IR Assignment 1

Homework Submission Guidelines

1. **Due date: 15.11.18 at 23:55**
2. Homework must be done in groups of 2 students
3. Answers can be submitted either in English or Hebrew
4. HW submission should be done via moodle in the corresponding area (by **only** one of the students)
5. Late submission penalty (20% a day) for submitting after the assignment's due date
6. Questions / clarifications and more in the dedicated discussion sub-forum.
7. Total time of machine usage is **600** minutes. Use the "**Stop**" button to shutdown your machine when needed.

In this programming assignment, you will build an inverted index for the AP collection and retrieve documents using Boolean queries.

The files for the assignment are located in /data/HW1/
Inside the folder you will find the following files and directories:

a. **AP_Coll_Parsed**– a directory containing files of 242,918 documents from the AP dataset in a **trectext** format. Each file contains several documents, separated by <DOC> tags. Each document has a unique document ID, specified by the <DOCNO> tag, which comes right after the opening <DOC> tag. The text of the document to be indexed is contained within <TEXT> tags.
Note: The text was lowercased, and the following punctuation marks were removed:
!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~

Here is an example document:

```
<DOC>
<DOCNO> AP900101-0002 </DOCNO>
<FILEID>AP-NR-01-01-90 0005EDT</FILEID>
<FIRST>r w PM-SocialSecurity-Glance     01-01 0304</FIRST>
<SECOND>PM-Social Security-Glance,290</SECOND>
<HEAD>New Year Brings Social Security Changes</HEAD>
<HEAD>With PM-Social Security Bjt</HEAD>
<DATELINE>WASHINGTON (AP) </DATELINE>
<TEXT>
here are some changes in social security
benefits and taxes that take effect with the new year
benefits monthly benefit checks increase 47 percent to offset
the effects of inflation the average retired workers social
security check will rise from 541 to 566
…
</TEXT>
</DOC>
```

b. **BooleanQueries.txt** – a query file with 10 Boolean queries

Boolean Query Structure

A Boolean Query is composed of terms and the following logical operators:
AND, OR, NOT.
Consider an example Boolean Query:

((south AND african) NOT sanctions)

For this query, we want to retrieve all the documents containing both the terms "**south**" and "**african**" but not the term "**sanctions**".
Note that parentheses determine precedence.

Part 1 (33%) – Inverted Index

Your first task is to write a function that creates an inverted index for the AP collection.
Before implementing the function please read the following notes carefully.

- During index construction, specifically, for building the posting lists you should use successive integers as document internal identifiers (IDs) for optimizing query processing, as taught in class, but you still need to be able to get the original document ID when required.
- Name your function "InvertedIndex".

Part 2 (34%) – Boolean Retrieval Model

Your second task is to write a function that given an inverted index and a Boolean query retrieves a set of matching documents.
Before implementing the function please read the following notes carefully.

- Write the retrieval results to a file "Part_2.txt" as follows.
  Each line (10 lines total) contains the original IDs (not internal IDs) of the retrieved documents separated by space.
  Keep the same line order as in the "BooleanQueries.txt" file.
- Name your function "BooleanRetrieval".

Part 3 (33%) – Collection Statistics

Your third task is to write the following statistics to a file "Part_3.txt".

1. Write the top 10 terms with the highest document frequency (10%).
2. Write the top 10 terms with the lowest document frequency (10%).
3. Explain the different characteristics of the above two sets of terms (3%).

Submission Instructions:

1. Zip all files together and submit a file with name of the following format:
HW1_Student1ID_Student2ID.zip (20% grade penalty otherwise)

2. The Zip file must contain the following files: **InvertedIndex**, **BooleanRetrieval**, **Part_2.txt** and **Part_3.txt**