sminerport /
word2vec-skipgram-tensorflow

Code Issues Pull requests Actions Projects Security Insights

Word2Vec Skip-Gram model implementation using TensorFlow 2.0 to learn word embeddings from a small Wikipedia dataset (text8). Includes training, evaluation, and cosine similarity-based nearest neighbors

scottminer.netlify.app

MIT license

1 star 1 fork 1 watching 2 Branches 0 Tags Activity

Public repository

2 Branches 0 Tags Go to file t Go to file + Add file Code ···

sminerport Merge pull request #2 from sminerport/add-new-files ··· 8c79f97 · last year

src	hit it with the formatter	last year
.gitignore	update git ignore and update word2...	last year
LICENSE	Initial commit	last year
README.md	Update README.md	last year

README MIT license

Word2Vec (Word Embedding) with TensorFlow 2.0

This repository contains an implementation of the Word2Vec algorithm using TensorFlow 2.0 to compute vector representations of words. The Word2Vec model used is the Skip-Gram model, which is trained on a small chunk of Wikipedia articles (the text8 dataset).

Background

Word2Vec is a popular word embedding technique that represents words as vectors in a high-dimensional space. These embeddings can be used in various natural language processing tasks, such as sentiment analysis, document classification, and machine translation. The main idea behind Word2Vec is that words with similar meanings tend to occur in similar contexts.

For more information on Word2Vec, please refer to the following research paper: Mikolov, Tomas et al. "Efficient Estimation of Word Representations in Vector Space.", 2013

Getting Started

To run the Word2Vec implementation, simply clone this repository and execute the `word2vec.py` script using Python 3.

Prerequisites

- Python 3
- TensorFlow 2.0
- NumPy
- urllib
- zipfile

Text8 Dataset

This project uses the Text8 dataset for natural language processing tasks. The dataset is not included in the repository to keep the repository size small. You can download the dataset using the provided Python script or by downloading it manually.

Download using Python script

1. Run the `download_text8.py` script in the project folder:

```
python download_text8.py
```

This script will download the text8.zip file and save it in the text8_dataset folder. If the file already exists, it won't be downloaded again.

Manual download

1. Download the text8.zip file from the following URL: http://mattmahoney.net/dc/text8.zip
2. Create a folder named `text8_dataset` in the project directory and move the downloaded `text8.zip` file into it.

Implementation Details

- The text8 dataset of Wikipedia articles is downloaded and processed to create a vocabulary of words.
- Rare words with occurrences below the specified threshold are removed.
- A Skip-Gram model is trained on the dataset for a specified number of steps using Stochastic Gradient Descent (SGD) optimization and Noise Contrastive Estimation (NCE) loss.
- The model's performance is evaluated periodically by finding the nearest neighbors of a set of test words based on their vector representations.

Author

- Aymeric Damien - GitHub

License

Releases

No releases published

Packages

No packages published

Languages

- ● **Python** 100.0%