Started on Monday, 21 March 2022, 9:12 AM
State Finished
Completed on Monday, 21 March 2022, 9:32 AM
Time taken 19 mins 56 secs
Grade **5.00** out of 10.00 (**50**%)

Question **1**

Correct

Mark 1.00 out of 1.00

A neural network model with a single hidden layer, which is **predicting a word** given near-context of words is called:

a. continuous skip-grams

b. continuous bag of words ✔

c. paragraph vector - distributed bag of words

d. deep structured semantic model

e. embedding for language models

Your answer is correct.

The correct answer is:
continuous bag of words

Question **2**

Incorrect

Mark 0.00 out of 2.00

Word "**чикипарабум**" is met in 3 times out of 15 document of a collection.

One for the documents is

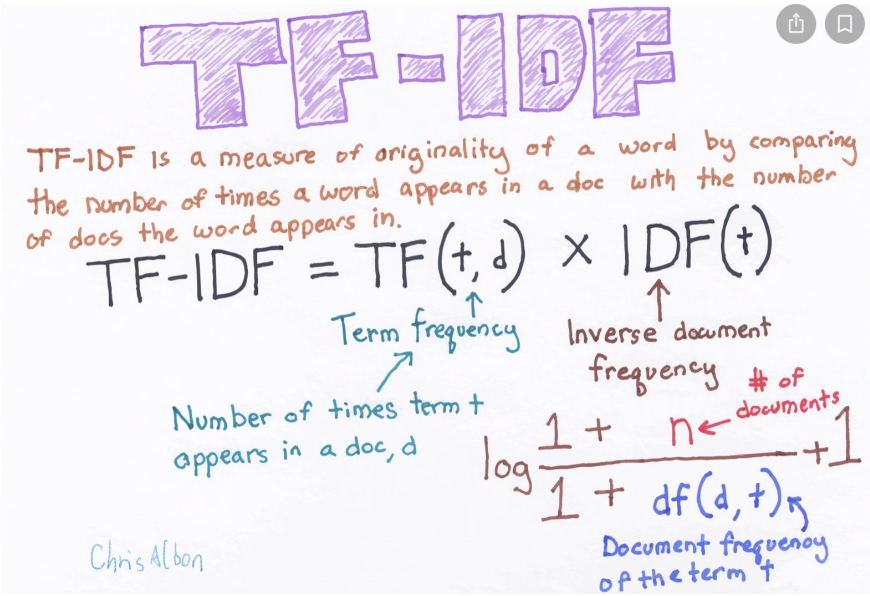
```
0 e чикипарабум рамамбахара чикипарабум бум
```

where {**e, рамамбахара**} are stop words, which we don't include in the vocabualary.

What is the value of TF-IDF for "**чикипарабум**" in this document, if we use this formula:

NB1 log base is 2.

NB2 Don't miss +1 in IDF



- a. 0.5
- b. 1
- c. 1.5
- d. 0
- e. 3.0 ✖

Your answer is incorrect.

With no stopwords our text is

```
0 чикипарабум чикипарабум бум
```

Or **BoW = {чикипара бум: 2, бум: 1, о: 1}**

Thus, TermFrequency = 2 / 4 = 0.5

IDF = log(1 + n / (1 + df)) + 1 = log(1 + 15 / (1 + 3)) + 1 = log(16/4) + 1 = log(4) + 1 = 2 + 1 = 3

TFxIDF = 0.5 * 3 = 1.5

The correct answer is:
1.5

Question **3**

Correct

Mark 1.00 out of 1.00

What is common for the following list of words?

- **fair**
- **unfair**
- **fairness**

a. lemma

b. part of speech

c. stem ✔

Your answer is correct.

PoS are Adj, Adj, Noun

lemmas are fair, unfair, fairness

The correct answer is:
stem

Question **4**

Correct

Mark 1.00 out of 1.00

Mark all techniques, which can be used to reduce dataset dimensions.

a. Latent Semantic Analysis ✔ Uses SVD inside

b. c Means clustering

c. k Means clustering

d. Random Projections ✔

Your answer is correct.

Clustering does not influence dimensionality, but reduces the number of items (from many to cluster count).

The correct answers are: Latent Semantic Analysis,
Random Projections

Question **5**

Correct

Mark 2.00 out of 2.00

You are doing LSA with SVD for a DTM, defined for D=9000 of documents and W=66000 words vocabulary.

You could find an accurate low-rank decomposition DTM=A*B. And the rank of both matrices is R=105.

Your matrices are dense and you use **double-precision** floating point values (recheck number of bytes!) to store them.

How many BYTES you need to store these matrices?

Answer: 6300000 ✔

E.g. for D = 300K and W=20K you will have 2 rectangular matrices D * R and R * W.

Thus, R * (D + W) * BYTES_IN_DOUBLE = (300K * 400 + 20K * 400) * 8B = 320K * 8 * 400 = 320K * 3200 = **1024 000 000**

The correct answer is: 63000000

Question **6**

Incorrect

Mark 0.00 out of 2.00

You study alien language of not-so-developed civilization. You want to propose them to build an inverted index for their (yet small) national library.

Their language **exactly** obeys **Zipf's law**, and the most frequent word's frequency is 70%. You agreed that long tail words which has **strictly less then 0.00013** (0.013%) freqency will not appear in the lexicon.

How many words will there be in their lexicon?

Answer: 70 ✖

probabilty border B = 0.00013

K's word probability (by Zipf's law) = 0.7 / K

We are looking for all numbers K which satisfy

B <= 0.7 / K

K <= 0.7 / B = 0.7 / 0.00013 = 5384.6...

Thus, maximum K is **5384**.

Let's check:

0.7 / 538**5 = 0.0001299907**

The correct answer is: 5384

Question 7

Incorrect

Mark 0.00 out of 1.00

For the word "**fax**" Which of the following typo corrections build using **IoU of bigrams** (Jaccard score) is the best (with respect to this metric).

NB For simplicity *don't include word borders*: **fax** bigrams are just {**fa**, **ax**}.

- a. faximile ✖ 2 / 7
- b. fox
- c. axe
- d. fix

Your answer is incorrect.

fax = {fa, ax}

axe = {ax. xe}

IoU = |{ax}| / |{fa, ax, xe}| = 1/3

The correct answer is:
axe

◄ IR07. Vector Model

Jump to...

IR08. Vector space modelling with ML ►