- Registration in QIS: everyone who wants to participate in the exams has to register in QIS
- Exam Date 1: Monday 18th July 2022 13:00 to 15:00 Hörsaaltrakt Bockenheim - H IV
- Exam Date 2: Tuesday 23rd August 2022 13:00 to 15:00 Hörsaaltrakt Bockenheim - H IV
- Present your outcomes of the following programming task in class (choose a time slot in Moodle)

TOPIC T2
LOCALITY SENSITIVE HASHING

Prepare some presentation slides to present the following steps:

1) Briefly describe the process of word-level(!) shingling in your own words (1 slide).

2) Briefly describe locality-sensitive hashing LSH for the Jaccard similarity (1 slide).

3) From a public repository (see list below), choose a (subset of a) dataset for a medical use case that can be represented as a set of bags of words (e.g. doctoral notes for multiple patients) and give a description of the dataset (1 slide).

4) Apply word-level shingling on each bag in the dataset chosen under 3) and obtain the characteristic matrix (as shown in Figure 3.2 in the MMDS book) such that such that the shingle size (i.e. amount of words in each single) can be flexibly set as a parameter. You may remove stop-words from each bag before shingling as needed. Report on the implementation (1 slide).

5) Choose a public implementation[1] of LSH (minhashing) as well as Jaccard similarity, and apply both on the characteristic matrix obtained in 4). Report on the implementation (1 slide).

6) Implement a simple web frontend (e.g. using streamlit or svelte) to select an arbitrary **pair** of bags of words from 3), set parameters (like shingle size) and display the shingles obtained for the two sets (from 4)) together with both the minhash based and the standard Jaccard similarity (from 5). Show some sample screenshots (1 to 3 slides).

7) Create your own repository at github.com and commit your source code. Put your github link and any other references you have used on a References slide (1 slide). Submit your programming task slides in Moodle.

8) Choose a date in Moodle for your slide presentation in class.

Datasets:

1. medical-nlp
   https://github.com/socd06/medical-nlp
2. MeDAL Dataset
   https://www.kaggle.com/datasets/xhlulu/medal-emnlp
3. n2c2 NLP Research Data Sets
   https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

1 You can also decide to implement the chosen algorithm on your own