IR Assignment 2

Homework Submission Guidelines

1. **Due date: 10.12.19 at 23:55**
2. Homework must be done in your assigned groups
3. **PDF submissions only! (20% grade penalty otherwise)**
4. **The file name is: HW2_Student1ID_student2ID**
5. Answers can be submitted either in English or Hebrew
6. HW submission should be done via moodle in the corresponding area (by **only** one of the students)
7. Late submission penalty (20% a day) for submitting after the assignment's due date
8. Questions / clarifications and more in the dedicated discussion sub-forum.
9. Total time of machine usage is **600** minutes. Use the "**Stop**" button to shutdown your machine when needed.

Dry part (70%)

Vector space model (5%):

The following matrix represents the word frequencies of four documents d1, d2, d3, d4. Columns represent the documents in the above order; rows represent the vocabulary of six indexed terms a,b,c,d,e,f in that order. (Use ln.)

	d1	d2	d3	d4
a	0	1	1	1
b	1	2	0	1
c	2	0	0	0
d	0	0	0	0
e	1	0	1	1
f	7	5	7	2

Assume that the fraction of corpus documents in which each term appears is 10%, 10%, 20%, 5%, 50%, 90% for the terms a, b, c, d, e, and f, respectively.

1. Compute the cosine similarity between d1 and d2 where terms are represented by the tf-idf scheme. (Describe the tf-idf scheme you have used and provide details of the computation. Use raw tf.) (5%)

Term Weighting and Ranking (10%):

1. What causes the short-documents bias effect when using cosine similarity? (4%)
2. Name two different examples where:
   a. The removal of stopwords reduces the recall. (3%)
   b. The removal of stopwords reduces precision. (3%)

Relevance feedback and evaluation (45%)

1. User 'A' submitted a query to a search engine and obtained an ordered result list. Then, the user provided feedback to the engine (4 – the document is highly relevant to the information need expressed by the query, 0 – the document is not relevant)

DocID	Relevance
5	4
2	1
1	1
3	3
4	0

   The total number of relevant documents in the collection is 10. Calculate the AP, precision and recall (at rank 5) (5%)

2. Suggest a version of Rocchio's model that utilizes graded relevance judgments. (10%)
3. Suggest a version of Rocchio's model that utilizes the rank of relevant documents in the list. (10%)
4. Propose a variant of AP that uses gradual relevance judgments (10%)
5. In which cases evaluation using MAP will yield the same results as evaluation using MRR? Mention at least 4 different cases (10%)

True/False questions (10%) :

Mark each of the following sentences as true or false and give a short **(but full)** explanation for why your answer is correct:

1. df_t is an inverse measure of the informativeness of term t. (1%)

2. Cosine similarity and Euclidean distance are equivalent for ranking documents in response to a query under some condition. (3%)

3. Vector space-based retrieval is always more effective than Boolean retrieval. (1%)

4. In the vector space model, the higher the value of the normalization factor for a document is, the lower are the chances of retrieval for that document. (1%)

5. The stemming process increases the number of unique terms in the index (1%)

6. Values of beta>1 in F-measure emphasize precision. (1%)

7. In Rocchio's model, q_0 might be closer to the centroid of the relevant documents than q_m. (2%)

Wet part – Intro to Indri (30%)

Part A: (/data/HW2/WET_PART_A)

1. The collection for Part A is located in **docs.txt**
2. Create an Indri index using the following parameters:

```
<parameters>
  <memory>1G</memory>
  <corpus>
    <path> docs.txt  path</path>
    <class>trectext</class>
  </corpus>
  <index>Your folder and index name</index>
</parameters>
```

If the index is created correctly you will find a manifest file **inside** the index directory which looks as follows:

```
<corpus>
   <document-base>1</document-base>
   <frequent-terms>0</frequent-terms>
   <maximum-document>5</maximum-
document>
   <total-documents>4</total-documents>
   <total-terms>212</total-terms>
   <unique-terms>140</unique-terms>
</corpus>
```

Run retrieval with the following parameter file:

```
<parameters>
   <memory>1G</memory>
   <index>Path to your index</index>
   <count>5</count>
   <trecFormat>true</trecFormat>
   <baseline>tfidf,k1:1.0,b:0.3</baseline>
</parameters>
```

1. Run a query "corporation" over the collection using the above parameter file
    a. How many documents did you retrieve?
    b. How many documents did you expect to retrieve? Perform and explain the change that is needed for getting the additional documents. (Examine the text of documents.)
2. Write a query that will return document D2 first; use up to 2 words; explain your choice.
3. Write a query that will return document D1 first; use up to 2 words; explain your choice.
4. By running the query: " Michael Jackson" you will retrieve document D4.
    a. Do you think D4 is relevant to the information need expressed by this query? Explain.
    b. Type a query for which D4 can be marked as relevant document; use up to 2 words; explain (refer to the ranking score assigned to D4 as a result of the two queries)

Part B:

1. The files for PartB are located in /data/HW2/WET_PART_B/
2. In the PartB folder you will find the following files and directories:
    a. "AP_Coll.tgz" compress file contains AP documents ("database")
    b. "queries.txt" – query file with 150 queries
    c. "qrels_AP" file – the AP relevance judgments
    d. "StopWords.xml" – the INQUERY 418 stopwords list
    e. "IndriBuildIndex.xml" – build index configuration file

3. Build 4 indexes using the given "database" directory and parameter file " IndriBuildIndex.xml".
   Report the time it takes to build each index (you can use **stopwatch** or use the "**time**" command to launch prior to IndriBuildIndex application):
    a. Index1: **Without** stopwords removal and **without** stemming.
    b. Index2: **With** stopwords removal and **without** stemming.
    c. Index3: **Without** stopwords removal and **with** stemming (Use "Krovetz" stemmer)
    d. Index4: **With** stopwords removal and **with** stemming

   (Note: Create first 4 index directories, each of which for an index version)
4. Which index version took less time to be created? Explain.

5. Run retrieval over the four indexes with the following parameter file (using tf.idf weights):

```
<parameters>
    <memory>1G</memory>
   <index>Your index Path</index>
   <count>1000</count>
   <trecFormat>true</trecFormat>
   <baseline>tfidf,k1:1.0,b:0.3</baseline>
</parameters>
```

6. In your irstudent directory, unpack the trec_eval file located in the 'parameters.tgz' file.
   Use the trec_eval application to evaluate the 4 retrieval results and complete the following table. Which retrieval result obtained the highest MAP value? Explain.

Stopword Removal	Krovetz Stemmer	MAP	P@5	P@10
Without	**Without**			
With	**Without**			
Without	**With**			
With	**With**			

Good Luck