a whole page. The problem of course relates well to the theme of the class because it's all about being able to use more natural language to search a page, similar to how you would search for the document that you want to find using google.

We plan to use BM25 as our retrieval function. We will likely need some sort of dictionary dataset to help accomplish stemming of words. This would of course be quite large though as it would have to contain every English word so we instead take an algorithmic approach to this stemming problem (Porter stemming algorithm). We will demonstrate our chrome extension working on various web pages to find the "document" that is most relevant to the query that we enter. (Recall that our extension will have to define a "document" as some sub-section of a single webpage). In order to create this extension, we plan to use Javascript as possible HTML and CSS.

In order to make the workload of our topic at least 100 hours as we have 5 people on our team we plan to add multiple search functions, work on creating an intuitive and aesthetic UI, and add customization to configure what a “document” is exactly defined as. We also plan to have a dedicated “PDF” mode. As stretch goals, we also might add a bookmark function as well as potentially an annotation system that would allow for the user to highlight and write in notes with their searches.

Idea so far:

- **Do the first example in the “Theme 1: Intelligent Browsing” section**
    - Index the current page and allow users to search over the page using a common retrieval function, such as BM25 (the current search capabilities are limited to exact keyword match)
    - Only problem is that we need to satisfy the N*20 hours of work requirement (N is how many people we have). Andrew suggested:
        - “I think there are lots of add-ons and things we can do to justify the time. For instance, we could provide a fancy UI, add different search algorithm options, maybe we also index and search links on the current page”

Answering the questions down here first before writing essay format above

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.

- NetIDs/names: Miguel (miguelf4), Riya Gyanmote (rgyanm2), Merian Croos (mcroos2)
- The captain is: Miguel (miguelf4)

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

- We chose the “Theme 1: Intelligent Browsing theme. Specifically, we want to create a chrome extension that indexes the current page to allow users to search over the page using common retrieval functions. This is very similar to the first example, but since our team consists of 5 members we that means we will have the opportunity to be more ambitious with the project.
- This is a problem, because currently the built-in search for chrome (ctrl+f) only does exact-word match filtering. For many purposes this isn’t very useful. Being able to search a single page with an intuitive “search query”-based type of search can help a lot. Since we only plan to be indexing one page at a time, this could prove especially useful for long pages, but this also means that we have to decide what exactly we define as a “document”. We plan to have this be a configurable option based on what users are looking at. For example we could treat a document as a single sentence, a paragraph, or in the case of a PDF a whole page.
- The problem of course relates well to the theme of the class because it’s all about being able to use more natural language to search a page, similar to how you would search for the document that you want to find using google.

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

- We plan to use BM25 as our retrieval function. We'll probably need some sort of dataset to group words that have the same root together (run, running, ran)

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

- We will demonstrate our chrome extension working on various web pages to find the “document” that is most relevant to the query that we enter. (Recall that our extension will have to define a “document” as some sub-section of a single webpage,)

Which programming language do you plan to use?

- We plan to use Javascript + html + css to create our chrome extension.

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.

- Blah blah add stuff about adding multiple search functions, making the ui look nice, being able to configure what a “document” is defined as. Having a dedicated “PDF” mode. And maybe more if you guys can think of anything.