In the name of Allah

دانشگاه شیراز

Text Mining Instructor: Dr. Fakhrahmad

Assignment_3: Preprocessing and Language Modeling Due Date: **1402/04/15**

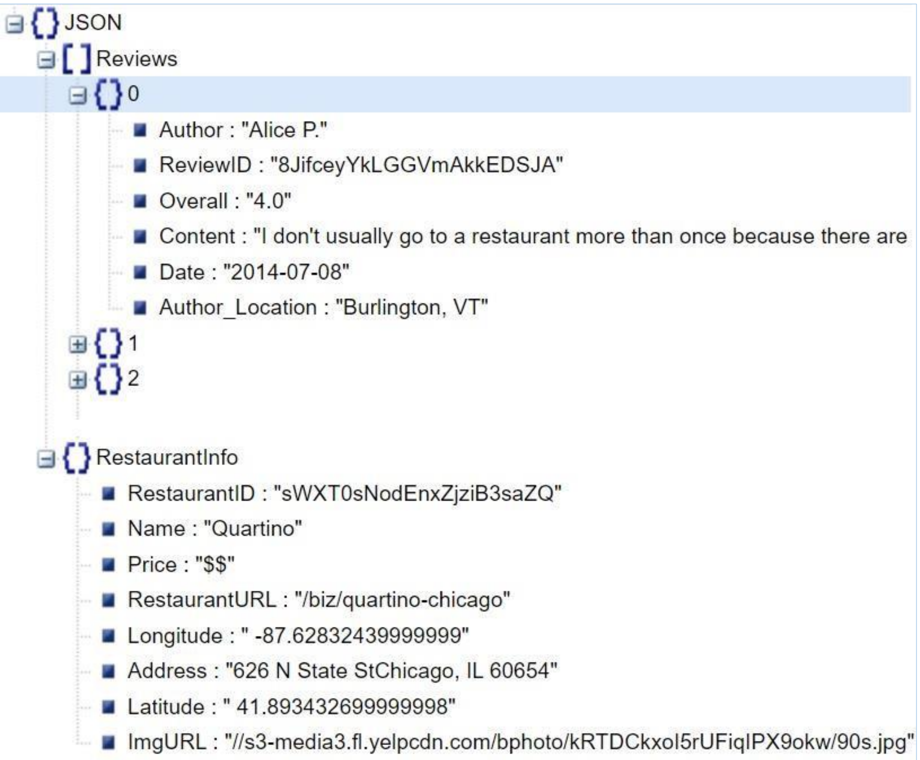
Preprocessing is an important task and critical step in Text mining, Natural Language Processing (NLP) and information retrieval (IR). In the area of Text Mining, data preprocessing is used for extracting interesting and non-trivial knowledge from unstructured text data. As the first practice, we investigate some of the key steps of preprocessing namely Tokenization, Normalization, Stemming and Stopword Removal.

The next practice is devoted to Language Modeling (LM) which is one of the most important parts of modern Natural Language Processing (NLP). There are many sorts of applications for Language Modeling like: Machine Translation, Spell Correction, Speech Recognition, Question Answering, Sentiment analysis etc. Each of those tasks require use of language model. Language model is employed to represent the text to a form understandable from the machine point of view.

The dataset of this assignment is about business reviews which contains 38688 documents. You can download this data set using the following link:

- http://www.mediafire.com/file/5silqtc9n78kfhn/Dataset.zip/file

Each document is represented as a JSON file. The file is included an array of reviews and an object about the business. To view JSON files, you can use available tools like the following one: - http://jsonviewer.stack.hu/



```
{} JSON
  [] Reviews
    {} 0
      Author : "Alice P."
      ReviewID : "8JifceyYkLGGVmAkkEDSJA"
      Overall : "4.0"
      Content : "I don't usually go to a restaurant more than once because there are
      Date : "2014-07-08"
      Author_Location : "Burlington, VT"
    {} 1
    {} 2
  {} RestaurantInfo
    RestaurantID : "sWXT0sNodEnxZjziB3saZQ"
    Name : "Quartino"
    Price : "$$"
    RestaurantURL : "/biz/quartino-chicago"
    Longitude : " -87.62832439999999"
    Address : "626 N State StChicago, IL 60654"
    Latitude : " 41.893432699999998"
    ImgURL : "//s3-media3.fl.yelpcdn.com/bphoto/kRTDCkxol5rUFiqIPX9okw/90s.jpg"
```

* Consider that some JSON file may not follow the above definition. Therefore, please handle the exceptions when you are parsing the files.

Task 1: Preprocessing

The preprocessing phase applies several steps to review documents in order to clean the data and extract the desired tokens. The following preprocess should be performed:

- **Tokenization**: Each of the review documents are divided into pieces, called tokens.
- **Normalization**: In the Normalization step, you will remove punctuation marks, convert all the tokens into lowercase, and then recognize the integers and decimals and replace them with the token "NUM".
  - **Stemming**: This step is done to extract the base form of the tokens by removing affixes from them.
- **Stopword Removal**: It is simply removing the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words. These words have no significance or very little impact in NLP tasks.

Task 2: Language modeling

In this section, you should estimate Unigram and Bigram language models using the Maximum Likelihood Estimation (MLE) approach. These models are built based on the review documents from the dataset. When estimating the Bigram language model, employ the Linear Interpolation Smoothing method. This approach uses a weighted combination of the bigram and unigram probabilities. Note that in the Linear Interpolation Smoothing method, the parameter λ plays a key role (use λ=0.9). The Expectation Maximization (EM) algorithm can be used to find the best value for this parameter.

From the resulting Bigram language model, find the top 10 ranked words that are most likely to follow the word "decent".

Note that you are asked to

- Prepare a report in PDF format about your program and the strategy you chose for each task.
- Represent the result of the question from Task 2.
- Submit your answer before the deadline (1402/04/15).
- Zip all files and submit as TM-HW3-(Your Student Number).zip

Hint

If you have any question about the assignment, please email to **dianati.shiraz@gmail.com.**