(6 points) Use the following text to construct distributional thesaurus for *eyes*, *here* and *valley*. Use a 5-word window including open- and closed-class words, ignore case and line-breaks and weight contexts by frequency.

The eyes are not here
There are no eyes here
In this valley of dying stars
In this hollow valley

Compute the following

1. Dice coefficient between (eyes, valley)
2. Overlap coefficient between (here, valley)

(a) In a corpus of 10000 documents you randomly pick a document, say D, which has a total of 250 words and the word 'data' occurs 20 times. Also, the word 'data' occurs in 2500 (out of 10000) documents. What will be the tfidf entry for the term 'data' in a bag of words vector representation for D.

Solution:

Using normalized term frequency $tf(\text{`}data', D) = \frac{20}{250}$. Idf is $idf = ln(\frac{10000}{2500})$. So $tfidf = \frac{2}{25} \times ln\,4$.

(b) You have the following three documents - **D1, D2, D3**:

D1: Natural language processing is becoming important since soon we will begin talking to our computers.

D2: If computers understand natural language they will become much simpler to use.

D3: Speech recognition is the first step to build computers like us.

Answer the following with respect to the above set of 3 documents after text normalization (stop word removal and lemmatization) has been done on all 3 documents.

i. What is the vocabulary $\mathcal{V}$?

Solution:
There are multiple answers possible based on which words are treated as stop words and whether some degree of chunking is done - for example one may decide to chunk 'natural language processing' and/or 'first step'. Here we take a rather simple approach. In the context of the given documents we retain words that are likely to provide useful content and discard the rest. We also we do not chunk.

$\mathcal{V}$ =(become, build, computer, first, important, language, like, natural, processing, recognition, simple, speech, step, talk, understand, us, use)

ii. What are the number of bigrams and trigrams in **D2**?

Solution:

D2 after normalization looks as follows:

```
computer understand natural language become simple use
```

Bi-grams and trigrams are extracted by sliding windows of size 2 and 3 respectively over the sentence. So, if n is the length of the sentence then No. of bigrams=$(n-1)$ where $n \geq 2$ and No. of trigrams=$(n-2)$ when $n \geq 3$. So, we get 6 bigrams and 5 trigrams.

Of course, this answer will differ if your normalized document has more or less words.

iii. What will be the BoW document vector for document **D3** if we are using a *tf* based document vector?

Solution:

Document **D3** after normalization:

```
speech recognition first step build computer like us
```

Bag-of-words vector with normalized tf:

$\frac{1}{8}(0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0)$

All words occur only once in **D3**. So, if you use plain *tf* then the factor $\frac{1}{8}$ will not be present.

(c) Suppose you have the following two 4-dimensional word vectors for two words w_1 and w_2 respectively:

$\mathbf{w}_1 = (0.2, 0.1, 0.3, 0.4)$ and $\mathbf{w}_2 = (0.3, 0, 0.2, 0.5)$

What is the cosine similarity between $\mathbf{w}_1$ and $\mathbf{w}_2$? Are the words w_1 and w_2 similar or dissimilar?

Solution:

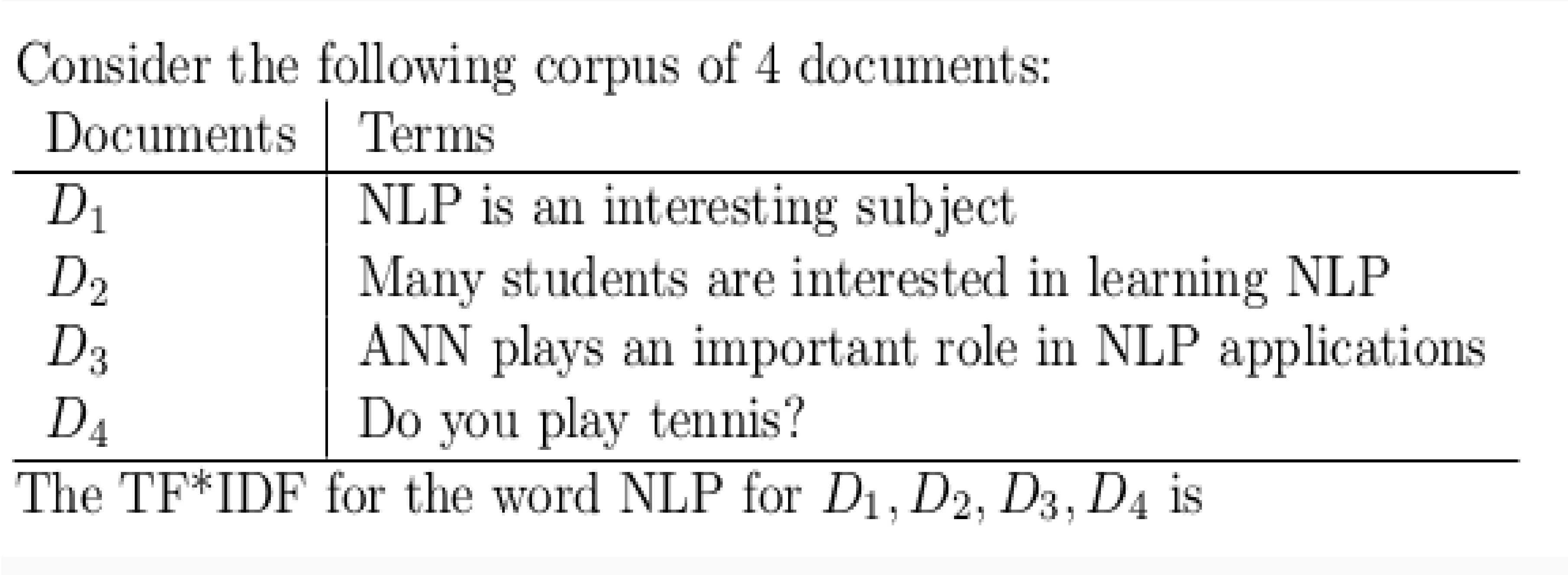
We can calculate cosine similarity from:

$$cosine(\theta) = \frac{\langle \mathbf{w_1}, \mathbf{w_2} \rangle}{\| \mathbf{w_1} \| \| \mathbf{w_2} \|}$$

The expression is easier to calculate if we scale both vectors by multiplying by 10 - this does not change the cosine. We get,

$$\begin{aligned} cosine(\theta) &= \frac{6+6+20}{\sqrt{4+1+9+16}\sqrt{9+4+25}} = \frac{32}{\sqrt{30 \times 38}} \\ &= \frac{32}{2\sqrt{285}} \\ &\approx \frac{16}{17} \end{aligned}$$

The value is very close to 1 so the words w_1, w_2 are very similar.



Consider the following corpus of 4 documents:

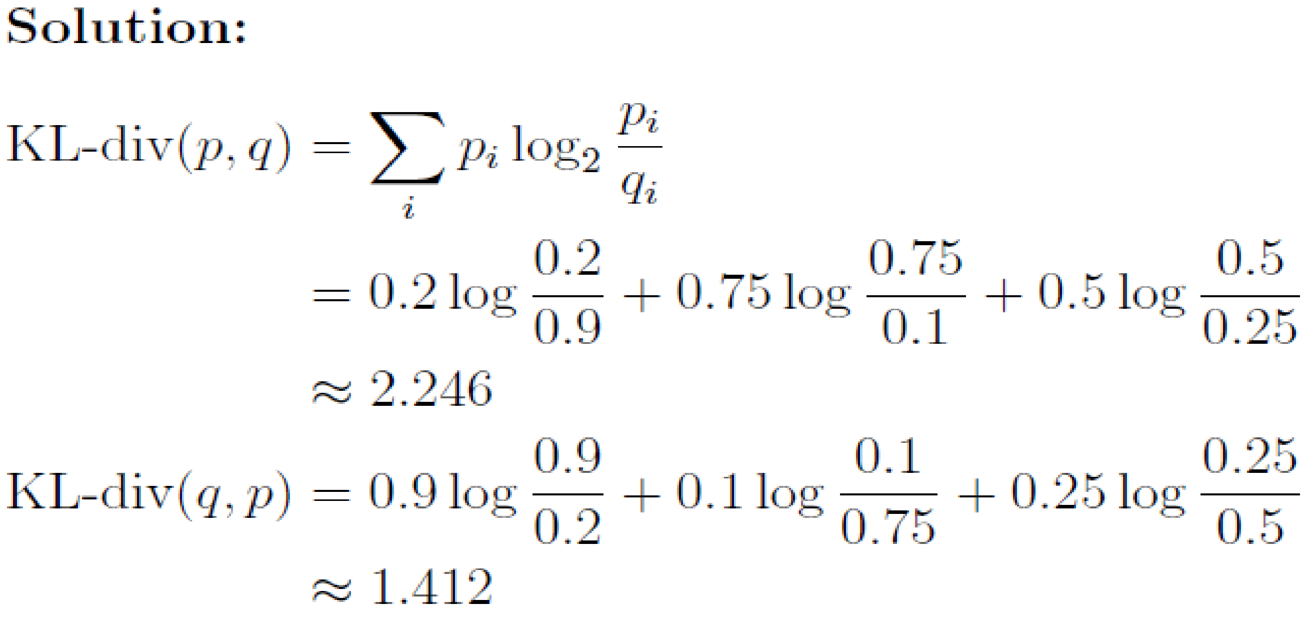
Documents	Terms
D_1	NLP is an interesting subject
D_2	Many students are interested in learning NLP
D_3	ANN plays an important role in NLP applications
D_4	Do you play tennis?

The TF*IDF for the word NLP for D_1, D_2, D_3, D_4 is

1. [1/5, 1/7, 1/8, 0] $\log_{10}(3/4)$
2. [1/6, 1/7,1/8, 0] $\log_{10}(3/4)$
3. [1/5, 1/7, 1/8, 0] $\log_{10}(4/3)$
4. None of the these

Ans: 3

Consider two probability distribution for two words be p and q. Compute
their similarity scores with KL-divergence.
p = [0.20, 0.75, 0.50]
q = [0.90, 0.10, 0.25]
Note: Use base 2 in logarithm.



Solution:

$$\begin{aligned}
\text{KL-div}(p, q) &= \sum_i p_i \log_2 \frac{p_i}{q_i} \\
&= 0.2 \log \frac{0.2}{0.9} + 0.75 \log \frac{0.75}{0.1} + 0.5 \log \frac{0.5}{0.25} \\
&\approx 2.246 \\
\text{KL-div}(q, p) &= 0.9 \log \frac{0.9}{0.2} + 0.1 \log \frac{0.1}{0.75} + 0.25 \log \frac{0.25}{0.5} \\
&\approx 1.412
\end{aligned}$$

Consider the following word co-occurrence matrix given below. Compute
the cosine similarity between (i) w1 and w2, and (ii) w1 and w3.

	w4	w5	w6
w1	2	9	4
w2	1	5	6
w3	3	0	1

Solution:

$$\text{cosine-sim}\left(\overrightarrow{p}, \overrightarrow{q}\right) = \frac{\overrightarrow{p} \cdot \overrightarrow{q}}{\|\overrightarrow{p}\| \cdot \|\overrightarrow{q}\|}$$

$$\text{cosine-sim}\ (\text{w1, w2}) = \frac{2 \times 1 + 9 \times 5 + 4 \times 6}{\sqrt{2^2 + 9^2 + 4^2} \times \sqrt{1^2 + 5^2 + 6^2}} \approx 0.897$$

$$\text{cosine-sim}\ (\text{w1, w3}) \approx 0.315$$

- **Understanding Pointwise Mutual Information in NLP**
- How to understand whether two (or more) words actually form a unique concept. If this is the case, we can reduce the dimensionality of our task, since that couple of words (called bigram- or n-grams in the general case for n words) can be considered as a single word. Hence, we can remove one vector from our computations.
- Namely, consider the expression ‘social media’: both the words can have independent meaning, however, when they are together, they express a precise, unique concept.
- Since if both words are frequent by themselves, their co-occurrence might be just a chance. Namely, consider the name ‘Las Vegas’: it is not that frequent to read only ‘Las’ or ‘Vegas’ (in English corpora of course). The only way we see them is in the bigram Las Vegas, hence it is likely for them to form a unique concept. On the other hand, if we think of ‘New York’, it is easy to see that the word ‘New’ will probably occur very frequently in different contexts. How can we assess that the co-occurrence with York is meaningful and not as vague as ‘new dog, new cat…’?

The answer lies in the Pointwise Mutual Information (PMI) criterion. The idea of PMI is that we want to quantify the likelihood of co-occurrence of two words, taking into account the fact that it might be caused by the frequency of the single words. Hence, the algorithm computes the (log) probability of co-occurrence scaled by the product of the single probability of occurrence as follows:

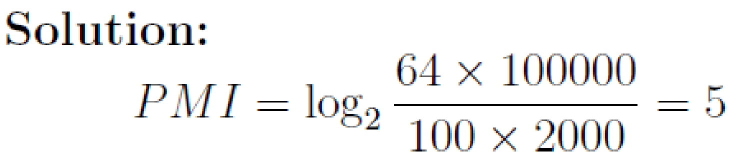
$$PMI(a,b) = \log(\frac{P(a,b)}{P(\mathrm{a})\mathrm{P}(\mathrm{b})})$$

Now, knowing that, when ‘a’ and ‘b’ are independent, their joint probability is equal to the product of their marginal probabilities, when the ratio equals 1 (hence the log equals 0), it means that the two words together don’t form a unique concept: they co-occur by chance.

- On the other hand, if either one of the words (or even both of them) has a low probability of occurrence if singularly considered, but its joint probability together with the other word is high, it means that the two are likely to express a unique concept.

Example:

- What is the value of *PMI* (w1, w2) for *C(w1)* = 100, C(w2) = 2000, *C(w1,w2)* = 64, N = 100000?
- N: total number of documents.
- *C(wi)*: number of documents, wi has appeared in.
- *C(wi, wj):* number of documents where both the words have appeared in.
- Note: use base 2 in logarithm.



Solution:

$$PMI = \log_2 \frac{64 \times 100000}{100 \times 2000} = 5$$