**Memorial University of Newfoundland** **Fall Semester 2018**
**Jacob House and Noah Gallant** **Computer Science 4750**

# Course Project Proposal

Motivation to study computerized text summarization — referred to by the term *automatic abstracting* by those in the field — stemmed from curiosity about the mechanisms used in an automatic abstraction PowerShell script found years ago on the Internet [**PSScript**]. Further research has shown automatic abstraction's usefulness in many interesting fields, including but not limited to the legal and medical professions, scholarly research, and search engine result sorting and summarization.

Three previously-studied methods of NLP-based automatic abstraction have been considered: nested trees, evolutionary algorithms, and graphs for evaluation and summarization, as well as basic sentence extraction.

The nested tree structure method uses both inter-sentence and inter-word dependencies. A document is represented as a nested tree composed of both document trees, which have nodes representing sentences, and sentence trees, which have nodes for words. The relationship between sentence and word nodes is defined by this inter-sentence and inter-word dependency relationship. Performing text summarization involves trimming the tree of less important information until the desired compactness is reached [**art3**].

Evolutionary algorithms (EA) encompass many techniques, however the most common of these uses algorithms to assign weights based on text features. This provides more accurate weighting of important and unimportant information. Examples of such features include assigning a score based on sentence location (e.g. awarding 1 to the first sentence, $\frac{4}{5}$ to the second, and so on, until we reach the fifth sentence which scores $\frac{1}{5}$; remaining sentences receive a score of 0), font style (e.g., scoring capitalized text higher), numerical content (e.g., scoring sentences with quantitative numerical information higher), and word similarity, which assigns a weight based on how often a word appears in different sentences. An EA fitness function then combines the weights to produce a complete text summarization [**art1**].

Graph-based text summarization assigns each sentence a node and edges are used to connect both sentence nodes that have common words and sentences appearing next to each other in the text. The summarization is done by evaluating the nodes with the most edges as important sentences with higher fitness [**art4**].

The Implementation of this design will largely follow the graph based method outlined above. This will involve a singleton class encompassing the entire text, this singleton will contain nodes representing every sentence in the text. Each sentence node will be connected to its adjacent sentences in the text with an edge; sentence nodes that share common words will also be connected with an

edge. To make the detection of similarities between sentences easier, the words from each sentence will be stored as a Trie-DFA. A summary of the text will be made by returning the sentences with the largest number of edges. The number of sentences returned for the summary will depend on the provided summary constraints.

This method of forming a summary will extract the importance of the text by finding the similarities between sentences and is therefore expected to be more ideal for larger amounts of texts as a single paragraph is comprised of very few sentences which would result in very few comparisons between sentences.

Pre-processing is also a requirement for this implementation. This involves separating the text into individual sentences to be added to nodes. Splitting a text into sentences is more complicated than simply breaking a sentence at a period, as periods do not always terminate sentences. In many cases, periods are used for abbreviations or in other special situation, and marking these as individual sentences could fragment the summary.

Some Concerns with this implementation include the difficulty of duplicating the same topic, which through our research is a problem that is common, but not yet perfectly solved. This implementation also will not synthesize summary sentences; rather it only returns a subset of existing sentences, which may appear broken and unformatted.

*Note that only four of the references included below have been cited in the above text. References not cited will be included in the final report and are present to demonstrate that sufficient relevant literature is available.*