IR Project Outline **version 3.0**

Throughout the semester you will work in groups to hand in three jupyter notebooks, one for each milestone. You need to submit one notebook for each milestone per group. The structure of the notebooks should follow this template. We will communicate the submission links to you throughout the semester.

All cells in the template are mandatory. Do not forget to write your reflections!

Milestone 1 - Data **Due date: 02.05.2023**

The overall goal of this milestone is to process a raw dataset that you will use as the basis for your domain-specific information retrieval system. The dataset that you will use is called the 'IR Anthology'. It is a collection of information retrieval publications over the past several decades.

What you need to do

You will download (i.e., from `https://files.webis.de/teaching/ir-ss23/`) and process the raw documents in the dataset into a format that is compatible with Milestone 2, and create *topics* that represent several information needs that you devise. **You must create one topic per group member**. The processed dataset will consist of:

1. the document collection in `.jsonl`-format, a form consistent with `ir_datasets`, e.g., like so

```
{"doc_id": "0001", "text": "How quickly daft jumping zebras vex."}
{"doc_id": "0002", "text": "Quick fox jumps nightly above wizard."}
{"doc_id": "0003", "text": "The jay, pig, fox, zebra and my wolves quack!"}
```

   NB: The `doc_id` and `text` fields are necessary for Milestone 1. Please use the `id` field from the raw documents as the `doc_id`.
2. your custom topics for your dataset in TREC XML-format, e.g., like so:

```
<topics>
  <topic number="1">
    <title>fox jumps above animal</title>
    <description>What pangrams have a fox jumping above some animal?</description>
    <narrative>Relevant pangrams have a fox jumping over an animal (e.g., an dog). Pangrams
        containing a fox that is not jumping or jumps over something that is not an animal are
         not relevant.</narrative>
  </topic>
  <topic number="2">
    <title>multiple animals including a zebra</title>
    <description>Which pangrams have multiple animals where one of the animals is a zebra?</
        description>
    <narrative>Relevant pangrams have at least two animals, one of the animals must be a zebra
        . Pangrams containing only a zebra are not relevant.</narrative>
  </topic>
</topics>
```

For a valid submission, your notebook must register the dataset into `ir_datasets`. You must register it using the name `ir anthology-<team>`, where `<team>` is the name of your team **in TIRA**.

Some more resources that you might find helpful:

- introduction to Python
- introduction to jupyter

What you will hand in

You will upload a docker image containing a jupyter notebook that performs these steps to TIRA. The output result of this process will form the input for Milestone 2. We show you how to do this in the first two tutorials (i.e., this link).

Milestone 2 - Methods **Due date: 30.05.2023**

The overall goals of this milestone are to (1) create relevance assessments for the documents from Milestone 1; and (2) create a baseline information retrieval system that produces a run file using the topics and relevance assessments you have created that you use to evaluate your retrieval system.

What you need to do

- Relevance Assessments: You will create binary relevance assessments (i.e., a `qrels` file) for the output of Milestone 1 (run file). For a valid submission, you must include your relevance assessments into the dataset you created in Milestone 1 and register it into `ir_datasets`. The format of your `qrels` file follow the TREC style:

  ```
  qid 0 docno relevance
  ```

  Here, `qid` is the query number, `0` is the literal 0, `docno` is the id of a document in your collection, and `relevance` is how relevant is docno for qid.
- Baseline Retrieval System: You will develop a baseline information retrieval system that will produce a run file in the same format as the output of Milestone 1, that is standard TREC run file format:

  ```
  qid Q0 docno rank score tag
  ```

  Here, `qid` is the query number, `Q0` is the literal Q0, `docno` is the id of a retrieved document from your collection, `rank` is the position (1 to maximum 1000) in the ranked list for docno and qid, `score` is the score computed by your retrieval system for this docno-qid pair, and `tag` is the identifying name of the retrieval system.
- You will then evaluate the effectiveness of this baseline information retrieval system using the relevance assessments you created in the previous step. For a valid submission, you must persist the run in TIRA. See this notebook for how to do this.

To implement your retrieval system, you could use one of the following libraries:

- pyterrier
- pyserini

What you will hand in

You will upload (1) docker images with an updated version of your dataset containing qrels registered in ir_datasets and upload it to TIRA as before in Milestone 1 , and (2) a second docker image containing a jupyter notebook with the implementation of your baseline retrieval system that produces a run file to TIRA.

Milestone 3 - Analysis **Due date: 20.06.2023**

The overall goal of this milestone is to produce a more effective retrieval system than that of Milestone 2. The final goal is that your system is effective for topics it has not seen.

What you need to do

You will need to modify or extend the baseline system you created in Milestone 2. The system should not only be more effective for topics that you have created relevance assessments for; but also for new topics. You will be presented with an award for developing the most effective system out of all other teams using the same dataset. For a valid submission, you must persist the run in TIRA.

What you will hand in

You will upload a docker image containing a jupyter notebook containing the implementation of your baseline retrieval system that produces a run file to TIRA.