Farasa python package project

Project Description:

Python is one of the major programming languages and is becoming the de facto language for Machine Learning. NLTK package is one of the core packages for NLP with support for many languages; Its support for Arabic is limited. Farasa has many modules that provides the state-of-the-art processing for Arabic and its dialects.
Farasa is a suite of tools for Arabic Natural Language Processing. The suite is composed of several modules such as: Segmenter, Part of Speech (POS) Tagger, Diacritizer, Spell Checker and Named Entity Recognizer. These tools are available as API. The aim of this project is to build a wrapper around Farasa APIs that would allow these tools to be accessible directly from a python code. This will facilitates the integration of Farasa in other NLP tasks. This will greatly help researchers and developers to master Arabic processing with ease.

Duties/Activities:

- Register a new project/package.
- Develop interfaces for the API access points
- Package and publish code for PyPI
- Manage versions of the package

Required Skills:

- Python
- Code management/version control
- Understanding Arabic is a plus

Preferred Intern Academic Level:

- Junior/Senior levels

Learning Opportunities:

- This internship will help you to learn how to develop code and deploy it. Also It will help the intern to get familiar with code management and versioning systems.

Expected Team Size: *it is preferable to have team projects*

- 1 to 2

Mentors

Name: Ahmed Abdelali email: aabdelali@hbku.edu.qa
Name: Hamdy Mubarak email: hmubarak@hbku.edu.qa
Name: Kareem Darwish email: kdarwish@hbku.edu.qa
Name: Sabit Hassan email: sahassan2@hbku.edu.qa

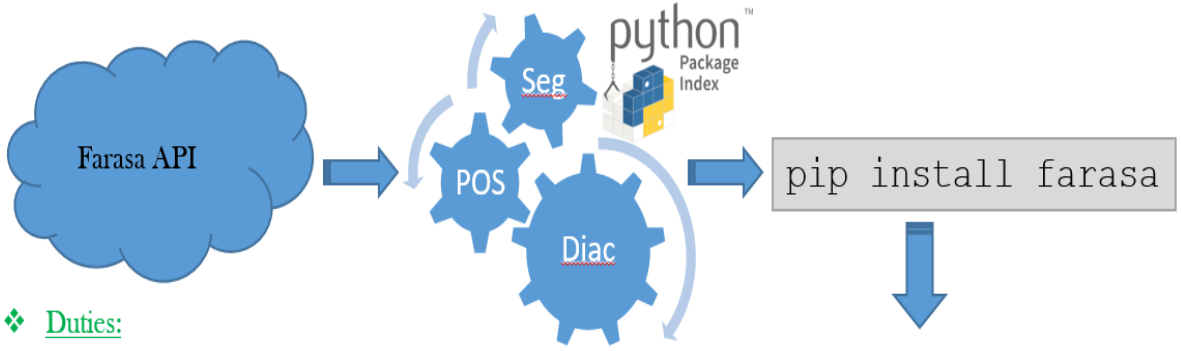
Farasa Python Package Project

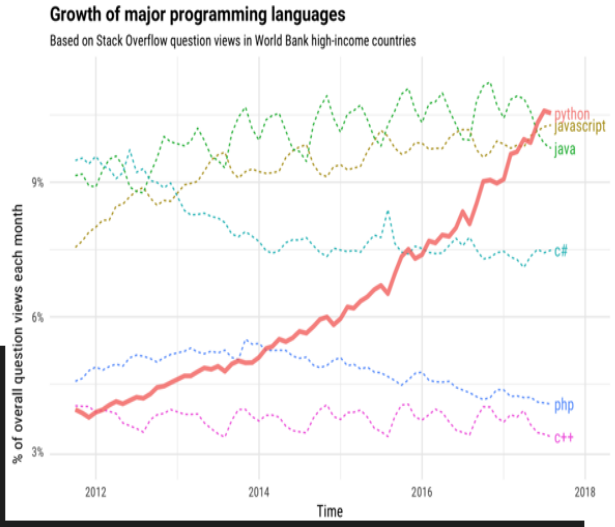
Mentors: Ahmed Abdelali, Hamdy Mubarak, Kareem Darwish, Sabit Hassan {aabdelali,hmubarak,kdarwish,sahassan2}@hbku.edu.qa

Motivation: Farasa is a suite of tools for Arabic Natural Language Processing. The suite is composed of several modules such as: Segmenter, Part of Speech (POS) Tagger, Diacritizer, Spell Checker and Named Entity Recognizer. These tools are available as API.

Python is one of the major programming languages and is becoming the de facto language for Machine Learning. NLTK package is one of the core packages for NLP with support for many languages; its support for Arabic is limited. Farasa has many modules that provides the state of the art processing for Arabic and its dialects.

Objective: The project aims to build wrapper around Farasa APIs that would allow these tools to be accessible directly from a Python code, and facilitates integration in other NLP tasks. This will greatly help researchers and developers to master Arabic processing with ease.





❖ Duties:

- Register a new project/package.
- Develop interfaces for the API access points
- Package and publish code for PyPI
- Manage versions of the package

❖ Required Skills

✓ Python
✓ Code management/version control
✓ Understanding Arabic is a plus

```
1  #!/bin/env python
2  import farasa
3
4  text = 'هذا نص باللغة العربية'
5  seg_text = farasa.segment(text) ==> هذا نص ب+ال+لغ+ة ال+عربي+ة
6  pos_text = farasa.pos(text) ==> هذا/PRON نص/NOUN ب+/PREP ة+ لغ+ ال/DET+NOUN+NSUFF+FS
7  diac_text = farasa.diacitize(text) ==> هَذا نَصٌّ بِاللُّغَةِ العَرَبِيَّةِ
```