HW1 Write up answers

Word Vectors: Distributed Representations of Words

PangFa Chou | choupa@oregonstate.edu

Task 1.1:

For tokenization, I use the lemmatization function in nltk package to help me tokenize the text.

Eg. The tokenization gets a string:

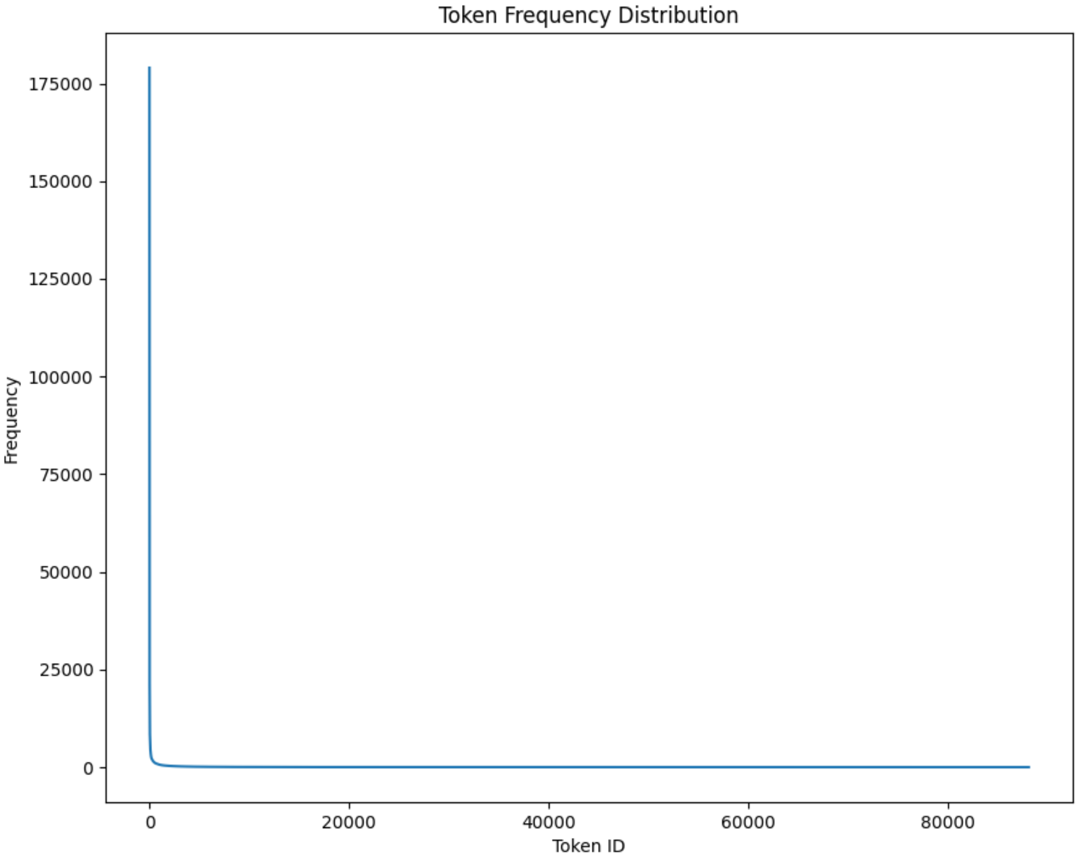
"Google IPO faces Playboy slip-up The bidding gets underway for Google's public offering, despite last-minute worries over an interview with its bosses in Playboy magazine."

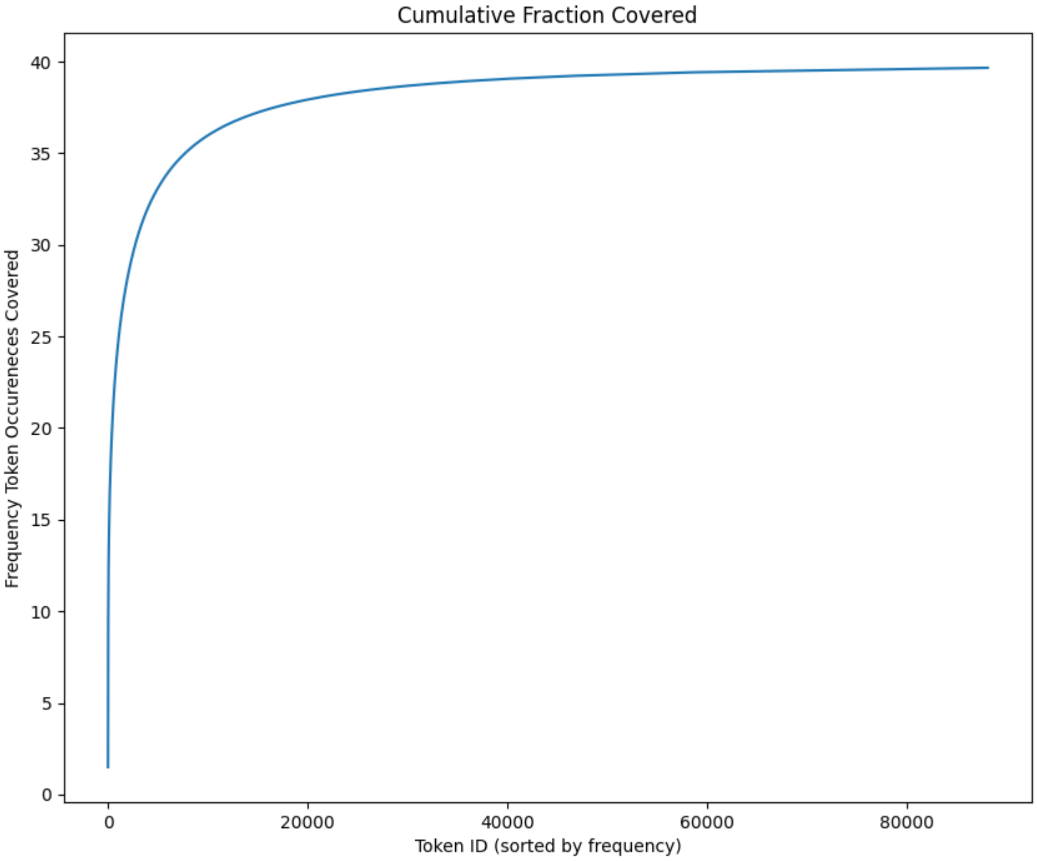
It will remove the punctuation and tokenize this text into this list:

['Google', 'IPO', 'faces', 'Playboy', 'slip', 'up', 'The', 'bidding', 'gets', 'underway', 'for', 'Google', 's', 'public', 'offering', 'despite', 'last', 'minute', 'worries', 'over', 'an', 'interview', 'with', 'its', 'bosses', 'in', 'Playboy', 'magazine']

Task 1.3:

I set my freq at 47 so it could catch 90% of the tokens from datasets. The other tokens are considered as 'UNK'. I set 47 as the threshold because the datasets will conform to the long-tail property, which means most of the tokens won't be used while some tokens are used very frequently. Since statistics distribution in long-tail property usually suggests 80% as the majority, I added up a little bit more to 90% so it could cover more tokens. Therefore, after I sorted the freq dictionary, I found that when freq equaled 47 I had 90% of tokens coverage. That's why I set my freq at 47.


Token Frequency Distribution
Frequency
175000
150000
125000
100000
75000
50000
25000
0
Token ID
0
20000
40000
60000
80000



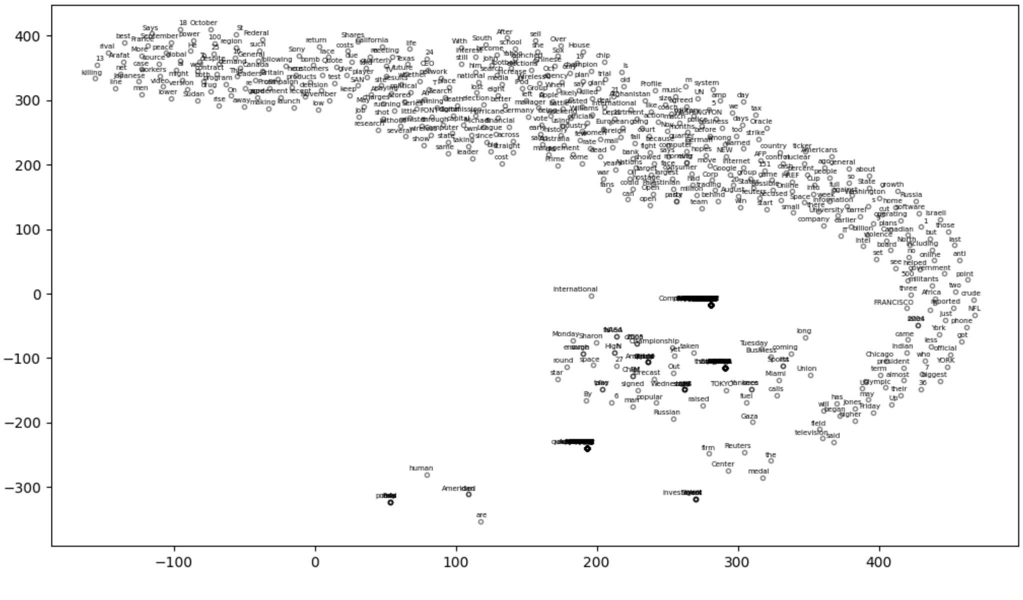
Task 2.1:

The minimum and maximum of PMI can be infinite/ negatively infinite. The higher the PMI is, the more confidence we have to believe that these two tokens have a relationship. On the other hand, the lower the PMI is, the less possibility that there is a relationship between two tokens. Therefore, intuitively I think the PPMI shows the co-occurence level of two tokens because it defines those unrelated tokens as 0.

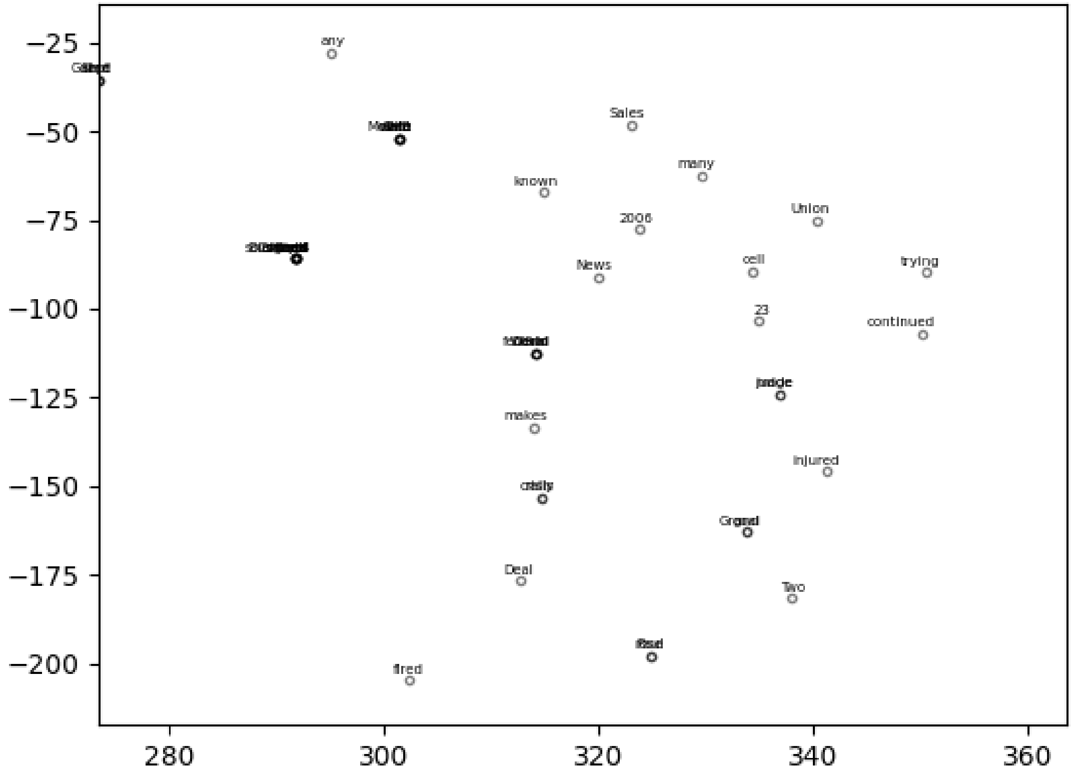
Task 2.2:

I use the entire sentence as a context.

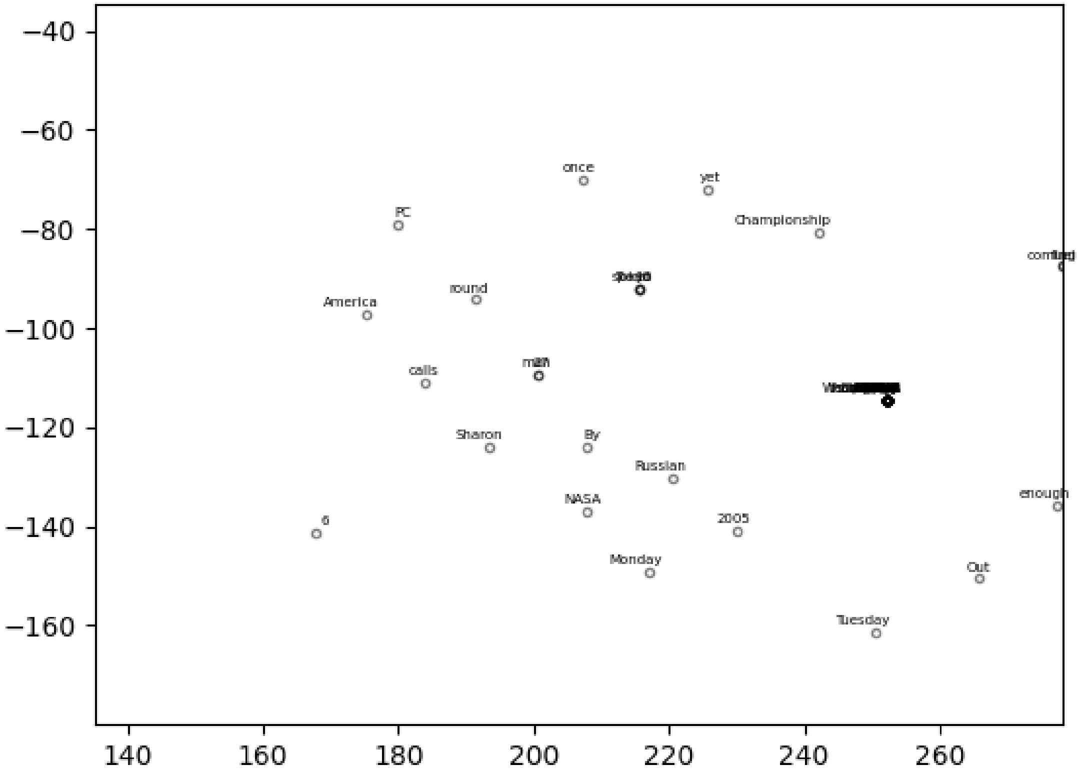
Task 2.4:



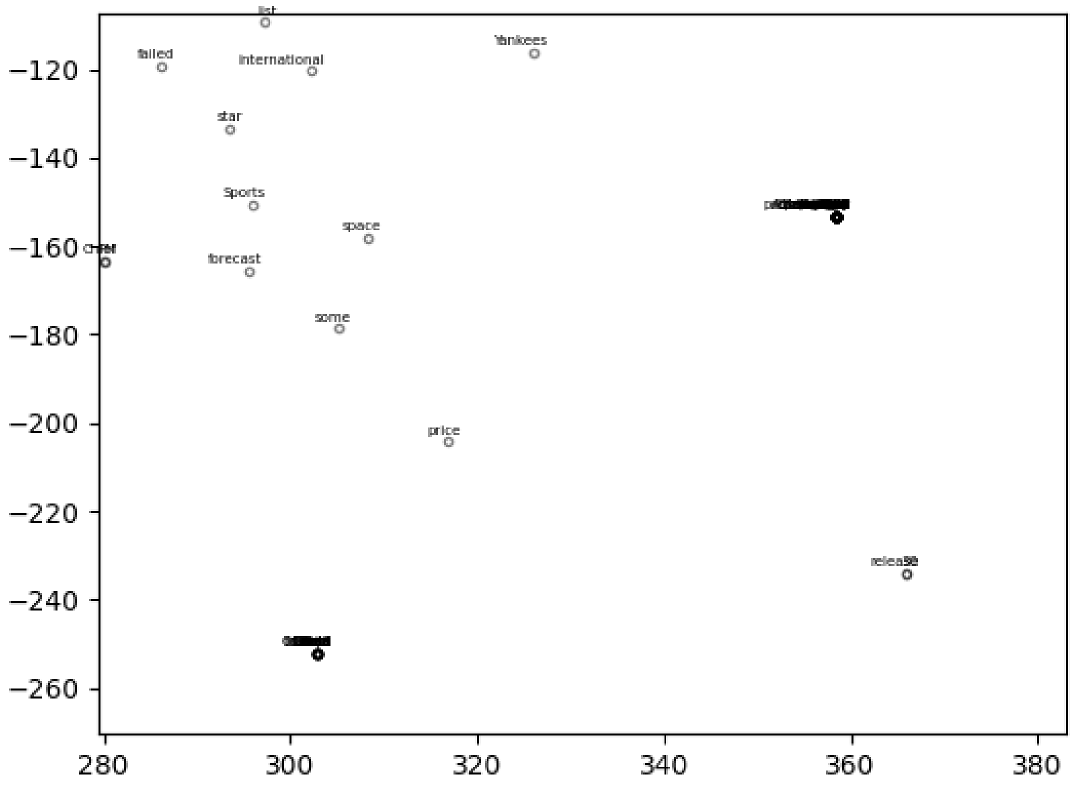
This is the plot of common tokens.



This is the cluster that is related to reports industry like News.

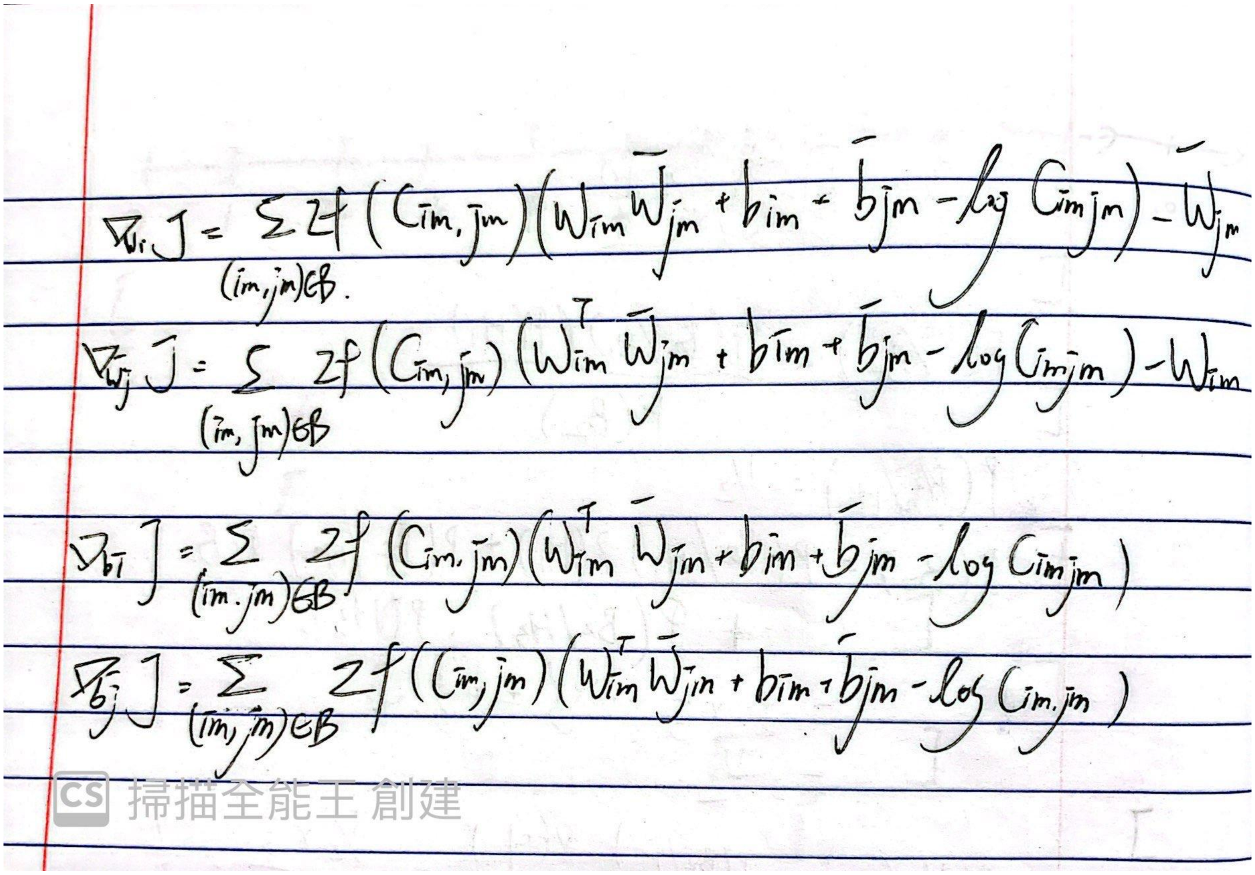


This is the cluster related to great achievement like Olympic and NASA.



This is related to sports.

Task 3.1:



$$\nabla_{w_i} J = \sum_{(i_m,j_m)\in B} 2f(C_{i_m,j_m})(w_{i_m}\bar{w}_{j_m} + b_{i_m} + \bar{b}_{j_m} - \log C_{i_m j_m}) - \bar{w}_{j_m}$$

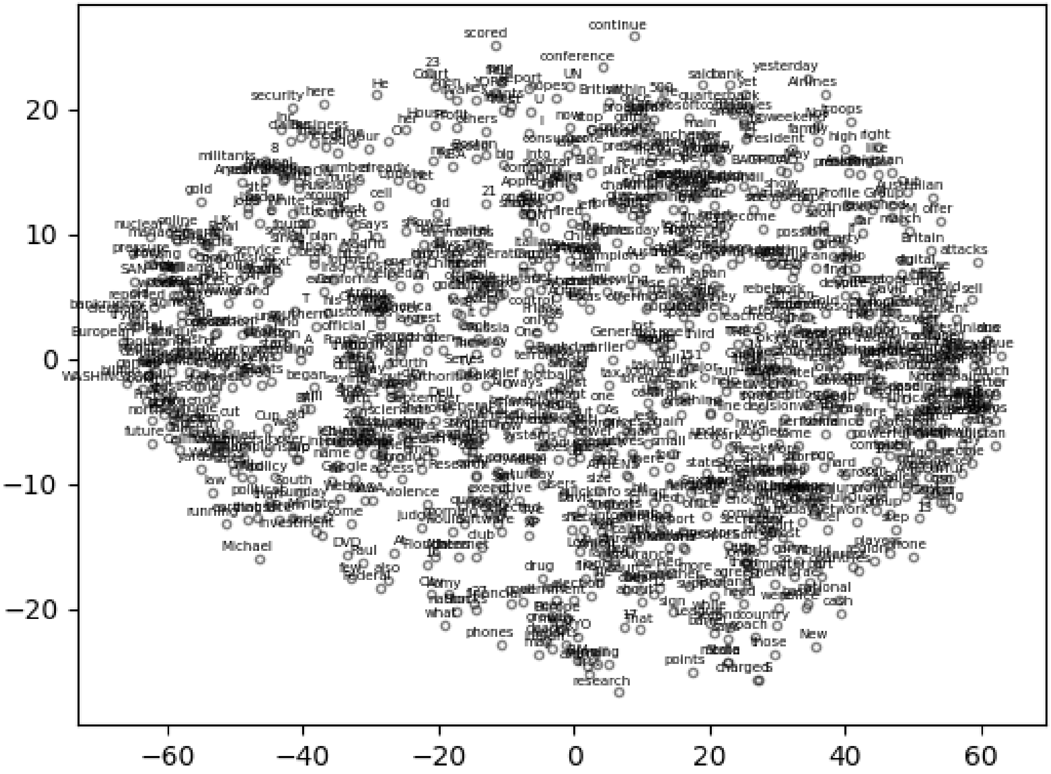
$$\nabla_{w_j} J = \sum_{(i_m,j_m)\in B} 2f(C_{i_m,j_m})(w_{i_m}^T \bar{w}_{j_m} + b_{i_m} + \bar{b}_{j_m} - \log C_{i_m j_m}) - w_{i_m}$$

$$\nabla_{b_i} J = \sum_{(i_m,j_m)\in B} 2f(C_{i_m,j_m})(w_{i_m}^T \bar{w}_{j_m} + b_{i_m} + \bar{b}_{j_m} - \log C_{i_m j_m})$$

$$\nabla_{\bar{b}_j} J = \sum_{(i_m,j_m)\in B} 2f(C_{i_m,j_m})(w_{i_m}^T \bar{w}_{j_m} + b_{i_m} + \bar{b}_{j_m} - \log C_{i_m,j_m})$$

Task 3.3:

We can see that the loss is reduced to about the power of 10-7 on average over the 5 epochs.



This is the TSNE plot for visualizing GloVe vectors.

Task 4.1:

1 .roses : red :: violet : ?

[('blue', 0.61), ('yellow', 0.53), ('colored', 0.523), ('orange', 0.483), ('purple', 0.482), ('reddish_orange', 0.481), ('mottle', 0.479), ('reddish_pink', 0.475), ('pinkish_orange', 0.474), ('purpley', 0.473)]

2. summer : hot :: winter : ?

[('cold', 0.582), ('Hot', 0.525), ('toasty', 0.492), ('chilly', 0.488), ('warm', 0.482), ('frigid', 0.479), ('hotter', 0.478), ('hottest', 0.471), ('wet', 0.468), ('CHEFS_Chefs', 0.455)]

3. depress : negative :: happy : ?

```
[('positive', 0.587), ('postive', 0.529), ('glad', 0.526), ('ecstatic', 0.499), ('satisfied', 0.498), ('pleased',
0.489), ('optimistic', 0.457), ('positve', 0.456), ('overjoyed', 0.456), ('Said_Hirschbeck', 0.449)]
```

These are common analogies that people would describe with those nouns. For example, summer is hot and winter is cold.

Task 4.2:

1. mountain : tall :: valley : ?

```
[('feet_tall', 0.512), ('taller', 0.501), ('tall_skinny', 0.484), ('inches_tall', 0.444), ('Tall', 0.436),
('pronounced_nuh', 0.429), ('6feet', 0.414), ('wheatish_complexion', 0.412), ('freakishly_tall', 0.409),
('tall_slender', 0.408)]
```

2. love : like :: hate : ?

```
[('RUSH_Yeah', 0.454), ('everywhere', 0.437), ('stupid', 0.421), ('maybe', 0.416), ('think', 0.415),
('Mr._DIONNE', 0.414), ('do', 0.412), ('bother', 0.411), ('quite_frankly', 0.41), ('anymore', 0.41)]
```

3. cheetah : fast :: turtles : ?

```
[('clams', 0.404), ('crabs', 0.401), ('kingfish_bite', 0.399), ('shellfish', 0.38), ('inshore_lumps', 0.377),
('shell_clams', 0.374), ('fish', 0.373), ('menhaden', 0.373), ('slow', 0.372), ('Fast', 0.367)]
```

These examples don't really work. They only work in some specific context. For example, Love is like, and hate is supposed to be dislike, but it gives me some words like everywhere, stupid, which is not what I thought, but it fits in some contexts.

Task 4.3:

1.

man : party :: woman : ?

```
[('Party', 0.601), ('parties', 0.587), ('LOUDON_NH_Brad_Keselowski', 0.472), ('partys', 0.465),
('Alberta_Wildrose_Alliance', 0.465), ('caucus', 0.456), ('Democratic_Party', 0.454), ('Democratic_Pary',
0.452), ('Akhilesh_Pratap_Singh', 0.449), ('leader_Wojciech_Olejniczak', 0.449)]
```

woman : party :: man : ?

```
[('Party', 0.575), ('parties', 0.526), ('partys', 0.521), ('faction', 0.466), ('pary', 0.451), ('Democratic_Party',
0.446), ('mad_hatter_tea', 0.443), ('Monster_Raving_Looney', 0.434), ('Partys', 0.432),
('Ignatius_Shixwameni', 0.43)]
```

At first I thought it should be a gender neutral word for both men and women, but it turns out when we put a keyword “party” in man, it is more likely to have content about political things.

2.

man : careless :: woman : ?

```
[('judicially_imprudent', 0.529), ('carelessly', 0.511), ('carelessness', 0.502), ('thoughtless', 0.494),
('reckless', 0.477), ('unladylike', 0.472), ('negligent', 0.459), ('Careless', 0.452), ('inadvertent', 0.433),
('irresponsible', 0.433)]
```

woman : careless :: man : ?

```
[('reckless', 0.549), ('wreckless', 0.532), ('Careless', 0.524), ('carelessness', 0.523), ('judicially_imprudent',
0.514), ('undisciplined', 0.492), ('sloppy', 0.478), ('lackadaisical', 0.476), ('carelessly', 0.475), ('cavalier',
0.453)]
```

It only shows undisciplined in men’s similarity. I think it is biased because it means men should be trained to make no mistake.

Task 4.4

It is because the word2vec uses the words that people use in daily life, and sometime we talk about the similar perceptions or topics together, and they are all reflected on the word2vec, leading to the inaccurate co-occurrence relationship. The inaccurate co-occurrence result may misjudge some contexts.