Assignment 1

December 5, 2024

This assignment will cover lecture 4 and 5.
Deadline: 11:59 PM on 12.12.2024

General Rules

- The most important rule is that you always **show all your workings**.
- You can either do the exercise in groups of 2-3 or submit the exercise by yourself.
- You can do the assignment and submit it anyway. Just make sure everything is **readable**.
  - It is good practice if you write the assignment in LaTeX and submit a PDF as it will help you in your future reports/thesis/assignments especially if you want to pursue master's or PhD.
- If you do it in a group **everyone in the group must submit the same work**. You should include the names in your submission. If there are multiple files kindly upload a .zip.
- If you have any question regarding exercise please post it on course forum: `https://wuecampus.uni-wuerzburg.de/moodle/mod/forum/view.php?id=3252043`
  - If you find any bugs in the assignments, please report it on forum.

1 Distributional Semantics

1.1 Raw Co-occurrence Vectors

Given the following raw co-occurrence counts of words with contexts *jealous* (c_1) and *gossip* (c_2):

	c_1 (*jealous*)	c_2 (*gossip*)
w_1	2	5
w_2	3	0
w_3	4	0
w_4	0	4

1. **Compute the TF-IDF Weighted Co-occurrence Matrix**

   Use the following formulas:

$$\text{tf}(w, c) = \log\left(\frac{\text{freq}(w, c)}{\max_{w'} \text{freq}(w', c)} + 1\right)$$

$$\text{idf}(c) = \log\left(\frac{|V|}{|\{w \in V : \text{freq}(w, c) > 0\}|}\right)$$

   Where $|V| = 4$.

2. **Represent Each Word as a TF-IDF Vector**

3. **Compute the Euclidean Distance Between:**

   (a) w_1 and w_2

   (b) w_2 and w_3

4. **Discussion:**

   Based on the Euclidean distances computed, evaluate whether Euclidean distance is an appropriate measure for capturing semantic similarity between word vectors in this context.

1.2 Prediction Based Word Vectors

1. Why does Word2Vec use separate input vectors ($\mathbf{u}_w$) and output vectors ($\mathbf{v}_w$) for each word, and how does this benefit the model's performance?

2. What are the primary differences between the Skip-Gram and Continuous Bag-of-Words (CBOW) models in Word2Vec, and in what scenarios might one outperform the other?

3. How does negative sampling improve the efficiency of training Word2Vec models compared to using the full softmax function?

4. How does the choice of window size in Word2Vec affect the type of semantic relationships the model captures?

5. What strategies can Word2Vec employ to handle out-of-vocabulary (OOV) words, and what are the implications of these strategies?

2 Topic Modeling

Consider a simple corpus with the following characteristics:

- **Vocabulary (V):** { `apple`, `banana`, `cherry` }
- **Number of Topics (K):** 2
- **Number of Documents (M):** 2

The initial topic distributions over words (ϕ_k) and document distributions over topics (θ_m) are randomly initialized as follows:

$$\phi_1 = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right], \quad \phi_2 = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$$

$$\theta_1 = \left[\frac{1}{2}, \frac{1}{2}\right], \quad \theta_2 = \left[\frac{1}{2}, \frac{1}{2}\right]$$

Suppose the documents are:

- **Document 1:** `apple`, `banana`
- **Document 2:** `banana`, `cherry`

1. For each word in each document, compute the probability of assigning it to each topic using the current ϕ and θ values. Specifically, calculate

$$P(z_{mn} = k) \propto \phi_k[w] \times \theta_m[k]$$

   for each word w in document m.

2. Based on the probabilities computed earlier, assign a new topic to each word in each document. Assume you sample deterministically by choosing the topic with the higher probability.

3. Update the ϕ_k and θ_m distributions based on the new topic assignments. Compute the new probabilities:

$$P(w|k) = \frac{C(w,k)}{\sum_{w'} C(w',k)}$$

$$P(k|d) = \frac{C(k,d)}{\sum_{k'} C(k',d)}$$

   where $C(w,k)$ is the count of word w assigned to topic k across all documents, and $C(k,d)$ is the count of topic k in document d.