CSC 583: Programming Project 1

Andrew Zupon

1 Code Description

2 Indexing and Retrieval

This section describes how I built the various indexes used for retrieval, along with how I built the queries from the Jeopardy questions and categories.

First I will discuss how I preprocessed the Wikipedia documents. The first step was to gather the directory into a list of files, then loop through each file. Within each file, I joined the lines into one long string and then split on `\n\n[[`, since each Wikipedia page title is enclosed in double brackets and follows the preceding page after two empty lines.

I ended up building three different indexes: one that used stemming (`indexStems`), one that used lemmatization (`indexLemmas`), and one that used neither stemming nor lemmatization (`indexPlain`). The stemming index uses Lucene's StandardAnalyzer, while the latter two use the WhitespaceAnalyzer.

3 Measuring Performance

4 Changing the Scoring Function

5 Error Analysis

6 Improving Retrieval