# Task 1 - Language Model

1. Given the following documents and queries, build the language model according to the document collection.

***d1 =*** *York is a city in northeast England that was founded by the ancient Romans.*
***d2 =*** *The state of New York is often referred to as New York State to distinguish it from New York City.*
***d3 =*** *Which cities were founded by the Romans?*

***q1 =*** *York*
***q2 =*** *York England*
***q3 =*** *York university*

Use MLE for estimating the unigram model and estimate the query generation probability using the Jelinek-Mercer smoothing:

$$\hat{P}(t|M_d) = (1 - \lambda)\hat{p}mle(t|M_d) + \lambda\hat{p}mle(t|_C), \qquad \lambda = 0.5$$

For each query, rank the documents using the generated scores.

d1 = 14 terms.
d2 = 20 terms.
d3 = 7 terms.
Total = 41 terms.

***q1 = York***

$P(q1|d1) = 0.5 * \left(\frac{1}{14} + \frac{4}{41}\right) = 0.0844948$
$P(q1|d2) = 0.5 * \left(\frac{3}{20} + \frac{4}{41}\right) = 0.1237805$
$P(q1|d3) = 0.5 * (0 + \frac{4}{41}) = 0.0487805$

Here we can see that the relevant documents is as follow: *d2>d1>d3*

***q2 = York England***

$P(q2|d1) = \left(0.5 * \left(\frac{1}{14} + \frac{4}{41}\right)\right) * \left(0.5 * \left(\frac{1}{14} + \frac{1}{41}\right)\right) = 0.0040481$
$P(q2|d2) = \left(0.5 * \left(\frac{3}{20} + \frac{4}{41}\right)\right) * \left(0.5 * \left(0 + \frac{1}{41}\right)\right) = 0.0015095$
$P(q2|d3) = \left(0.5 * \left(0 + \frac{4}{41}\right)\right) * \left(0.5 * \left(0 + \frac{1}{41}\right)\right) = 0.0005949$

After the calculation we see that the ranking of the documents is as follow: *d1>d2>d3*

***q3 = York university***

$$P(q3|d1) = \left(0.5 * \left(\frac{1}{14} + \frac{4}{41}\right)\right) * (0.5 * (0 + 0)) = 0$$

$$P(q3|d2) = \left(0.5 * \left(\frac{3}{20} + \frac{4}{41}\right)\right) * (0.5 * (0 + 0)) = 0$$

$$P(q3|d3) = \left(0.5 * \left(0 + \frac{4}{41}\right)\right) * (0.5 * (0 + 0)) = 0$$

University does not occur in any of the documents, therefore the ranking is the same for all the documents: *d1=d2=d3*

2. **Explain the language model, what are the weaknesses and strengths of this model?**
   The language model is a model to distinguish words from a sentence by their relevance.
   The advantages of the mode is that it is intuitive, simple and mathematically precise. On the other hand, the weaknesses is that it can be difficult to know what the user wants, and it is difficult to improve the relevance.

3. **Explain what smoothing means and how it affects retrieval scores. Describe your answer using a query from the previous subtask.**
   The term smoothing is for adjusting the maximum likelihood estimator due to the language model so that it becomes more accurate and we will get the relevant documents in return. It removes all irrelevant document from the selection (documents that does not contains the query and are therefore "noise" in the results).
   Seen by use of the Jelinek-Mercer smoothing method where a lambda ($\lambda$) is used to even out the results. If we look at the document d3 and take the queries q1 and q2 in closer view, we see that none of the queries exists in d3, but q2 creates more noise in the results and the value of the calculation is lower and therefore the document is less relevant to us.

## Task 2 - Evaluation of IR Systems

1. **Explain the terms Precision and Recall, including their formulas. Describe how differently these metrics can evaluate the retrieval quality of an IR system.**

   **Precision** in information retrieval is a term used to describe the accuracy of the returned results of a query, but is also used to determine the overall precision of an IR system by averaging the precision score for a large amount of queries. The term is defined as the fraction of documents in a corpus that are classified as *relevant* to a query. This means that a perfect precision score infers that every document retrieved by a query was relevant, but it does not tell us anything about whether *all* relevant documents were retrieved.

$$Precision = \frac{|\{Relevant\ documents\} \cap \{Retrieved\ documents\}|}{|\{Retrieved\ documents\}|}$$

**Recall** in information retrieval is a term used to describe the amount of relevant documents that were retrieved. The term is defined as the fraction of successfully retrieved documents that are relevant to a query. This implies that a perfect recall score would retrieve all relevant documents to a query, but does not tell us anything about the amount of irrelevant documents that were also retrieved.

$$Recall = \frac{|\{Relevant\ documents\} \cap \{Retrieved\ documents\}|}{|\{Relevant\ documents\}|}$$

*Precision* and *recall* are both metrics that describe document relevancy. They do, however, differ in the sense that *recall* is a metric that describes the fraction of the total amount of relevant documents that were retrieved, while *precision* describes the fraction of the returned documents that were relevant.

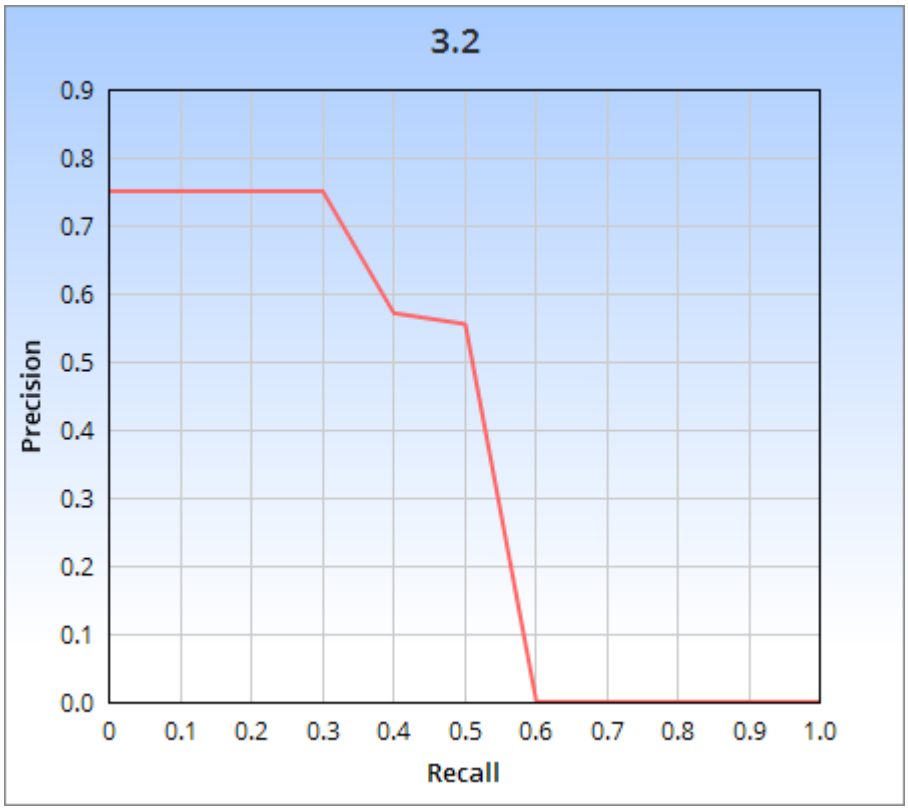
**2. Given the following set of retrieved documents ret = {A, B, C, D, E, F, G, H, I} and a set of all relevant documents rel = {B, C, D, G, I, L, M, N, Q, R}, provide a table with the calculated precision and recall at each level.**

| *ret* | *Relevant?* | *Precision* | *Recall* |
|---|---|---|---|
| A | No | 0 | 0 |
| B | Yes | *12=0.5* | *110=0.1* |
| C | Yes | *23=0.667* | *210=0.2* |
| D | Yes | *34=0.75* | *310=0.3* |
| E | No | *35=0.6* | *310=0.3* |
| F | No | *36=0.5* | *310=0.3* |
| G | Yes | *47=0.571* | *410=0.4* |
| H | No | *48=0.5* | *410=0.4* |
| I | Yes | *59=0.555* | *510=0.5* |

## Task 3 - Interpolated Precision

1. **What is interpolated precision?**
   Working with arranged retrieval we have recall and precision. When the recall is the share of the relevant documents thats sourced, the precision is how accurate the system has been extracting those documents.
   The interpolated precision at a certain level r is defined as the highest precision found for any recall level r' >= r.

2. **Given the example in Task 2.2, find the interpolated precision and make a graph.**




## Task 4 - Relevance Feedback

1. **What is the purpose of relevance feedback? Explain the terms Query Expansion and Term Re-weighting. What separates the two?**

   The purpose with relevance feedback is to assess the result of a given query, and judge its relevance to execute a new query.

   Query expansion is, as the name implies, an expansion of a query. This is done by evaluating the original query and expanding it to find more documents with better relevance. This can be done, in example, by finding synonyms of words and searching for those too.

Term re-weighting is to change the weight of terms so you can find documents that are relevant more easily. Term re-weighting can be a part of a query expansion, but re-weighting itself isn't seen as a query expansion, since you only use the original terms.

**2. Explain the difference between automatic local analysis and automatic global analysis.**

The difference between automatic local analysis and automatic global analysis is that local analysis only uses documents that are returned from the query, while global analysis uses information from all documents in the collection.