

Student ID :
Name-Surname :

Natural Language Processing midterm exam (2019-2020 Fall)

(Please fit your answers to the related blanks.)

1. (20P) Assume that, you have two different corpora that you want to work on them. Your aim to find the similarity between these two corpora. How can you do that?

2. (80P) Fill in the blanks, each question has 5P.

- a. Before we work on the text, we need to pre-processing that converts a list of words to a more uniform sequence, we called this process as _____.
- b. In any text, we call the set of lemmas as _____.
- c. _____ is the analysis of the internal structure of words and forms.
- d. _____ means that parsing words by their types or categories.
- e. Language is not a bag of words, and it has some grammatical rules. We call that rules as _____.
- f. When working on some NLP applications, such as spelling corrections and next-word predictions, we consider words as countable data by using _____ models.
- g. The _____ contains two or more languages edition of the same source text. Wikipedia and the United Nations documents are an example of this.
- h. The _____ corpus contains a native language of the source text and its translations.
- i. The _____ is a parsed sentence structure that it shows syntactic or semantic structures. PENN is one of the famous examples of that.
- j. In general, _____ has no meaning alone and we get rid of them before works on the text in NLP researches.
- k. _____ analysis checks the spelling of the word according to the dictionary, and detects illegal inputs.
- l. The words are considered as the points using _____.
- m. We can compute the _____ value between two words by Cosine function.
- n. _____ analysis is a famous text classification method that aims to discover topics of the documents.
- o. The _____ is used to determine certain meaning of the words, and it needs a semantic analysis in word level.
- p. A collection of texts is important for NLP, and it is called _____.