سوال ۲) برای اسناد موجود در پیکره آزمون، بردار بازنمایی اسناد را با استفاده از روش‌های زیر به‌دست آورید.

الف) استفاده از بردار کلمات آموزش دیده توسط word2vec مدل skip-gram روی مجموعه آموزش و محاسبه بردار جملات پیکره آزمون با استفاده از میانگین وزن‌دار بازنمایی کلمات سند با استفاده از TF-IDF هر یک از کلمات.

ب) محاسبه بردار اسناد براساس مدل doc2vec آموزش داده شده.

Hi, I hope you are doing alright.

Part A of this question is utterly vague. Could you clarify it with one concrete example?

Here are some of my interpretations:

Let's say train corpus contains three documents: "Hello world",
"Hello to world world world " ,
" Hello hello hello hello hello"

And let's say our Word2Vec model embeds words as follows:

Hello : [0.5 , 0.5]
world: [0.1 , 0.1]
to: [0.2 , 0.3]

And we what to represent one of the test documents like: " to to to world"

First question 1)

we know tf-idf means : Term frequency and inverse document frequency
Are the following calculation correct?

Term frequency("to") = number of "to" occurrence in our test document =3
Term frequency("world") = number "world" of occurrence in our test document =1

Document frequency("to") = number of documents containing "to" in ALL train corpus =1
Document frequency("world") = number of documents containing "world" in ALL train corpus =2

First question 2)

For representing this test document (“ to to to world”) is the following method correct?

Document vector =

(first word tf-idf value * first word Word2Vec model vector) + … + (last word tf-idf value * last word Word2Vec model vector)

= ((tf-ifd(“to”)* [0.2 , 0.3]) + (tf-ifd(“to”)* [0.2 , 0.3]) +(tf-ifd(“to”)* [0.2 , 0.3]) + (tf-ifd(“world”)* [0.1 , 0.1])) * (**1 /** (tf-ifd(“to”)+ tf-ifd(“to”)+ tf-ifd(“to”)+ tf-ifd(“world”))

First question 3)

Please explain your answer to question #2. Why it is reasonable to represent a document in this way?!