Natural Language Processing

Assignment - 1

Question - 1 :

pre_process: In this method, a string is taken as an input. Punctuations are separated from the words by a space by using regular expressions. Then the string is stripped of spaces at the ends by using *strip()*. Then the string is split into tokens using the *split()* method where the separator is a space. This list of tokens is returned at the end of the function.

Question - 2:

*to_feature_vector_dictionary: T*his method, takes two lists a the input, a character document consisting of lines of the character and extra features that contains the context of the lines of the characters. The method assigns Parts-of-speech tags to the words and counts the combination(a tuple of a word and its POS) and returns a dictionary.

Question - 3:

Character in Train data	**Character closest ranked in heldout data**
CHRISTIAN	MAX
CLARE	MAX
HEATHER	MAX
IAN	CHRISTIAN
JACK	MAX
JANE	MAX
MAX	STACEY
MINTY	MAX
PHIL	MAX
RONNIE	STACEY, TANYA
ROXY	MAX
SEAN	MAX
SHIRLEY	MAX
STACEY	MAX
TANYA	STACEY

Character in Train data	Character farthest ranked in heldout data
CHRISTIAN	MINTY, CLARE
CLARE	IAN
HEATHER	CHRISTIAN
IAN	ROXY
JACK	MINTY
JANE	MINTY
MAX	IAN
MINTY	IAN
PHIL	CLARE
RONNIE	IAN
ROXY	IAN
SEAN	IAN
SHIRLEY	IAN
STACEY	IAN
TANYA	IAN

From the feature vectors we can observe that the characters whose feature vectors are the closest use similar words frequently and vice versa.

From the data frame *df* it can be observed that the appearance of characters in the same scene does not necessarily mean their feature vectors are similar.

Question - 4:

In context of the line, the lines that are in the same scene and has a cosine similarity of greater than or equal to 0.8 are included. The pre-processed context is passed as extra features into *to_feature_vector_dictionary.*

Question - 5:

I have improved the vectorisation by converting the vector matrix into Tf-iDF form using *TfidfTransformer*. I have changed the attribute *sublinear_tf* in *TfidfTransformer* to bool but it has yielded worse results.

Question - 6:

I have made the necessary changes to the code that performs on the test_data to accommodate the changes made in the previous questions.