CS410 Project Progress Report

Group Name: TextRulez

Group Members:

- Timothy Crosling - tgc3
- Avi Nayak - anayak5
- Srishti Sharma - srishti9
- Deepthi Abraham - deepthi7 (Team Captain)

Recap: Project Vision

When reading webpages it is often important to identify the key topics first. Our proposal is to create a tool which will allow users to identify the top topics from a webpage and provide easy access to information related to these topics. This will leverage an inverted index along with text analysis to identify key topics from a website corpus and the ability to cross-reference these topics with the internet (e.g., Wikipedia) to provide supplemental information to the user.

Progress Summary:

We have made steady progress over the last few weeks, namely in developing the Chrome Extension code to which we can base the project. We have successfully created a chrome extension that creates both a popup and sidebar capability to scrape text from a web page. We have also experimented with regular expressions and tokenization in Javascript to do basic pre-processing of the webpage text content - including code to create an inverted index in Javascript.

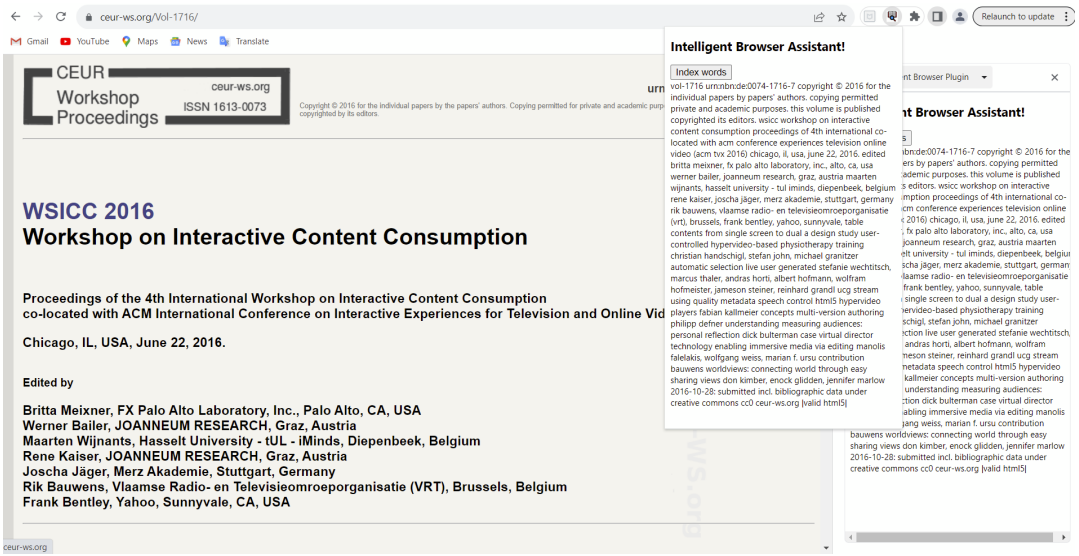

Figure 1: Screenshot of Intelligent Browser Assistant popup and sidebar

Our main focus area right now is extending this code-base to implement a topic estimation model using python. We have experimented with py-script to run this on the client browser, but have had some difficulty in building a local runtime library with the appropriate NLP python libraries. As a result, we are now experimenting with a Google runtime service leveraging JSON for the browser extension to call on demand.

In terms of the NLP text estimation, we have experimented successfully using SpaCy to perform Named Entity Recognition. While this does not fully realize the scope we originally envisaged - the code is fast and we believe could give a basic implementation if integrated into the web extension. We are also investigating a more complete LDA (Latent Dirichlet Allocation) model using gensim - but this poses challenges both in implementation complexity and runtime with each model taking 10-15 seconds (too long for a web service).

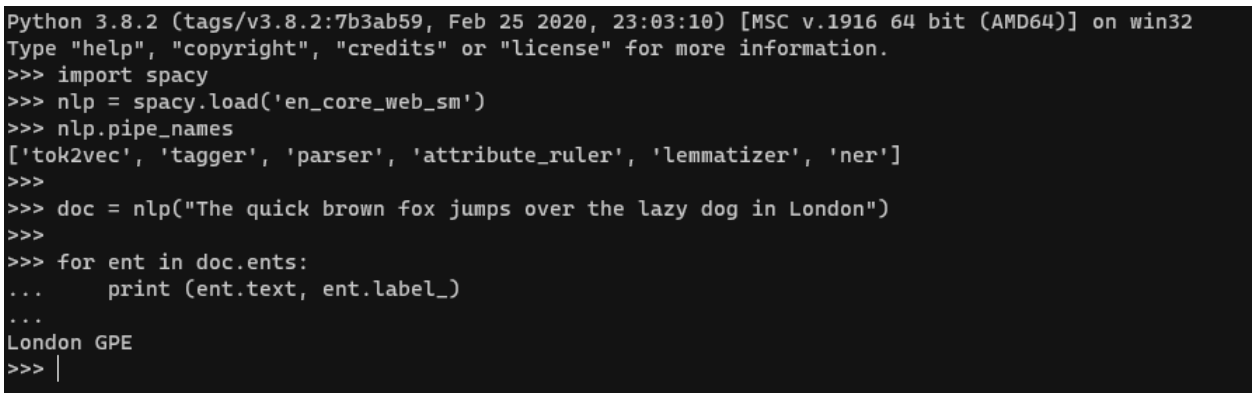

Figure 2: Named Entity Recognition prototype using SpaCy

Next Steps:

There are four key steps we need to complete this project:

(1) Cleanse the text input from the webpage
(2) Implement NER or LDA text analysis module
(3) Load into Cloud Run service on Google Cloud
(4) Augment the Chrome Extension to submit text to the Cloud Run service and process output JSON
(5) Display results to the user within the Chrome Extension popup
(6) Testing and visual interface improvements

Aside from working through the technology we have not yet faced any major challenges or issues.