Information Retrieval (CS317) Fall 2019

Assignment 3

Due: **22 Nov** 2019

Question 1 (WordToVec)

You will need genism library for this question. You can install gensim using following steps.

1. Go to pip folder in Python installation path (Find python installation path by running "where python" in command prompt (cmd) in windows)
   For example: cd C:\Program Files\Python\Python36-32\Scripts
2. Run "pip install gensim"

You can get help on using gensim for this homework from following link

https://radimrehurek.com/gensim/

Download reviews dataset (File named Question1) and train wordtoVec using genism on this dataset. You will have to read input file and then preprocess (tokenization, etc) data using nltk (as done in HW1) or genism. Train wordtoVec on this dataset using following command

genism.model.Word2Vec(preProcessedData, size, window, min_count, workers)

This command will return a trained model which should be saved in some variable for later use.

Parameters of this model are as follows:

a. PreProcessedData: This should be list of preprocessed(tokenized) sentences (list of lists)
b. Size = number of dimensions of word vectors
c. Window = number of context words
d. Min_count = words that have frequency lower than min_count are ignored while training
e. Workers = number of threads

Use most_similar function (trainedModel.wv.most_similar()) of trained model to find most similar words to following words.

a. Clean
b. Unclean
c. Amazed
d. friendly

Question 2 (Sentiment Analysis)

The purpose of this assignment is to build a sentiment classifier using the Naive Bayes classification techniques and a bag of words model.

You can use scikit-learn (machine larning tool for python) for using implementations of classification algorithms.

Install scikit using "pip install scikit-learn"

Perform sentiment analysis on attached dataset for question 2. The data set contains movie reviews with labels 0 (negative) and 1 (positive). Split the data into train and test set by using "train_test_split(DataSet)" of scikit. By default it will split into 75% training and 25% test data.

Implement following 2 models for sentiment analysis and report accuracies of both models.

1. Bag of words based on raw counts
2. Bag of words based on TfIDF

Read following links about using Vectorizer (Bag of words based on raw counts) and transformer (Bag of words based on TfIDF) for converting list of sentences to vectors

1. **https://scikit-learn.org/stable/modules/feature_extraction.html**
2. **https://scikit-learn.org/stable/auto_examples/text/plot_document_classification_20newsgroups.html#sphx-glr-auto-examples-text-plot-document-classification-20newsgroups-py**

You can use scikit learn for implementation of Naïve Bayes as explained in above links.

The data is in CSV format so you can read it in dataframe using pandas library. Use BeautifulSoup library for parsing html.

Submission

Submit your zipped code file on slate.