Lab Class NLP:IX

Exercise 1 : Word Mover Distance

The lecture introduced the Word Mover Distance (WMD) for measuring word vector similarity. WMD finds the minimum cumulative transportation cost to move all words from one sentence to words in another sentence in an embedding space.

You are given the sentences A, B, and C and the 3-dimensional word vectors $[d_1, d_2, d_3]$ for all occurring words:

- Sentence A: "The **cat climbed** the tree."
- Sentence B: "The **feline scaled** the tree."
- Sentence C: "The **kitten ascended** the tree."

	the	cat	climbed	tree	feline	scaled	kitten	ascended
d_1	0.1	0.4	0.7	1.0	0.35	0.75	0.45	0.77
d_2	0.2	0.5	0.8	1.1	0.55	0.85	0.57	0.87
d_3	0.3	0.6	0.9	1.2	0.65	0.95	0.67	0.97

(a) Calculate the WMD between Sentence A and Sentence B.

(b) Calculate the WMD between Sentence A and Sentence C.

(c) Which sentence B or C is more similar to Sentence A?

Exercise 2 : Sentence Embeddings

The lecture introduced sentence embeddings as a way to represent sentences in a continuous vector space. One approach to generating sentence embeddings is to average the word embeddings $\mathbf{w}$ of all words in a sentence s:

$$\mathbf{s}_{emb} = \frac{1}{|s|} \sum_{\mathbf{w}_i \in s} \mathbf{w}_i$$

You are given the sentences A, B, and C and the 3-dimensional word vectors $[d_1, d_2, d_3]$ for all occurring words:

- Sentence A: "The **`cat climbed`** the tree."
- Sentence B: "The **`feline scaled`** the tree."
- Sentence C: "The **`kitten ascended`** the tree."

	`the`	`cat`	`climbed`	`tree`	`feline`	`scaled`	`kitten`	`ascended`
d_1	0.1	0.4	0.7	1.0	0.35	0.75	0.45	0.77
d_2	0.2	0.5	0.8	1.1	0.55	0.85	0.57	0.87
d_3	0.3	0.6	0.9	1.2	0.65	0.95	0.67	0.97

(a) Calculate the sentence embeddings for Sentence A, B, and C using vector averaging.

(b) Which sentence embedding pair is more similar $(\mathbf{A}, \mathbf{B})$ or $(\mathbf{A}, \mathbf{C})$? For your answer, calculate the Manhattan distance D between the embeddings of the sentences in each pair.

(c) What are the limitations of the vector averaging approach for generating sentence embeddings? Name at least two.

(d) Another approach to measuring the similarity between two embedding representations is to calculate the cosine similarity between them:

$$sim_{cosine}(\mathbf{s}_1, \mathbf{s}_2) = \frac{\mathbf{s}_1 \cdot \mathbf{s}_2}{\|\mathbf{s}_1\| \cdot \|\mathbf{s}_2\|}$$

(d1) Interpret the following cosine similarity values between two sentence embeddings:

a) $sim_{cosine}(\mathbf{s}_1, \mathbf{s}_2) = -1$
b) $sim_{cosine}(\mathbf{s}_1, \mathbf{s}_2) = 0$
c) $sim_{cosine}(\mathbf{s}_1, \mathbf{s}_2) = 1$

(d2) Calculate the cosine similarity between Sentence A and Sentence B, and between Sentence A and Sentence C. Which sentence is more similar to Sentence A according to the cosine similarity measure?