START OF QUIZ
Student ID:
36304153,Kang,David

Question 1

Topic: Lecture 2
Source: Lecture 2

As we increase the size of a corpus, the frequency of Hapax Legomena generally increases. Would the frequency of function words like ”the” or ”is” also increase? Why or why not? (1)

Question 2

Topic: Lecture 1
Source: Lecture 1

Why is the .split() method useful when working with sentences or phrases? (1)

Question 3

Topic: Lecture 3
Source: Lecture 3

When we nest deep structures in dictionaries, we lose their O(1) benefits. Can you think of a better way to represent complex data sets? (1)

Question 4

Topic: Lecture 1
Source: Lecture 1

What method would you use to check if a string contains only numeric digits (including decimals) without using any additional libraries? (1)

Question 5

Topic: Lecture 3
Source: Lecture 3

When would we want to represent linguistic data in a list, instead of a dictionary or a set? (1)

Question 6

Topic: Lecture 2
Source: Lecture 2

Is it possible for a corpus of a low-resource language to follow Zipf's law? What factors might influence the degree to which the law applies in such languages? (2)

Question 7

Topic: Lecture 4
Source: Lecture 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 8

Topic: Lecture 4
Source: Lecture 4

Attributive adverbs are a type of adverb that provides ”flavour” to speech verbs (example: ”she said quickly”; ”he spoke loudly”). They are often frowned upon in formal writing, because they can be replaced with other verbs: ”blurted” or ”shouted”, in the example. Write a quick function that finds them in the Brown corpus, and reports how many sentences in 1000 have them. (2)

Question 9

Topic: Long
Source: Lecture 2

Imagine you are working with a corpus in a language you don't know, and you need to identify the stopwords in it. You cannot use machine learning but can perform basic statistical analysis. How would you approach identifying stopwords? What metrics would help you confirm that you've identified them correctly? (3)

END OF QUIZ