Felicitaciones! ¡Aprobaste!

Calificación recibida 100 %
Calificación del último envío 100 %
Para Aprobar 80 % o más

Ir al siguiente elemento

1. Why did we remove certain stop words? **1 / 1 punto**

● These words might make it more difficult for our model to learn as they are not always informative.

○ These words were not in our vocabulary, so they could not be used.

✓ Correcto
Correct. In general, it is advisable to test multiple approaches to see if these words may have some value. In our use case, removing them did not hinder performance.

2. What is the difference between Continuous Bag of Words (CBOW) and Skip-Gram (SG) approaches to Word2Vec? **1 / 1 punto**

● CBOW uses context to predict a given word, while Skip-Gram uses a given word to predict the context.

○ They are effectviely the same, they are just fed data in a different order.

○ CBOW uses multiple words to predict the context, while Skip-Grams use multiple words to predict a single word.

○ Skip-Gram uses context to predict a given word, while CBOW uses a given word to predict the context.

✓ Correcto
Correct! These models are similar in structure, but have inverted inputs and outputs.

3. Why is using a heatmap in our results not effective?

1 / 1 punto

☑ Half of the computations in the similariy matrix would be wasted.

✓ **Correcto**

Correct. In the heatmap, since it is symmetric, we only care about half of the results.

☑ It needs a lot of compute power, and it would be very difficult to read hundreds of words on both dimensions in a plot.

✓ **Correcto**

Correct. It is simply not feasible to have a thorough understanding of a model in this way.

□ The heatmap similarity scores are an approximation.

4. What are some downsides of using the approach we have, specifically in the case of new data? What about the data it already has?

1 / 1 punto

☑ Words not in our Word2Vec model are not handled directly, and will result in an error.

✓ **Correcto**

Correct. This is why our generated recipes are unique on each call!

☑ Word2Vec, like all NLP models, will be biased depedning on our training data. For example, our recipes are primarily Western-oriented. Recipes of traditional regions in less-popular areas in the world might not always be represented in this model.

✓ **Correcto**

Correct. Bias must always be considered when training a model on large amounts of any kind of data!

☑ All distance metrics have to be evaluated every time a new value is introduced. This is expensive and can be slow.

✓ **Correcto**

Correct. Alternative embedding approaches to visualize our data might be more useful, for example, a clustering algorithm such as HDBSCAN.

5. What approaches would you use to improve or change this model?

1 / 1 punto

aumentar las dimenciones del embeding

Correcto

I think a dimensionality reduction algorithm, such as an Autoencoder or PCA might be useful to deal with adding data to an exiting graph-like structure. Furthermore, it might also be clever to add in some artificial data such that less common words won't exactly throw an error, even if the model is less confident about these words (do you agree with this second point? Food for thought!).