Project Proposal – Document Similarity

Anand Bahety
Cody Dunne

This project idea is centered on finding similarity between different parts of documents. The idea is to scan several sentences in different paragraphs and try to find some similarity or relationship between them. The use of inexact string matching algorithms can potentially be useful for this. These algorithms try to match and align the characters based on some scoring functions. However, for English text we need to align words and try to find meaningful relationship between sentences instead of just checking them character by character. This includes checking words for synonyms; antonyms etc. and designing the scoring function accordingly.

We will try to use local/global alignment inexact string matching algorithms to align sequences of words instead of characters with a scoring function based on data from the WordNet database. WordNet is a freely available lexicographic database of words and their relationships. There are many existing word pair scoring functions and libraries based around WordNet already, such as WordNet::Similarity. Ours will build upon those already existing, and we plan to take into account the scoring of synonyms, antonyms, differing parts of speech, and the hierarchical network of word relationships WordNet provides.

The first challenge in the creation of our inexact matching algorithm is modifying Smith-Waterman or BLAST to run with a dynamic programming table representing words instead of individual characters. An interesting complication will be the complexities of word rearrangements and the various types of gaps that must be scored intuitively. For example, differing numbers of adjectives or adverbs may not change the meaning of a sentence but will be hard to code into the algorithm. Also, the use of prepositions or pronouns instead of nouns many be confusing.

The motivation for this project comes from various other projects such as developing a versioning system and detecting plagiarism in the documents of different authors. In the given amount of time, we will try to implement the above idea for comparison of paragraphs with a few lines to see if this idea works in practice. Several natural language processing concepts can be applied to achieve better results but may not be used here due to limited scope and time available for the project.