This document is my progress report for the course project for Text Information Systems.

Team

The registered team name is “Wonder from Down Under”. This is a single person team with captain and sole member **Alan Anderson**, with netID **alansa3**

Recap on project scope and planned implementation

The topic I have chosen is to write a browser extension to allow the user to search the text in the page using a more sophisticated algorithm than simple keyword ‘exact match’.

Often, a google search will yield pages with lengthy text, and there is merit in being able to quickly locate the points in the text relevant to the query. For instance, if we wanted to learn about the economic impacts of the War of the Roses, we might well find that useful information is buried in a long document on the War of the Roses. We could just search for the keyword ‘economic’, but we might get a better ranking of relevant sections by performing a BM25 search for a larger set of relevant terms.

I am implementing my project as a Chrome extension using Javascript, using the BM25 algorithm. I will compare the results of applying the BM25 algorithm with the results of a simple ‘exact match’ keyword search, to see if the top-ranked parts of the document are more relevant.

In my previous submission, I identified the following tasks. I have split my original task 3 into two components.

1. Developing and documenting this proposal
2. Learning how to build a browser extension
3. Implementing the solution:
    a) Implementing a framework browser extension to highlight or otherwise elevate selected text
    b) Implementing the BM25 algorithm to identify which text to highlight.
4. Developing use cases for testing and running tests and comparisons
5. Improving the usability of the extension (if time allows).

1. Completed tasks

I will preface this section by noting that my wife gave birth to our first child last week, so I have not been able to complete as much of the project as I had hoped. (This complication is why I chose to be in a solo team: I did not want to inconvenience other team members with my the disruptions to my personal schedule.)

In the previous submission, I developed and documented the proposal, completing task 1.

Since then, I have focused on tasks 2 and 3.

Task 2 was learning how to build a browser extension. I have not previously written any browser extensions, and have limited Javascript experience. I therefore assumed that I would require ~5 hours to

get up to speed on this with some online tutorials, and that my development of my solution would be slower than for an experienced developer of extensions and Javascript programmer.

While completion of Task 2 is obviously a subjective question, I would classify this task as complete.

- I reviewed the 'Getting Started' guides for Chrome Developers on the Chrome web site.
- I built the initial 'hello world' plug-in per their tutorial.
- I built and experimented with the three subsequent tutorial examples:
  - Reading time
  - Focus mode
  - Tabs manager.
- I realized that my knowledge of CSS style sheets was inadequate for the task, so undertook a brief tutorial on this topic.

Task 3 I have divided into part (a), the implementation of a framework browser extension to highlight the most relevant text, and part (b), the implementation of the actual BM25 algorithm to identify the most relevant text.

I have completed part(a) of this task to a rudimentary level of a plug-in that searches for and highlights specific text collected from a pop-up. However, I would like to make further refinements in due course to improve usability by collecting and linking to the top X sections of relevant text in ranked order, rather than just highlighting them on the page.

2. Tasks yet to be completed

The next major task is 3(b): the implementation of the actual BM25 algorithm. One this is completed, I will have a working prototype of my project.

I will then start to experiment with some use cases and testing on them (Task 4). My preliminary view is that Wikipedia pages would be a good target for this. I expect there will be a level of iteration between Tasks 3 and 4: testing on real pages with real queries will help to make the right design choices. E.g. should we be applying the algorithm to whole sentences or whole paragraphs?

Continuing Task 4, once I have some useful test cases, I can evaluate the difference between my extension and the results obtained by using the simple 'Ctrl-F' text search to see if it yields improved targeting.

Finally, to the extent time permits, I will try to improve the usability of the extension by adjusting the format in which results are delivered.

3. Challenges/issues

As I have significant coding experience in other contexts (although not in javascript), I am confident that I will be able to implement the BM25 algorithm and get a basic prototype working. I anticipate the following issues and challenges will be the more significant ones:

Finding appropriate samples to test the use cases:

- The examples that came to mind when I developed this proposal related to Wikipedia entries about history, and other historical texts online. However, I have not yet tested this use case against actual online texts.
- I intend to start with Wikipedia pages.
- The extent to which finding appropriate samples is a challenge will become clearer once I have developed and tested the prototype against some initial documents/pages.

Tailoring the BM25 approach to page format and content:

- I am uncertain whether to rank whole paragraphs or just sentences in the documents, or some other break-down of text.
- I will decide this once I see some results from my test cases.
- However, I'm conscious that the right answer may be different for different types of documents, depending on how the writer has structured the information on the page. This may mean that my extension should be tailored to a more specific sort of page (e.g. Wikipedia articles).
- Again, this is a decision that can be made once I have done some testing of the prototype.

Making the browser extension user-friendly:

- Because of my lack of experience programming extensions or coding in Javascript, it will take me longer to figure out how to deliver results in the most user-friendly format.
- The usefulness of the improved search algorithm would be compromised in practice if the results are not easily reviewed and navigated.
- I will do my best in a reasonable period of time, noting that this element is more 'nice to have' than the delivery of the core functionality of BM25 from the perspective of the learning objectives of this course.