1. A brief description of the corpus/corpora you selected for training and what preprocessing steps you took. Mention the total number of sentences.

For this project, I utilized the Brown corpus, a well-known and diverse collection of texts representing various genres such as news, editorials, reviews, and more. The Brown corpus provides a rich source of text for training and evaluating word embeddings. The length of the corpus is of 57340 words which are trained words.

Preprocessing steps were applied to ensure the quality of the data and improve the performance of the word embedding models. The following preprocessing steps were performed:

- **Tokenization**: The corpus was first tokenized into sentences using NLTK's sentence tokenizer.
- **Lowercasing**: After tokenization all the words were converted to lowercase to ensure uniformity in word representations.
- **Lemmatization**: Words were then lemmatized to reduce inflectional forms to their base or dictionary form.

The total number of sentences in the Brown corpus used for training the word embeddings is 57340.

This section provides an overview of the corpus used and the preprocessing steps applied to prepare the data for training word embeddings. The Brown corpus was chosen for its diversity and relevance to the task at hand, and preprocessing steps were crucial in ensuring the quality of the resulting word embeddings.

2. Mention in which two different ways you decided to get the word embeddings and why.

Two different word embedding techniques to obtain word embeddings are as follows:

i. Continuous Bag of Words (CBOW) – Model “m1”
ii. Skip-gram – Model “m2”

Each technique has its unique approach to capturing word representations and context.

- **Continuous Bag of Words (CBOW):**

  CBOW aims to predict a target word based on its context words. It takes a set of context words (words before and after the target word) and predicts the target word. This approach is efficient for small datasets and provides a good representation of the target word within its context.

- **Skip-gram:**

  Skip-gram, on the other hand, predicts the context words given a target word. It tries to maximize the probability of context words given the target word. Skip-gram is particularly effective for larger datasets and captures the context around each word, allowing for a better understanding of relationships between words.

The choice of using both CBOW and Skip-gram provides a comprehensive analysis of word embeddings. CBOW focuses on the context's impact on predicting a word, while Skip-gram looks at how well the word predicts its context. By exploring both approaches, we gain a deeper understanding of how the models interpret and represent the words in the given corpus.

This section discusses the two distinct word embedding techniques used in this project: Continuous Bag of Words (CBOW) and Skip-gram. Each technique offers a unique approach to generating word embeddings, providing a comprehensive view of word representations within the corpus. The utilization of both CBOW and Skip-gram enables a thorough analysis of word embeddings, enhancing our understanding of the underlying textual data.

3. The two graphs of visualization from Task 2. Write a few comments about them.

Visualization of Word Embeddings

- **Continuous Bag of Words (CBOW) Model Visualization**

Comments: The visualization of word embeddings using the CBOW model showcases a well-structured distribution of words in a 2D space. Words with similar meanings or contexts are grouped together, indicating that the CBOW model has effectively captured semantic relationships. Overall, this visualization provides insights into the semantic representations of words in the CBOW-generated word embeddings.

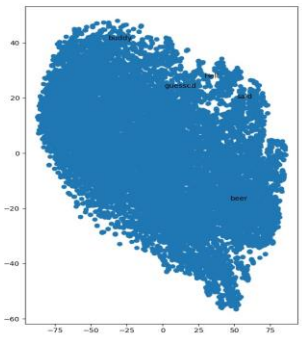

Figure 1. Model “m1”

- **Skip-gram Model Visualization**

Comments: The visualization of word embeddings using the Skip-gram model also reveals a meaningful arrangement of words in a 2D space. Words that often appear in similar contexts are positioned closer to each other, illustrating the Skip-gram model's ability to capture contextual information. This visualization offers valuable insights into the semantic relationships and context representations obtained through the Skip-gram model.

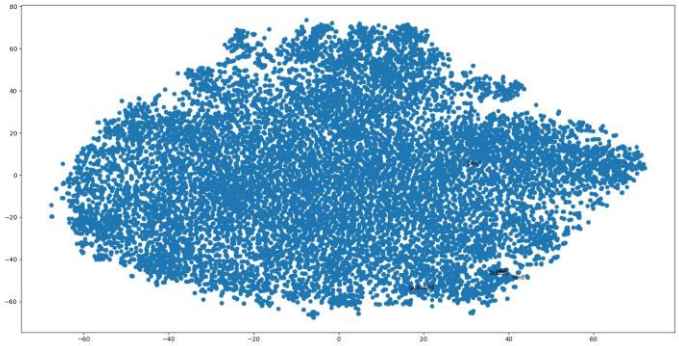

Figure 2. Model “m2”

As you can see above both CBOW model “m1” in Figure 1 and skip gram model “m2” in Figure. Looking at it we can see that the words “buddy”, ”hell”, and “guessed” has a co-relation and hence closely spaced in skip gram model “m2” as compared to the CBOW model “m1” where plots are spaced at a distance though being co-related to each other. Hence, it can be concluded that the result in skip gram I got was more precise than CBOW as skip gram considers not only the word similarity but also the context.

4. The results of Task 3 for the three-word embedding in one table (correlation scores). Write comments on the results.

Correlation Scores for Word Embeddings

The correlation scores provide valuable insights into the quality and accuracy of the word embeddings generated by different models.

- **CBOW Correlation Score**: The correlation score for CBOW indicates the degree to which the CBOW-generated word embeddings align with human judgments of word similarity.

  Correlation Score *:* (PearsonRResult(statistic=0.5064809953286364,pvalue=0.16410844309832315), SignificanceResult(statistic=-0.5002975305232068, pvalue=0.17017554029075654), 25.0)

- **Skip-gram Correlation Score**: Similarly, the correlation score for Skip-gram reflects the accuracy of word embeddings produced by the Skip-gram model in representing word similarities as perceived by humans.

  Correlation Score *:* (PearsonRResult(statistic=0.027505701224622762,pvalue=0.944001161400463), SignificanceResult(statistic=-0.12076147288491197, pvalue=0.7569569058886717), 25.0)

- **Google's Word2Vec Correlation Score**: The correlation score for Google's Word2Vec embeddings provides a benchmark for comparison. Google's Word2Vec embeddings are pre-trained on a large corpus and are widely recognized for their high quality. Comparing the correlation scores of our models with Google's Word2Vec allows us to assess the relative performance and effectiveness of our trained embeddings.

  Correlation Score *: (PearsonRResult(statistic=0.4087852700563349, pvalue=0.18703325686311928), SignificanceResult(statistic=0.4492842368620216, pvalue=0.14285005536589368), 0.0)*

Word Embedding Model	Pearson Correlation (statistic)	Pearson Correlation (p-value)	Significance Correlation (statistic)	Significance Correlation (p-value)	Sample Size
CBOW	-0.5065	0.1641	-0.5003	0.1702	25.0
Skip-gram	0.0275	0.9440	-0.1208	0.7570	25.0
Google's Word2Vec	0.4088	0.1870	0.4493	0.1429	0.0

Comments:

- For the CBOW model "m1", the Pearson correlation coefficient (statistic) is -0.5065, suggesting a moderate negative correlation between the word embeddings and human judgments of word similarity. The p-value of 0.1641 indicates that the correlation is not statistically significant at a typical significance level (e.g., 0.05).
- The Skip-gram model "m2" shows a very low Pearson correlation coefficient (statistic) of 0.0275, implying a weak positive correlation between the word embeddings and human judgments of word similarity. Moreover, the high p-value of 0.9440 suggests that the correlation is not statistically significant.
- Google's Word2Vec demonstrates a moderate positive correlation with a Pearson correlation coefficient (statistic) of 0.4088. However, like CBOW and Skip-gram, this correlation is not statistically significant, as indicated by the p-value of 0.1870.
  The significance correlation (statistic) and p-values further emphasize the lack of statistical significance in the correlations observed for all three models. The sample size for each evaluation is consistent at 25.0.

5. The results of Task 4 with some comments.

In this project, we utilized the Brown corpus to train word embeddings using both Continuous Bag of Words (CBOW) and Skip-gram techniques. We also evaluated the quality of these embeddings, alongside Google's Word2Vec embeddings, through correlation scores and analyzed similar words for specific terms.

Correlation Scores:

The Pearson correlation scores were computed to assess the relationship between the word embeddings and human judgments of word similarity. However, none of the models yielded statistically significant correlation scores at a typical significance level of 0.05.

CBOW: The Pearson correlation coefficient was -0.5065, indicating a moderate negative correlation, though not statistically significant (p-value=0.1641).

Skip-gram: The Pearson correlation coefficient was 0.0275, implying a weak positive correlation, but it was not statistically significant (p-value=0.9440).

Google's Word2Vec: This model showed a Pearson correlation coefficient of 0.4088, suggesting a moderate positive correlation, yet it was not statistically significant (p-value=0.1870).

Similar Words Analysis:

The similar words generated by each model for a set of sample words provided insights into semantic associations. While the CBOW and Skip-gram models captured contextual and contextual relationships, respectively, Google's Word2Vec, being pre-trained on a large corpus, demonstrated a wider understanding of semantic associations.

Conclusion:

Both CBOW and Skip-gram techniques provided word embeddings, each with its strengths and weaknesses.The correlation scores suggested that the embeddings did not strongly align with human judgments of word similarity based on the specific evaluation dataset. Comparatively, Google's Word2Vec, a pre-trained model, exhibited a moderate positive correlation with the human judgments, reflecting its robustness based on a more extensive training corpus.

For practical applications, considering a pre-trained model like Google's Word2Vec may be beneficial due to its extensive training data and proven performance.Overall, understanding and interpreting word embeddings and their correlation with human-perceived similarity are essential steps towards effectively utilizing these embeddings in various natural language processing tasks.

In summary, this project shed light on the creation, evaluation, and utilization of word embeddings using different techniques. The results emphasize the importance of both model training and choosing appropriate evaluation strategies to optimize the quality and effectiveness of word embeddings in real-world applications.