IR Assignment 3

Homework Submission Guidelines

1. **Due date: 15.06.23 at 23:55**
2. The assignment can be done in pairs
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 submissions will not be accepted
6. Questions / clarifications and more in the dedicated discussion sub-forum in Piazza

Dry part (80%)

Language Models (30%):

1. (5%) Rank the following documents with respect to the query "information retrieval" using the query likelihood model and Dirichlet smoothing with mu=100
    a. Doc1: information retrieval course is fun
    b. Doc2: information information information information information information computer
    c. Doc3: information retrieval retrieval methods

2. (5%) Propose a variant of **JM smoothing** method in which λ will depend on **query length**.

3. (10%) Show that the ranking induced using the -CE (negative Cross Entropy) score between a query and a document is equivalent to that induced using the query likelihood method. Specify your assumptions.

4. (10%) KL divergence is an asymmetric divergence measure, which measures how different two probability distributions are from each other. In our course we use it to measure how different the probability distribution M_q is at modeling M_d. Suggest a **symmetric** and **non-negative** measure for comparing query and document language models based on KL divergence.

Relevance Models (20%):

Given the query q: "$cat\ dog$" and the following list of initially retrieved documents:
d_1: $cat\ dog\ cow\ pig\ horse$
d_2: $the\ cat\ and\ the\ dog\ are\ playing\ together$
d_3: $cat\ cat\ cat\ cat\ cat$

1. (5%) Write the RM1 formula. Explain what the formula means.

2. (5%) Induce a query model using the RM1 relevance model assuming that p(q|d) is constant and that an MLE is used for the document language model. Explain your calculations.

3. (10%) Induce a query model using the RM3 relevance model. Use $\beta = 0.3$ and the same assumptions as before.

Passage Retrieval (20%)

The passage retrieval task is ranking passages of documents by their relevance to the information need expressed by a query.
A passage is any sequence of text in a document which is usually much shorter than the entire document length.

The idea is to estimate the relevance according to the probability of generating a passage g given the query q, expressed as $p(g|q)$. Instead of directly estimating this probability, we use Bays rule and passages can be ranked using the query-likelihood approach: $p(q|g)$. Note that we assume that the prior $p(g)$ is uniform and thus can be removed.
$p(q|g)$ can now be estimated using the standard language-model-based approach. For each passage a language-model is inferred.
Your task is to suggest **3** different approaches (**that were not taught in class**) to estimate $p(q|M_g)$, where M_g is the passage model.

Tips:
1. In your solution you should address the vocabulary mismatch problem between the terms used in the query and in the short relevant passages.
2. Use the document that contains the passage and the collection for smoothing.
3. Each suggestion should result in a **valid** language model.

Elaborate and detail all of your notations, free parameters, equations, etc…

Positional Language Models (10%)

Rank the following documents with respect to the query “onion soup onion” using the ‘best position strategy’. Use Dirichlet smoothing with $\mu = 100$ and MLE for the query model. For the propagation function use the Gaussian kernel with σ=5.

D1: onion vegetable vegetable

D2: corn onion soup

D3: potato onion

Wet part – Query expansion (20%)

1. Files can be copied using:
sftp -r irlabsharedstorage.irlabuser@irlabsharedstorage.blob.core.windows.net:HW3

Enter password: yUvF4gAs1+PQIYzKyB6DbZNz2J1/6XSZ

2. Inside the folder you will find the following files and directories:
   a) "indriRunQuery.xml" – A retrieval parameter file. We use the Dirichlet smoothed unigram language model as our retrieval method.
   b) "Dinit.res" – an initially retrieved document list using a Dirichlet smoothed unigram language model. We retrieve the top 1000 documents for 10 ROBUST queries.
   c) "qrels_10_Queries " file – the ROBUST relevance judgments.
   d) "query_relDoc" directory – Each file in the directory is in the format: "**queryId_document_name**.txt". Each file contains the text of one relevant document (document name) for a given query (query id).
   e) ROBUSTIndex – The collection index. We used Krovetz stemming
   f) 10_ROBUST_Queries – 10 queries

3. Fill in the empty cells in Table 1 for "Dinit" columns using trec_eval evaluation tool (**10**%)
4. Expand each query using the provided relevant document's text to achieve the best MAP, P@5 and P@10 values as possible.
   a) You can expand each query by up to **2 words**.
   b) The original query words cannot be removed.
   c) Explain your expansion method – be creative (**10%**).

Table 1	Dinit			Best expansion		
Query	**MAP**	**P@5**	**P@10**	**MAP**	**P@5**	**P@10**
301						
302						
303						
304						
305						
306						
307						
308						
309						
310						
Average						

Submission Instructions:

1. A **PDF** file containing all answers to the questions (Dry and Wet parts).
2. The name of the file as follows:
   HW3_Student_1_EMAIL_Student_2_EMAIL.pdf