Mid Semester Examination

Assignment

Course Name: Natural Language Processing **Code: CS 563**

Marks: 9+6 **Duration:** 24 hours

Make reasonable assumptions as and whenever necessary. Carefully read the instructions circulated in the group on February 22nd, 2022.

(Q1). Correcting typographical errors for words where error has been induced by character replacement, can be modelled as a sequence labelling problem. Given a sequence of typed characters (which might contain mistakes), the problem is to predict the actual intended characters in the sequence.

Consider the passage **P1** containing typographical errors and its corrected version **P2**:

> **P1:** *star wars is **ploying** at **thi** regal lloyd **center** and imax multnomah st portland **ang** also at **tho centupy** eastport plaza **wuuld** any of **thoss** times **wurk** for **yoz***
>
> **P2:** *star wars is **playing** at **the** regal lloyd **centre** and imax multnomah st portland **and** also at **the century** eastport plaza **would** any of **those** times **work** for **you***

i. In total there are 10 spelling mistakes in the passage P1 (highlighted).

ii. You have to build an automated spelling mistake corrector using Hidden Markov Model (HMM). Use the *bigram* model.

iii. Once you have trained an HMM model, test it on the passage P1, and report what percentage of spelling mistakes from the passage your model was able to correct (document this result in the report).

iv. Use the dataset provided in the following link to train your model: https://drive.google.com/file/d/12OVHD2ulmg3m6KE659qDWy8hfC05k5dx/view?usp=sharing

v. The dataset consists of two columns. The first column is a sequence of input characters that sometimes contain typographical errors. The second column consists of the correct or intended sequence of characters. The word ending is demarcated by an underscore '_' sign. A small sample of the dataset is given below.

t	t
e	o

m	m
o	o
r	r
r	r
o	o
w	w
_	_
a	a
y	t
_	_

Documents to submit:

i. Codes with appropriate documentation

ii. Report accuracy; precision, recall, and F1-score on the test set provided. Submit the test set predictions.

iii. In your documentation, highlight the cases for which the HMM model easily corrects, and the cases where the HMM model fails.

(Q2). Consider the PoS tagging assignment given in the class. Assume a trigram version of HMM model, with additional context dependency (i.e. while calculating the emission probability, it should also take the previous word).

(Q2a). Determine the PoS tagging sequence for the following sentence using the trained HMM model:

That former Sri Lanka skipper and ace batsman Aravinda De Silva is a man of few words was very much evident on Wednesday when the legendary batsman , who has always let his bat talk , struggled to answer a barrage of questions at a function to_F promote.

(Q2b). Modify the Viterbi decoding algorithm, and consider the most probable three paths rather than only one (at every time point). Determine the best state sequence for the above example .

(Q2c). Explain with proper intuition the reasons behind the same or different sequences obtained (*in terms of path probability*).