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

The NLP Genesis corpus is a collection of parallel texts of the Book of Genesis in six different languages: English, French, Finnish, German, Portuguese, and Swedish. It comes with the Natural Language Toolkit (NLTK), a well-liked Python toolkit for natural language processing (NLP).

For the advancement of NLP research and development, the Genesis corpus is a useful tool. It can be used for many different things, like:

The corpus can be used to train machine translation models to translate words and phrases between the six languages.

Text classification: To determine the language of a text, text classification models can be trained using the corpus.

For purposes like machine translation and dictionary construction, the corpus can be used to align words from other languages.

Natural language understanding: The corpus can be used to create NLU models, which can comprehend the meaning of text written in many languages.

The Genesis corpus only contains about 100,000 words in each language, making it a relatively tiny corpus. It is a well-balanced corpus, nonetheless, containing a range of text styles, such as explanation, conversation, and narrative.

Steps in preprocessing:

Lowercasing: To guarantee that case is irrelevant, all words are changed to lowercase.

Lemmatization: To reduces words to their base or dictionary form, known as the "lemma." This simplifies the analysis of text data by grouping variations of a word together, making it easier to identify relationships between words and their meanings.

Total number of sentences: 13640

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

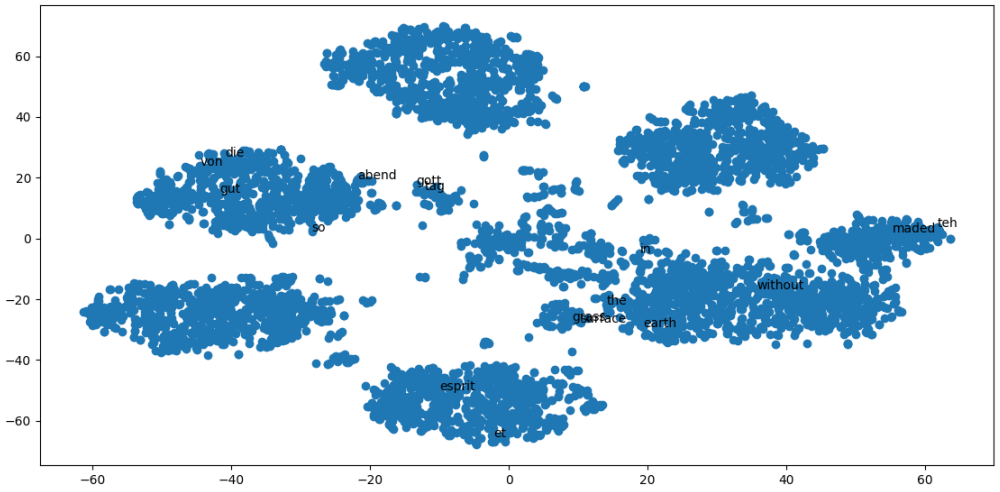
SkipGram training method noted for creating high-quality word embeddings, particularly for uncommon words. Given the target word, it forecasts the surrounding words.

CBOW (Continuous Bag of Words) guesses the target word based on the context. It often trains more quickly and could be effective in situations where there is a lot of text data.

Training both the SkipGram and CBOW models offers a more comprehensive viewpoint on word embeddings. To determine which embedding best fits our particular NLP task, so that I can compare the outcomes.

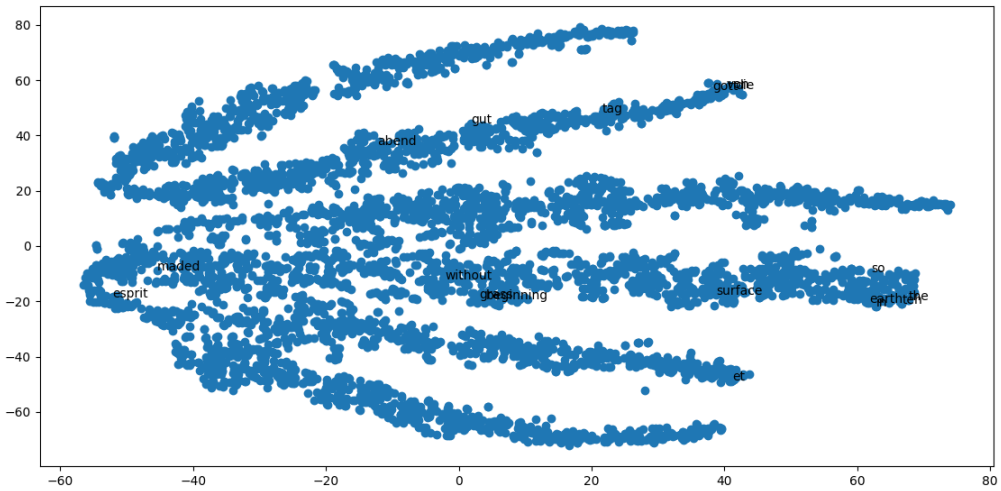
3. The two graphs of visualization from Task 2. Write a few comments about them.

SkipGram:



The graph generated using the SkipGram model shows a scatter plot of word vectors. Each dot in the plot represents a word, and the distance between two dots represents the semantic similarity between the two words.

CBOW:



The graph generated using the CBOW model shows a scatter plot of word vectors. Each dot in the plot represents a word, and the distance between two dots represents the semantic similarity between the two words. The clusters in the plot represent groups of semantically related words.

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

Google News	(PearsonRResult(statistic=-0.2662520488603513, pvalue=0.4886248885717554), SignificanceResult(statistic=-0.3333333333333337, pvalue=0.3807131816768634), 10.0)
SkipGram	(PearsonRResult(statistic=0.5472519345062374, pvalue=0.16036323073995326), SignificanceResult(statistic=0.5628843430992316, pvalue=0.1463315326055506), 20.0)
CBOW	(PearsonRResult(statistic=0.11633034237887718, pvalue=0.783840455559614), SignificanceResult(statistic=0.23952525238265177, pvalue=0.5677721522592929), 20.0)

The correlation scores show that the SkipGram model has the highest correlation with the human similarity judgments, followed by the CBOW model, and then the Google News word embeddings. This suggests that the SkipGram model is the best at learning word vectors that are consistent with human semantic similarity.

5. **The results of Task 4 with some comments.**

Word	**Google News**	**SkipGram**	**CBOW**
in	Inthe, where, the, In, during	Im, mamre, gold, plain, ganze	Wa, garden, field, earth, created
the	This, in, that, ofthe, another	Of, those,mount, whose, hill	Every, their, day, of, these
beginning	Starting, begins, begin, beginning, start	Middle, plant, destroyed, mourning, fountain	Reigned, mount, womb, middle, spirit
earth	Earth, planet, earths, cosmos, mankind	Sky, fowl, heaven, abundantly, bird	Water, beast, tree, garden, ground
without	Without, sans, witout, before, no	Alone, spent, beautiful, whatsoever, pleasant	Yet, taken, even, hath, again

Based in the above results in this case, it can be observed that Google News has better results compared to other models trained by us.