Semantic analysis

Nil Biescas, Joan Lafuente and Xavi Soto

March 8, 2024

1 Exercise 1

In the first task, we explored for different words their spatial representation according to its word embedding, showing the closest words. The embeddings were obtained using word2vec, which learns to encode the words based on their context, following the distributional hypothesis, where similar words appear in similar contexts. In Figure 1 we study the word embedding of the word Alice. The words "carroll," "lewis," "carol," and possibly "booth" and "frank," when compared to the word embedding of "Alice," are likely to be closely related due to their connections with literature, authors, or characters from the story of "Alice in Wonderland" by Lewis Carroll. "Carroll" directly refers to the author, while "lewis" is his first name. "Carol" might be a misspelling or variation of "Carroll." The connections of "booth" and "frank" to "Alice" are less direct; they may not be as closely related in a word embedding space unless there's specific context linking them to the story or author.

"Bob" is the outlier in this set as it does not have a direct connection to "Alice in Wonderland" or Lewis Carroll in a widely recognized manner. It might be closer to other common English names in the embedding space rather than being closely related to the theme of "Alice in Wonderland."

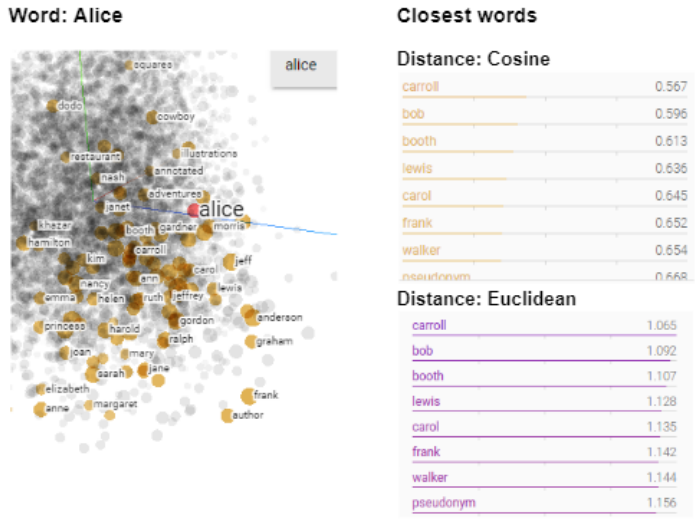

Figure 1: Visualization of the Word2Vec embedding near the word "Alice" and similar words using Cosine and Euclidean distances.

In Figure 2, we use the word bank to observe the other words closer to it. We observe that the plural, banks, is very similar to the singular form as it is usually used in similar contexts. We also see the name of a banking institution such as IMF (International Monetary Fund) which is a very important entity in the field of banking. Furthermore, we note that changing the type of distance to euclidean does not change the order of similarity. Since Euclidean distance considers magnitude and cosine similarity focuses on direction, changing between these metrics can change

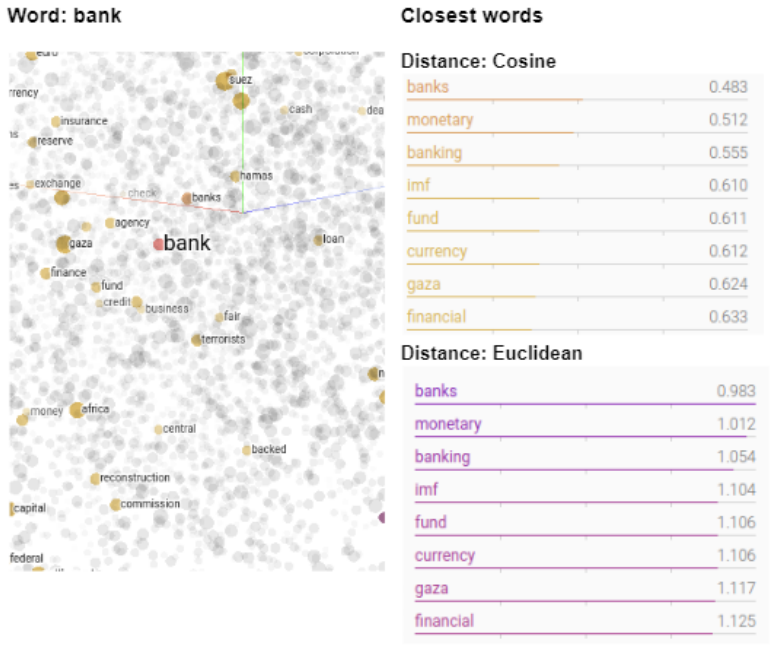

Figure 2: Visualization of the Word2Vec embedding near the word "Bank" and similar words using Cosine and Euclidean distances.

the similarity rankings among a set of vectors. For example, two vectors with a small angle between them (similar direction) but different magnitudes may be considered similar using cosine similarity but not using Euclidean distance. The case here is that changing the metric does not change the order of similarity, suggesting that the vectors being compared are relatively uniform in magnitude, making their spatial orientation the primary determinant of similarity. In other words, the differences in direction between the vectors are more pronounced than the differences in their magnitudes.

In practical applications, the choice between Euclidean distance and cosine similarity depends on what aspect of the data is most relevant to the task at hand—magnitude, direction, or both.

2 Exercise 2

In this second task, we have explored the vector representation of different words, checking the length of the vectors and whether the word was inside the vocabulary. To compute the word embeddings we have used the spacy word2vec model *en_core_web_md*, we have used the same model in the next exercises. We can see that no matter whether the word is in the vocabulary, the length of the vector will be 300. But if we check the different vectors, for the words that are out of the model's vocabulary (OOV) are 0 in all the dimensions of the vector.

To visualize it, here we show the first six dimensions for one word in the vocabulary and one OOV:

- Football : [-1.863, 4.282, -7.157e-01, 1.235, 5.460, 6.024, ...]
- Franfurteria: [0, 0, 0, 0, 0, 0, ...]

As we can see, Frankfurteria is not in the model vocabulary. Additionally, we checked if the word flowers is an OOV, as it is not, we obtained its embedding, which you can see the first six dimensions:

- Flowers: [-3.163, -3.120, -7.435, 2.046, 3.106, -6.577, ...]

Checking that the word flowers is in fact inside the model vocabulary.

Finally, we checked the representation not just for a word but a whole sentence, to see the vector representation of it.

- **Sentence**: I love football
- First dimensions: [-5.558e-01 4.049e-01 -3.531 -4.515, -3.779, 1.584, ...]

If we check the size of the vector, it is also 300, so no matter the length of the sentence that we are trying to represent, the size will always be the same. This happens as the vector representation of a sentence is computed as the average representation of all the words in it, so in this case it is the average of the representation of: "I", "love" and "football".

3 Exercise 3

In this third task, we have explored the similarity between different words and sentences. The first sentences that we compared were "I visited Scotland" and "I went to Edinburgh", obtaining a similarity of: 0.7532. This high similarity between the sentences was expected, as both sentences are expressing that the writer was in Scotland at some point.

Moreover, we defined 2 new very similar sentences and 2 very dissimilar sentences.

The similar sentences that we defined are:

- I went on a vacation to the beach
- I will go to the beach on the holidays

The similarity between them is 0.9165, this happens because both sentences are talking about a trip to the beach on a holiday period, expressing a similar meaning. The similarity is not higher because one of the sentences is expressing a future action and the other a past one.

The dissimilar sentences that we defined are:

- I will go to the beach on the holidays
- The weather was nice yesterday

The similarity between this last two sentences is 0.4597, which is much lower than in the previous case. This was expected as in this case as each sentence talks about a very different topic, about the past weather or about a future trip.

After comparing the embeddings of some sentences, we explored the word embeddings of some words. To do this, we needed to reduce the dimensionality of the vectors, to be able to visualize it better. For this task, we have used PCA and selected the two components that explained more variance. At this point, we plotted the vectors of different words, where we can visually see the similarity depending on how close they are from each other.

In Figure 3 can be seen the results. In the legend can be seen the different words that were used. In the plot we see 2 main groups of words: animals and food, separated from each other, while the words inside that groups are very close. It can be seen that despite the clear separation, words like cat and dog are also very separated from the other animals, which we can assume that it is because they can be differentiated from the others as they can be considered as pets.

If we define a new set of words in Figure 4, we observe a distinct division among the primary clusters when examining the selected words. These words are derived from six distinct categories: animals, flowers, food, clothing, technology, and furniture. It is evident that words from identical categories are proximate within this space. For instance, while 'oak' is not a flower like 'rose', 'tulip', or 'maple', its embedding captures the connection between flowers and trees. This explains why, in terms of spatial proximity, 'oak' is closer to flowers rather than to any other category examined here.

Finally, we also decided to visualize them on three dimensions, to look if some of the similar words were distanced when rising the number of dimensions. We used the first three components obtained with PCA for this. It can be seen in Figure 3, it can also be seen in an interactive way in the notebook. From what we can see is that those words of similar categories continue to be close, like "smartphone" and "laptop", and those from different ones are more separated as we have a third dimension.

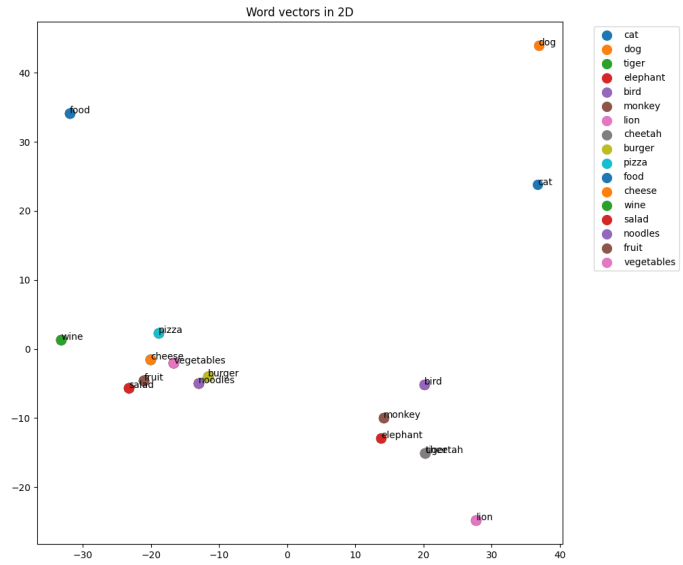

Figure 3: Visualization of the word embeddings for some words using the first two dimensions of the PCA.

4 Exercise 4

In this last exercise, we have explored the similarity of words with sentences and how we can use this information to filter non-related sentences.

The first step was to compute the similarity of some specific sentences with the word perfume. The sentences used were:

- "I purchased a science fiction book last week."
- "I loved this fragrance: light, floral and feminine."
- "I purchased a bottle of wine."

After computing the similarity of each sentence, in table 1 we can see that the only sentence that is talking about a perfume is the one with a higher cosine similarity of 0.51. The rest of the sentences that are talking about other topics have a low similarity.

Sentence	Similarity to perfume
I loved this fragrance: light, floral and feminine	0.51
I purchased a bottle of wine	0.43
I purchased a science fiction book last week	0.28

Table 1: Similarity of some sentences to the word "perfume".

4.1 Filtering Reviews

In this exercise, our objective was to filter out those reviews from Amazon that were not associated to the "music" property. In order to do that, we followed a similar approach to the previous exercise, and we computed the similarity of "music" word embedding to each one of the reviews embeddings. Once we did that, we filtered out those reviews that had less than a 0.5 similarity.

As we realised that some words had different representation if they started with an uppercase, we decided to add a preprocessing step before computing the reviews embeddings. In this preprocessing we converted all the words to lowercase as well as removing words that did not provide

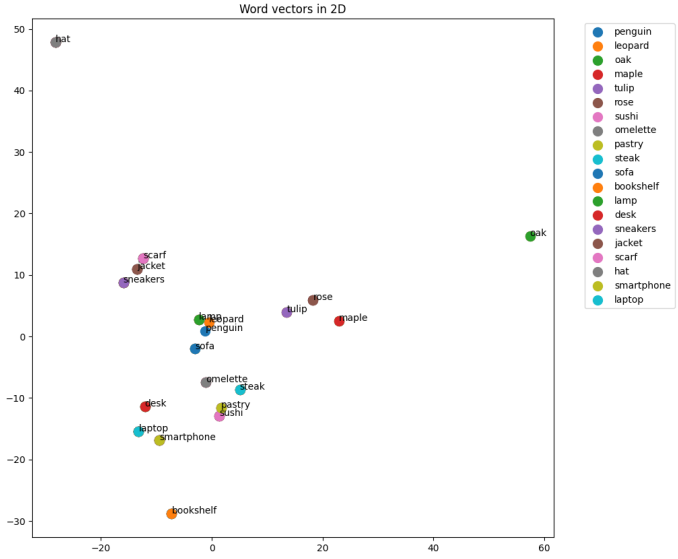

Figure 4: Visualization of some words embeddings for the selected words using the first two dimensions of the PCA.

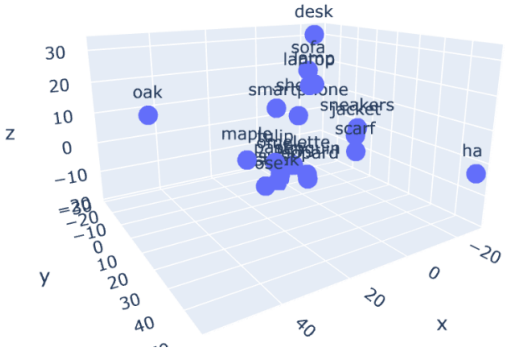

Figure 5: Visualization of some words embeddings for some selected words using the first three dimensions of the PCA.

meaning to the sentence, the stop words. For instance, the review: *Really happy with this purchase. Great speaker and easy to set up.* after the prepossessing becomes: *really happy purchase. great speaker easy set up.* Removing stopwords, which do not add significant meaning to the sentences, eliminates unnecessary noise and improves the overall quality of the sentence representation as they are not taken into account when doing the average of the word embeddings of the sentence.

The ten most similar reviews to "music", using preprocessing, can be seen in Table 2.

Similarity	**Review**
1.00	Music
0.80	Sound is terrible if u want good music too get a bose
0.80	Sound is terrible if u want good music too get a bose
0.80	Clear music
0.80	Clear music
0.78	Love the music the stories
0.77	Music mainly but still checking other features
0.74	The sound is amazing and many collections of all music
0.74	The sound is amazing and many collections of all music
0.67	Great product we listen to music all the time

Table 2: Reviews Ranked by Similarity to the word "music".