Word2Vec representation

Project ID

Project ID: 41

Github Link

https://github.com/Anirudh5/Word2Vec

Team Members

1. Samyak Jain
2. Raghu Vamshi
3. Aayush Sanghavi
4. Anirudh Sharma

Main Goal(s) of the Project

- Generate vector representations for words using both skip gram and CBOW algorithms.
- Analysing and comparing results from generic (like Wikipedia or Google News) as well as domain specific (medical or legal) datasets with varying window sizes.

Problem Definition

Word2vec is a group of related models that are used to produce word embeddings. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space. The problem is to generate such word embeddings using skip gram and CBOW models on different types of datasets.

Expected results

- Examples of linguistic regularities captured by the model. For example:
    - vector('Paris') - vector('France') + vector('Italy') ~ vector('Rome')
    - vector('king') - vector('man') + vector('woman') ~ vector('queen')
- Examples of clusters of similar words.
- Examples of difference in linguistic regularities captured by the domain specific vectors and generic vectors.

Tasks distribution

Team 1	Team 2
Members: ● Samyak Jain ● Anirudh Sharma	Members: ● Raghuvamshi Reddy ● Aayush Sanghavi
Tasks: 1. Train the CBOW model on both datasets 2. Scraping of domain data from websites and writing parsers for the scraped data.	Tasks: 1. Train the skip-gram model on both datasets 2. Analysing the linguistic regularities captured by both the word to vec algorithms.

Project milestones and proposed timeline

Date	**Milestone**
30th March	1. Complete scraping of data 2. Write data parsers 3. Begin coding of CBOW and skip-gram algorithms
6th April	1. Finish coding and debugging of models 2. Begin training of models on both the datasets
13th April	1. Finish training of both the models on both the datasets 2. Get preliminary results on trained models
20th April	1. Plot appropriate analysis graphs 2. Document final results