Preface

T-75.4400 Information Retrieval
Group 15: Tablet ergonomics

Helsinki, April 15, 2014,

Sadi Hossain 69423U

Patrik Vilja 401531

Peter Vilja 401544

Contents

1. Introduction

The documents are indexed with lucene. The queries searches results from the index instead of searching results from the documents itselfs. All the words in the index can be stemmed if it improves

During the implementation we had some difficulties to understand what we exactly wanted to calculate and how the results should be presented. First we seemed to get same results for VSM and BM25 but later noticed that also the order of results counts.

We decided to use d3.js JavaScript library to draw the recall curves. Search application writes the results to csv files and then JavaScript reads the files and draws the curves from the results. The library was new for us and took some time to learn. In the end we spend much more time actually implementing the search than showing the result with d3.js.

Our solution uses stop words and Porter stemmer to shorten the words in the query. Stop words is word which occurs many time in different contexts and thats why they are not easing the search. For example “and”, “are” and “by” are stop words which are ignored while searching. Lucene uses stop words by default in StandardAnalyzer. It automatically ignores stop words in search, this means that if the query contains stop word search won’t find any results with the stop word. Porter stemmer shortens the words to the base form. Now while indexing the documents we can stem the words in the document and while searching documents we can stem the words in the query. Stemming can either increase or decrease the effectiveness of query.

2. Compared Techniques

2.1 VSM

Asiaa VSM:stä.

2.2 BM25

Asiaa BM25:sta.

2.3 Stemmers

Mikä on stemmeri ja mihin sitä käytetään.

3. Evaluation

Jaaa sitten on hirvee määrä tekstiä evaluoinnista.

4. Conclusions

Mitä opittiin? kaikkee jännää.

5. References

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. [e-book] Cambridge: Cambridge university press, 2009. Available through: http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf [Accessed: 14 Apr 2014].

6. Contributions