max planck institut informatik

Problem 1 (Dump infobox scraping). In the Wiki-centric systems, infoboxes often store core information about entities. Figure 1 is the infobox of the page "Podrick Payne" from *Game of Thrones* in Wikia. Figure 2 shows the part which describes the infobox of the page "Podrick Payne" (selection text). Each infobox includes a list of attributes along with lists of values. For example, the infobox in figure 2 has attributes: *Title, Season, First seen, Last seen, etc.* Each attribute has value(s), like *Title:* `Podrick Payne`, *Status:* `Alive`.

Figure 1: Infobox of page "Podrick Payne" in Wikia.

Given the dump file of the universe *Game of Thrones* from Lab #01 (Download here):

1.1. Write a program that takes as input a page title, for example, *Podrick Payne*, and returns a list of attributes along with their values (suggested datastructure: map). For pages without infoboxes, return `null`. For simplicity, all attributes and values can be considered as strings. For attributes without values empty strings can be used.

1.2. Run the extraction on infoboxes of all entities in the dumpfile and print out the top 20 most frequent attributes. Save the result in the following format:
`attribute 1` [tab] `#occurrences`
`attribute 2` [tab] `#occurrences`
...

Problem 2 (Web scraping #1). Write a program that takes as input the URL of an (English) Wikipedia page, and outputs a list of all categories. For example, for `https://en.wikipedia.org/wiki/Saarland`, the output should be `[Saarland; NUTS 1 statistical regions of the European Union; States and territories established in 1957; 1957 establishments in West Germany]`.

Problem 3 (Web scraping #2). In HISPOS, there is information about courses of each semester. For example, this page[1] shows the information of the core course **Information Retrieval and Data Mining (IRDM)**.

Write a program that takes as input a `url` of a course website in HISPOS, and returns the basic information of this course. Basic information refers to everything included in the table as shown in Figure 3. Again, produce a key-value map as in Problem 1.1.

Your submitted files should include the Python code files and a text file containing the results of problem **1.2**.
Please submit all necessary files, which are compressed into a zip file named:
Lab03_MatriculationNumber_Name.zip
to the email address: **cxchu@mpi-inf.mpg.de** with title of the email: **[IE]Lab03_MatriculationNumber_Name**

Deadline: 23:59 02.11.2019 (Saturday)

[1] `https://www.lsf.uni-saarland.de/qisserver/rds?state=verpublish&status=init&vmfile=no&publishid=120264&moduleCall=webInfo&publishConfFile=webInfo&publishSubDir=veranstaltung`

```
    <title>Podrick Payne</title>
    <ns>0</ns>
    <id>4704</id>
    <revision>
      <id>382327</id>
      <parentid>382326</parentid>
      <timestamp>2018-11-28T02:47:58Z</timestamp>
      <contributor>
        <username>Shaneymike</username>
        <id>3035448</id>
      </contributor>
      <text xml:space="preserve" bytes="27577">{{Heraldry
|image = House-Payne-Main-Shield.PNG
|link = House Payne
}}
{{Character
| Title=Podrick Payne
| Image=GOT Season 7 15.jpg
| Season=[[Season 2|2]], [[Season 3|3]], [[Season 4|4]], [[Season 5|5]], [[Season 6|6]], [[Season 7|7]], [[Season 8|8]]
| First=&quot;[[The Night Lands]]&quot;
| Last=
| Mentioned=
| Appearances=30 episodes &lt;small&gt;([[#Appearances|see below]])&lt;/small&gt;
| Aka=Pod
| Status= [[:Category: Living individuals|Alive]]
| Death=
| Place=
| Allegiance=[[House Payne]] &lt;small&gt;(by birth)&lt;/small&gt;&lt;br&gt;[[Brienne of Tarth]]&lt;br&gt;[[House Stark]]&lt;br&
| Family=[[Ilyn Payne]] - distant cousin
| Actor=[[Daniel Portman]]}}
{{Quote|There has never lived a more loyal squire.|[[Tyrion Lannister]] to Podrick|Breaker of Chains}}
```


Figure 2: Infobox of the page “Podrick Payne” in the dump.

Basic Information

Type of Course	Lecture / Exercise/problem-solving class	**Long text**	
Number	120264	**Short text**	
Term	WiSe 2019/20	**Hours per week in term**	
Expected no. of participants		**Max. participants**	
Turnus		**Assignment**	no enrollment
Credits			
Additional Links	https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/teaching/ws1920/irdm19/		
Language	english		

Figure 3: Basic information of the course IRDM.