Advanced NLP Assignment-1

T H Arjun, CSD, 2019111012

All the assignment requirements were met such as:

- Making the SVD Model
- Making the CBOW Model
- Comparison with pre-trained Gensim Model on Google News Corpus (word2vec)
- Top 10 words for 5 words TSNE Plots for all three models
- Top 10 words for "camera" for all three models
- Report with comparisons

Report is included as `Report.pdf`

Reproducing results

Downloading the data

If you want to reproduce the results without doing all the steps from scratch then you can download the zip file from the following link which has all the models that were produced by the notebooks and you could run the visualisation step only, or get word embeddings. I have also uploaded the pytorch checkpoint (that was trained for 50 epochs) of the cbow model here that can be loaded for further use here.

Installing the Python Packages

```
pip install -r requirements.txt
```

Steps

- Run the notebook `preprocess.ipynb` to make a dataset from the given dataset that is cleaned. The cleaning process includes lowercasing, removing stopwords, emojis, punctuation and tokenising to sentences. This will create `corpus.json` inside `data` folder
- Run the notebook `SVD.ipynb` to create the SVD Model. The parameters at the top of the notebook can be changed according to compute available
- The last notebook would have created `svd_model.json` and `svd_vocab.json`, now run `visualize.ipynb` with `CBOW=False` to produce visuals for SVD.
- Run the notebook `cbow.ipynb` to create the CBOW Model. The parameters at the top of the notebook can be changed according to compute available
- The last notebook would have created `cbow_model.json` and `cbow_words.json`, now run `visualize.ipynb` with `CBOW=True` to produce visuals for CBOW.
- Run the notebook `gensim.ipynb` to visualise with pre-trained vectors.

Running on HPC Cluster

The notebooks require a lot of computing. Especially the CBOW Model. Training took 3 days 21 hrs on ADA Cluster with a 1080Ti for 50 epochs (pytorch cuda). Thus if you want to run the code for many epochs I recommend using papermill to run the notebooks using sbatch. A sample sbatch script has been included as `run.sh`