AI539 Spring 2024 Homework I

Rigved Naukarkar – naukarkr@oregonstate.edu

July 31, 2024

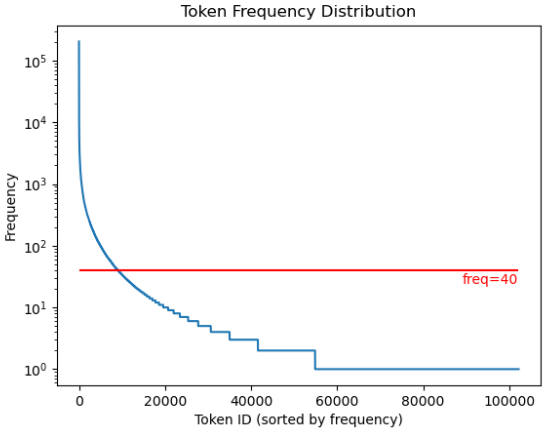
Task 1.1

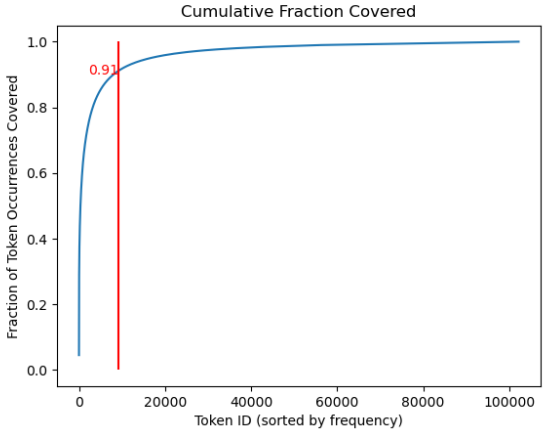
Implemented the tokenizer with regex such that all the alphanumeric chars are only left in the vocabulary.

- Example Input - Alternative tokenization's include sub-word tokenization (jumped -> [jump, ed]) or pre-processing with lemmitization to reduce words to their roots (jumped -> [jump], are -> [be]).
- Output - ["Alternative", "tokenizations", "include", "subword", "tokenization", "jumped", "jumped", "or", "preprocessing", "with", "lemmitization", "to", "reduce", "words", "to", "their", "roots", "jumped", "jump", "are", "be"]

Task 1.3

Implemented a red cutoff line where the frequency is a minimum of 40. Tried to get the percent of words covered above 90%. With this frequency, I got 91% of words covered. More in the graphs below.





Task 2.1

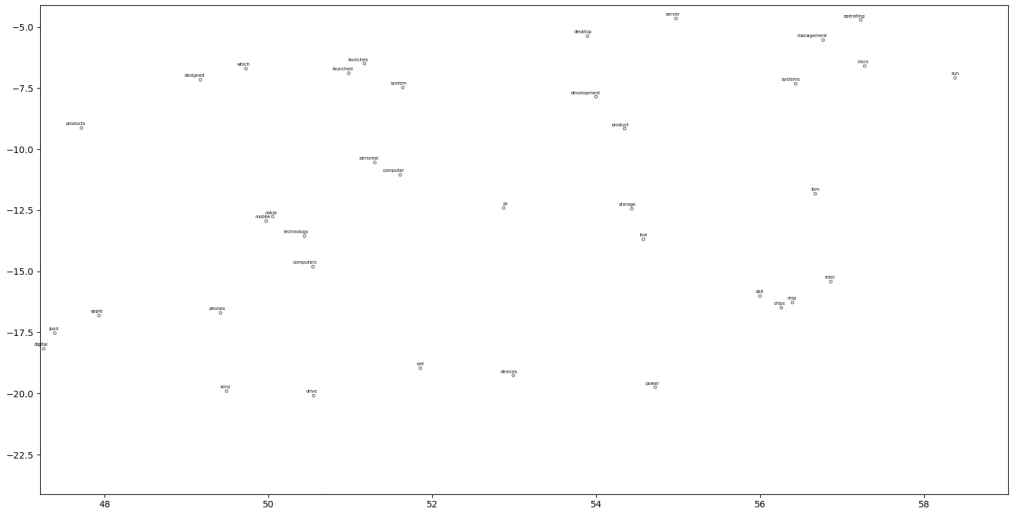
The PMI function can range from $-\infty$ to $+\infty$. If two tokens have positive PMI, then it means that the words are more similar and occur together in more contexts than expected. If both words are independent of each other's context, then the PMI becomes $\log(1) = 0$. If we obtain negative PMI, it means that $P(w_i|w_j) < P(w_j)$, indicating that there are more occurrences of w_j than of w_i given w_j, implying they are less similar.

Task 2.2

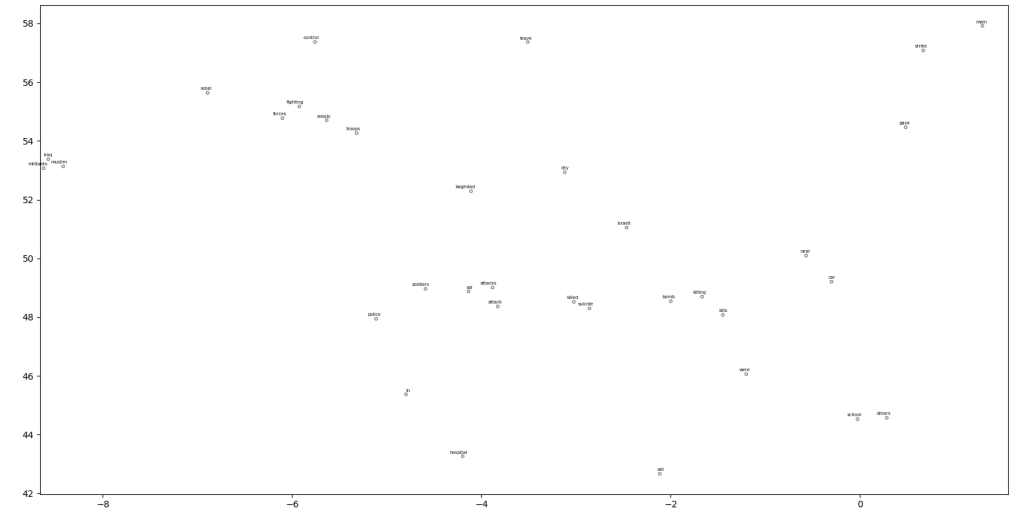
The context size is defined as the size of 4 words around the center word. Total of 8 words around each word are considered for the context size.

Task 2.4

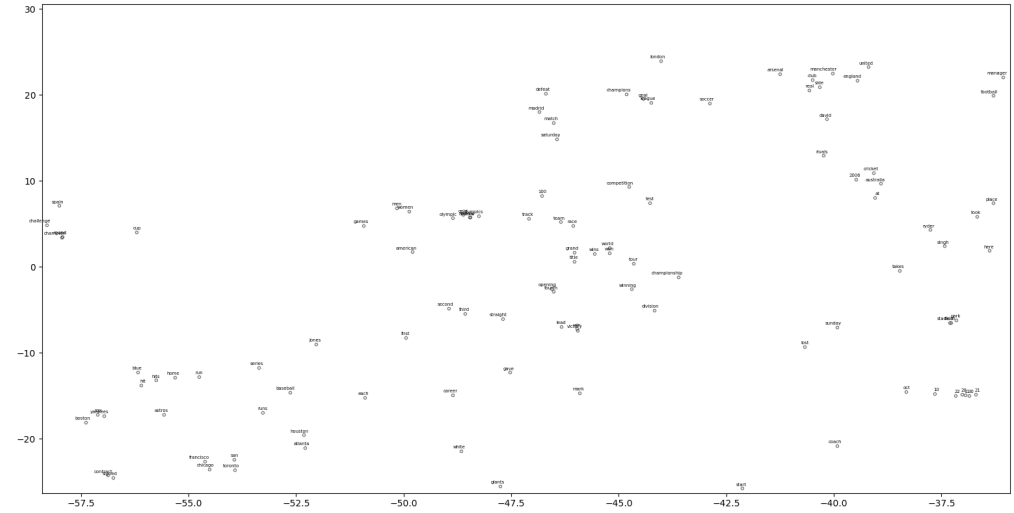
The below cluster has all the technology-related words. In the center we see "PC", then "computer", "apple", "intel", "dell" etc.



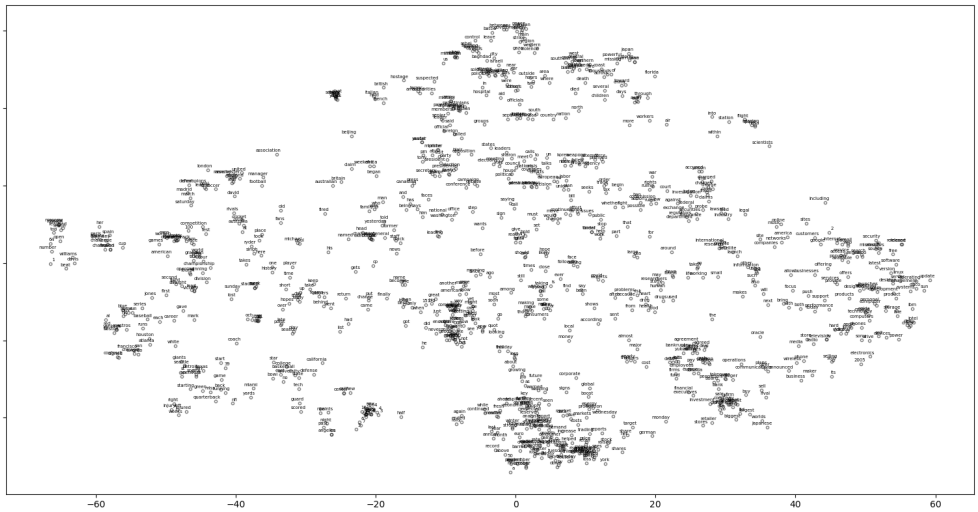
The below cluster has war-related words like "killing", "attacks", "hospital", "soldiers" etc.



The below cluster has games-related words like "football", "olympics", "basketball", "competition", "winning" etc.



The whole plot:



Task 3.1

$$\nabla_{w_{i_m}} J_B = 2 \sum_{(i_m, j_m) \in B} f(C_{i_m j_m}) \cdot \left(w_{i_m}^T c_{j_m} + b_{i_m} + b_{j_m} - \log(C_{i_m j_m}) \right) \cdot c_{j_m}$$

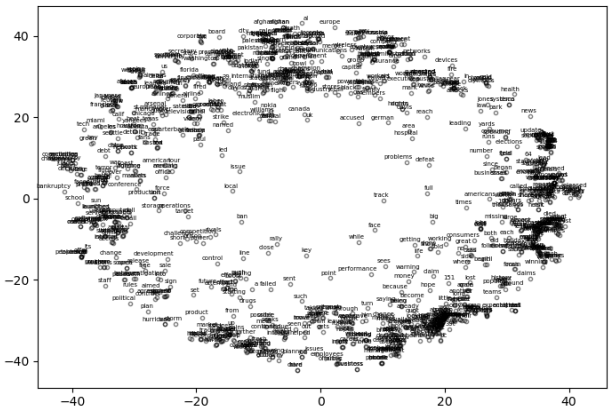
$$\nabla_{c_{j_m}} J_B = 2 \sum_{(i_m, j_m) \in B} f(C_{i_m j_m}) \cdot \left(w_{i_m}^T c_{j_m} + b_{i_m} + b_{j_m} - \log(C_{i_m j_m}) \right) \cdot w_{i_m}$$

$$\nabla_{b_{i_m}} J_B = 2 \sum_{(i_m, j_m) \in B} f(C_{i_m j_m}) \cdot \left(w_{i_m}^T c_{j_m} + b_{i_m} + b_{j_m} - \log(C_{i_m j_m}) \right)$$

$$\nabla_{b_{j_m}} J_B = 2 \sum_{(i_m, j_m) \in B} f(C_{i_m j_m}) \cdot \left(w_{i_m}^T c_{j_m} + b_{i_m} + b_{j_m} - \log(C_{i_m j_m}) \right)$$

Task 3.3

The loss decreased quickly initially until it reached 0.07. I did not see much improvement, leading to 0.05 at the end of 5 epochs. So, I increased the epochs and made the learning rate diminish a little to achieve loss convergence at 0.043 finally.



Task 4.1

Below are three more analogies that made sense to me. Tuna is a fish and Owl is a bird. Computers are made by intel and cars are made by chevy. Stanford is a good research university and UConn has a good basketball team.

```
>>> analogy("fish","tuna","bird")
fish : tuna :: bird : ?
[('owl', 0.519), ('seabird', 0.499), ('birds', 0.484), ('squirrel', 0.478), ('avian', 0.478), ('parrot', 0.468), ('falcon',
 0.467), ('pelican', 0.466), ('seagull', 0.466), ('rusty_blackbird', 0.464)]
```

```
>>> analogy("computer", "intel", "car")
computer : intel :: car : ?
[('Chevy', 0.46), ('Celica', 0.454), ('suv', 0.453), ('sedan', 0.45), ('SEATs', 0.444), ('Corvette', 0.441), ('Mercedes_SL#
', 0.44), ('vette', 0.435), ('Mercedes_E##_AMG', 0.435), ('SUV', 0.432)]
```

```
>>> analogy("research", "stanford", "basketball")
research : stanford :: basketball : ?
[('uconn', 0.541), ('bball', 0.537), ('baseketball', 0.536), ('bas_ketball', 0.521), ('usc', 0.505), ('lebron', 0.5), ('rou
ball', 0.498), ('byu', 0.489), ('nba', 0.485), ('fvc', 0.485)]
```

Task 4.2

These 3 look wrong to me. This one has mammal:cow::reptile:cows. Could be mammal:cow::reptile:lizard.

```
>>> analogy("mammal", "cow", "reptile")
mammal : cow :: reptile : ?
[('cows', 0.569), ('goat', 0.521), ('pig', 0.513), ('piglet', 0.483), ('dairy_cow', 0.471), ('bovine', 0.471), ('goats', 0.
469), ('cattle', 0.468), ('bovines', 0.464), ('snake', 0.454)]
```

This one has mountain:everest::canyon:scotty. Could be mountain:everest::canyon:grand

```
>>> analogy("mountain","everest","canyon")
mountain : everest :: canyon : ?
[('scotty', 0.414), ('bioshock', 0.408), ('omg_i', 0.407), ('height_##px', 0.406), ('piton', 0.406), ('playboat', 0.405), (
'joshua', 0.404), ('@_donlemoncnn_@', 0.402), ('bungie', 0.399), ('steph', 0.395)]
>>>
```

This one has planet:earth::satellite:satellites. Could be planet:earth::satellite:voyager

```
>>> analogy("planet","earth","satellite")
planet : earth :: satellite : ?
[('Satellite', 0.579), ('satellites', 0.563), ('sattelite', 0.504), ('geostationary_communications', 0.501), ('geostationar
y_satellite', 0.497), ('Insat_4B', 0.488), ('TECSAR', 0.484), ('iPSTAR_broadband', 0.481), ('remote_sensing_satellites', 0.
478), ('Hotbird', 0.47)]
```

Task 4.3

Below one has racial bias. Leading to saying, for example, that no black people exist in America.

```
>>> analogy("america", "white", "africa")
america : white :: africa : ?
[('black', 0.615), ('colored', 0.547), ('brown', 0.529), ('blue', 0.507), ('Coloured', 0.483), ('gray', 0.482), ('blues_gra
', 0.481), ('browns_grays', 0.474), ('browns_beiges', 0.473), ('kangas', 0.452)]
```

```
>>> analogy("africa","white","america")
africa : white :: america : ?
[('black', 0.617), ('wrote_Newitz', 0.524), ('blue', 0.499), ('yellow_stripe', 0.484), ('Anglo_Saxon_Protestant', 0.484), (
'brown', 0.474), ('red', 0.468), ('grays_browns', 0.467), ('whites', 0.462), ('confederate_flags', 0.46)]
```

Below one has political bias. This one shows a political party for India. Instead it could show the president of India.

```
>>> analogy("america", "obama", "india")
america : obama :: india : ?
[('bjp', 0.567), ('modi', 0.564), ('pakistan', 0.561), ('bihar', 0.556), ('gujarat', 0.544), ('advani', 0.538), ('manmohan'
, 0.535), ('singh', 0.535), ('kalmadi', 0.527), ('mccain', 0.526)]
```

```
>>> analogy("india","bjp","america")
india : bjp :: america : ?
[('dems', 0.624), ('repubs', 0.608), ('gop', 0.588), ('hillary', 0.587), ('repub', 0.572), ('barack_obama', 0.572), ('obama
', 0.564), ('boehner', 0.563), ('john_mccain', 0.561), ('hitler', 0.554)]
```

Task 4.4

The biases are because of the data used for training. Maybe biased news headings are used to train it. Maybe the data was regional.

If this word2vec is used in the live system, the response of, for example, chatbot might also be biased. Like recently the Google Gemini was generating only black-skinned human images.