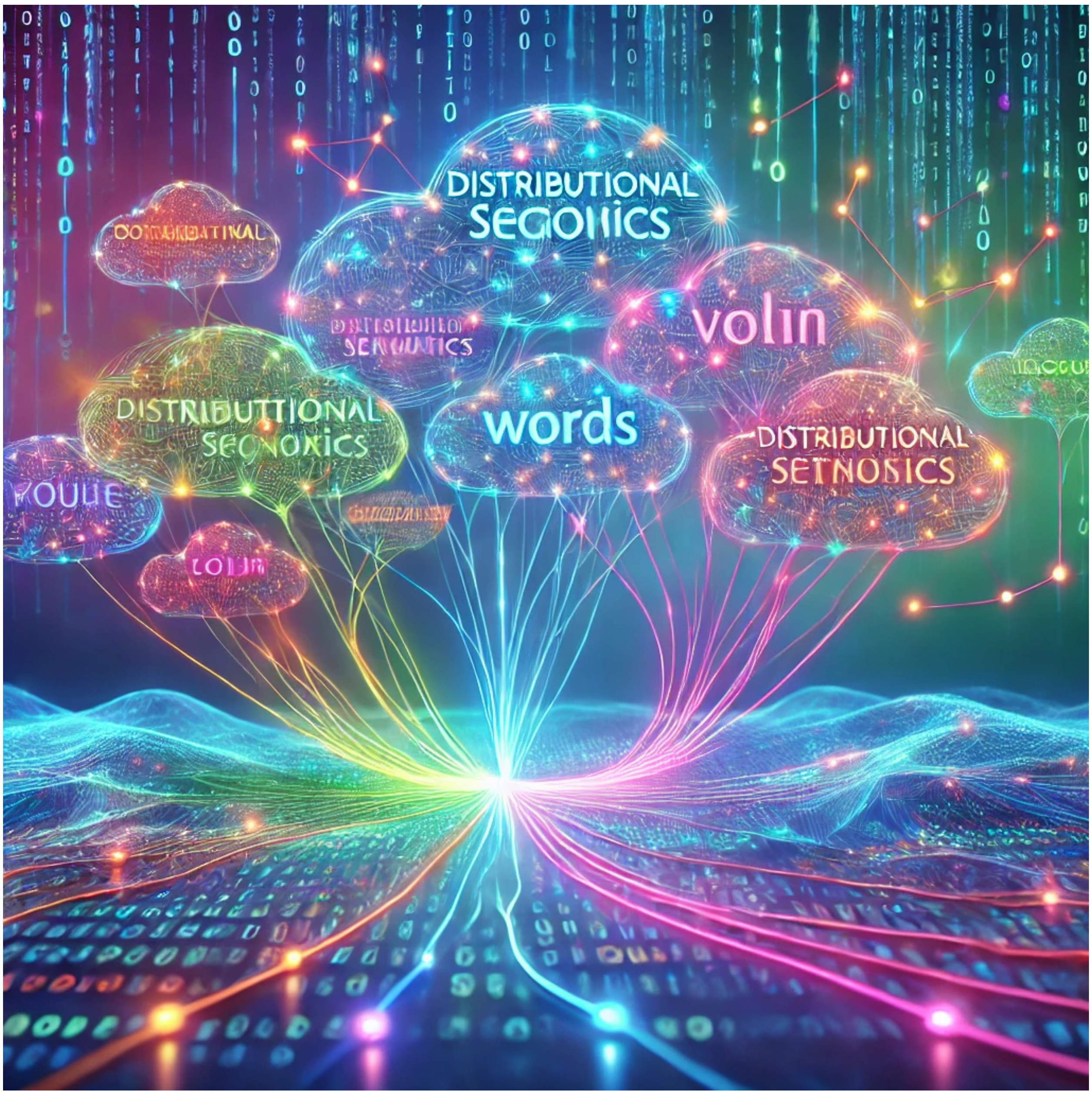
Introduction to Natural Language Processing, Assignment 2

Enrique Mesonero Ronco   Sergio Sánchez García   Ismael Cross Moreno

December 10, 2024



Contents

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:
   $$\mathrm{tf}(w,c) = \log\left(\frac{\mathrm{freq}(w,c)}{\max_{w'} \mathrm{freq}(w',c)} + 1\right)$$
   $$\mathrm{idf}(c) = \log\left(\frac{|V|}{|\{w \in V : \mathrm{freq}(w,c) > 0\}|}\right)$$
   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 the relationships between the words.

1.2 Prediction Based Word Vectors

- Why does Word2Vec use separate input vectors (u_w) and output vectors (v_w) for each word, and how does this benefit the model's performance?
- 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?
- How does negative sampling improve the efficiency of training Word2Vec models compared to using the full softmax function?
- How does the choice of window size in Word2Vec affect the type of semantic relationships the model captures?
- What strategies canWord2Vec 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:

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

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

Documents

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

Steps to Solve

1. Compute Topic Assignment Probabilities

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. Assign New Topics

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 Distributions

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.