HW4: Vector Space Retrieval and Error Analysis

Fernando Garza - 146368
April 26, 2014

1 INTRODUCTION

The purpose of this homework was to build a model which retrieves a vector space of documents, which returns the "best" document for a given query, and improve the result by several methods and techniques.

2 VECTOR SPACE RETRIEVAL

The system consists of several phases:

- **Document Reader**: a collection reader was given which took the documents and imported them into the system.
- **Vector Annotator**: this is the analysis engine of the system, which tokenizes the documents and updates the CAS, code was implemented for this class.
- **Retrieval Evaluator**: this is a class that finds the similarity between queries and documents, and further calculates the error. Each of the documents and queries are represented as strings in a text file, within the system they are vectors. The similarity is calculated by testing for the cosine similarity between the query and each of the documents. Then, the system calculates the mean reciprocal rank between the best selected documents and the gold standard document for each of the queries.

The baseline results returned a MRR of 0.67.

3 Error Analysis

Although the baseline is more or less good, it is not the best result. A feature added for this was to remove stopwords; a file with stopwords was included with the archetype. Also, punctuation signs and marks were added to the stopwords list, which improved the MRR to 0.877. Only one query was not given the correct answer as rank 1.

Also, setting tokens to lower case was implemented, but didn't improve the MRR.

Other techinques such getting lemmas from tokens were thinked to be implemented, but overfitting could occur, and thus in the end was not included.