START OF QUIZ
Student ID:
84574284,Cheng,Yushun

Question 1

Topic: Lecture 1
Source: Lecture 1

Why is strip() such a useful function? (1)

Question 2

Why do we not care about the extra space required to create a reverse index? (2 reasons) (1)

Question 3

Topic: Lecture 3
Source: Lecture 3

Although lexicons are often good starting points, they are often less capable than ML methods. What are some reasons (at least 2) that lexicons are insufficient for state-of-the-art training. Briefly explain. (2)

Question 4

In French, negation is often indicated by "ne ... pas" (ie, "je ne parle pas" - "I am not speaking"; "tu ne conduis pas" - "You are not driving", etc.). However, in speech, one of the two is often dropped: "je ne parle." or "tu conduis pas.". Using this information, how would you determine whether a corpus was composed of written or spoken French? You don't need to write the code, but explain the logic that you would use to come to this conclusion. (2)

Question 5

When would we *not* want to lowercase text prior to training a model? Give a concrete example. (1)

Question 6

Topic: Lecture 2
Source: Lecture 2

What are 2 benefits of providing a corpus reader with your corpus? (1)

Question 7

Topic: Lecture 2
Source: Lecture 2

Do you think it's possible for a language not to follow a Zipfian curve? What consequences might that have on communication (if, let's say, if the curve were linear)? (2)

Question 8

Topic: Lecture 4
Source: Lecture 4

We discussed two alternative methods for noise reduction: removing all words above a certain frequency, or only removing those from a curated lexicon. Name an advantage to both. (1)

Question 9

Topic: Coding
Source: Coding

Imagine that we have an encrypted data set in a language we don't know, but it is written in the Latin script (ie, the script of English, French, etc.). What are some tests that we could run to try to determine the original language? Please list any assumptions you make. Assume that machine learning is not an option. (3)

END OF QUIZ