Para la aplicación de word2vec me valí del siguiente tutorial:

https://www.kaggle.com/pierremegret/gensim-word2vec-tutorial

El cual toma como corpus un dataset de los diálogos del programa “the simpsons”.

Trate de re implementar mucha de la funcionalidad del programa para hacerle ver que entendí el código.

I. Implementacion usando corpus de los dialogos de “The simpsons”

Dentro de la carpeta adjuntada en drive se encuentra lo siguiente:

1. *simpsons_dataset:*

   Es el csv que contiene el corpus a usar

2. *word2vec_simpson.ipynb:*

   Toma simpsons_dataset, preprocesa el corpus y sirve como texto para entrenar al modelo “word2vec_simpsons” y lo guardo en word2vec_simpsons.model

3. *word2vec_simpsons.model:*

   Es el modelo guardado generado por el programa.

4. *output_simpsons_corpus.png*

   Utilizo algunas de las funciones del objeto modelo de gensims para demostrar la similitud entre ciertas palabras

II. Implementacion usando corpus de los dialogos de “The big bang theory”

Dentro de la carpeta adjuntada en drive se encuentra lo siguiente:

1. *carpeta “Transcripts”:*

   Contiene los diálogos en crudo de la serie

2. *createDataset.py:*

   Crea un script que toma todos los transcripts de la serie y genero el dataset datasetTBBT.csv

3. *datasetTBBT*

   Es el dataset que contiene el corpus de la serie

4. *word2vec_tbbt.ipynb:*

   Toma datasetTBBT, preprocesa el corpus y sirve como texto para entrenar al modelo “word2vec_tbbt” y lo guardo en word2vec_tbt.model

5. *word2vec_tbt.model:*

   Es el modelo guardado generado por el programa.

6. *output_tbbt_corpus.png*

   Utilizo algunas de las funciones del objeto modelo de gensims para demostrar la similitud entre ciertas palabras.

Al visualizar las similitudes entre palabras del modelo entrenado noto que todas son bastantes altas. Debido a que no he cursado ninguna materia relacionada con machine Lear Ning no puedo resolver o explicar este problema.

Espero pueda explicarme profesor.

Gracias