Spotle Masterclass Project

Implementation of Word2Vec

Deliverables

This project would need you to develop a Word2Vec CBOW model

Deliverables

- Create a Word2Vec, CBOW model
- We will give you the dataset that you are required to use for this project work. You need to use the dataset for training and testing purposes.

Deliverables

- Following are the deliverables at each step
  1. Save the model (i.e. vector of all words) that you have created in a file.
  2. Your program should have the functionality of uploading the saved model.
  3. Your program should have the functionality that would return vector of a given word.
  4. Your program should provide functionality for comparing cosine and Euclidean distance between two given words.
  5. Your program should have the functionality that would return ‘n’ closest words of a given word for any given number n.

Reference

You may refer to the following information where original work by Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean and David Meyer have been documented. This may help you implementing your required Word2Vec model.

Efficient Estimation of Word Representations in Vector Space
https://arxiv.org/pdf/1301.3781.pdf

How exactly does word2vec work?
http://www.1-4-5.net/~dmm/ml/how_does_word2vec_work.pdf

Also watch Document Clustering - Part 1 and Document Clustering - Part 2 lecture videos again.

Marking System

- This project work would carry **100** marks.
- Out of the 100 marks for this project **70** marks will be given for the concept of this project implementation, describing the methodologies to be used and the reason for application of the methodologies, Name and short description of the Python libraries to be used, algorithm/ pseudocode etc. And **30** marks will be given for writing the Python code.

Usage Of Dataset/ T&C

The dataset to be used for this project is TripAdvisor Hotel Review Dataset.

The dataset contains around 20K reviews and ratings of hotels from TripAdvisor site.

Use review text for training of your Word2Vec CBOW model.

As suggested by the owner of the dataset, add the following citation/ acknowledgement at the end of your report where you have used this dataset.

“Barkha Bansal. (2018). TripAdvisor Hotel Review Dataset [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1219899"

Disclaimer: In the course of your project work you must not violate any copyright. In case you have violated any copyright in the course of your work Spotle.ai, Spotle Learn, NumeroSeven Technologies, Any part, division, instructors, mentors, employee, management of the organization will not be responsible for the violation. You will be the sole responsible person in the event of any copyright or contract violation.