Part 1 - Train your CBOW embeddings for both datasets

➢ Describe your decisions (preprocessing, class structure) in the lab report.

- **Preprocessing Steps:**

    1. Sentence Splitting: The text from the Sci-fi dataset was split into sentences using regular expressions. Individual reviews from the Hotel Reviews dataset were extracted from the 'Review' column of the dataframe.

    2. Tokenization: Each sentence, split from both the Sci-fi or Hotel Reviews dataset, was tokenized. This process involves extracting words, converting them to lowercase, and eliminating any punctuation, which ensure vocabulary consistency.

    3. Stopword Removal and Stemming: Commonly used words, or "stopwords," were filtered out from the datasets. Also, words were stemmed using the PorterStemmer to reduce them to their root form, which enhanced the quality of word embeddings.

    4. Filtering Low-Frequency Words: Words were filtered out if they appeared fewer than 10 times. This is beneficial to speed up the efficiency of the training process and removes noise from the datasets.

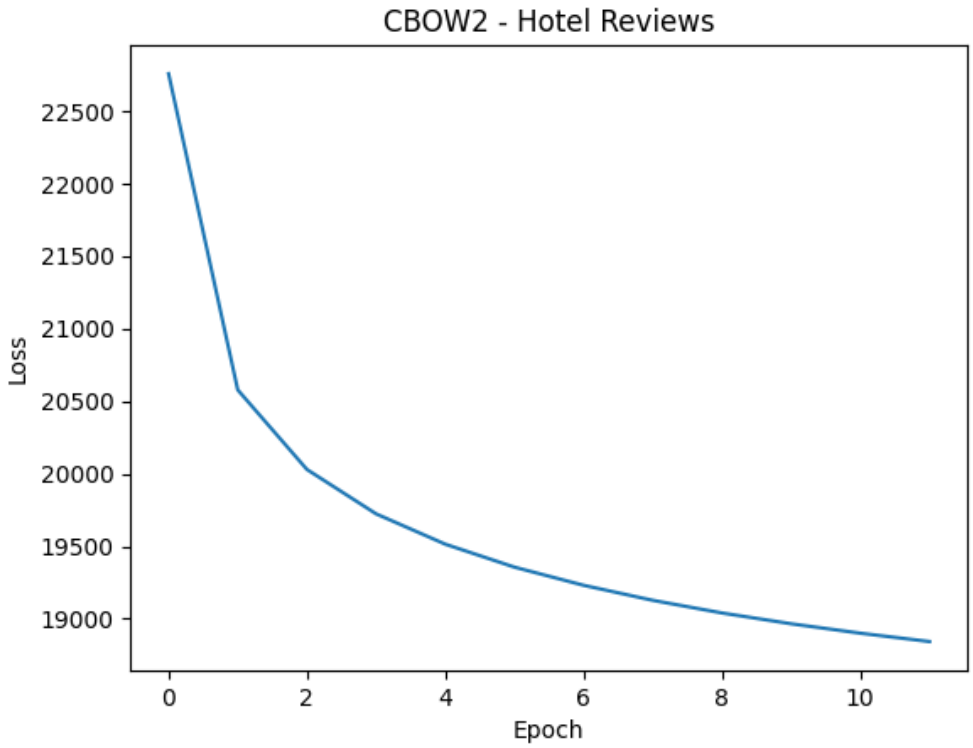
    5. (Using the WordNetLemmatizer() function instead of the stemmer was also an approach we considered. But the stemmer achieved generally better results as it is more aggressive. As such we opted out of lemmatizing words.)

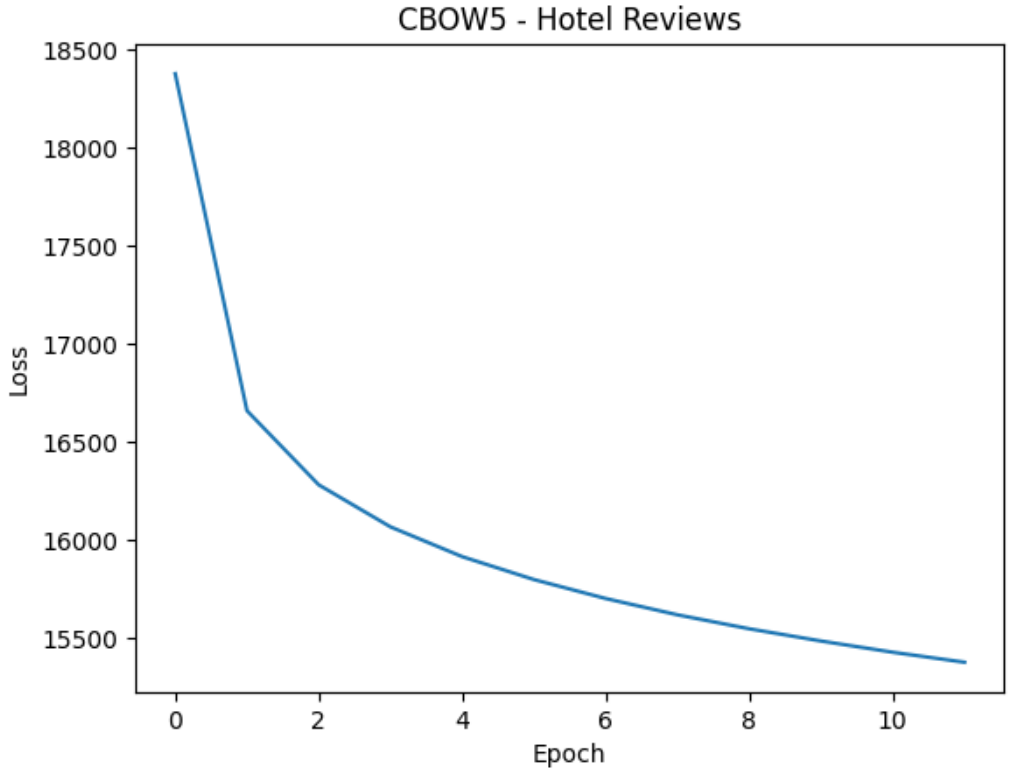
- **Class Structure**

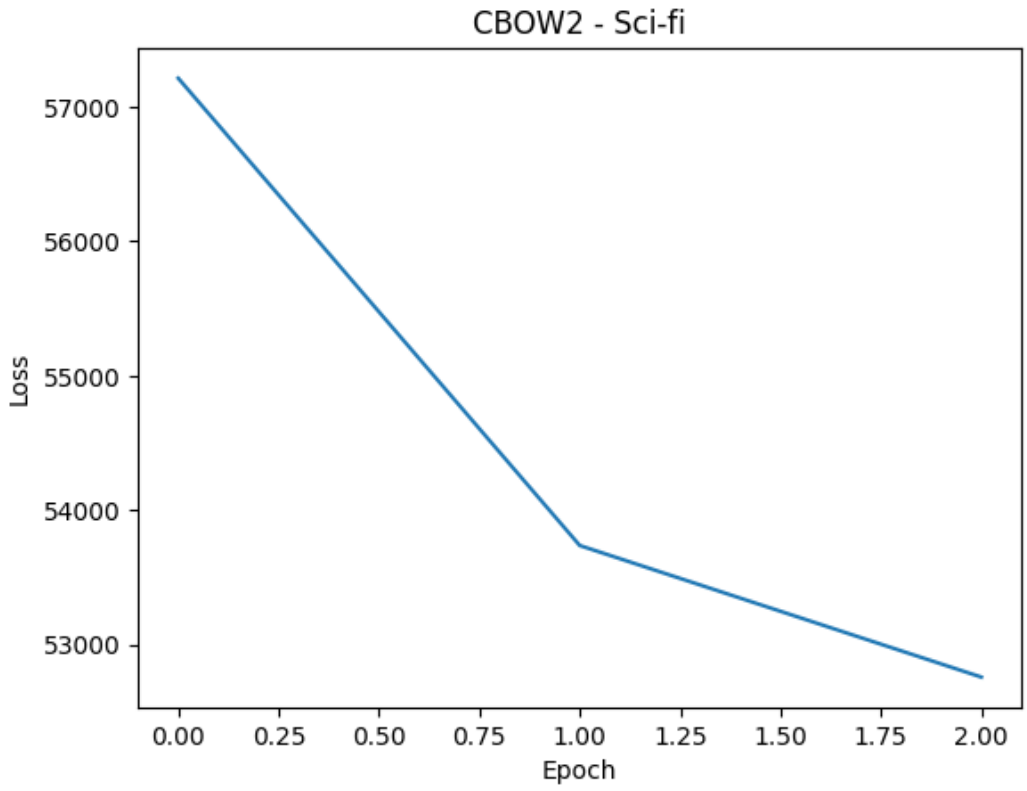
    We created a vocabulary comprised unique words for each dataset. Two dictionaries were constructed to map words to indices and map indices to words, respectively. Additionally, context-target pairs were generated by looping through each word in a sentence and considering words within a certain window size as their context.

➢ Results and Observations:

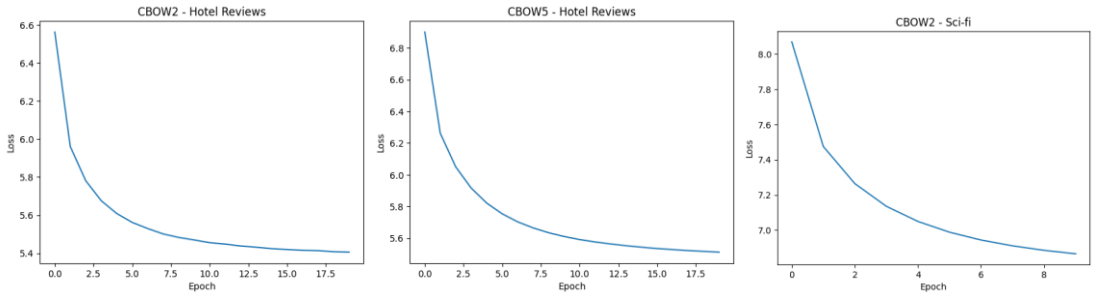
Three CBOW's loss plots showed the training progress. When the epochs increase, the loss greatly reduces, which means the CBOW model is learning to predict target words from their contexts. Besides, for the same dataset of hotel reviews, it could lead to a lower initial loss and potentially faster convergence when increasing the context window from 2 to 5. Also, different datasets such as hotel reviews and sci-fic, might vary in terms of loss and the speed of training progresses.


CBOW2 - Hotel Reviews
Loss
Epoch
22500
22000
21500
21000
20500
20000
19500
19000
0
2
4
6
8
10


CBOW5 - Hotel Reviews
Loss
Epoch
18500
18000
17500
17000
16500
16000
15500
0
2
4
6
8
10



These were total values of the losses. On other runs we calculated the average epoch loss: *average_epoch_loss = total_loss / len(loader)* and trained with more epochs. Likewise, the losses have the same tendencies and are still converging. Training embeddings does generally come with a higher loss, but this may also indicate that the current parameters are still yet to be optimized.



Part 2 - Test your embeddings

- **2. For the hotel reviews dataset, choose 3 nouns, 3 verbs, and 3 adjectives. Make sure that some nouns/verbs/adjectives occur frequently in the corpus and that others are rare. For each of the 9 chosen words, retrieve the 5 closest words according to your trained CBOW2 model. List them in your report and comment on the performance of your model: do the neighbours the model provides make sense? Discuss.**

Note: The received words/neighbours may differ from run to run in some cases and this are only the results from a specific run (e.g., the word "vote" was not even in the vocabulary on the last test run):

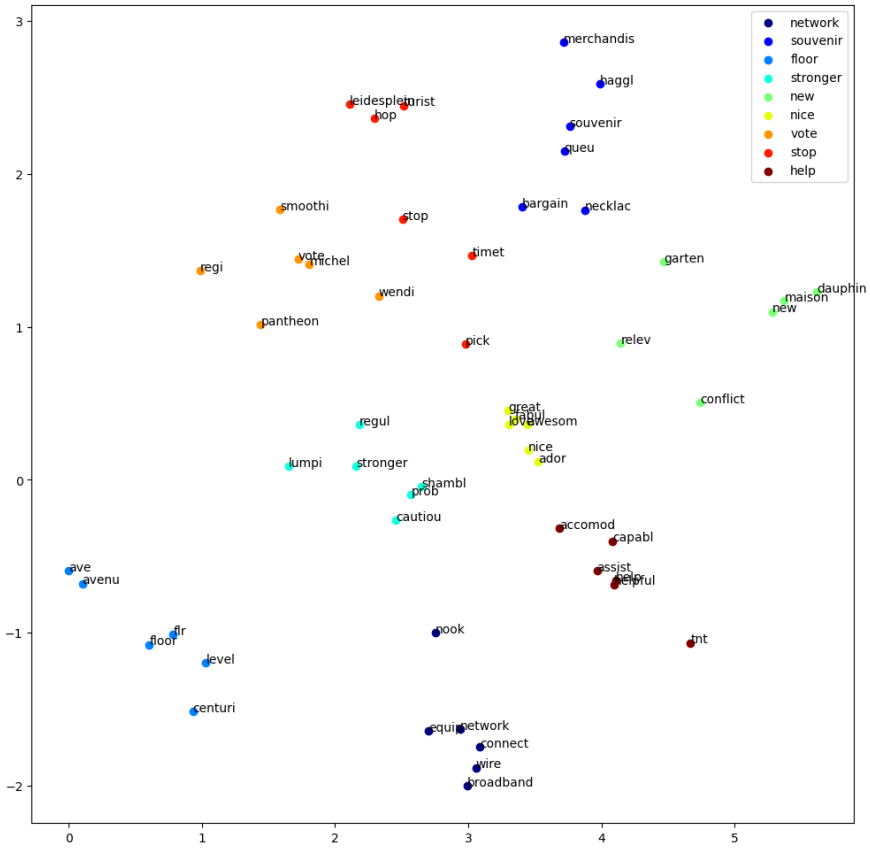
For the hotel reviews dataset, we chose 3 nouns, 3 verbs, and 3 adjectives representing low, medium, and high occurrence frequencies:

Nouns:
network: wire, broadband, connect, nook, equip
souvenir: haggl, queu, necklac, merchandis, bargain
floor: avenu, level, flr, ave, centuri

Verbs:
vote: smoothi, regi, michel, wendi, pantheon
stop: turist, leidesplein, pick, timet, hop
help: helpful, accommod, assist, capabl, tnt

Adjectives:
stronger: regul, prob, cautiou, lumpi, shambl
new: dauphin, maison, relev, conflict, garten
nice: love, fabul, great, awesom, ador

For the high-frequency words like 'floor', 'nice' and 'help', our model showed strong performance, since the closest words for them were semantically and contextually related. However, the closest words for low-frequency have less clear relationship. For example, the meaning of 'shambl' is not quite similar to 'stronger' and other closest word like 'regul'.



➢ 3. Repeat what you did in 2. for the Sci-fi dataset.

For the Sci-fi dataset, we also chose 3 nouns, 3 verbs, and 3 adjectives representing low, medium, and high occurrence frequencies:

Nouns:
skylight: cue, thish, slumber, deceas, undercut
fish: ice, fig, losel, fruit, yearn
space: lawyer, feder, cirissin, messagecraft, chooka

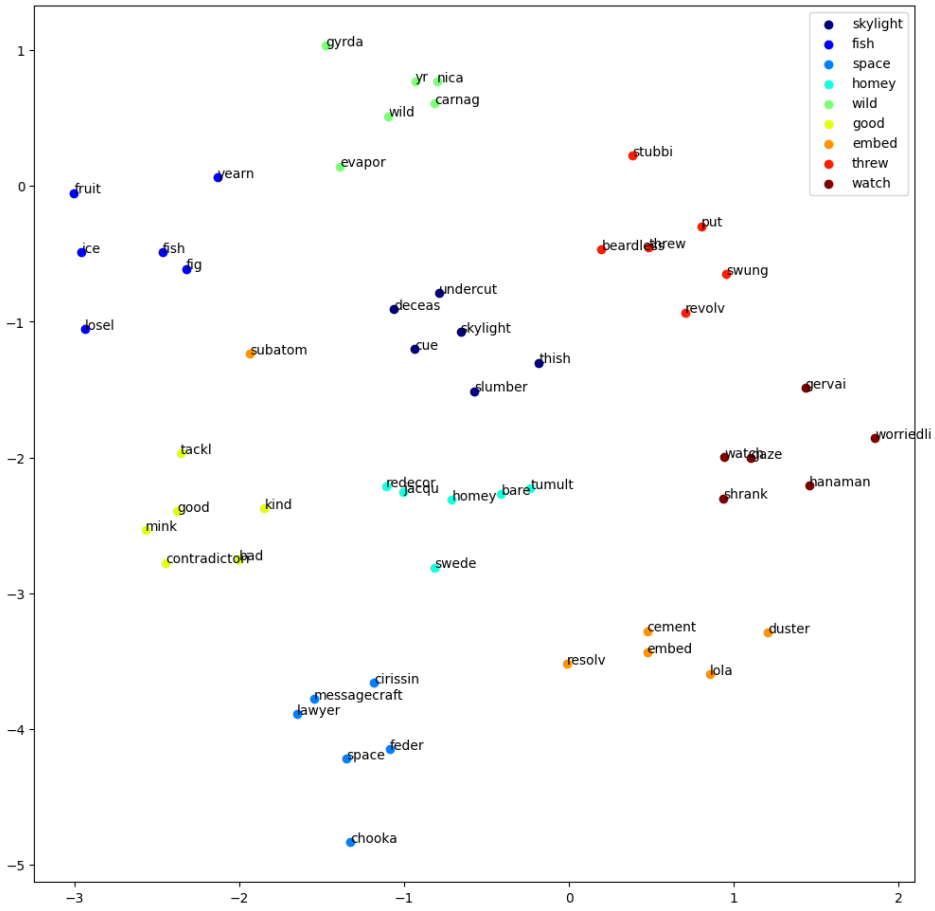
Verbs:
embed: cement, lola, subatom, resolv, duster
threw: put, stubbi, swung, beardless, revolv
watch: gervai, hanaman, shrank, gaze, worriedli

Adjectives:
homey: jacqu, bare, redecor, tumult, swede
wild: yr, nica, evapor, gyrda, carnag
good: contradictori, kind, tackl, bad, mink

In our CBOW2 model for the Sci-fi dataset, some high and medium frequency word like ‘fish’ is associated with strongly semantic relationships like ‘ice’ and ‘fig’, indicating potential scenarios

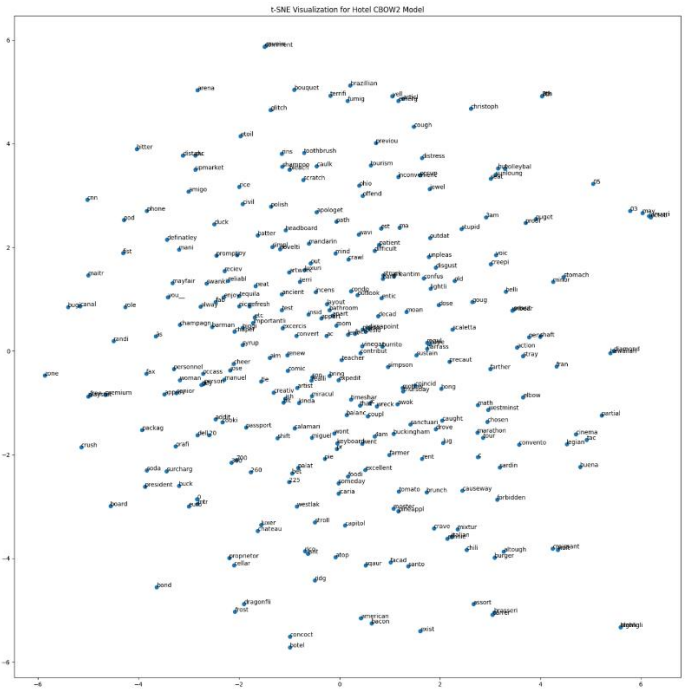
where fish live in the ice and terrestrial environments. Besides, the neighbors of ‘good’ includes both ‘bad’ and ‘kind’, suggesting this model can link words with similar but also opposite meaning.

Conversely, the highly frequent word ‘space’ ‘s nearest neighbors are not as semantically close as we expect, although it is rather important term in scientific fiction. Similarly, the connections between "embed" and its neighbors, or "threw" and its neighbors, also seem a bit distant. This might show the model could do further fine-tuning.

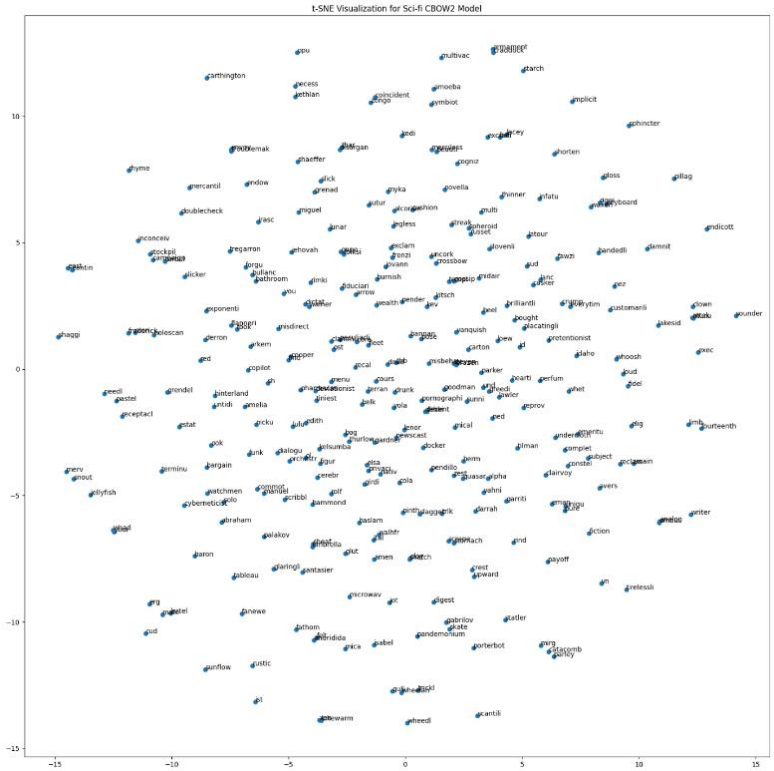


➢ 4. How does the quality of the hotel review-based embeddings compare with the Sci-fi-based embeddings? Elaborate.

Compared to Sci-fi embeddings, the hotel review-based embeddings can capture higher semantic relevance and more straightforward context. The t-SNE results also show that the hotel dataset has distinct clusters around common themes than the sci-fi dataset, implicating better semantic separation.



In contrast, the Sci-fi embeddings were drawn from domain-scientific and imaginative concepts, potentially leading to capture more abstract, diverse, and even less reasonable relationships. In the t-SNE visualization, there are rare words displayed and there are not quite clear distinct clusters like the hotel review's.



- **5. Choose 2 words and retrieve their 5 closest neighbours according to hotel review-based embeddings and the Sci-fi-based embeddings. Do they have different neighbours? If yes, can you reason why?**

  Comparison:
  Hotel (CBOW2) area: ['neighborhood', 'section', 'indoor', 'embarcadero', 'vicin']
  Hotel (CBOW5) area: ['clubhous', 'mediterranean', 'thatch', 'attend', 'path']
  Sci-fi (CBOW2) area: ['shelf', 'oldtim', 'harmlessli', 'satellit', 'land']

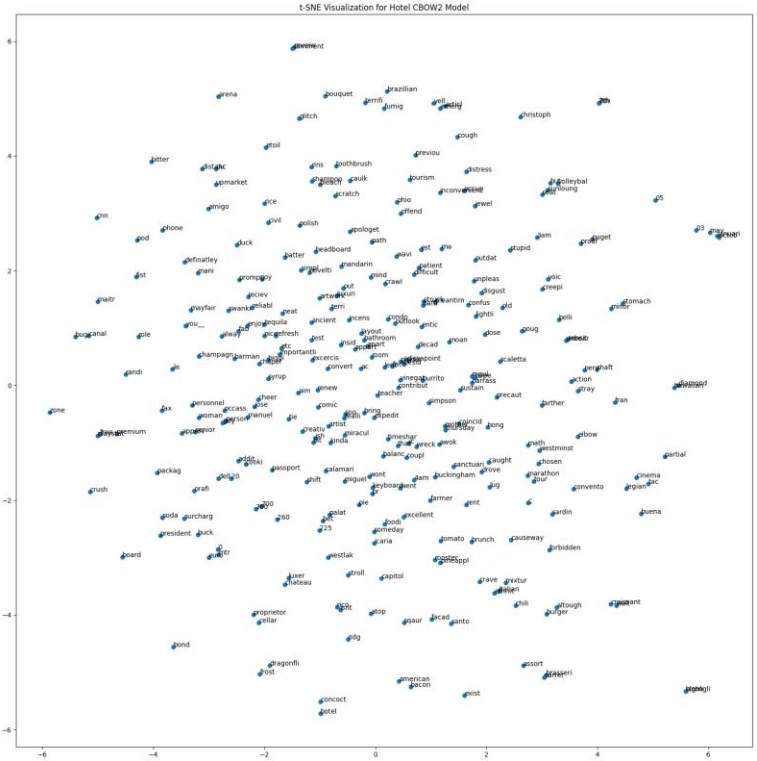
  Hotel (CBOW2) star: ['boutiqu', 'hous', 'prais', 'chain', 'unaccomod']
  Hotel (CBOW5) star: ['presidenti', 'minut', 'min', 'compar', 'transform']
  Sci-fi (CBOW2) star: ['quench', 'saucer', 'chrysler', 'wonderingli', 'sun']

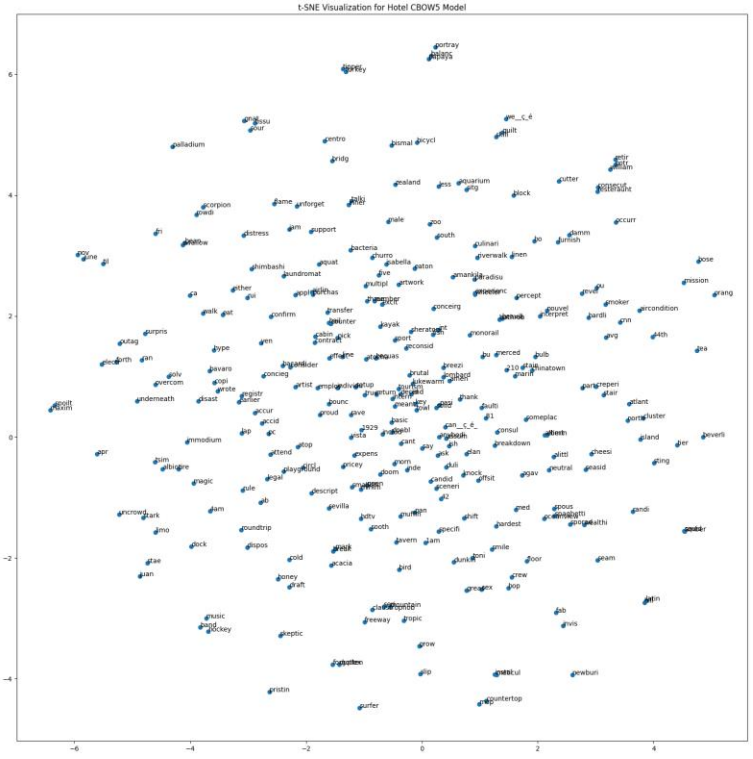
  When we chose the word embeddings for ‘area’ and ‘star’, there are quite distinct neighbors across hotel reviews and Sci-fi contexts. These differences arise from contextual nuances.

  The neighbors like ‘neighborhood’ and ‘vicin’ for ‘area’ in hotel reviews reflect locality. For the same word ‘area’ in Sci-fi, ‘shelf’ and ‘satellit’ suggest diverse interpretations. ‘Star’ in hotel context has ‘boutiqu’ and ‘hous’, potentially showing hotel categories. In Sci-fi, the neighbors of ‘star’ are like ‘sun’ and ‘saucer’, which mirror celestial topics.

- **6. What are the differences between CBOW2 and CBOW5? Can you "describe" them?**

  For loss function trends, the CBOW5 model, with a larger window to capture a broader context around the target word, starts with a much lower loss and a faster drop than CBOW2 in response to training data, suggesting superior initial performance. Additionally, the words seem to have a more dispersed pattern and fewer overlapping words in CBOW5 visualization than in CBOW2 visualization, which means that a wider context can capture relatively more distant and more nuanced relationships.


t-SNE Visualization for Hotel CBOW2 Model


t-SNE Visualization for Hotel CBOW5 Model