Classification with NN model

In task2, I vectorized the words using Word2Vec and created a simple neural network classifier for training. I created two models using CBOW and Skipgram and trained them with 80% of the input data and validated them with the rest of the input data. After training, I could get the accuracy of 0.85 on both models. There were a little bit of variations in the results after each training and validation on both models but there were no big differences between them.
I worked on this task with python on IPython environment because it can show the process visually. You can open and run it in Jupyter Notebook or Google colab. To run the file, you need to download a file from this link and place it in the same folder with classification.ipynb file. If you just want to see the code, you can see it in pdf file - classification_ipython.pdf.

Below are the detailed informations for each step.

Step1. Read a file.
Read the CSV file with 'utf-8' encoding and make a dataframe form using Pandas library. I dropped the 'hid' columns because it's not necessary for classification. The dataframe after reading the file is like below.

	chunk	has_space
0	Landmark Center, 8th Fl	0
1	Contact: The C3 team at MakemeC3@cic.us -- Add...	0
2	A powerful tool for developers, the MySQL Data...	0
3	Easy access to T, Hubway, and parking	0
4	Check out our Private Offices	1

Figure 1. A snippet of the dataframe made with input data

Step2. Data Preprocessing
Preprocess the data in 'chunk' column before training for more precise result.

no.	Steps	Example	
		Before preprocessing	After preprocessing
1	Remove leading and trailing whitespaces, newline and tab characters.	\r\t\t\t\t\t\t\t\t\tWorkbars	workbar s
2	Convert text to lowercase	Landmark Center	landmark center
3	Remove the characters [\, ', +, "]	$500+ / month	$500 / month
4	Replace email address forms with "emailadd"	MakemeC3@cic.us	email add
5	Replace phone number forms with "phonenumber"	450-345-3458	phonenumber
6	Replace rent fee forms with "rentfee"	$500 / month	rentfee
7	Remove none-alphabet characters	8th	th

Step3. Tokenizing
Tokenizing the sentences in the 'chunk' column before applying Word2Vec.

no.	Steps	Used library and methods
1	Tokenize the string into words using nltk library.	nltk, word_tokenize

2	Remove non-alphabetic tokens, such as punctuation	-
3	Filter out stop words using nltk library.	nltk, stopwords
4	Lemmatize the words.	nltk, WordNetLemmatizer

	chunk	has_space	clean_chunk	token
0	Landmark Center, 8th Fl	0	landmark center th fl	[landmark, center, th, fl]
1	Contact: The C3 team at MakemeC3@cic.us -- Add...	0	contact the rentfee team at emailadd additiona...	[contact, the, rentfee, team, at, emailadd, ad...
2	A powerful tool for developers, the MySQL Data...	0	a powerful tool for developers the mysql datab...	[a, powerful, tool, for, developer, the, mysql...
3	Easy access to T, Hubway, and parking	0	easy access to t hubway and parking	[easy, access, to, t, hubway, and, parking]
4	Check out our Private Offices	1	check out our private offices	[check, out, our, private, office]
5	\r\r\r\r \r By Michael Carne...	0	by michael carney written on june rentfee rent...	[by, michael, carney, write, on, june, rentfee...
6	\r\t\t\t\t\t\t\t\t\t\t\tWorkbars coworking spac...	1	workbar s coworking spaces provide the right b...	[workbar, s, coworking, space, provide, the, r...
7	◯We went from 3,000 sq ft to 13,000 sq ft. Tha...	0	we went from rentfee sq ft to rentfee rentfee...	[we, go, from, rentfee, sq, ft, to, rentfee, r...
8	Common space / kitchen, available for use day ...	0	common space kitchen available for use day and...	[common, space, kitchen, available, for, use, ...
9	Workbar Union $350 / month full-time open wor...	1	workbar union rentfee full time open workspace...	[workbar, union, rentfee, full, time, open, wo...

Figure 2. A dataframe after data preprocessing and tokenizing.

Step4. Word Embedding using Word2Vec.
Make a word to vector model and applying pre-trained Google news Dataset.

Step1	Implementing Word2Vec model using CBOW (Continuous Bag of Words) model, with pre-trained Google Word2Vec (Google news Dataset). (Note. I didn't include the Google news Dataset file since it's a very large file. But if you want to run the code, then you need to download it from this link and place it in the same folder with classification.ipynb file.)
Step2	Building the vocabulary table on our data.
Step3	Training of the word2vec model.
Step4	Repeat Step1 to Step4 with Skipgram model instead of CBOW.

Step5. Training with Neural network.
Make a neural network classifier model and train it.
The characteristics of classifier is like below.

• Number of Dense layers	2
• Activation Function	Relu and sigmoid for the last dense layer
• Dropout	0.7
• Optimizer	Adadelta
• Loss	Binary Cross Entropy

6. Results

Before training the neural network model, I have divided our dataset into training and test sets - 80% of the input data is for training the neural network model, the rest of them is for validating the model. Below are evaluation metrics.

• Accuracy	$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$ Where $\hat{y}_i$ is the predicted value of the i-th sample, y_i is the corresponding true value and n is the number of samples.
• Confusion matrix	True Negative (TP) False Positive (FP) False Negative (FN) True Positive (TP)
• Precision	TP / (TP + FP)
• Recall	TP / (TP + FN)
• F1 score	2*(precision * recall) / (precision + recall)

The results on both models.

	CBOW	Skipgram
accurcy	0.850000	0.850000
confusion matrix	[[11 2] [1 6]]	[[11 2] [1 6]]
F1 score	0.800000	0.800000
precision	0.750000	0.750000
recall	0.857143	0.857143

	tokenized chunk	has_space	predicted has_space_CBOW	predicted has_space_skipgram
0	[prestigious, kendall, square, business, addre...	0	0	0
1	[let, s, grow, our, business, together, a, a, ...	0	0	0
2	[a, powerful, tool, for, developer, the, mysql...	0	0	0
3	[on, site, gym, come, soon]	0	0	0
4	[workshop, brookline, rentfee, full, time, mem...	1	1	1
5	[if, you, would, like, to, host, an, event, in...	0	1	1
6	[learnlaunch, campus, unlisted, pricing, for, ...	1	1	1
7	[the, majority, of, our, host, location, be, i...	1	1	1
8	[locate, in, a, historic, brick, and, beam, bu...	1	1	1
9	[cambridge, coworking, community, rentfee]	0	0	0
10	[your, workbar, membership, be, a, flexible, a...	1	0	0
11	[then, there, wework, the, co, work, juggernau...	1	1	1
12	[special, pricing, on, multi, month, menbershi...	0	0	0
13	[photo, credit, ed, wonsek]	0	0	0
14	[we, go, from, rentfee, sq, ft, to, rentfee, r...	0	0	0
15	[unlimited, access, most, flexible]	0	0	0
16	[coworking, be, a, concept, create, to, satisf...	0	0	0
17	[cove, rentfee, night, weekend, access, rentfe...	1	1	1
18	[article, by, evona, w, niewiadomska]	0	0	0
19	[common, space, kitchen, available, for, use, ...	0	1	1

Figure 3. True labels (has_space) and predicted labels