Project Title: Building a corpus of Arabic Web Tables

Project Description:
Many HTML tables on the Web contain data in tabular format (aka web tables) which can be useful for various applications including data integration, data cleaning, and data transformation. Here are two examples, one taken from Wikipedia (https://tinyurl.com/ybhph3dy) and the other from a regular website (https://tinyurl.com/ycxz7o4m).

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

Required Skills: Good analytical skills. Good command of a programming language (C++, Java, or python).

Learning Opportunities: Developing research and coding skills in data analytics and data curation. Intensive coding.

Expected Team Size: 2

Mentor: Mourad Ouzzani**, mouzzani@hbku.edu.qa**

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