Skip-gram with Negative Sampling – Assignment Report

---

1. Introduction

In this project, I learned how to create word vectors using the Skip-gram model with Negative Sampling.
The model was trained on Wikipedia text.
Because training was slow on my laptop, I used Google Colab to run the final training and tests.

---

2. Method Dataset

I used the Wikipedia dump (enwik9).
The text was cleaned and split into words.

Model

Skip-gram model

Negative Sampling

Vector size: 300

Tool used for training: Gensim in Google Colab

---

3. Comparison with Gensim

My trained vectors were compared with Gensim's pretrained Word2Vec model.
We used cosine similarity to check how close the vectors are.

Both models used 300-dimensional vectors, so comparison was correct.

---

4. Word Analogy Test

I tested this:

king – man + woman = ?

The model predicted queen,
which shows that the word vectors learned good meaning.

---

5. Bias Detection

I checked if some words are closer to “he” or “she”.

Results

Word	Bias (he − she)
Doctor	-0.171
nurse	-0.134
engineer	-0.058
teacher	-0.139

Meaning:
Negative values mean the word is closer to “she”.
This shows that word vectors can learn gender bias from text.

---

6. Conclusion

In this project:

I implemented Skip-gram with Negative Sampling

Trained word vectors using Wikipedia data

Compared them with Gensim vectors

Tested them using:

cosine similarity

word analogy

bias detection

The results show that word embeddings learn word meaning well,but they can also learn bias from data.but they can also learn bias from data.