D5: Databases and Information Systems
Automated Knowledge Base Construction, SS 2022
Lab #01: Data Familiarization

max planck institut
informatik

Problem 1 (Named-entity Recognition). Download the data dump folder containing several text files from the course website. In each text file, the first line has the title of the Wikipedia page, the second line gives the Wikipedia pages' URL, followed by source text of the Wikipedia page. Use the Spacy library (`https://spacy.io/api`) and helpful wiki markups present in the source text to do the following:

1. Identify all the unique named-entities
2. Calculate the frequency of each named-entity

The output from each source text should be separately saved in an excel file with .csv format and use title as the filename (e.g., Bill_Gates.csv, Germany.csv, ...). Place all the csv files into a folder called "Problem1". The total number of rows in an excel file should equal the count of unique named-entities extracted and also contain the following 3 columns:

Title	**Named-entity**	**Frequency**
Bill Gates	Melinda Gates	7
Bill Gates	Microsoft	20
...	...	...

Problem 2 (POS Tagging, Tokenization). Using the same data dump given in the previous problem, do the following on each source text:

1. Identify the 5 most frequent verbs and adjectives
2. For each unique named-entity extracted in problem 1, provide all the sentences where it is mentioned

The extracted verbs and adjectives should be mentioned in their root form (e.g., asks -> ask, largest -> large, ...). Result of 2.1 should be saved in "Problem2_1.csv" file containing the following columns:

Title	**POS Type**	**POS**	**Frequency**
Bill Gates	verb	ask	3
Bill Gates	adjective	large	8
Germany	verb	derive	1
...	...	...	

For problem 2.2, the output from each source text should be separately saved in an excel file with .csv format and use title as the filename (e.g., Bill_Gates.csv, Germany.csv, ...). Place all the csv files into a folder called "Problem2_2". Each csv file should contain the following columns:

Title	**Named-entity**	**Sentence**
Bill Gates	Melinda Gates	As of 2007, Bill and Melinda Gates were the second-most generous philanthropists in America.
Bill Gates	Melinda Gates	Bill and Melinda Gates have said that they intend to leave their three children $10 million each as their inheritance.
...	...	...

Problem 3 (Wikidata, SPARQL). This exercise is designed to gain familiarity with the highly prominent Wikidata knowledge base. Wikidata is queryable using the SPARQL (pronounced "sparkle") query language on the Wikidata endpoint service (`https://query.wikidata.org`). Check the examples of

using SPARQL on Wikidata here.[1]

Using SPARQL, extract the following information:

1. All “characters” in The Lord of the Rings (Q15228)
2. All “male characters” in The Lord of the Rings (Q15228)
3. All fictional universes (Q559618), sorted by the number of “fictional characters” (P1080)

The SPARQL queries should be saved in a text file inside “Problem3” folder. Each text file should contain the SPARQL query followed by the link to Wikidata query service of the answer query. For example:

Extracting all instances of house cats:

```
SELECT ?item ?itemLabel
WHERE
{
    ?item wdt:P31 wd:Q146.
    SERVICE wikibase:label {bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en".}
}
```

You can try out this example query at: `https://query.wikidata.org/#SELECT%20%3Fitem%20%3FitemLabel%0AWHERE%0A%7B%0A%3Fitem%20wdt%3AP31%20wd%3AQ146.%0ASERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22%5BAUTO%20LANGUAGE%5D%2Cen%22.%20%7D%0A%7D`

Please submit all necessary files, which are compressed into a zip file named:
Lab01_MatriculationNumber_Name.zip
to the email address: **akbc-assignments@mpi-inf.mpg.de** with title of the email: **[AKBC]Lab01_MatriculationNumber_Name**

Deadline: 23:59 02.05.2022 (Monday)

[1] https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples