WORD2VEC WITH SKIP-GRAM ARCHITECTURE

DATASET-

- Size of the Data – 10,000 from the given dataset
- Preprocessing the texts – Involves conversion to lower case, removal of special characters and punctuation, filtering words based on minimum frequency, creating a mapping from word to index and vice versa, noting down the vocabulary size.
- Preparation of Training Data – Creating the list of tokens for each word and pairs and labels which will be used for training. Each word is then converted to their id.

ARCHITECTURE-

Skip-Gram –

- Window Size – 10 words, to increase maximum word matching
- Negative Samples – 5, to reduce elimination of words for smaller data set.
- Embedding Dimension – 50, sufficient for smaller no of training data

Word2Vec Model –

- Embeddings – Created for context and target embeddings.
- Output Layer – Dense with sigmoid activation.
- Optimizer – Adam
- Loss – Binary Cross Entropy

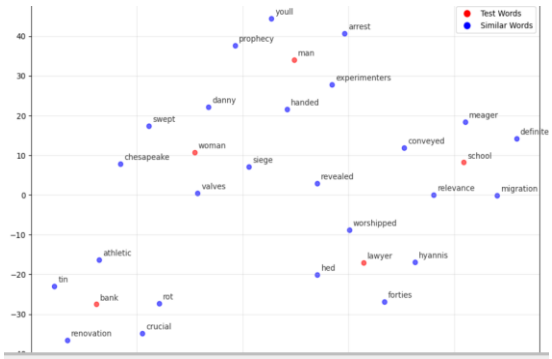
Training –

- Training the model on words and pairs of data.
- Number of epochs -5, with early stopping patience of 2.
- Data split into 90% training data, and 10% validation data.

Similarity –

- Cosine Similarity for test words.
- Sample over top 5 words.

RESULTS –




```
Word Similarities:

Similar words to 'man':
  youll: 0.5880
  experimenters: 0.5463
  arrest: 0.5047
  handed: 0.4760
  prophecy: 0.4693

Similar words to 'woman':
  valves: 0.5224
  swept: 0.4934
  siege: 0.4916
  danny: 0.4821
  chesapeake: 0.4703
```

- All the test words are marked in red while the words matched with highest similarity are in red.
- All the test words are well separated and the matched words for each words form a cluster around the test word. Indicating the formation of clusters of similar words.
- Observed cosine value is between 0.45 to 0.60 for each test word.

ANALYSIS –

- The observed range of cosine similarity for each word is in the range of 0.45-0.60. This indicates a decent amount of maturity of the model in predicting similar words.
- The similar words do not hold any special type of relation (eg: synonym, antonym, etc) which indicates similarity only implies higher co-occurrence from the given dataset.
- Formation of well defined clusters and retention with dimensionalty reduction implies the similar words are well separated from less similar words.

DIFFERENT TRIALS-

- Trials included experimentation with embedding vector size, decided on 50 given the size of dataset of 10,000 sentences. Due to the availability of resources.
- Trails were made for the window size, no of negative values and min count. The no of epochs was taken 5, with early stopping patience of 2 and sigmoid activation. Part of the metrics were decided based on the size of the model and dataset.
- Trial were conducted for checking the metric of optimization between precision, recall and accuracy. No notable differences were observed.

NOTES-

- The experiment was setup on HuggingFace Spaces using Nvidia T4 Small.
- The code for creating the model and training it were adjusted for the GPU using generative idea as there was a mismatch in the tensorflow version required for GPU.