INF 551 – Fall 2018

Homework #1: Firebase and JSON

Due: September 16, Sunday (end of day)
100 points

In this homework, we provide you with an JSON data set on nobel prizes: "prize.json". The data set contains 585 entries. Each entry describes the detail of an award such as: category, laureates, and award year. For each laureate, it lists his/her first and last name, motivation (contribution & acknowledgement), etc.

1. [60 points] Write a Python script (with REST requests embedded) called "load.py". The script will do two things:
    - Load the dataset into Firebase. You may need Python "requests" package as shown in class.
    - Create an inverted index for the motivation content of laurates. The index has an entry for each unique (non-stop) word in the content and the value of entry is a list of corresponding ids (value of id attribute). Content is tokenized by white spaces. You should discard all stopwords listed here (https://www.ranks.nl/stopwords ,choose: Default English stopwords list), e.g., this, that, a, an, etc.

      For example, consider a motivation of a prize (whose id is 941): "for decisive contributions to the LIGO detector and the observation of gravitational waves". Unique words include: decisive, contributions, etc. You need to lowercase all tokens. So "LIGO" will be stored as "ligo".

      An example index:

      ```
      {"index": {
        "ligo": [941, 942, ...]
        ...
      }
      ```

      which says "ligo" appears in price with ids: 941 ,942, …

    Execution format:

    - python load.py prize.json

- [40 points] Write a Python script called "search.py". The script takes a list of keywords and return a list of ids of prizes whose **motivation attribute** contains some keywords in the list. The search needs to be executed using the data stored in your Firebase database. Note that the search is NOT case-sensitive. For example,
    - python search.py "ligo waves"

should return the ids of prizes whose motivation contain at least one keyword in the input list.

Submissions: Name your 2 scripts as below and submit to Blackboard by the due time. **DO NOT** place them in a folder or zip file.

- <FirstName>_<LastName>_load.py
- <FirstName>_<LastName>_search.py

Note: Please use Python 2.7 (installed by default on EC2) for the coursework.