Machine Learning for NLP
Lab Report of Exercise 2

October 30, 2022

1 Data Inspection, Preprocessing and Class Structure (Part 1)

1.1 Data Inspection

The Tripadvisor data set consists of 20'491 reviews. Each review is between 7 and 1'931 words long. On average, a review has length of 104 words.

The SCIFI data set is a SCIFI story with approx. 1'227'345 sentences and 15'540'723 words. On average, each sentence has 12 words.

In Figures 1 and 2, the most common words of the Tripadvisor and SCIFI dataset are displayed, respectively.

1.2 Preprocessing Steps

The following preprocessing steps were applied:

(i) **Lower-Casing:** All words are lower-cased. Reasoning: The task of creating word embeddings is of semantical nature. Hence, capitalization does not benefit and only introduces additional noise to the data.

(ii) **Split Sentences Into its Subsets ('Sub-Sentences'):** Each sentence is split into its subsets. Reasoning: The proximity of words across different subsets (i.e., end word of one subset and start word of the next subset) are arbitrary and do not contain any semantic meaning. Hence, all sentences are split into their subsets based on the punctuation symbols [.,!?].

(iii) **Remove Extraneous Symbols:** All extraneous symbols that are not punctuation are removed from the subsets. Reasoning: Special characters do not support the model in learning the semantic relationship between words. Since they introduce noise to the data, numerical symbols are removed.

(iv) **Remove Numerical Symbols:** All numeric symbols are removed from the subsets. Reasoning: Numbers do not support the model in learning the semantic relationship between words. Since they introduce noise to the data, numerical symbols are removed.

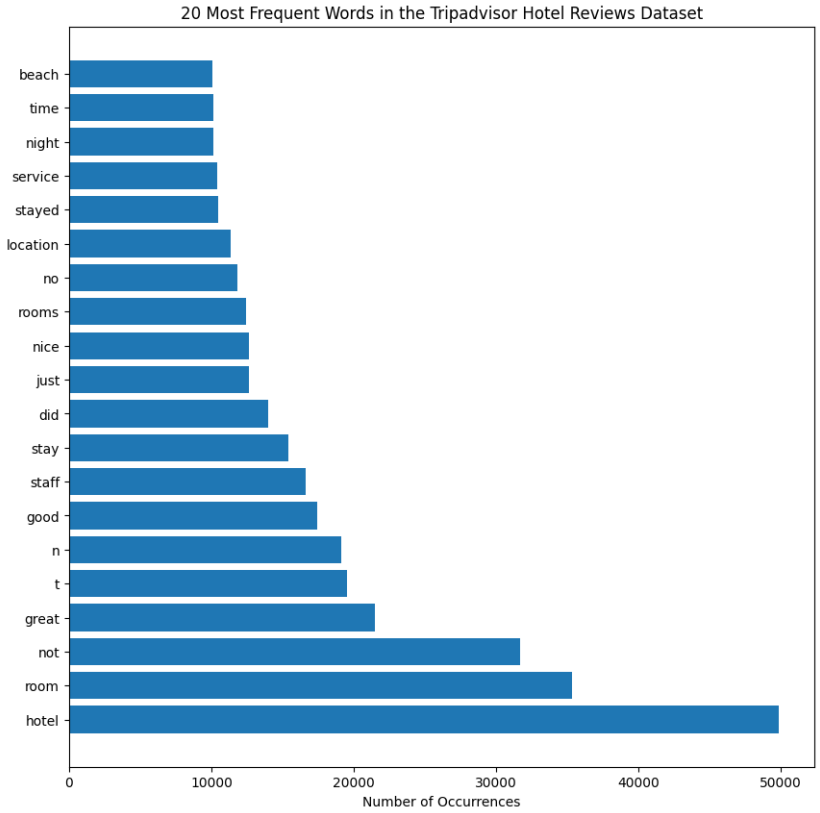

Figure 1: Top 20 Most Common Words in the Tripadvisor Dataset

1.3 Class Structure

For this exercise, an object-oriented approach was followed. This enhances readability and encapsulates the behavior of different functionalities. To achieve this, the following classes were created in order to process the data:

- **Vocabulary:** The Vocabulary class processes text and extracts the vocabulary for mapping.
- **CBOWVectorizer:** The Vectorizer class coordinates the Vocabularies and puts them to use.
- **TextDataset:** The TextDataset class encapsulates the data sets and their vectorizers. It implements the __getitem__() and __len__() methods which are used by the PyTorch DataLoader class to construct batches.

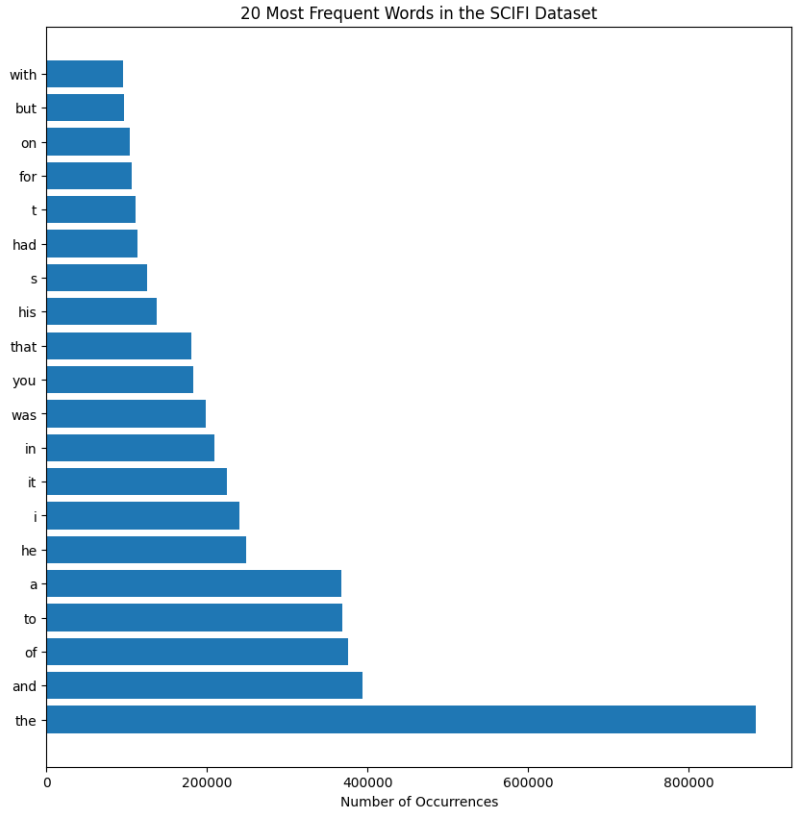

Figure 2: Top 20 Most Common Words in the SCIFI Dataset

2 Testing of the Embeddings (Part 2)

In the following, we are going to discuss the embeddings of the two trained CBOW2 models.

2.1 Hotel Reviews Dataset: Analysis of 9 words

For the hotel reviews dataset, the following words were chosen.

- **Nouns:** hotel, beach, flowers
- **Verbs:** walk, recommend, isolate
- **Adjectives:** good, clean, mixed

The five nearest neighbours according to CBOW2 are displayed in Table 1. The last column indicates the number of occurrences of the word in the corpus.

The following points can be observed:

- **Semantic Embeddings Rather Poor:** Looking at the words which are the nearest neighbours in the embedding space, the model performs rather poorly. For almost all words, the semantic relationship of the given word and its nearest neighbours is far apart or non-existent.
- **Part of Speech for Nouns Reasonable:** For the three chosen nouns (hotel, beach, email), the nearest neighbours are also nouns in many cases. Hence, the part-of-speech was correctly

	Word	*NN1*	*NN2*	*NN3*	*NN4*	*NN5*	# in Corpus
Nouns	hotel	*belize*	*shampoonothing*	*parachute*	*horseferry*	*godawful*	49'886
	beach	*shrugged*	*shans*	*moldey*	*charactor*	*swank*	10'072
	email	*reward*	*temperature*	*endless*	*throughout*	*bourdon*	380
Verbs	walk	*faux*	*grigio*	*aris*	*cocierges*	*workmanship*	6'256
	recommend	*dicussion*	*quads*	*uncooperative*	*doughy*	*bidding*	4'865
	send	*gucci*	*cheesy*	*lafeyette*	*trendy*	*get*	246
Adjectives	good	*playpen*	*otur*	*fluxuated*	*pinot*	*foreginers*	17'424
	clean	*nancy*	*overdue*	*meeted*	*sport*	*asheville*	9'599
	mixed	*roomier*	*laquered*	*nazis*	*tonic*	*estrella*	333

Table 1: Top 5 Nearest Neighbours in the Tripadvisor Hotel Reviews Dataset (CBOW2).

learned by the CBOW2 model for the nouns. However, for the verbs and adjectives, this cannot be said because the nearest neighbours are often from a different part-of-speech.

Overall, the nearest neighbours of the chosen words do not make much sense. Hence, the semantic embeddings can be considered as poor.

The following points might be reasons for the poor performance:

- **Small Context Size:** As the CBOW2 model only has a context size of two (i.e., 2 words to the left and to the right of the target word), this might not be enough to appropriately capture the semantic context of a word. One could use a larger context size (e.g., CBOW5) to check if a larger context improves model performance.
- **Large Batch Size:** In order to improve the learning speed of the model, a batch size of 512 was chosen. Arguably, this led to a decrease in performance. Hence, in future research a smaller batch size could be applied.
- **Low Embedding Dimension:** As instructed in the exercise statement, an embedding dimension of 50 was chosen. Most probably, a 50-dimensional space is not enough to capture the semantic meanings of the English language. As such, the model is not able to appropriately capture/store differences in the semantic meaning of words. It is expected that a higher embedding space would lead to better model performance.

2.2 SCIFI Dataset: Analysis of 9 words

For the SCIFI dataset, the following words were chosen.

- **Nouns:** time, way, literature
- **Verbs:** see, think, publish
- **Adjectives:** little, long, conventional

The five nearest neighbours according to CBOW2 are displayed in Table 2. The last column indicates the number of occurrences of the word in the corpus.

The following points can be observed:

- **Semantic Embeddings Rather Poor:** Looking at the nearest neighbours in the embedding space, the model performs even poorer than the Hotel Reviews-based model. One of the few neighbours that is somehow semantically close to the given word is *juniors* to **little**. These are semantically similar, even though the part-of-speech is different.

	Word	*NN1*	*NN2*	*NN3*	*NN4*	*NN5*	# in Corpus
Nouns	time	*widemos*	*freshly*	*tregasid*	*tactless*	*levantman*	32'907
	way	*fmmy*	*karksen*	*essed*	*helena*	*zpt*	21'081
	literature	*superenergetic*	*nothiflfe*	*viewfinder*	*alnjost*	*wasters*	366
Verbs	see	*paymenc*	*rudely*	*overwick*	*incbnsequentiality*	*concerting*	21'211
	think	*sunken*	*sumptuously*	*endross*	*resample*	*imew*	17'003
	publish	*undersecretary*	*tiiis*	*dummkopf*	*scatters*	*interceptor*	182
Adjectives	little	*juniors*	*suppareddi*	*copter*	*allallu*	*willa*	17'267
	long	*porpoise*	*hynm*	*captivating*	*binds*	*ducats*	16'490
	conventional	*symmetry*	*supplies*	*ooks*	*dustmote*	*tainly*	173

Table 2: Top 5 Nearest Neighbours in the SCIFI Dataset (CBOW2).

- **Part of Speech for Verbs (Partly) Reasonable:** For the three chosen verbs (see, think, publish), the nearest neighbours are partly of the same part-of-speech. However, for the nouns and adjectives, this cannot be said because the nearest neighbours are often from a different part-of-speech.

The following points might be reasons for the poor performance:

- **Small Context Size:** As the CBOW2 model only has a context size of two (i.e., 2 words to the left and to the right of the target word), this might not be enough to appropriately capture the semantic context of a word. One could use a larger context size (e.g., CBOW5) to check if a larger context improves model performance.
- **Large Batch Size:** In order to improve the learning speed of the model, a batch size of 512 was chosen. Arguably, this led to a decrease in performance. Hence, in future research a smaller batch size could be applied.
- **Low Embedding Dimension:** As instructed in the exercise statement, an embedding dimension of 50 was chosen. Most probably, a 50-dimensional space is not enough to capture the semantic meanings of the English language. As such, the model is not able to appropriately capture/store differences in the semantic meaning of words. It is expected that a higher embedding space would lead to better model performance.

2.3 General Comparison of the Embedding Quality Based on Hotel Reviews and SCIFI

Overall, it can be observed that the Tripadvisor Hotel Reviews-based embedding quality is better than the SCIFI-based embedding quality.

Possible reasons for this could be found in:

- **Preprocessing the Texts Into Subsets:** The preprocessing step of splitting each sentence into its subsets/sub-sentences might have harmed the quality of the SCIFI dataset. This is because in a written story, part-of-speeches such as conjunctive adverbs are inserted into the text and separated with commas. Splitting each sentence into its subsets, therefore, might create sub-sentences with no meaning.
- **Short Sentence Length:** As mentioned in Section 1.1, the average sentence length of the SCIFI dataset is only 12 words. On the other hand, the Tripadvisor dataset has an average sentence length of 104 words. Shorter sentences might, therefore, also lead to poorer embedding quality.

2.4 In-Depth Comparison of the Embeddings Based on Hotel Reviews and SCIFI for 2 Words

In this section, the Tripadvisor Hotel Reviews-based and SCIFI-based embeddings for 2 specific words are compared.

For this, the following words are chosen:

- room
- time

	Word	***NN1***	***NN2***	***NN3***	***NN4***	***NN5***	**# in Corpus**
Hotel Review-Based	room	*celler*	*thirties*	*utilising*	*departure*	*rolled*	35'367
	time	*furthermore*	*acts*	*gasshem*	*mamosa*	*srambled*	10'135
SCIFI-Based	room	*whosever*	*authorides*	*highpoints*	*yourse*	*storij*	10'873
	time	*widemos*	*freshly*	*tregasid*	*tactless*	*levantman*	32'907

Table 3: Top 5 Nearest Neighbours in the Tripadvisor Hotel Reviews and SCIFI Dataset.

It can be observed that the nearest neighbours are totally different for the two embedding spaces. These differences make sense. Possible explanations might be:

- **Different Domains:** The two embeddings were generated by two entirely different datasets from different domains (Hotel Reviews vs. SCIFI story).
- **Sentence Style:** The style of sentences (reviews vs. "eloquent" written stories) may contribute to the different embeddings. For example, there are many dialogues in the SCIFI datasets.
- **Part-of-Speech:** The part-of-speeches encountered might be different in the two datasets. For example it could be reasonable to assume that the SCIFI story dataset contains more verbs than the Tripadvisor Hotel Reviews dataset. This might also lead to differences in the embeddings.