Progress Report

We have roughly sectioned our project into three sections (one for each member): text collection from Coursera, fuzzy match algorithm development and implementation, and Firefox extension creation. Below is the progress for each.

Text Collection

Completed tasks:
Research limitations of the Coursera API -- the API offers only summary information, and is therefore not sufficient for retrieving subtitle text. Corpus can be constructed by collating paginated returns from empty search requests. This is painful, but doable.

Assembling timestamp chunks -- this is done by assembling the srt subtitle text data from downloaded form coursera into a corpus. Each entry in the srt is too small to provide reasonable context in search results, so they are chunked into larger documents. This can also be accomplished by hitting the search endpoint and combining the paginated results of an empty query.

Pending tasks:
Assembling timestamp chunks -- get results from the search endpoint rather than the downloaded srt file. This isn’t critical, but it would sidestep the need to serve up the documents from a local webserver.

Issues:

No blockers right now. It would have been nice if the Coursera API was at all useful.

Algorithm Development

Completed tasks:
Research of algorithms and algorithm selection - we have decided to implement an approximate string matching algorithm (rather than a ranking algorithm). The similarity measure between query and document (caption file) will be similar to Levenshtein distance, with some modifications. We have written a formula for the approximate string matching (still in progress).

Pending tasks:
Implementation of algorithm.

Issues:
Unsure of how to parse all substrings in the caption file efficiently.

Extension Creation

Completed tasks:
Research regarding implementing a Firefox extension (also known as “add-on”).
Research regarding basic methodology for retrieving a user query and displaying a “results” page to hold the results of our selected algorithm.
At this point, we have a basic Firefox extension which can be loaded and which provides a right-click context menu that retrieves a selected text and sends to a new “results” tab.

Pending tasks
- Integrating the retrieved query into an API or other method for obtaining search results based on that query
- Displaying the results on the results page
- Allowing for manually entering the input query on the results page for subsequent searches

Issues
- Current methodology for input query retrieval is to copy a selected text to the user’s clipboard and paste it in the results tab. There is likely a more efficient solution, but the current implementation will suffice for basic functionality development.