Report

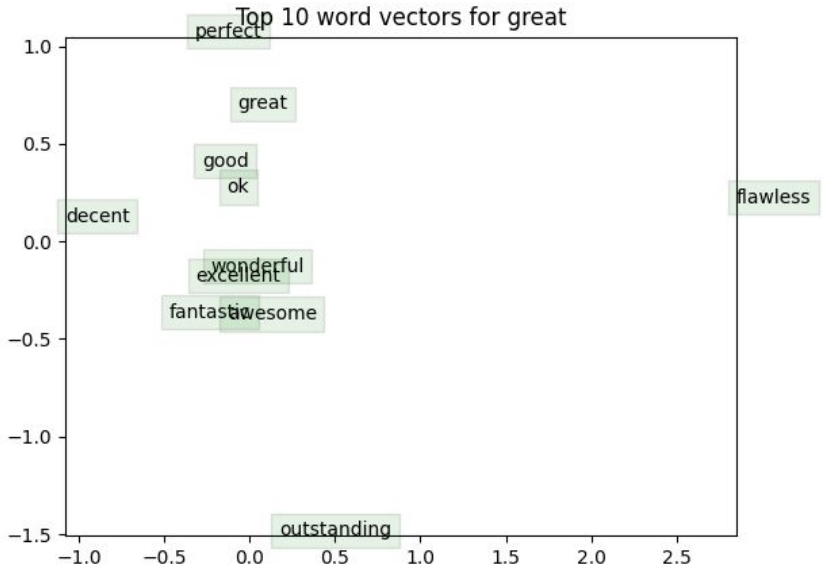
Jashn Arora

2018114006

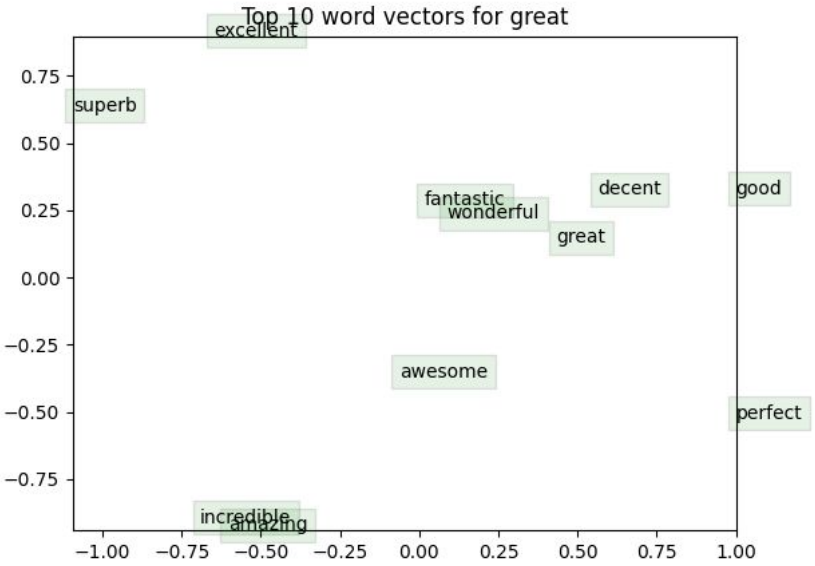
- Both Skip Gram and CBOW models were trained with negative sampling, with following parameters:
  - Embedding length : 100
  - Window Size : 3
  - Min Frequency of words: 3
  - Number of negative samples for each positive sample: 10
  - No. of epochs: 2
  - Dataset: 50k reviews

Q1 Display the top-10 word vectors for 5 different words (a combination of nouns,verbs, adjectives etc) using the above pre-trained models (1,2) using t-SNE (or such methods).

➢ **Word 1: Great**

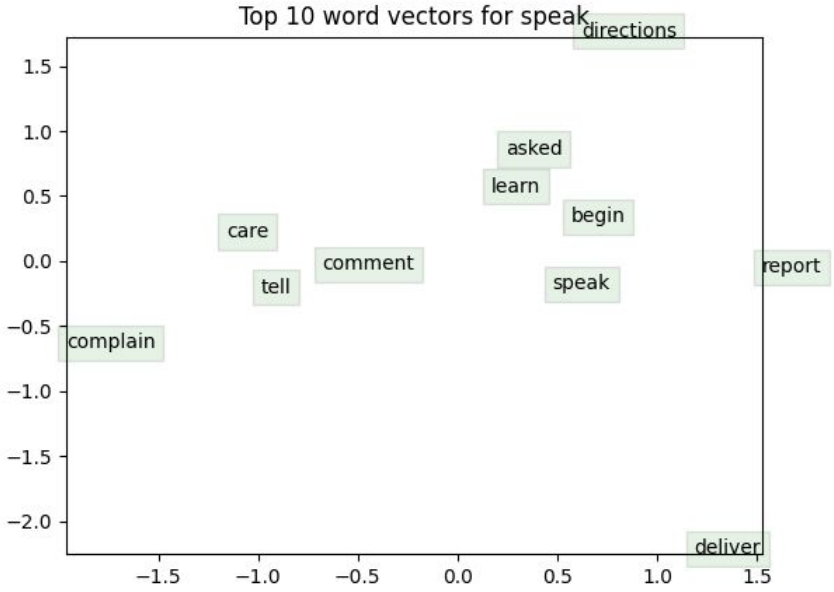


For CBOW

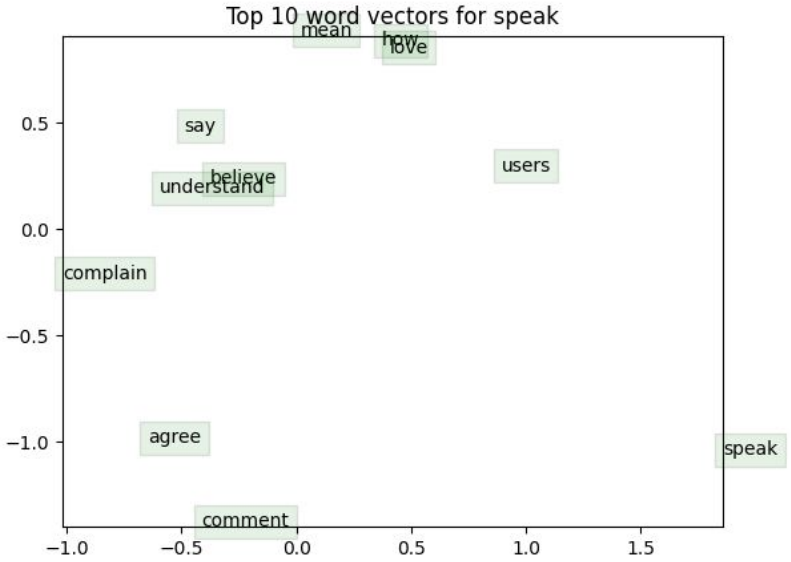


For Skip Gram

➢ Word 2: Speak

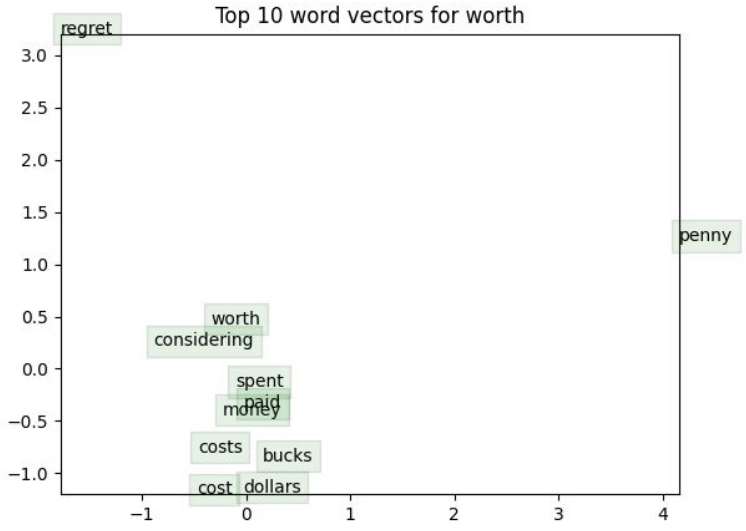


For CBOW

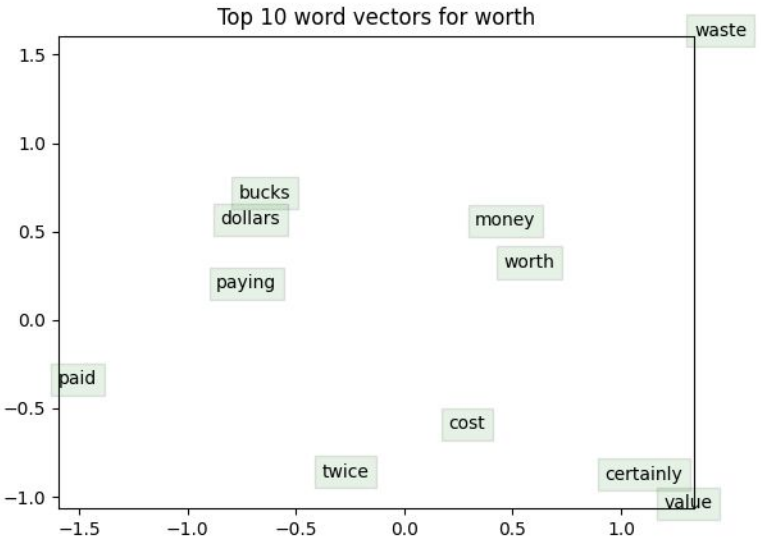


For Skip Gram

➢ Word 3: Worth

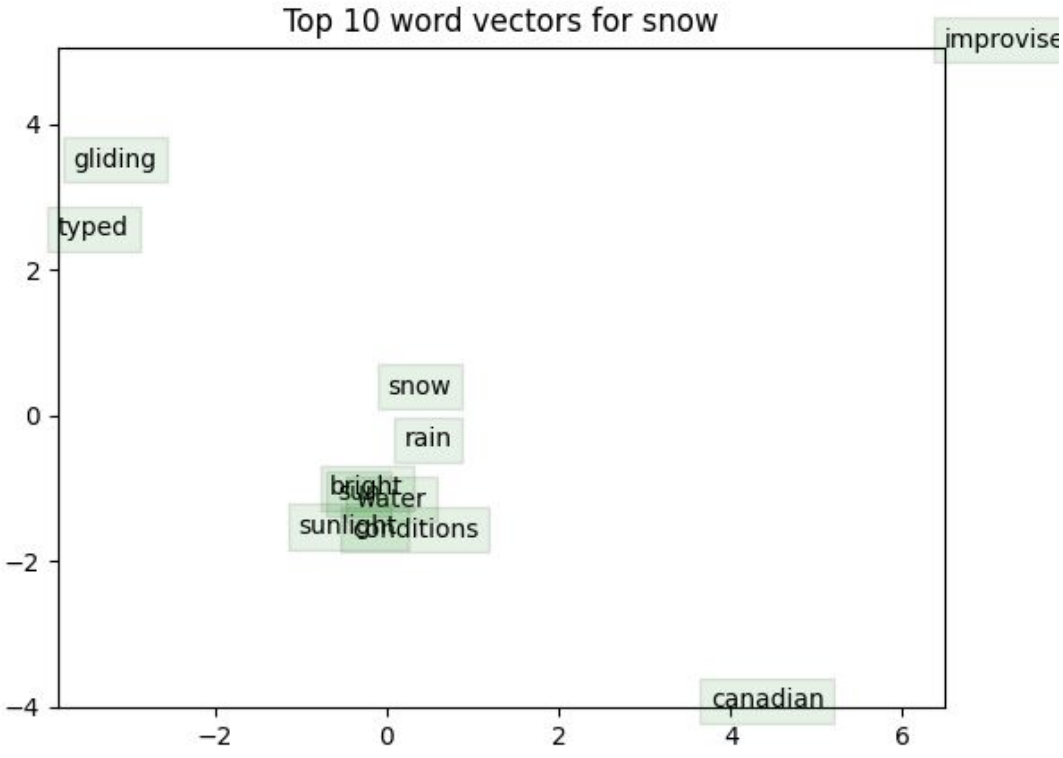


For CBOW

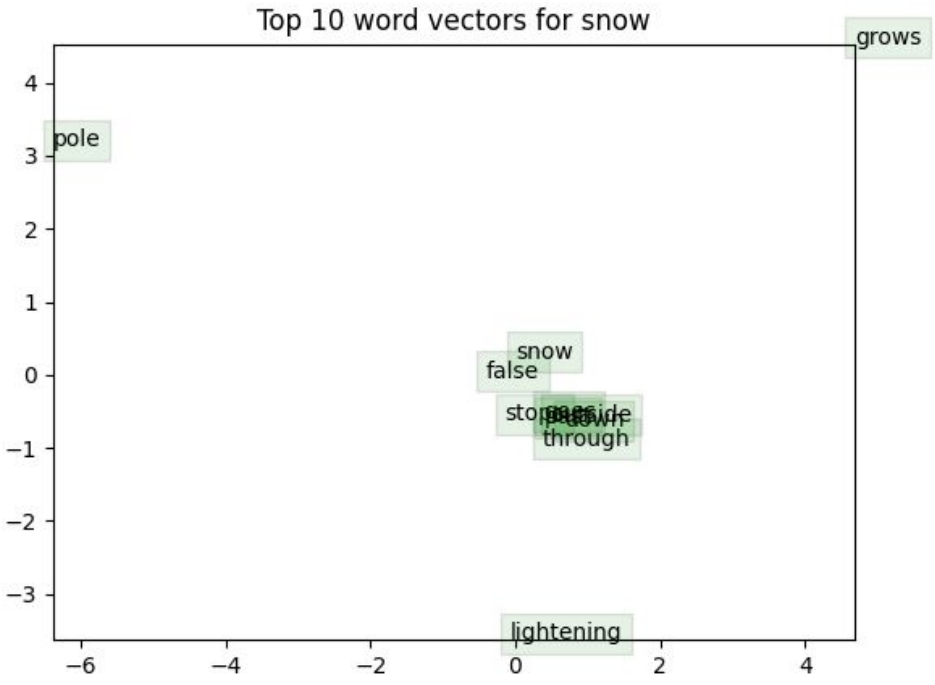


For Skip Gram

➢ Word 4: Snow

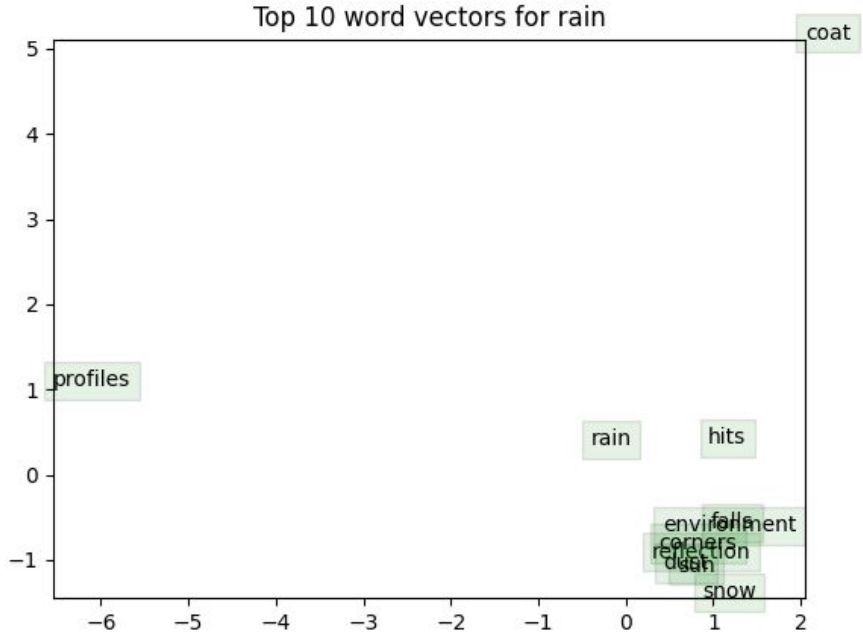


For CBOW

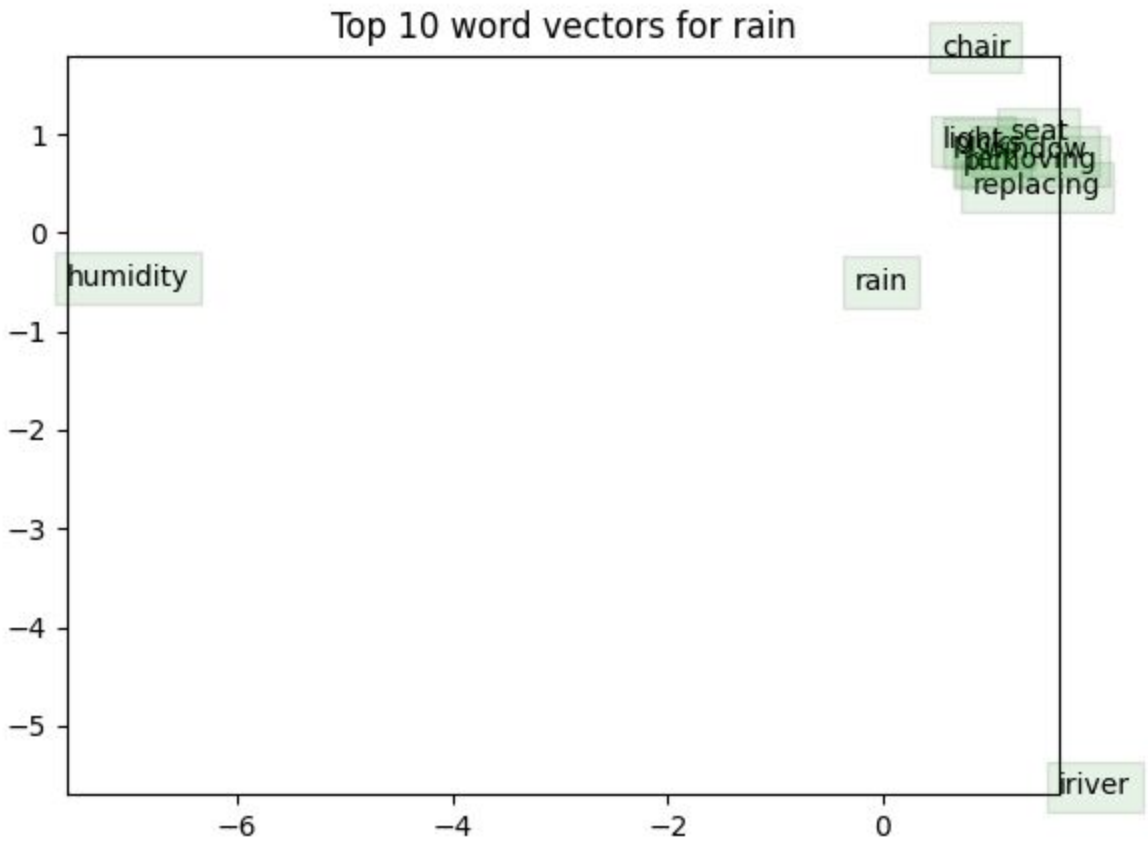


For Skip Gram

➢ Word 5: Rain



For CBOW



For Skip Gram

Q2 What are the top 10 closest words for the word 'camera' in the embeddings generated by your program. Compare them against the pre-trained word2vec embeddings that you can download off the shelf (can use gensim).

➢ **For CBOW:**

Top 10 closest words to word 'Camera' for my model are:

1. Body
2. Dslr
3. Bag
4. Canon
5. Slr
6. Lens
7. Rebel
8. Lense
9. Nikon
10. Eos

Top 10 closest words to word 'Camera' for model trained using gensim:

1. Cameras
2. Canon
3. Rebel
4. Camcorder
5. Scope
6. Lens
7. Nikon
8. Slr
9. Elph
10. Western

Comparing two lists , my model was able to capture words like slr, canon, nikon, lens, rebel but even the words that are not present in the list of words generated by model trained by gensim, also can be said to be similar to the word 'camera'. As

gensim choses best hyperparameters for training, a model trained by gensim is expected to be better than a model trained by me.

➢ **For Skip Gram:**

Top 10 closest words to word 'Camera' for my model are:

1. Lens
2. lense
3. Nikon
4. Slr
5. Zoom
6. Rebel
7. Body
8. Tripod
9. Canon
10. kit

Top 10 closest words to word 'Camera' for model trained using gensim:

1. Cameras
2. Cam
3. Rebel
4. Camcorder
5. Digicam
6. Lens
7. Lense
8. Elph
9. Canon
10. Minolta

Comparing two lists , my model was able to capture words like lense, lens, rebel, canon but even the words that are not present in the list of words generated by the model trained by gensim, also can be said to be similar to the word 'camera'. As gensim choses best hyperparameters for training, a model trained by gensim is expected to be better than a model trained by me.

*Note

One drive link containing embeddings and other evaluation code is:

https://iiitaphyd-my.sharepoint.com/:f:/g/personal/jashn_arora_research_iiit_ac_in/EotCDRThfSJLm6PeTMEizEMBxmQpkOvcOuxXR8VF7wDpQg?e=RxEPKN