Title

IR: Assignment 1

Due

14 Aug 2018 5:00 PM

Number of resubmissions allowed

Unlimited

Accept Resubmission Until

14 Aug 2018 5:00 PM

Status

Not Started

Grade Scale

Points (max 100.00)

Instructions

The goal of this assignment is to test the effectiveness of using a thesaurus to improve on retrieval of results.

Step 1:

Create a testbed to support your IR experiments. The testbed must have real data from the Web or an offline source.

Your testbed must include the following:

- Collection of exactly 30 English text **documents**, each containing at least 200 words and at most 2000 words, stored in the format used in the provided *simplesearch* code.
- Set of 3 **queries** known to match documents in the collection. Queries must be stored in a text files named query.1, query.2 and query.3
- Set of **relevance judgements** for every query+document combination. Relevance judgements must be in the range 0-2, where 0 represents irrelevant, 1 represents somewhat relevant and 2 represents highly relevant. You have to manually inspect each document to determine relevance to each query! Relevance judgements must be stored in files named relevance.1, relevance.2 and relevance.3, each containing a list of integers, one per line, corresponding to the relevance judgements of each document in sequence with the query in question.

All files must be stored in a *testbed* directory.

Step 2:

Use your testbed and the sample *simplesearch* engine (under Resources) as the basis for this assignment.

Modify the sample search engine to include the use of an English thesaurus. Use any text thesaurus you can find online.

Evaluate your search engine relative to the original system using Recall, Precision, MAP and NDCG, based on the data in your testbed. For both metrics, calculate a "before" value based on the original system and an "after" value based on the system with thesaurus support included. You can do the calculations by hand, on a spreadsheet or in code.

Write a short report on the design of your implementation (up to 1 page), followed by the detailed relevance metrics calculations/results.

This assignment must be done individually - no groupwork is allowed.

NOTE: You MUST use queries that result in different results for the 2 cases. If your query terms are not in the thesaurus, find other query terms.

Submit all your data, code and your report (as a PDF file) in a ZIP file.

Marking Guide:

- Testbed - documents, queries, judgements: 40
- System design/implementation: 20
- Recall, Precision, MAP and NDCG calculations: 40