Task 1: Preparation and Training

The task involved training **Word2Vec** and **GloVe** models using the news corpus from the **NLTK** library, implementing a function to dynamically modify the window size during training (with a default **window size of 2**), and experimenting with different models **(Skip-gram, Skip-gram with negative sampling, and GloVe)** to analyze their performance.

Task 2. Model Comparison and Analysis

Model	Window Size	Average Training Loss	Total Training Time	Syntactic Accuracy	Semantic accuracy
Skip-gram	2	12.21	7 mins 42 secs	0.00%	0.00%
Skip-gram (NEG)	2	3.34	6 mins 52 secs	0.00%	0.00%
GloVe	2	28.17	2 mins 18 secs	0.00%	0.00%
GloVe (Gensim)	2	-	-	15.50%	19.74%

Dataset: fit.vut.cz/person/imikolov/public/rnnlm/word-test.v1.txt

According to the table, the **GloVe** model is much faster compared to the Skip-gram models in terms of training time. However, **GloVe** also has the highest training loss.

Additionally, the models were tested for semantic and syntactic accuracy using the Word Analogies dataset, specifically the capital-common-countries for semantic tasks and past-tense for syntactic tasks.

The low accuracies for the **custom-trained models (Skip-gram, Skip-gram with Negative Sampling, and GloVe)** are expected due to the limitations of the corpus used. The dataset is not rich enough for meaningful word analogies, especially for specific tasks like syntactic and semantic analysis. In contrast, the **pre-trained GloVe model from Gensim** performs reasonably well on both tasks.

Model	Skip-gram	Skip-gram (NEG)	GloVe	GloVe (Gensim)	Y_true
MSE	31.63	30.85	31.41	27.86	0.00
Spearman Correlation	0.0030	0.1599	-0.0057	0.6019	-

Dataset: [WordSim353 - Similarity and Relatedness]

The models were also evaluated for their correlation with human judgments of word similarity **(Spearman Correlation)** and **Mean Squared Error (MSE)**.

The results in the table show that the **pre-trained GloVe model (from Gensim)** performs much better than the custom-trained models. This is likely because it was trained on a much larger and diverse corpus, which helped it create better word vector representations.

Task 3. Search similar context- Web Application Development

The goal of Task 3 was to create a simple web application that allows users to input search queries and retrieve the top 10 most similar words based on their vector representations. For this task, the **Skip-gram model (without negative sampling)** was chosen. The web application was built using **Flask**.

Below are the screenshots of the sample web application:

1. The homepage.
2. The page displayed when the user clicks search without entering a word, showing a notification prompting them to enter a word.
3. The result page for the sample word "global," displaying the top 10 most similar words.


Word Similarity Search
127.0.0.1:5000
Word Similarity Search
Enter a word
Search
© 2025 Word Similarity Search. All rights reserved.


Word Similarity Search
127.0.0.1:5000
Word Similarity Search
Enter a word
Search
Please enter a word to search.
© 2025 Word Similarity Search. All rights reserved.


Word Similarity Search
127.0.0.1:5000
Word Similarity Search
Enter a word
Search
Top 10 Similar Words to "global"
automotive,
<phx>
jump,
scientific
(shows)
henan
startup
combined,
brazilian
3,816,580
© 2025 Word Similarity Search. All rights reserved.