İstanbul Bilgi University

Department of Computer Engineering

Spring 2018-2019

CMPE 346 : Natural Language Processing

Term Project - Phase I

11^{th} March, 2019 23:59

1. The term project is composed of several phases.
2. First phase is about collection and analysis of a textual data.
3. You might work as group of two people or as individual.
4. You need to decide about the group members and inform via e-mail to tahsin.yazkan@bilgiedu.net.
5. For each phase of the term project, you need to make a submission before the deadline.
6. After the submission of the regarding step, you need to be present during the determined lab section and make a demonstration to explain your work and to show your progress.
7. The choices of the programming language and the development environment are up to you. However, it is highly encouraged to use Pyhton and NLTK package.
8. NLTK is quite popular and widely used open source package for text analysis and natural language processing purposes.
9. You can refer to the sample programs that we have done during the lab sessions.
10. The official website of the NLTK[1] contains practical tutorials, samples and documentation which might be really beneficial for the project.
11. Please check announcements and file uploads regularly for the upcoming steps of the project.

[1] http://www.nltk.org/

CONTINUED

Steps of Phase I

1. For the first step of this phase, you are expected to collect data or to find a appropriate dataset which can be about movies, hotels, restaurants etc. The data will be having the reviews themselves along with complementary information, such as movie name, movie year, country of the movie, score of the reviews, the class/label of the reviews(e.g. very positive, positive,neutral, negative) etc. The data could be in tabular form and could be stored in a file format of your decision(e.g. csv, excel etc.). For this phase, it will only be about the reviews themselves.

   Example[2]:

   "Magnolia is Paul Thomas Anderson's first big movie that is as wild as it is weird. It is a quick cut, but long and slow narrative between around 10 major characters' lives. I wish it were shorter and more fast paced, but alas Anderson fails to cut down his films to a more manageable size. However, I thoroughly enjoyed Magnolia. Its unique shots, story twists, and excellent writing keep it in check. Beautiful music and heartfelt writing collide as the intertwined lives of these various figures in society mesh for the most original film I have seen in a long time. There is no other film quite like Magnolia. It is like the intense crescendo of harrowing events like Requiem for a Dream with the scattered perspective narrative of Pulp Fiction. It even has moments of the surreal comedy like the Coen Brothers' The Big Lebowski or Fargo."

2. First, tokenize the text into words

   Example:

   ['Magnolia', 'is', 'Paul', 'Thomas', 'Anderson', "'s", 'first', 'big', 'movie', 'that', 'is', 'as', 'wild', 'as', 'it', 'is', 'weird', '.', 'It', 'is', 'a', 'quick', 'cut', ',', 'but', 'long', 'and', 'slow', 'narrative', 'between', 'around', '10', 'major', 'characters', "'", 'lives', '.', 'I', 'wish', 'it', 'were', 'shorter', 'and', 'more', 'fast', 'paced', ',', 'but', 'alas', 'Anderson', 'fails', 'to', 'cut', 'down', 'his', 'films', 'to', 'a', 'more', 'manageable', 'size', '.', 'However', ',', 'I', 'thoroughly', 'enjoyed', 'Magnolia', '.', 'Its', 'unique', 'shots', ',', 'story', 'twists', ',', 'and', 'excellent', 'writing', 'keep', 'it', 'in', 'check', '.', 'Beautiful', 'music', 'and', 'heartfelt', 'writing', 'collide', 'as', 'the', 'intertwined', 'lives', 'of', 'these', 'various', 'figures', 'in', 'society', 'mesh', 'for', 'the', 'most', 'original', 'film', 'I', 'have', 'seen', 'in', 'a', 'long', 'time', '.', 'There', 'is', 'no', 'other', 'film', 'quite', 'like', 'Magnolia', '.', 'It', 'is', 'like', 'the', 'intense', 'crescendo', 'of', 'harrowing', 'events', 'like', 'Requiem', 'for', 'a', 'Dream', 'with', 'the', 'scattered', 'perspective', 'narrative', 'of', 'Pulp', 'Fiction', '.', 'It', 'even', 'has', 'moments', 'of', 'the', 'surreal', 'comedy', 'like', 'the', 'Coen', 'Brothers', "'", 'The', 'Big', 'Lebowski', 'or', 'Fargo', '.']

3. Secondly, take out the stop words of English language from the tokenized text.

   ['Magnolia', 'Paul', 'Thomas', 'Anderson', "'s", 'first', 'big', 'movie', 'wild', 'weird', '.', 'It', 'quick', 'cut', ',', 'long', 'slow', 'narrative', 'around', '10', 'major', 'characters', "'", 'lives', '.', 'I', 'wish', 'shorter', 'fast', 'paced', ',', 'alas', 'Anderson', 'fails', 'cut', 'films', 'manageable', 'size', '.', 'However', ',', 'I', 'thoroughly', 'enjoyed', 'Magnolia', '.', 'Its', 'unique', 'shots', ',', 'story', 'twists', ',', 'excellent', 'writing', 'keep', 'check', '.', 'Beautiful', 'music', 'heartfelt', 'writing', 'collide', 'intertwined', 'lives', 'various', 'figures', 'society', 'mesh', 'original', 'film', 'I', 'seen', 'long', 'time', '.', 'There', 'film', 'quite', 'like', 'Magnolia', '.', 'It', 'like', 'intense', 'crescendo', 'harrowing', 'events', 'like', 'Requiem', 'Dream', 'scattered', 'perspective', 'narrative', 'Pulp', 'Fiction', '.', 'It', 'even', 'moments', 'surreal', 'comedy', 'like', 'Coen', 'Brothers', "'", 'The', 'Big', 'Lebowski', 'Fargo', '.']

[2] https://www.rottentomatoes.com/m/magnolia/reviews/?type=user

4. After that, it is needed to apply stemming techniques to find the words' roots. For this step, you are expected to use at least two different stemmming algorithms (e.g. Porter, Lancaster or Snowball etc.) and display the resulting lists of stems. Furthermore, in case of spotting differences between these applied algorithms' results, you need to elaborate and point out the differences of the approaches.

   Example: By using Porter Stemmer from nltk package we might have the following result

   ['magnolia', 'paul', 'thoma', 'anderson', "'s", 'first', 'big', 'movi', 'wild', 'weird', '.', 'It', 'quick', 'cut', ',', 'long', 'slow', 'narr', 'around', '10', 'major', 'charact', "'", 'live', '.', 'I', 'wish', 'shorter', 'fast', 'pace', ',', 'ala', 'anderson', 'fail', 'cut', 'film', 'manag', 'size', '.', 'howev', ',', 'I', 'thoroughli', 'enjoy', 'magnolia', '.', 'it', 'uniqu', 'shot', ',', 'stori', 'twist', ',', 'excel', 'write', 'keep', 'check', '.', 'beauti', 'music', 'heartfelt', 'write', 'collid', 'intertwin', 'live', 'variou', 'figur', 'societi', 'mesh', 'origin', 'film', 'I', 'seen', 'long', 'time', '.', 'there', 'film', 'quit', 'like', 'magnolia', '.', 'It', 'like', 'intens', 'crescendo', 'harrow', 'event', 'like', 'requiem', 'dream', 'scatter', 'perspect', 'narr', 'pulp', 'fiction', '.', 'It', 'even', 'moment', 'surreal', 'comedi', 'like', 'coen', 'brother', "'", 'the', 'big', 'lebowski', 'fargo', '.']

5. Display the frequency distribution information of the stemmed text.

   Example:

   FreqDist with 80 samples and 111 outcomes

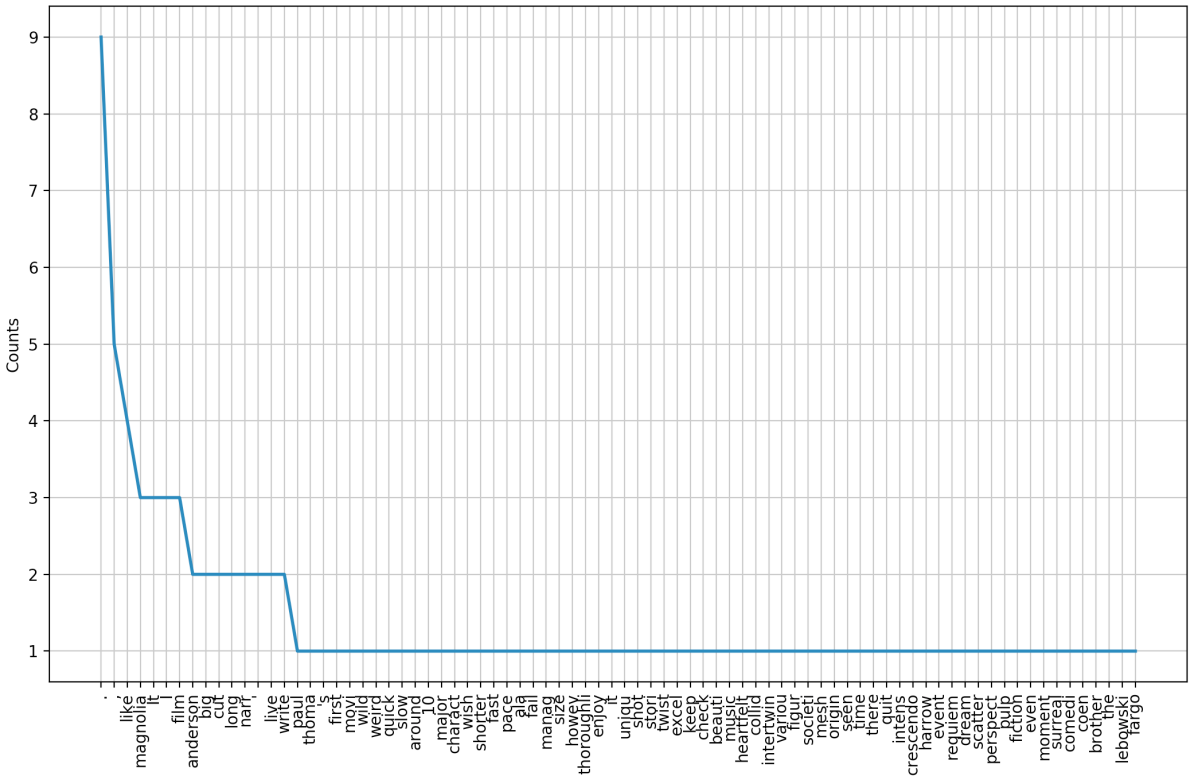
6. Display the most frequent 10 stems.

   Example:

   [('.', 9), (',', 5), ('like', 4), ('magnolia', 3), ('It', 3), ('I', 3), ('film', 3),('anderson', 2), ('big', 2), ('cut', 2)]

7. Visualize the frequency distribution using graphical plots.

   Example:



CONTINUED

8. List all the words from the text which have more than 10 letters.

   Example:

   [”characters””, ’intertwined’, ’perspective’]