Natural Language Processing

LAB 4

NLP WORD REPRESENTATIONS

Students:
Neil de la Fuente, 1630223
Daniel Vidal, 1634599
Joan Samper, 1631430

Professor:
Xim Cerdá

$8^{th} March 2024$

Exercice II - Word2vec

2. Show the word vector for “football”, how long is it?

Word vector: [-1.8633e+00 4.2826e+00 -7.1575e-01 1.2353e+00 5.4601e+00 6.0244e+00 -2.2204e+00 1.7759e+00 2.5588e+00 -2.0682e+00 4.4509e+00 -2.2267e+00

...

-1.7003e+00 -3.6575e+00 2.4663e+00 -2.7056e+00 -2.0104e+00 2.9346e+00 -2.3386e-01 -1.1868e+00 2.6620e+00 -1.1154e+00 -7.8072e-02 8.8475e-02]

Length: 300

3. Show the word vector for “frankfurteria”, how long is it?
Only words in the model’s vocabulary have vectors, the rest are called out-of-vocabulary (OOV)

Word vector: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.

...

0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

Length: 300

The word "Frankfurteria" has all zeros because it’s an out-of-vocabulary word that doesn’t have any associated embedding

4. Check whether the word “flowers” is in the model vocabulary

The word ’flowers’ is in the vocabulary, and has an associated embedding.

5. Create a sentence including the word ”football”, and show the sentence vector

Sentence: Football is a great sport

Sentence Vector: [-0.627118 2.986326 -2.07868 -0.14277804 4.8246202 1.05336 -0.15930007 3.9424198 0.403222 -0.3256761 9.86448

...

-2.56306 1.9427601 0.76902 -0.74228007 1.602906 1.3250601 -1.117224 1.9765482 0.63135993 -0.727106 -2.9633138 3.307632]

6. How long is the sentence vector? How is it calculated?

Length: 300

The length remains 300 even having more than 1 word because the model automatically do the average between the embedding words to obtain an embedding of the sentence with size 300.

Exercice III - Similarity

1. Define the two utterances "I visited Scotland" and "I went to Edinburgh"

2. Calculate the similarity between these two sentences

Using the cosine similarity

$$\text{cosine similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

between the sentence embeddings, we obtain a similarity of 0.753

Green Exercice: Define two similar sentences and calculate their similarity, then define two different sentences and calculate their similarity.

Sentence one: "I do not like football"

Sentence two: "I hate soccer"

Cosine similarity: The cosine similarity for sentences: 'I do not like football' and 'I hate soccer' is 0.84.

Even if the sentences seem to be the same maybe the reason why they don't have 100% similarity is because football, in some places, can be interpreted as American football, which is different from soccer.

Sentence 3: "I really hate football"

Sentence 4: "My sister loves watching tennis matches"

Cosine similarity: The cosine similarity for sentences: "I really hate football" and "My sister loves watching tennis matches" is 0.382. This value seems reasonable since the sentences express very different things even if they talk about sports and personal tastes.

3. Consider the following words [cat, dog, tiger, elephant, bird, monkey, lion, cheetah, burger, pizza, food, cheese, wine, salad, noodles, fruit, vegetables]

4. Calculate the word vector for every word

5. Apply a PCA, consider the first two components, and represent the words in the feature space:

PCA PLOT 2D:

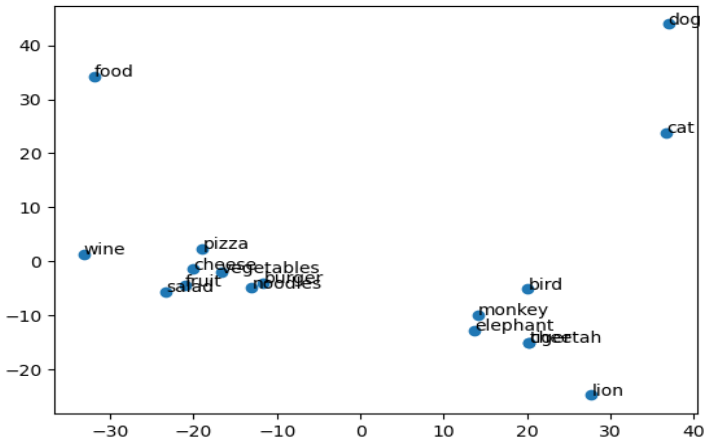

Figure 1: Plot of the words embeddings reduced to 2 dimensions using PCA.

The amount of variance explained by each of the selected components is:
component 1: 0.243
component 2: 0.125
Total variance explained: 0.367

PCA PLOT 3D:

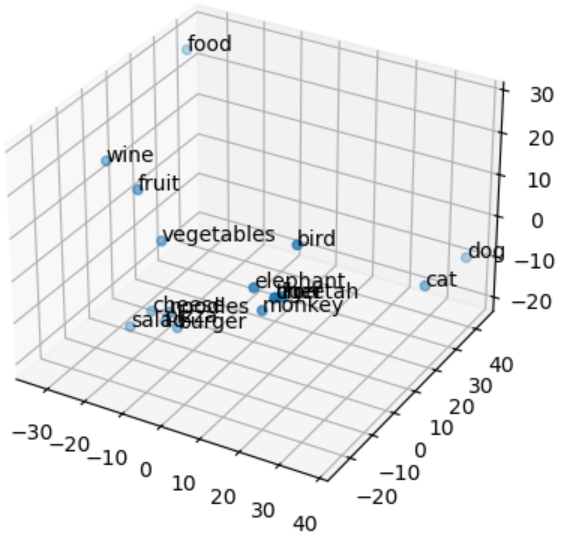

Figure 2: Plot of the words embeddings reduced to 2 dimensions using PCA.

The amount of variance explained by each of the selected components is:
component 1: 0.243
component 2: 0.125
component 3: 0.092
Total variance explained: 0.459

In the PCA plot, we can observe that words that are similar semantically, are also near in the plot of its embedding. For example, the words related to food like pizza, cheese, fruit, vegetables, etc, are near. Also, words related to animals like birds, monkeys, elephants, etc are also near in the plot. This allows us to see that the semantic embeddings are correct and have a geometric meaning, and we can work with distances to compare words and train models.

Green Exercice: Define a new set of words (at least 20 different words), and represent them in the feature space.

list of words used: Apple, Banana, Grapes, Pear, Orange, Melon, Tomato, Pineapple, Raspberry, Watermelon, Information, Data, Bit, Computer, Mouse, Tower, Screen, Music, Network, Phone

PCA PLOT 2D:

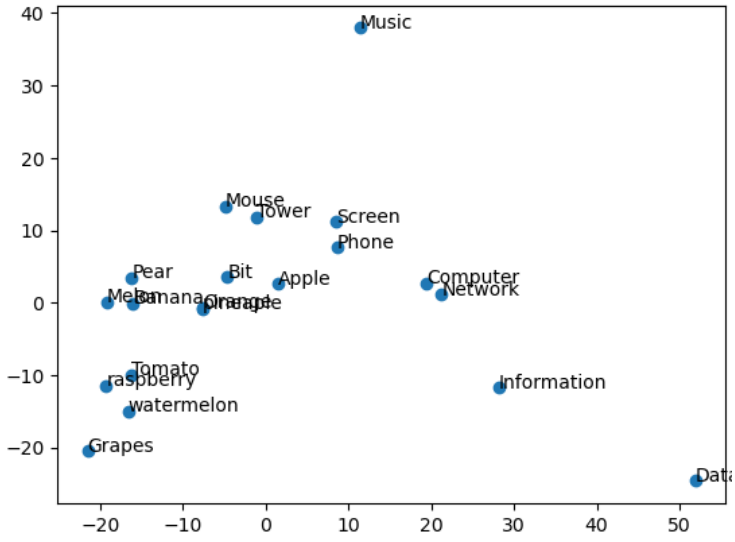

Figure 3: Plot of the words embeddings reduced to 2 dimensions using PCA.

The amount of variance explained by each of the selected components is:
component 1: 0.242
component 2: 0.124
Total variance explained: 0.367

PCA PLOT 3D:

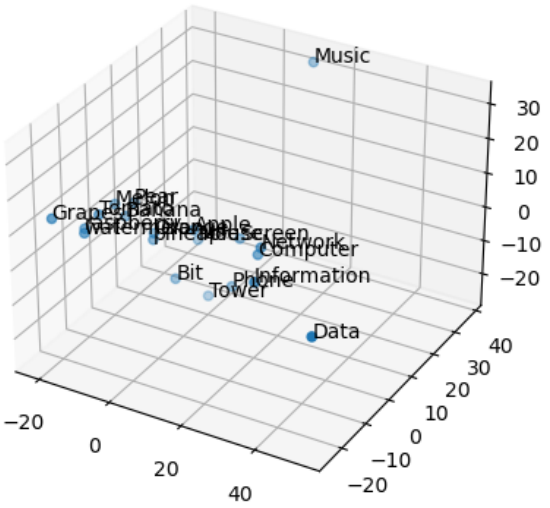

Figure 4: Plot of the words embeddings reduced to 2 dimensions using PCA.

The amount of variance explained by each of the selected components is:
component 1: 0.242
component 2: 0.124
component 3: 0.091
Total variance explained: 0.459

In the PCA plots we can appreciate that the food like banana, grapes, pear, melon, etc are close in the embedding space, also the words related with technology like mous, screen, phone, etc have similar embeddings. We can see that apple and bit are between the food and the technology, because they are polisemic words with more than one meaning, for example apple can be associated to the fruit or to the tech company. The words music and data are far away from other words, but this could be because of lose of information because of the PCA compression.

Exercise IV - Categorizing text with semantic similarity

1. Define a set of sentences, e.g., "I purchased a science fiction book last week. I loved this fragrance: light, floral and feminine. I purchased a bottle of wine."

Sentences

1. "Wandering through the cobbled lanes of the city, I was captivated by the delicate aroma emanating from a quaint fragrance boutique.",
2. "Her dresser was adorned with an array of scent bottles, among which a handcrafted essence from a remote village in France was her most cherished possession.",
3. "The art of scent crafting intrigued me, merging hints of amber, bergamot, and jasmine to create an aroma that felt intimately mine.",
4. "At the dawn of spring, the garden was alive with the sweet fragrances of cherry blossoms and lilacs, enveloping the air with nature's own bouquet.",
5. "To discover an essence that resonates with your soul is to capture the essence of cherished memories, each note a narrative of journeys, feelings, and aspirations."

2. Define a keyword, e.g., perfume

Keywords: "Perfume"

3. Calculate the similarity between each sentence and the keyword

$$\text{cosine similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

Similarity between sentence 1 and keyword 'perfume' is: 0.276

Similarity between sentence 2 and keyword 'perfume' is: 0.321

Similarity between sentence 3 and keyword 'perfume' is: 0.384

Similarity between sentence 4 and keyword 'perfume' is: 0.297

Similarity between sentence 5 and keyword 'perfume' is: 0.263

4. Could we filter out the sentences which are not related with the keyword?

We can filter the sentences by setting a threshold value to consider that the sentences that have a similarity value with the keyword lower than the threshold are not related to the keyword.

For example, we can set this threshold to **0.3** so that the related sentences that remain are:

- Sentence 1 with 0.321 similarity
- Sentence 2 with 0.384 similarity

Load the Alexa's review dataset, and filter out the reviews which are not associated with the "music" property

	rating	date	variation	verified_reviews	feedback
0	5	31-Jul-18	Charcoal Fabric	Love my Echo!	1
1	5	31-Jul-18	Charcoal Fabric	Loved it!	1
2	4	31-Jul-18	Walnut Finish	Sometimes while playing a game, you can answer...	1
6	3	31-Jul-18	Sandstone Fabric	Without having a cellphone, I cannot use many ...	1
8	5	30-Jul-18	Heather Gray Fabric	looks great	1
9	5	30-Jul-18	Heather Gray Fabric	Love it! I've listened to songs I haven't hear...	1
10	5	30-Jul-18	Charcoal Fabric	I sent it to my 85 year old Dad, and he talks ...	1
11	5	30-Jul-18	Charcoal Fabric	I love it! Learning knew things with it eveyda...	1
12	5	30-Jul-18	Oak Finish	I purchased this for my mother who is having k...	1
13	5	30-Jul-18	Charcoal Fabric	Love, Love, Love!!	1

Figure 5: First 10 elements of the filtered Alexa's Dataset without the music element

There are 2684 elements without the word "music" in it's review in the Alexa's dataset.