max planck institut informatik

Problem 1 (Entity, Category, Entity Type). Download the dump file of the universe *Game of Thrones* from Wikia (here - https://s3.amazonaws.com/wikia_xml_dumps/g/ga/gameofthrones_pages_current.xml.7z). From the dump file, extract the following information:

1. List of all page titles in the dumpfile
2. List of all entity page titles (subset of (1))
3. List of all category page titles (subset of (1))
4. List of all categories from the dumpfile (Note: not all categories have their own pages) (optional)

All results should be saved in one excel file (.xls, .xlsx, or .csv) with 4 columns:

Page Titles	**Entities**	**Category Pages**	**Categories**
Eddard Stark	Jon_Snow	Wars	Characters
Arya Stark	House_Tyrell	Great houses	Kingsguard
Kingsguard	Gregor_Clegane	Kingsguard	Organizations
...	...	...	...

Problem 2 (Entity mentions). From the dump file in the previous problem, extract all entity mentions. The dump file is coded in wiki markup language. Each mention is linked to an entity using free links[1]. Based on this syntax, extract all entity mentions, along with their linked entities and corresponding relative frequency. For example:

```
Jon   Jon_Snow   1.0
Aerys Targaryen   Aerys_II_Targaryen   0.9   Aerys_Targaryen_(disambiguation)   0.1
```

In this example, `Jon` and `Aerys Targaryen` are mentions, `Jon_Snow`, `Aerys_II_Targaryen` and `Aerys_Targaryen_(disambiguation)` are entities, followed by frequency scores. Each field is separated by a tab.

The result should be saved in a text file, with above formatting.

Problem 3 (Wikidata, SPARQL). In this exercise, we get to know about Wikidata, a prominent existing knowledge base. To retrieve data from Wikidata, the Wikidata Query Service is useful, a SPARQL endpoint including a powerful Web-GUI. SPARQL (pronounced "sparkle") is an RDF query language, that is, a semantic query language for databases. You can find examples of using SPARQL on Wikidata here.[2]

Using SPARQL, extract the following information:

- All "characters" in The Lord of the Rings (Q15228)
- All "male characters" in The Lord of the Rings (Q15228)
- All fictional universes (Q559618), sorted by the number of "fictional characters" (P1080)

The SPARQL queries should be saved in a text file, followed by the link to Wikidata query service of these queries, for example:

Extracting all instance of house cats:

[1] https://en.wikipedia.org/wiki/Help:Wikitext#Free_links
[2] https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples


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

Link: https://query.wikidata.org/#SELECT%20%3Fitem%20%3FitemLabel%20%0AWHERE%20%0A%7B%0A%20%20%3Fitem%20wdt%3AP31%20wd%3AQ146.%0A%20%20SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22%5BAUTO_LANGUAGE%5D%2Cen%22.%20%7D%0A%7D

Please submit the code files (in Python) and output files, which are all compressed into a zip file named:
Lab01_MatriculationNumber_Name.zip
to this email address: **cxchu@mpi-inf.mpg.de** with title of the email: **[IE]Lab01_MatriculationNumber_Name**

Deadline: 23:59 19.10.2019 (Saturday)