Assignment – 1:

Task 1:

I have taken Gutenberg corpus, this corpus is a collection of texts available from Project Gutenberg, which is a digital library offering free access to a wide range of literary works, including many classic books. This corpus takes information/words from the books in Project Gutenberg. For example, the initial text taken from Jane Austen's novel "Emma".

Preprocessing steps taken:

1. Lowercased: All words in the sentences are converted to lowercase to ensure uniformity and avoid case sensitivity in word embeddings.

This is the code that do's this lowercasing:

```
sentences_list_lower = [[''.join([w.lower() for w in s]) for s in b] for b in sentences_list]
```

This step ensures that the model considers words in a case-insensitive manner.

2. Training Word Embeddings:

- CBOW Model: Trained using Word2Vec with Continuous Bag of Words (CBOW) approach.
- Skip Gram Model: Trained using Word2Vec with Skip-Gram approach.
- Pretrained Word Embeddings: Loaded Google News pre-trained Word2Vec embeddings.

This is the code that does this:

```
cbowModels = Word2Vec(sentences, vector_size=120, window=5, min_count=1, workers=4, sg=0)

skipgramModels = Word2Vec(sentences, vector_size=100, window=5, min_count=1, workers=4, sg=1)

wv = api.load("word2vec-google-news-300")
```

3. Saving Word Embeddings:

- The trained word embeddings are saved in Word2Vec format.

This is the code that does this:

```
cbowModels.wv.save_word2vec_format("cbowstep2.txt")

skipgramModels.wv.save_word2vec_format("skipgramstep2.txt")
```

4. Similarity Evaluation:

- Word embeddings are evaluated for similarity using the `evaluate_word_pairs` method.

This is the code that does this:

```
cbowEvaluation = cbowModels.wv.evaluate_word_pairs("10pairs.txt")

SkipgramEvaluation = skipgramModels.wv.evaluate_word_pairs("10pairs.txt")

googlenewsEvaluation = wv.evaluate_word_pairs("10pairs.txt")
```

These preprocessing steps set the foundation for training and evaluating word embeddings and exploring their semantic relationships.

Length of the sentences: 98552

Task 2:

I chose CBOW (Continuous Bag of Words) and Skip-Gram to get the word embeddings.

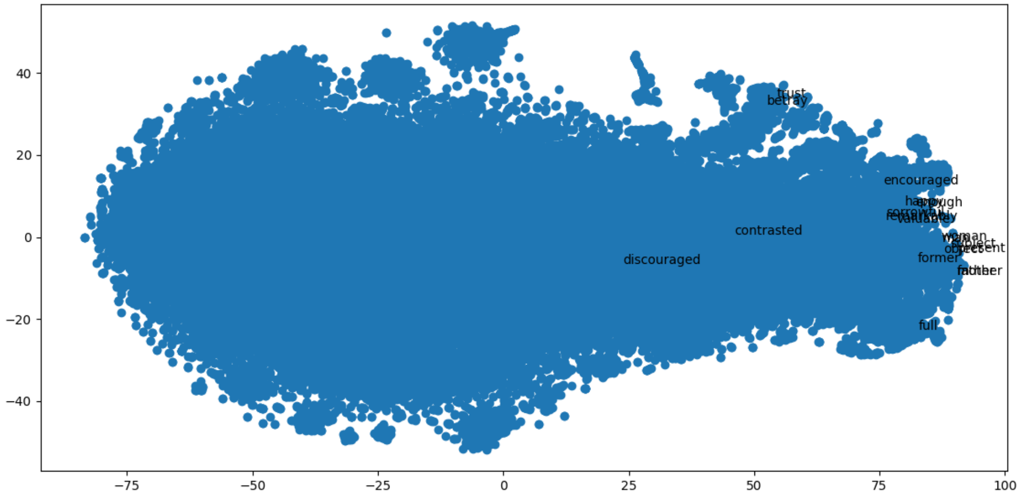
Reason I used?

- CBOW: It predicts the current word based on the context, which is suitable for scenarios where the meaning of a word is influenced by its context.

- Skip-Gram: It predicts the context words given the current word, which is beneficial when a word has multiple meanings or contexts.

Task 3:

Image generated from CBOW:



Discouraged: is positioned at 25, 2 and it is unique word out of the 20 words provided.

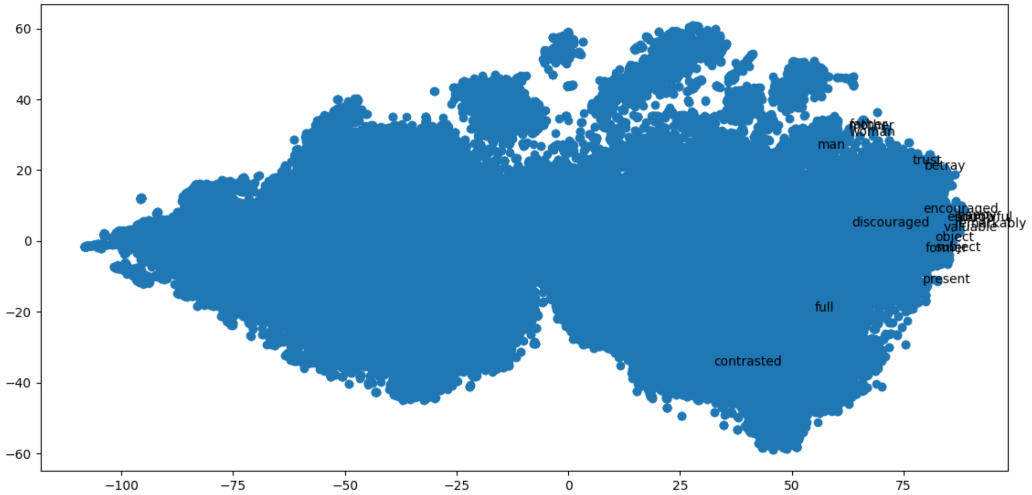
Contrasted: is positioned at 50, 0 and this is also a unique word out of the 20 words provided.

Trust & Betray: both are positioned at around 55, 35 and these 2 words are near to each other, meaning they often appear in similar contexts within sentences. The proximity reflects a semantic similarity, indicating that these words are likely to co-occur or share contextual relationships in various contexts.

Full: is positioned at 83, -20 and this is somewhat unique word.

Encouraged, enough, woman, man, happy, sorrowful, present, former, father, mother, valuable, remarkably, object, subject, interest: these all word is almost near to each other, x-axis co-ordinates varies from 75-100 and y-axis coordinates varies from -15 to 15, meaning they often appear in similar contexts within sentences. The proximity reflects a semantic similarity, indicating that these words are likely to co-occur or share contextual relationships in various contexts.

Image generated from Skip gram:



Contrasted: is positioned at 30, -30 and this is also a unique word out of the 20 words provided.

Trust & Betray: both are positioned at around 75, 20 and these 2 words are near to each other, meaning they often appear in similar contexts within sentences. The proximity reflects a semantic similarity, indicating that these words are likely to co-occur or share contextual relationships in various contexts.

Full: is positioned at 55, -20 and this is somewhat unique word.

Man, father, mother, woman: these all word is almost near to each other, x-axis co-ordinates varies from 55-70 and y-axis coordinates varies from 30 to 35, meaning they often appear in similar contexts within sentences. The proximity reflects a semantic similarity, indicating that these words are likely to co-occur or share contextual relationships in various contexts.

Discouraged, Encouraged, enough, happy, sorrowful, present, former, valuable, remarkably, object, subject, interest: these all word is almost near to each other, x-axis co-ordinates varies from 70-80 and y-axis coordinates varies from -10 to 15, meaning they often appear in similar contexts within sentences. The proximity reflects a semantic similarity, indicating that these words are likely to co-occur or share contextual relationships in various contexts.

Task – 4:

	CBOW	**Skip gram**	**Google News Evaluation**
Pearson Result	0.1021403486003611	0.4032604537644426	0.03949856266160449

The Pearson correlation coefficient is a measure of the linear correlation between two variables. In the context of word embeddings and similarity evaluation, a Pearson result closer to 1 indicates a stronger positive linear relationship, while a result closer to -1 indicates a strong negative linear relationship.

In the provided results:

- CBOW has a Pearson result of 0.1021, indicating a weak positive linear correlation between the model's predicted word similarities and the human-annotated similarities.

- Skip gram has a Pearson result of 0.4033, suggesting a moderate positive linear correlation.

- Google News embeddings have a Pearson result of 0.0395, indicating a weak positive correlation.

A higher Pearson correlation generally implies that the model's predicted similarities align better with human judgments.

In conclusion, the evaluation of word embeddings trained on the Gutenberg corpus reveals that Skip gram exhibits a stronger positive correlation with human-annotated word similarities compared to CBOW. This suggests that Skip gram, when applied to literary texts from the Gutenberg corpus, may better capture semantic relationships in the evaluated word pairs. However, the choice between CBOW and Skip gram may be context-dependent, and further fine-tuning could enhance the performance of these models on literary data. The comparison with Google News embeddings indicates that pre-trained embeddings might differ in semantic representations, emphasizing the importance of domain-specific training for optimal results in literary analysis.

Task – 5:

These results showcase the semantic similarities captured by CBOW, Skip Gram, and Google News embeddings for selected words:

1. Discouraged:

   - CBOW: Captures words with negative sentiments like "unpunished" and "dismayed."

   - Skip Gram: Emphasizes words related to emotional states and actions such as "apprehending" and "weakened."

   - Google News: Shows a mix of synonyms and closely related terms, e.g., "discouraging" and "dissuaded."

2. Managed:

   - CBOW: Associates with words like "rail" and "floats," indicating diverse contextual connections.

   - Skip Gram: Reflects words related to actions and reactions, e.g., "scolded" and "grin."

   - Google News: Displays a range of terms related to attempts and accomplishments.

3. Full:

   - CBOW: Links with words conveying abundance and positivity, such as "pure" and "warm."

   - Skip Gram: Highlights synonyms and terms with positive connotations, e.g., "pleasant" and "precious."

   - Google News: Includes variations of "full" and related terms.

4. Enough:

   - CBOW: Associates with words expressing quantity and intensity, like "too" and "quite."

   - Skip Gram: Indicates a range of intensities, from "impossible" to "remarkably."

   - Google News: Features variations of "enough" and related terms.

5. Man:

   - CBOW: Links with gender-specific terms and societal roles, e.g., "woman" and "gentleman."

   - Skip Gram: Reflects gender-related terms and additional associations like "lad" and "lion."

   - Google News: Shows gender-related terms and age-specific associations.

These comments highlight the diverse contextual relationships captured by each model, providing insights into the nuances of word embeddings generated by CBOW, Skip Gram, and Google News pre-trained embeddings.