DS 504
Natural Language Processing
Assignment 1
Deadline - 4.9.2022 23.59 Hrs

Question 1:

Write codes for spelling correction using the **bi-gram** Probabilistic model for - i) non-word error, and ii) real-word error. Real word errors are misspelled words that can be found in the dictionary and non-word errors are misspelled words that cannot be found in the dictionary. Use the text, Emma, by Jane Austen (can be accessed from Gutenberg corpus in nltk) as corpus.

Use 20% of the data as a test set (do not use it to train the model) to check the performance. For every statement in the test set, randomly choose a word and introduce a non-word/real word error in it randomly within 2 edit distance. Report your accuracy for both the cases. Submit the test set generation and evaluation codes too along with the model codes.

Some example text to check the code for non-word error -

a) `They had determined that thirr marriage ought to be concluded.`
b) `He began to thenk.`
c) `I think there is a litlle likeness between us.`
d) `Her fathar fondly replied.`
e) `He kaw his son every year.`

Some example text to check the code for real-word error -

a) `They had determined that there marriage ought to be concluded.`
b) `He began to pink.`
c) `I think there is a brittle likeness between us.`
d) `Her gather fondly replied.`
e) `He raw his son every year.`

Question 2:

Given a sequence of characters with no spaces in between, recover the correct sequence of words. For example, given a text “itdoesnotmattertotheboard”, the extracted sequence should be: - “It”, “does”, “not”, “matter”, “to”, “the”, “board”. Please use the brown corpus for modeling the n-gram model. You can use the **tri-gram** model for this problem.

Hint : Using the tri-gram model, enumerate all the possible sequences of words and choose the one with the highest Probability.

Use 20% of the data as a test set (do not use it to train the model) to check the performance. For every statement in the test set, combine it and check if your model can split it correctly. Report your accuracy. Submit the test set generation & evaluation codes too along with the model codes.

Some example strings to test your system: -

a) “Myunscientificfrienddoesnotbelievethathumanstatureismeasurablein termsofspeed”
b) “Myhostwentoverandstaredoutthewindowathispeacocks”
c) “wheninthecourseofhumaneventsitbecomesnecessaryforonepeople”

Question 3:
Calculate the perplexities of a bi-gram model, tri-gram model and quad-gram model for the corpus ‘webtext’. The webtext corpus includes content from a Firefox discussion forum, conversations overheard in New York, Pirates of the Caribbean movie scripts, personal advertisements, and wine reviews.

Divide the corpus into two parts - 70% for training the language models and 30% for testing and calculating perplexity. Use the same split for all three language models.

Submit the evaluation code too along with the model codes.