Model Building

Throughout the assignment, I had a lot of difficulty in creating the various models from scratch as the size of the TedTalk inputs was larger and complex. Because of this, I decided to use the Python packages for the models if they existed. This included the Word2Vec and TF-IDF models that are available in gensim and sklearn respectively. The PPMI model calculations were easier to work with once the overall DataFrame was created, so that one was created from scratch. However, there were some difficulties in creating the co-occurrence matrix since the dataset was large and crashed my Google Colab.

Analysis

When looking at the output of the three different models, the ***Word2Vec*** model seemed the best in capturing the meaningful word relationships. This is determined by the overall higher cosine similarity scores from the different word pairs.

It might have been from the way that it was coded, but the PPMI model did not do a very good job at finding meaningful word relationships. This could be due to a handful of things with the first being the lack of co-occurrence. When building the model, the script would continue to crash the code when trying to build the rolling window for finding co-occurrence instances. This resulted in the word pairs needing to be occur in the text exactly (for example ('climate' , 'change') had but ('change' ,'climate') did not. The other reason for the poor word relationship output is that the value from the PPMI comes from a single value, where the other two models used cosine similarity to determine the relationship. Because of this, the PPMI model is excluded from any further analysis.

While both are used to represent text data, the key difference between Word2Vec and TF-IDF is that Word2Vec is a neural network-based method that captures semantic relationships between words by considering their context within a corpus. This contrasts to TF-IDF as it is a statistical measure that simply calculates the importance of a word based on its frequency within a document and its rarity across the entire document collection, without considering context.