Laboratorio – Word Embeddings
CC3092 - Deep Learning y Sistemas Inteligentes

En este laboratorio trabajaremos con el siguiente conjunto de datos, que pueden descargar de:
https://www.kaggle.com/datasets/imnoob/reviews-dataset-cars-and-hotel?select=reviews_data.txt

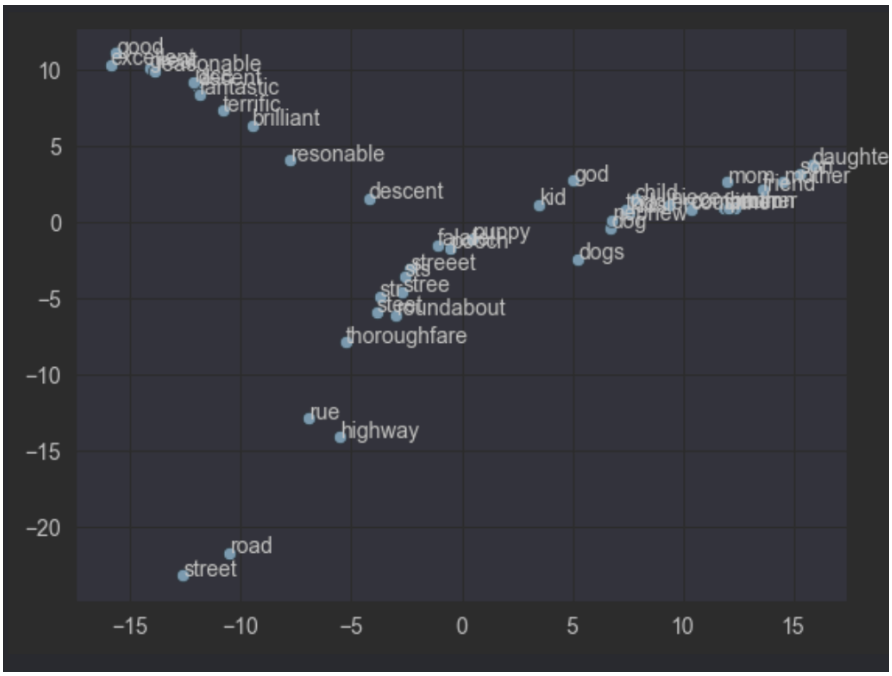
Trabajo en parejas:

Parte Práctica:

Utilizando el código proporcionado como base, realice las siguientes actividades:

a. Emplee la función **Word2Vec** de la librería **gensim** para generar embeddings a partir de los documentos del dataset. Elija los mejores valores a su opinion para cada uno de los parametros de la funcion (vector_size, window, min_count, workers).
b. Usando los embeddings generados, aplica la función **most_similar** para encontrar las 10 palabras más similares a cada una de las siguientes:
    - Street
    - Good
    - Dog
    - Mother
    - Bed
c. Extrae los embeddings de estas 55 palabras (las 5 palabras iniciales más sus 10 palabras similares cada una).
d. Aplica **PCA** para reducir la dimensionalidad de los embeddings a 2 componentes principales.

e. Crea un gráfico de dispersión (scatter plot) con estos 2 componentes y describe los resultados obtenidos. Su gráfico debería verse similar a este ejemplo:



Parte Teórica:

El constructor del modelo Word2Vec acepta dos parámetros importantes:

a. vector_size
b. window

Describa la función que cumple cada uno de estos parámetros dentro del algoritmo y disctua sobre las consecuencias de utilizar valores muy altos o muy bajos para cada uno de estos parámetros. Indique en qué situaciones podría ser apropiado ajustar de determinada manera (alto o bajo) estos valores o si, en ciertos casos, nunca es recomendable hacerlo.