**NLP HW1**

**Part 1: POS**

Ques 1:

1. Where is the queen of the Kingdom?
2. Surviving in Hawaii is extremely very enjoyable
3. She liked eating on the couch.
4. Shall we visit a trending restaurant?

Ques 2:

1. [(S (NP (NNP ‘Buffalo’) (NN ‘buffalo’)) (VP (VBZ ‘buffalo’) (NP (DT ‘the’) (NN ‘sheep’))))]
2. [(S (NP (DT ‘The’) (NN ‘metal’) (NNS ‘stands’)) (VP (VBP ‘stand’) (PP (IN ‘near’) (NP (DT ‘the’) (NN ‘door’ )))))]

Ques 3:

1. I had to watch the broth boiling in between my mission to find my missing watch.
2. The soaring high temperatures are drying the freshwater bodies. At this rate, the local water bodies will dry up, leaving us with the driest of lands.
3. Light is the path that lights our passions, making us crawl out of the dark void to the lightest room in the vicinity.
4. Everyone's complimenting my boss on his novel approach to current circumstances, but only I know from which underrated novel he copied the speech.

**Part 2: Coding inputs/outputs**

Programming part (i):

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Above two are the examples mentioned in the Assignment

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

From above example, program is not case sensitive.

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Above example explains that when words are same only the common words are considered.

Text

Description automatically generated

When there is input for normalization other than Y/y/N/n the output is given as above.

Programming part (ii):

For 2nd program also the above points are involved. [like not case sensitive, for multiple words only common are considered, invalid input]

Also only words are calculated for length, for accuracy spaces are removed.

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

The above two are the examples given.

Text

Description automatically generated

Text

Description automatically generated

The above example although both sentences mean same but still common bigram pos are different.

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

Programming part (iii):

In 3rd program, the threshold values are taken with normalization as ‘YES’. If the normalization is ‘NO’, only common words are extracted which cannot be compared to threshold values.

Examples are extracted from the link given in the assignment. Description given for the examples are:

Example 1: Writer A has included a word-for-word passage from his/her source without any indication that it is a direct quotation. [Direct Plagiarism]

Example 2: Writer A has included a verbatim passage from his/her source and has failed to indicate it's a direct quotation. [Direct Plagiarism]

Example 3: Writer B has borrowed with slight variations an uncited phrase from his/her source. As it's written, the passage would not constitute a paraphrase (even if Writer B had acknowledged his/her source) because it contains keywords from the original source that do not appear in quotation marks. [Mosaic Plagiarism]

Example 4: Writer B has borrowed an unacknowledged key word from his/her source, and has not indicated the omission of words with an ellipsis. [Mosaic Plagiarism]

Text

Description automatically generated

1.

Text

Description automatically generated

2.

The above is the outputs of the example1. The example is directly plagiarized, in this situation for unigram type the threshold value suits is 0.7. But for Bigram 0.1 suits best.

Text

Description automatically generated

3.

Graphical user interface, text, application, email

Description automatically generated

4.

In example 3 & 4 the texts are mosaic plagiarized, in these cases the threshold value suits are 0.4 for unigram and for bigram its 0.05.

From my observations [above examples and some other examples] I have concluded the below information.

|  |  |  |
| --- | --- | --- |
| Plagiarism Type | Threshold Value | Type |
| Direct Plagiarism | 0.7 | Unigram |
|  | 0.1 | Bigram |
| Mosaic Plagiarism | 0.4 | Unigram |
|  | 0.05 | Bigram |