PROJECT PROPOSAL

WHAT ARE THE NAMES AND NETIDS OF ALL YOUR TEAM MEMBERS? WHO IS THE CAPTAIN?

Team Name: Washington Football Team

Name	NetID	Role
Devin Burke	devinb3@illinois.edu	Captain
Joshua Ray	ray18@illinois.edu	Team Member

WHAT TOPIC HAVE YOU CHOSEN? WHY IS IT A PROBLEM? HOW DOES IT RELATE TO THE THEME AND TO THE CLASS?

The topic our project will cover is **intelligent browsing**. Specifically, we will create a Google Chrome browser extension that summarizes the webpage the user is currently on. It will pull the most relevant/descriptive sentences from the page and display those as a bulleted summary to the user. The number of sentences shown will vary depending on the length of the document.

The problem this browser extension will solve is efficient comprehension of the document with minimal time investment. This project relates to the topic and class because it performs text retrieval using algorithms and techniques learned in the course to augment a user’s experience and make browsing more intelligent.

DESCRIBE ANY DATASETS, ALGORITHMS OR TECHNIQUES YOU PLAN TO USE.

We will implement the vector-space model using bag-of-words representation of each sentence of the document. We will then create a similarity matrix using dot-matrix and measuring the cosine similarity between each sentence, based on the weight of overlapping words. In order to implement the benefit of IDF, we will use a stop-word collection to minimize the impact of common words. We will display the estimated time saved by reading our summary versus reading the entire document, which will be based on a statistical average words-per-minute read provided by reputable sources.

HOW WILL YOU DEMONSTRATE YOUR APPROACH WILL WORK AS EXPECTED?

We will implement the Cranfield methodology of pre-defining a list of documents, relevance-judgments about the summarizing sentences, and measures as to success thresholds. Using these, we will implement tests that utilize mean average precision (MAP) to measure our success.

WHICH PROGRAMMING LANGUAGE DO YOU PLAN TO USE?

As this is a Google Chrome extension, we will use JavaScript. We will implement the algorithms and calculations within JavaScript without the aid of an existing NLP library, and these calculations will be performed client-side. The results will be shown in HTML5/CSS3. Thus, this browser extension should be very fast because it does not need to communicate with an external server, which is part of what makes it unique.

PLEASE JUSTIFY THAT THE WORKLOAD OF YOUR TOPIC IS AT LEAST 40 HOURS.

Task	Assigned To	Est. Hours
Create dot-matrix, cosine similarity, bag-of-words representation, mean average precision, and stop-word functionality in JavaScript	Devin Burke	15
Create pre-defined Cranfield methodology data sets, measures, and relevance judgments	Joshua Ray	5
Implement Cranfield data sets in JavaScript using mean average precision as a measure	Joshua Ray	5
Create algorithm to score and rank sentences based on implemented techniques, and adjust according to success based on MAP evaluation	Devin Burke	10
Determine methodology for filtering page document to only applicable content (i.e., filter out ads, navigation, etc.)	Joshua Ray	5
Create underlying Chrome browser extension and configurability of what percent of the document should be returned	Joshua Ray	15
Research and implement algorithm for calculating reading time-savings based on summary versus entire document	Devin Burke	5
		60