İstanbul Bilgi University

Department of Computer Engineering

Spring 2018-2019

CMPE 346 : Natural Language Processing

Term Project - Phase IV

13th May, 2019 23:59

1. The term project is composed of several phases.
2. Second phase is to investigate the language model of the collected data.
3. For each phase of the term project, you need to make a submission before the deadline.
4. 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.
5. 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.
6. NLTK is quite popular and widely used open source package for text analysis and natural language processing purposes.
7. You can refer to the sample programs that we have done during the lab sessions.
8. The official website of the NLTK[1] contains practical tutorials, samples and documentation which might be really beneficial for the project.
9. Please check announcements and file uploads regularly for the upcoming steps of the project.

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

CONTINUED

Steps of Phase IV

For this phase of the project, you are required to apply some well known machine learning algorithms to your labeled (each review or commentary has designated class or rating) dataset.

1. In order to feed the applied machine learning algorithms with our textual data as input, it is needed to change the form of representation of the text data. You are expected to use the ***Bag-of-Words*** model to encode text data to a list of vectors of features.
2. After having the text data with the proper form of representation, you are expected to apply ***Linear Regression, Naive Bayes, SVM*** and ***kNN*** algorithms.
3. For each applied algorithm, you are expected to list ***Accuracy, Precision, Recall*** and ***F1*** score points.
4. As the last step, you are required to create a function which takes two parameters that are randomly selected reviews from your dataset and compares the similarity of the sentences semantically. You are expected to use WordNet to compute the similarity of the sentences, since it is well-known and extensively used in the community.

Important Note: You may refer to the following resources that explain the above mentioned techniques for practical purposes.

https://scikit-learn.org/stable/modules/feature_extraction.html

https://medium.freecodecamp.org/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-2

http://zacstewart.com/2015/04/28/document-classification-with-scikit-learn.html

https://scikit-learn.org/stable/modules/naive_bayes.html

https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

https://scikit-learn.org/stable/modules/svm.html

https://scikit-learn.org/stable/modules/neighbors.html

http://www.nltk.org/howto/wordnet.html