Information Retrieval Engine

Assignment 1 - Report

Information Retrieval - 2016/2017

André Lopes - 67833

Raquel Rocha - 62196

October 5, 2016

Table of Contents

1 Introduction **2**

2 Modelling **2**

2.1 Class Diagram 2

2.1.1 Data Flow 2

1. Introduction

Information retrieval is the act of collecting information from various sources, which main principle is querying the information gathered. A big example of information retrieval is Google, where a user inputs a set of keywords and gets the respective results ranked by relevance - this is called an information retrieval search engine, which is the objective of this assignment.

This report explains the structure of this assignment, in terms of class estructure and data flow of the engine.

2. Modelling

In order to model the structure of the search engine, there was a need to better understand each component:

- Document Processor
    - Opens each document to collect data (in Corpus Reader)
- Corpus Reader
    - Reads the data in each document processed by the Document Processor, according to its format
- Tokenizer
    - Divides the data collected in terms, removing stopwords, stemming and other text transformations needed
- Indexer
    - Indexes each term with a reference to the documents it belongs (postings lists)
- Searcher
    - Retrieves the documents that contain the queries token from the index

2.1 Class Diagram

INSERT PICTURE HERE

2.1.1 Data Flow

Explicacao do class diagram (como q os diferentes modulos funcionam)