Exercise 2 - To Embed or Not to Embed

In this exercise, we used PyTorch to create props-specific word embeddings. To be specific, we trained CBOW embeddings on the Trip Advisor hotel reviews and Sci-fi stories datasets. Then we tested the embeddings with 2 and 5 context width and compared the results.

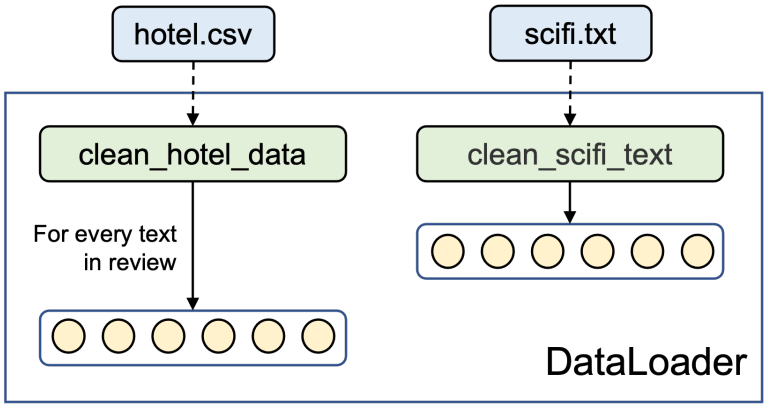
Data

- tripadvisor_hotel_reviews.csv
- scifi.txt

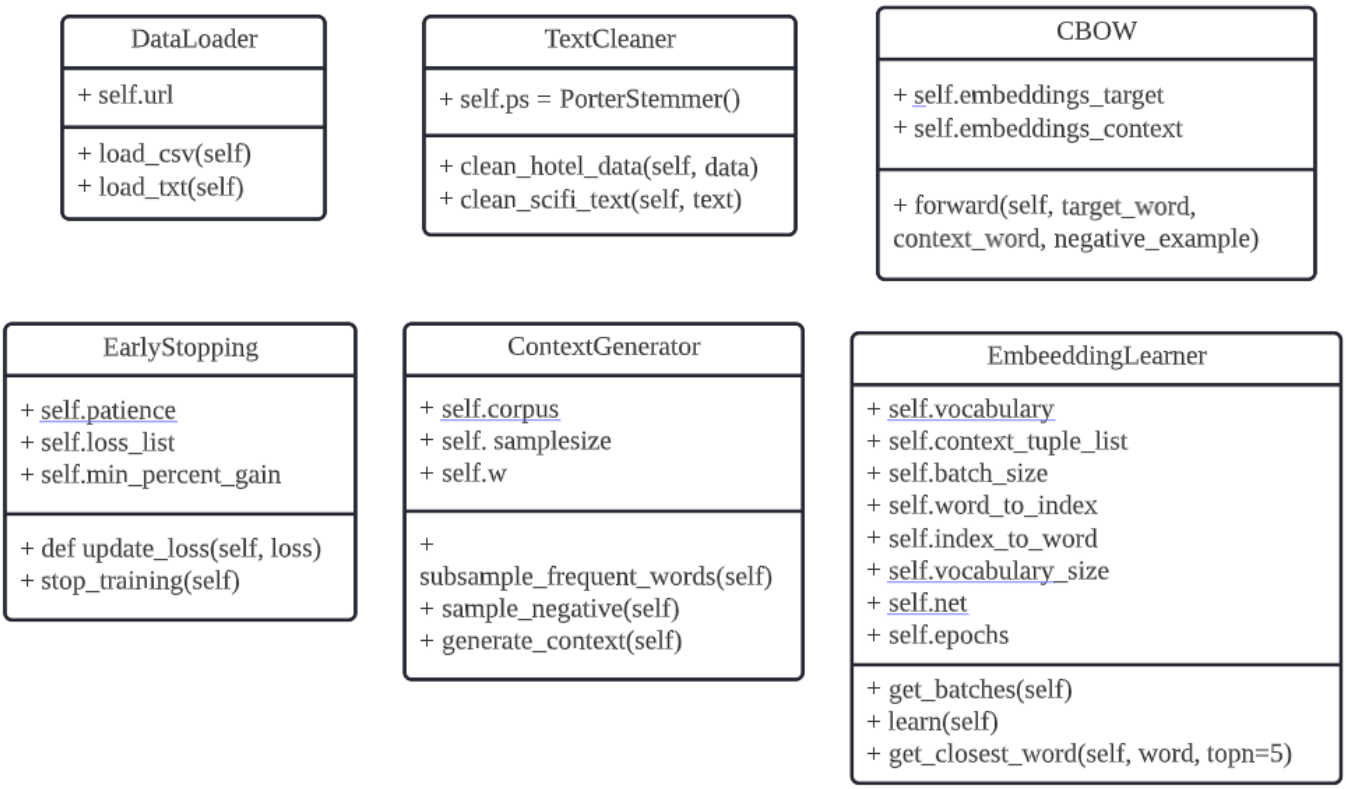
We performed data preprocessing with the following steps:

1. Lowercasing
2. Filtering Letters
3. Correcting Spellings
4. Stemming
5. Removing Stopwords

We included these steps in a cleaning function. For each dataset, we wrote a function to adapt the data file format. We designed a TextCleaner class to encapsulate the two functions. The DataLoader class is shown in figure 1, in which the yellow circles represent specific preprocessing steps.



The class structure of the whole program is designed as figure 2 below:



Results

- **CBOW2 for the Hotel Reviews dataset**

words	neighbor 1	neighbor 2	neighbor 3	neighbor 4	neighbor 5
bar	place 3.50	fine 3.71	nice 3.73	look 3.76	day 3.76
window	use 4.51	area 4.52	hotel 4.53	right 4.59	tri 4.60
mouth	andr 6.03	day 6.16	servic 6.17	beauti 6.18	pike 6.21
pay	locat 4.31	work 4.35	hotel 4.44	tri 4.48	close 4.49
teach	mind 7.31	level 7.93	food 7.95	separ 7.96	grace 8.07
jump	downtown 5.24	place 5.50	close 5.83	center 5.93	hotel 5.97
tireless	experi 4.87	nice 4.93	locat 4.95	travel 4.97	staff 4.99
rare	nice 5.39	use 5.53	great 5.56	staff 5.59	guest 5.66
nicest	book 5.75	wonder 5.88	food 5.93	staff 5.93	locat 6.02

From the results output by CBOW2, we can find that while most connections seem non-sense, there do have some meaningful neighbors.

For mouth, the connection can be **mouth** eat **pike**. For **teach**, **mind** and **level** can be used to describe the study process.

However, a great number of connections shown above are very weak. For example, we might suppose there is a connection between **window** and **hotel** because it sounds reasonable. But it's implicit. When we think of window, it doesn't remind us of the hotel at once. Besides, hotel itself does not provide us with useful information about window.

We noticed the average embedding distance is around 5.

- **CBOW2 for the Sci-Fi story dataset**

words	neighbor 1	neighbor 2	neighbor 3	neighbor 4	neighbor 5
mouth	brought 6.27	made 6.32	codiscover 6.49	came 6.60	movi 6.60
sport	get 6.42	kirk 6.52	ground 6.58	predatorycreatur 6.71	long 6.72
bedroom	made 5.80	swam 5.99	voluptu 6.21	bombshel 6.22	judi 6.33
trust	toward 5.99	kept 6.20	foreman 6.31	seiz 6.45	gaunt 6.51
fail	windi 5.93	sir 6.92	want 6.32	neverth 6.36	may 6.47
pay	list 6.07	innoc 6.66	question 6.70	knowledgeseek 6.70	lift 6.99
largest	woman 6.49	work 6.60	exactli 6.61	cornel 6.75	cover 6.95
clear	wordlessli 6.09	see 6.62	erect 6.73	lazili 6.74	dear 6.80
endless	seclud 6.48	thought 6.70	pubbound 6.80	seem 6.85	turkey 6.88

The quality of Sci-fi-based embeddings seems weakly worse than the hotel review-based embeddings. We made this conclusion by the word distance. The former has 6 words with an average distance to neighbors below 6, while the latter has 0. Also, it is more difficult to find meaningful connections in Sci-fi-based embeddings.

We guess this is because the topic of hotel reviews is concentrated, and the words are often used to express similar meanings. But in the scifi text, the topics may be various, and thus, more challenging to get the specific definitions.

To further verify this idea, we randomly chose two words- **mouth** and **pay** and checked their embedding neighbors on both datasets.

datasets	word	neighbor 1	neighbor 2	neighbor 3	neighbor 4	neighbor 5
hotel	mouth	special 5.85	money 5.88	highli 5.89	amaz 5.92	terribl 5.94
scifi	mouth	say 3.79	earth 4.04	alma 4.08	toward 4.08	one 4.20
hotel	pay	price 0.97	friendli 1.16	hotel 1.19	time 1.20	stay 1.20
scifi	pay	said 5.61	interest 6.00	fourth 6.03	chair 6.04	say 6.06

From the above table, we found that both words get different neighbors. By comparing the distances, we know that **mouth**'s embedding result is weakly better on the SciFi dataset, but for the hotel review dataset, **pay** gets far better embeddings. As we know, **pay** is often used in hotel reviewing conditions. By far, we can verify our suspicion:

1. Embedding performance is determined by the text topics.
2. More frequently shown words in more concentrated topics are more likely to get better embeddings because more information will be extracted from their neighbor dictionaries.

- **CBOW5 for the Hotel Reviews dataset**

words	neighbor 1	neighbor 2	neighbor 3	neighbor 4	neighbor 5
bar	citi 1.02	hotel 1.07	use 1.13	quit 1.16	thing 1.16
window	got 1.74	bathroom 1.77	old 1.79	larg 1.18	clean 1.83
mouth	special 5.85	money 5.88	highli 5.89	amaz 5.92	terribl 5.94
pay	price 0.97	friendli 1.16	hotel 1.19	time 1.20	stay 1.20
teach	saw 6.02	happen 6.09	upgrad 6.17	rude 6.18	inclin 6.24
jump	garag 5.90	soon 5.94	doe 5.95	coffe 6.00	walk 6.02
tireless	took 3.64	area 3.68	internet 3.76	access 3.80	use 3.82
rare	trunk 6.18	stain 6.56	felt 6.57	properti 6.60	remaind 6.63
nicest	pool 4.80	hour 4.82	singl 4.86	resort 4.86	morn 4.86

We trained CBOW5 on the Hotel Reviews dataset. We found that this operation largely improved the embedding quality. We are able to discover some very strong connections such as **{pay, price}, {window, large}, {window, clean}, {nicest, resort}, {nicest, pool}, {rare, stain}**, etc. Under this situation, we confirmed that these embeddings showed up not occasionally but in a more logical way.

For some of the words, we noticed that the neiborhood distances changed into small values, such as **window**, **pay,** and **tireless**, which means a better model performance. We can infer that using a larger context width 5 will get better embeddings than 2. Larger context dictionaries may provide more information because we can extract neighbor words from a wider range. This can be seen as an improvement in the long-term memory ability of the CBOW model.

- **CBOW5 for the Sci-Fi story dataset**

words	neighbor 1	neighbor 2	neighbor 3	neighbor 4	neighbor 5
mouth	say 3.79	earth 4.04	alma 4.08	toward 4.08	one 4.20
sport	depress 6.36	good 6.41	homicid 6.49	roar 6.49	hit 6.51
bedroom	peopl 4.85	got 5.02	wa 5.11	thi 5.13	came 5.19
trust	poor 7.94	go 7.99	come 8.09	cours 8.12	lift 8.21
fail	right 4.79	light 4.79	hi 4.79	day 4.80	came 4.83
pay	said 5.61	interest 6.00	fourth 6.03	chair 6.04	say 6.06
largest	man 4.71	say 4.73	put 5.12	step 5.19	head 5.29
clear	take 4.97	befor 5.29	even 5.36	sorri 5.40	came 5.46
endless	number 6.60	jackson 6.64	power 6.64	suggest 6.69	signific 6.70