

## Lab Class NLP:V

## Exercise 1 : Text Representation: Bag-of-Words

The lecture introduced Bag-of-Words (BoW) model as a simple text representation technique.

(a) What is the main assumption of the BoW model?

(b) Consider the following document collection  $D$  of 4 documents:

d1: not bad good film	d3: good film bad plot
d2: good film good plot	d4: not good bad film

Create a BoW representation of the document collection  $D$ . Which document is the most similar to the document  $d_1$  based on the BoW representations of the documents?

*Hint: you don't need to calculate the cosine similarity, just compare the resulting representations.*

(c) You want to use the BoW representation to train a model for sentiment analysis prediction (e.g., classifying movie reviews as positive or negative). Do you think the BoW representation is suitable this task? Use your BoW representation of the document collection  $D$  to support your answer.

Exercise 2 : Text Representation: Term Weighting  $tf \cdot idf$ 

The lecture introduced  $tf \cdot idf$  as a measure to evaluate the importance  $w$  of a term  $t$  in a document  $d \in D$  as:

$$w(t) = tf(t, d) \cdot idf(t, D)$$

(a) What is measured by  $tf(t, d)$  and  $idf(t, D)$  in the equation above? How are they calculated?

(b) Consider the following document collection  $D$  of 8 documents:

d1: bad bad fast cat	d5: job big big cat
d2: run unix cat job	d6: kill big big job
d3: big big big cat	d7: unix job run cat
d4: big cat big kill	d8: big cat big cat

(b1) Calculate the  $idf$  value for each term in the document collection  $D$ . Which term (or terms) have the highest  $idf$  value in this collection? Report the words and their  $idf$  values.

- (b2) The query  $q = \text{big cat}$  is run against the document collection  $D$ . Rank the documents according to the weighted sum of  $tf \cdot idf$  values for the query terms:

$$\sum_{t \in q} w(t) = \sum_{t \in q} tf(t, d) \cdot idf(t, D)$$

### Exercise 3 : Word Vectors

The lecture introduced distributional representations of words (word vectors) as embeddings of words in a latent space.

- (a) Which of the following statements about word vectors are true? (Select all that apply)

- Word embeddings are sparse vector representations of words.
- Embeddings are learned from the surrounding context in which words appear.
- Embeddings can only be learned for words, not subwords or sentences.
- Embeddings can capture semantic relationships between words.
- Embeddings can not capture syntactic relationships between words.
- The higher the dimensionality of the embedding, the more information they can capture.

- (b) Which of the following equations should hold for good word embeddings? (Select all that apply)

- $\mathbf{w}_{\text{boy}} - \mathbf{w}_{\text{girl}} \approx \mathbf{w}_{\text{brother}} - \mathbf{w}_{\text{sister}}$
- $\mathbf{w}_{\text{cat}} - \mathbf{w}_{\text{dog}} \approx \mathbf{w}_{\text{puppy}} - \mathbf{w}_{\text{kitten}}$
- $\mathbf{w}_{\text{berlin}} - \mathbf{w}_{\text{tokyo}} + \mathbf{w}_{\text{japan}} \approx \mathbf{w}_{\text{germany}}$
- $\mathbf{w}_{\text{bigger}} - \mathbf{w}_{\text{big}} + \mathbf{w}_{\text{cold}} \approx \mathbf{w}_{\text{colder}}$

- (c) What is the difference between static and contextualized word embeddings?

- (d) Most computational models of distributional similarity, including neural embeddings, such as word2vec, often embed antonyms (e.g. “good” and “bad”) close to each other in the vector space and assign them similar meanings<sup>1</sup>. Briefly explain the underlying reason for this phenomenon.

---

<sup>1</sup>You can visually inspect this phenomenon in the [TensorFlow Projector](#) tool.

#### Exercise 4 : Lexical and Distributional Semantics

Answer the following questions:

- (a) What is the name of the lexical semantic relationship between the word *chair* and the word *furniture*?
  
- (b) Suppose you are building a question-answering system. Using an example, explain why it is important for your system to be able to identify this kind of relationship (between *chair* and *furniture*).
  
- (c) What is meant by the *distributional hypothesis* in lexical semantics?
  
- (d) Name one lexical semantic relationship that is easy to identify using *distributional word representations*, and one that is hard to identify this way.
  
- (e) What are the characteristic features of a representation of a word in a distributional semantics? Give a formula which you can use to predict word meaning similarity from the distributional representations of the words.