Denham Preen

Prnden002

Sabelo Xulu

SOURCE – means needs to change before submitting

Information Retrieval of Medical Data

Assignment 2

Table of Contents

Introduction

The Dataset

Collection procedure

System

Algorithm

Solr

BM25

Results

Conclusion

queries, algorithms employed and evaluation results.  The report must be at most 10 pages.  The report is due on the final assignment deadline - submit both the report and the code you wrote in a single ZIP file.

**Mark Breakdown**

* Part 1: Report: appropriate dataset and queries 15%
* Part 2: Demonstration  25%
* Part 3: Report: algorithms employed and results 45%
* Originality/Creativity 15%

Introduction

This document describes the process behind building a search engine to find medical information from a dataset of SOURCE web documents specifically html pages relating to medical illnesses and diseases. The intention of the search engine is to search for symptoms and to ascertain websites that are relevant to those symptoms. This paper describes the process in generating the corpus and why the documents are relevant to the intended use of the application. The paper then goes on to explain how the application was implemented and the algorithm used to rank the results in the most relevant order to the user’s needs. The results from 3 users is tallied and a conclusion on the relevance drawn.

The Dataset

The corpus used compromises of SOURCE *number* documents. These documents were scraped from links on Wikipedia pages relating to medical documents. These documents were scraped using a chrome extension…Source. Naturally the magnitude of the domain of medical illnesses and diseases meant the documents we used cover a small subset all illnesses and diseases and so the queries we used to gather documents were aimed at identifying more common illnesses and ailments. Examples of queries used in building the corpus was “common colds”, “prevalent diseases” and “recent sicknesses”. Additionally, as we initially intended for the application to gather information specific to certain body parts we catered queries to different parts of the body. An example of this document search was “chest infections”.