Project Title: Building a corpus of Arabic Web Tables

Project Description:

Some HTML tables on the Web contains data in tabular format or web tables which can be useful for various applications including data integration, data cleaning, and data transformation. Below are two examples, one taken from Wikipedia (https://tinyurl.com/ybhph3dy) and the other from a regular web site (https://tinyurl.com/ycxz7o4m).

قائمة عن صندوق النقد الدولي (2014)[1]

الرتبة	البلد/الإقليم	ن.م.إ (مليون بالدولار الأمريكي)
	العالم[5]	77٬269٬168
	الاتحاد الأوروبي[1 n][5]	18٬527٬116
1	الولايات المتحدة	17٬348٬075
2	الصين	10٬356٬508
3	اليابان	4٬602٬367
4	ألمانيا	3٬874٬437

عواصم الدول العربيّة في قارة أفريقيا

اسم الدولة	العاصمة
مصر	القاهرة
ليبيا	طرابلس
السودان	الخرطوم
الجزائر	الجزائر
المغرب	الرباط
موريتانيا	نواكشوط
تونس	تونس
الصومال	مقديشو
جيبوتي	جيبوتي
جزر القمر	موروني

The goal of the project is to build a corpus of Arabic Web Tables. We will use the recently released Arabic Web Crawl, called ArabicWeb16 and found at https://sites.google.com/view/arabicweb16 as well as Arabic Wikipedia (A recent dump is available here https://archive.org/details/arwiki-20170220). After the corpus is built, we will make it available as an open source resource.

A second goal of the project is to select few tables and manually annotate them for specific data cleaning tasks such as entity resolution.

Duties/Activities:

- Read the related work on Web Tables extraction
- Collect useful open source code that can be leveraged for the project
- Adapt existing code to the Arabic context or write new code as needed
- Run web table extraction on the Arabic Wikipedia
- Run web table extraction on ArabicWeb16
- Analyze the different tables we obtained and compute different stats on them
- Build a simple web site from where the web tables can be downloaded

References and resources:

- Web Data Commons - Web Table Corpora - http://webdatacommons.org/webtables/index.html
- WikiTables: Public Site - http://downey-n1.cs.northwestern.edu/public/
- Arabic Web Crawl ArabicWeb16 - https://sites.google.com/view/arabicweb16
- Some related papers
    - **http://www.vldb2010.org/proceedings/files/papers/R118.pdf**
    - **http://www.dbai.tuwien.ac.at/staff/gatter/work/WWW_2007_Domain_Independent_Information_Extraction.pdf**
    - **http://sirrice.github.io/files/papers/webtables-vldb08.pdf**
    - **http://www.vldb.org/pvldb/2/vldb09-325.pdf**
    - **http://webdatacommons.org/webtables/goldstandard.html**

Required Skills:

- Good programming skills
- Java or python
- HTML/CSS

Preferred Intern Academic Level:

Junior or Senior undergrad or if you have good programming skills

Learning Opportunities:

- Learn new programming skills
- Develop a new resource that will be useful for many researchers
- Getting familiar with research related to web table extraction and data integration in general

Expected Team Size: 2~3

Mentor:

Name: Mourad Ouzzani email: mouzzani@hbku.edu.qa