Instructions!

This repository contains all the tasks for the course AI for the web, **DT506A**.

Dependencies and general info

Before starting to run the practical tasks, make sure that all the important dependencies need to be installed. All of these tasks have implemented on Ubuntu 20.4. They have not been tested in another environment. If you running on Ubuntu, please make sure to run the following:

```
pip3 install -r requirements.txt
```

Task1

Description

The task is about Implementing a simple parser that can be used to build an inverted index for a document collection.

Running instructions

```
python3 Task1.py file name
```

Task2

Description

The task is about extending the inverted index that has been created in the Task1 by a method for querying with a set of k-keywords.

Running instructions

```
python3 Task2.py file name
```

Task3

Description

This task is about computing the frequency for every word that occurs in the collection and also about estimating the constant factor α from Zipf's law.

Running instructions

```
python3 Task3.py file name
```

Task4

Description

The task is about extending the inverted index so that the inverted list are not containing several entries of the same document id and also about finding out the ten words with the largest number of documents.

Running instructions

```
python3 Task4.py file name
```

Task5

Description

This task is about creating two lists of length n and filling one with numbers from 0 to n-1 and the other in the same way and shuffle it from the first list. The purpose is to sum up all the indexes to the lists and observe the difference between the methods.

Running instructions

```
python3 Task5.py
```

Task7

Description

This task is about extending the inverted index from the prevuos tasks with an algorithm for merge-based queries with k-keywords and also assign scores to the inverted lists. Morever, the score system has to be modified by more or replaced with something else e.g. BM25.

Running instructions

> python3 Task7.py movies name benchmark file

Task13

Description

This task is about finding the longest inverted list and compute the amount of space needed for storing the index uncompressed and compressed.

Running instructions

> python3 Task13.py file name

Task14

This task is about implementing a function that computes the prefix edit distance.

Running instructions

> python3 Task14.py

Task15

Description

This task is about implementing q-gram index for a document collection and also about finding out how many 3-grams are there in the collection. It is also about finding out the most frequent 3-grams and how may words start with the letter K.

Running instructions

> python3 Task15.py file name

Task19

Description

This task is about implementing k-means for sparse term-document matrices.

Running instructions

> python3 Task15.py file name