Homework Assignment-3

Start Assignment

Due Tuesday by 11:59pm **Points** 45 **Submitting** a text entry box or a file upload
Available Nov 2 at 5:30pm - Nov 16 at 11:59pm 14 days

Assignment 3

Part 1 - 9

Import Semeval dataset (https://www.kaggle.com/azzouza2018/semevaldatadets) using Python.

Do the following for the Semeval dataset mentioned above using python:

- (1) Remove all URLs from the sentences

- (1) Remove sentences which have less than 4 tokens

- (1) Remove the tokens which starts with @

- (1) Remove all hashtags

- (2) Remove occurrences of characters that appear more than 3 times consecutively, keep max occurrence as 3. For example: Heeeeelooooo => Heeelooo, gooooo => gooo

- (3) Identify the slang words using a slang dictionary (e.g.: https://github.com/rishabhverma17/sms_slang_translator/blob/master/slang.txt) and output as follows in a CSV file:

    original tweet, slang word -> replacement word, ...

Part 2 - 8

Now for each text, apply data augmentation to generate variations (max 5 variations per text). Use the

following approach to generate variations:

- Replace nouns and verbs with their synonyms. Use wordnet for getting synonyms

(https://www.nltk.org/howto/wordnet.html).

Output the original text and augmented text in a csv file using the following format:

original text, augmentation1, augmentation2, augmentation3, augmentation4, augmentation5

Fill the empty cell with "None"

Part 3 - 6

- (3) Now print the size of your augmented dataset, ratio of original/augmented set and label distribution

of the augmented set.

- (3) Generate n-gram (unigram and bigram, trigram) for the augmented dataset. Print them to console

Part 4 - 8

(4) Now compute rank/frequency profile of words of the original corpus and augmented corpus.

To get the rank/frequency profile, you take the type from the frequency list and replace it with its rank, where the most frequent type is given rank 1, and so forth.

Print the rank/frequency profile of the words in csv files for both corpus as follows:

rank1, freq 1

rank 2, freq 2

.....

.....

(4) Output the percentage of corpus size made up by the top-10 words for both original and augmented corpus.

A Sample frequency list and associated rank list

Frequency list

Type freq

Apple 21

car 25

Sun 14

Cat 3

Rank list

Type rank

Car 1

Apple 2

Sun 3

Cat 4

The resulting rank/frequency profile

rank freq

2 21

1 25

3 14

4 3

Part 5 - 8

Find top-10 bi-grams and tri-grams from positive sentiment samples and top-10 bi-grams and tri-grams from negative sentiment samples using NLTK's significant collocation approach. Investigate the difference of these n-grams when you try:

- (4) PMI to select significant collocations (Review materials from lecture 6)
- (4) Maximum likelihood to select significant collocations (use likelihood ratio. Review materials from lecture 6)

Part 6 - 6

Compute word vectors from this data-set using the following approaches:

1) (2) Frequency based

2) (2) One-hot

3) (2) TF-IDF

Print the word vectors