Homework 4
Engineering and Error Analysis with UIMA

Fernando Aguilar - 114297

April 24th, 2014

1 Introduction

In this homework, an information retrieval system was developed, and a very brief error analysis was done. We used three similarity metrics in order to enhance the mean reciprocal rank (MRR).

2 Design

The information retrieval system has several processing phases (pipeline steps). Each phase is implemented in a UIMA analysis engine. The phases are described as follows:

2.1 Document Reader

This phase reads the data file provided for the homework. It reads each line of the data file and creates a CAS in order to process it in the following pipeline phases.

2.2 Vector Annotator

In this step, the component extract the text tokens in order to generate the term-frequency vector of the Sentence. Before generating the vectors, some preprocessing techniques are applied to the text:

- Strip and Trip white spaces
- Lowercase
- Remove stopwords
- Stemming

2.3 Retrieval Evaluator

This is the final phase, in which the MRR is calculated for each similarity metric proposed:

- **Cosine similarity** : Given two term frequency vectors U and V, the cosine similarity is defined as:

$$sim_{cos} = \frac{\|U \cdot V\|}{\|U\|\|V\|} \tag{1}$$

- **Jaccard distance**: Given two sets of terms (words) A and B, the Jaccard distance is defined as:

$$sim_{jac} = \frac{\|A \cap B\|}{\|A \cup B\|} \tag{2}$$

- **Dice coefficient**: Given two sets of tems (words) A and B, the Dice coefficient is defined as:

$$sim_{dice} = \frac{2\|A \cap B\|}{\|A\| + \|B\|} \tag{3}$$

- **Inverse Manhattan distance**: Given two term frequency vectors U and V, the inverse Manhattan (taxicab) distance is defined as:

$$sim_{Manh} = \frac{1}{\sum_{t \in C} |F(t,U) - F(t,V)| + \sum_{t}^{T(U) \setminus C} F(t,U) + \sum_{t}^{T(V) \setminus C} F(t,V)} \tag{4}$$

  Where C is the set of common terms between U and V, $T(X)$ is the set of terms of a term frequency vector X and $F(t, X)$ is the frequency of the term t in the term frequency vector X.

3 Error analysis

In the following text, we will show the MRR using different similarity metrics described above:

```
Cosine Similarity
-----------------
 (MRR) Mean Reciprocal Rank ::0.6666666666666666
Total time taken: 0.791

Dice Coefficient
----------------
 (MRR) Mean Reciprocal Rank ::0.611111111111111
Total time taken: 0.793

Jaccard Similarity
------------------
 (MRR) Mean Reciprocal Rank ::0.8333333333333334
Total time taken: 0.809

Inverse Manhattan distance
--------------------------
 (MRR) Mean Reciprocal Rank ::0.8333333333333334
Total time taken: 0.835
```

Assuming that the three retrievals with theirs documents are the only corpora used to evaluate the systems, we see that Dice coefficient is the similarity metric with the least MRR, while the Jaccard and Inverse Manhattan distance show the highest MRR. A possible explanation is that Inverse Manhattan integrate information about frequency and common terms. Unlike other metrics, the Jaccard similarity is well proven metric for bag-of-word models.