Assignment #3

Homework Submission Guidelines

1. **Due date: 9.1.2025**
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 submission penalty (**5% a day**) for submitting after the assignment's due date
6. Questions / clarifications and more in the dedicated discussion sub-forum in Piazza

Dry part (80%)

Language Models (35%):

1. (5%) Given the following documents:

   D1 : {a a a d}

   D2 : {a a b c}

   D3 : {a d d d}

   D4 : {b d d d}

   And a query q : {a d}. Rank the 4 documents with respect to the query (in descending order) using KL Divergence. Using Dirichlet for smoothing with the Dirichlet prior set to 1000. Detail your calculations.

2. (5%) Prove that JM (Jelinek Mercer) smoothed language model yields a valid probability mass function. Specify all the requirements and assumptions of this proof.

3. (5%) 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_d is at modeling M_q. Suggest a **symmetric** and **non-negative** measure for comparing query and document language models based on KL divergence.

5. (10%) Given a set of terms marked as relevant to an information need and a user's query (there is no overlap between the query terms and the set of relevant terms). Show, using formulas and explanations, how the language model that is induced from the original query can be expanded using these terms. The updated query model should place greater emphasis on the original query terms.

Relevance Models (15%):

1. 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: $cat\ cat\ cat\ cat\ cat\ cat$

(5%) Induce a query model using the RM3 relevance model. Use $\beta = 0.3$ and assuming that p(q|d) is constant and that an MLE is used for the document language model.

2. (5%) Mention 2 ways that can be used to prevent query drift in models that are based on pseudo relevance feedback. Limit your answer to 3 lines.

3. (5%) Suggest two **different** methods for measuring the similarity between two **short** queries. Elaborate and explain the methods and the reason that each was chosen.

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…

Diversification (10%)

The diversification problem in ad-hoc document retrieval refers to the challenge of providing a diverse set of relevant documents in response to the query. Instead of simply retrieving a list of most highly ranked documents by some retrieval method, the system aims to present a varied selection of documents that cover different aspects of the query and that do **not overlap** with each other as much as possible. This helps to offer more comprehensive results.

Your task is to propose a diversification method that considers the document's relevance score to the query and the level of redundant content compared to the documents already retrieved. You can select any retrieval method and must provide equations and explanations to support your solution.

<u>Wet part – Query expansion (20%)</u>

1. Files can be found on Moodle under: **Assignment_3/files/**

2. Inside the folder you will find the following files and directories:
   a) " Dinit_qld.txt" – an initially retrieved document list using a Dirichlet smoothed unigram language model. We retrieve the top 1000 documents for 10 ROBUST queries.
   b) "qrels_10_Queries " file – the ROBUST relevance judgments.
   c) "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).
   d) ROBUSTIndex – The collection index. We used Krovetz stemming and no stopword removal. Link is avaliable under **Assignment_3. (Unzip before use.)**
   e) 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. Conduct retrieval using the QL approach:

```
index_path = Path_to_RobustPyserini_index'
searcher = LuceneSearcher(index_path)
# specify custom analyzer for the query processing step to match the
way the index was built
analyzer = get_lucene_analyzer(stemmer='krovetz', stopwords=False)   #
Ensure no stopwords are removed from the query
searcher.set_analyzer(analyzer)
# Optionally, configure BM25 parameters (can be adjusted as needed)
searcher.set_qld(mu=1000)
```

   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_ID_Student_2_ID.pdf**