HW1 Report

bz957@nyu.edu

October 2018

https://github.com/abenpy/DS1011_NLP.git

1 Tokenization schemes of the dataset

1.1 Cleaning Dataset

Remove html special characters such as $' < br/ >'$ in the text data. There are 418 duplicates in the IMDB dataset. After removing the duplicates and splitting the dataset, tainning dataset size is 20000, validation dataset size is 4903, test dataset size is 24678.

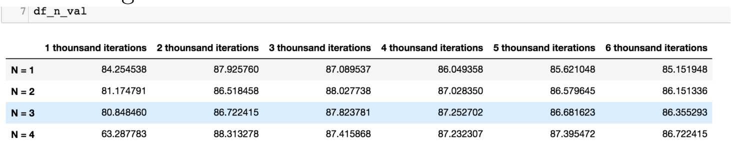
1.2 Tokenization sckemes

Step1: Using Spacy to token each sentence.
Step2: Lower words, remove punctuations and stopwords, stem words.
Step3: Use nltk everygram to get unigram, bigrams(1gram+2grams), trigrams(1gram+2grams+3grams) and fourgrams(1gram+2grams+3grams+4grams) output. Pickle them.

2 Model hyperparameters

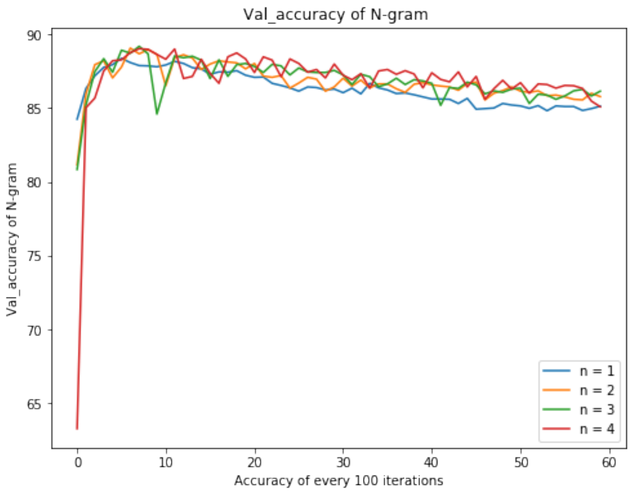
2.1 N for n-gram (n=1; 2; 3; 4)

According the length distribution of sentence, set MAX_SENTENCE_LENGTH = 450; 800; 1200; 1500 for n=1; 2; 3; 4. I got the best validation accuracy = 86.72 from n = 4 after 10 epochs. So the following tuning parameter will be based on 4grams.



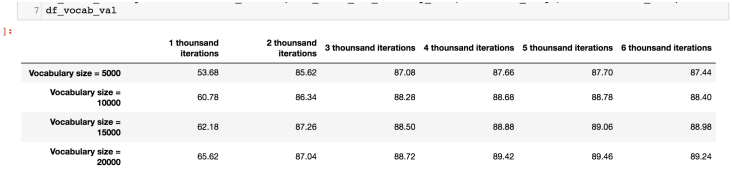
```
df_n_val
```

	1 thounsand iterations	2 thounsand iterations	3 thounsand iterations	4 thounsand iterations	5 thounsand iterations	6 thounsand iterations
N = 1	84.254538	87.925760	87.089537	86.049358	85.621048	85.151948
N = 2	81.174791	86.518458	88.027738	87.028350	86.579645	86.151336
N = 3	80.848460	86.722415	87.823781	87.252702	86.681623	86.355293
N = 4	63.287783	88.313278	87.415868	87.232307	87.395472	86.722415



2.2 Vocabulary size

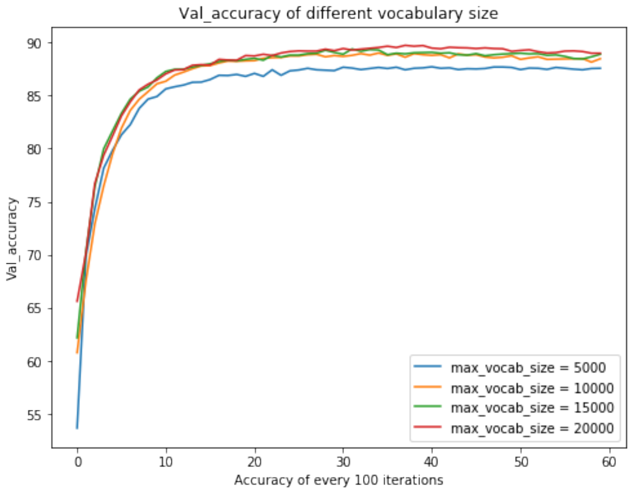
Vocabulary size = 20000 gets the best validation accuracy 89.24 after 10 epochs



```
df_vocab_val
```

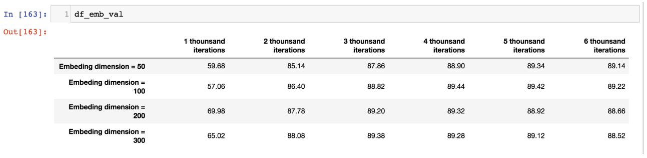
]:

	1 thounsand iterations	2 thounsand iterations	3 thounsand iterations	4 thounsand iterations	5 thounsand iterations	6 thounsand iterations
Vocabulary size = 5000	53.68	85.62	87.08	87.66	87.70	87.44
Vocabulary size = 10000	60.78	86.34	88.28	88.68	88.78	88.40
Vocabulary size = 15000	62.18	87.26	88.50	88.88	89.06	88.98
Vocabulary size = 20000	65.62	87.04	88.72	89.42	89.46	89.24



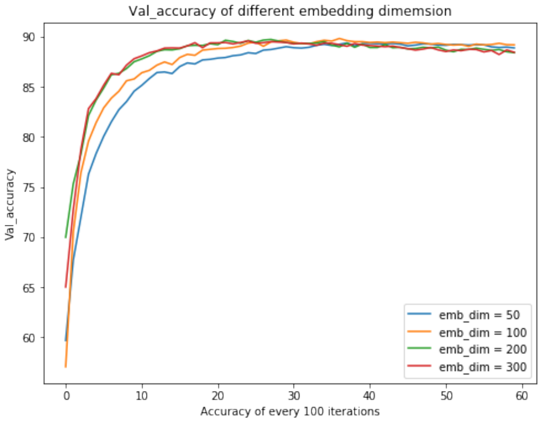
2.3 Embedding size

Embedding size = 100 gets the best validation accuracy 88.52 after 10 epochs



```
In [163]: df_emb_val
Out[163]:
```

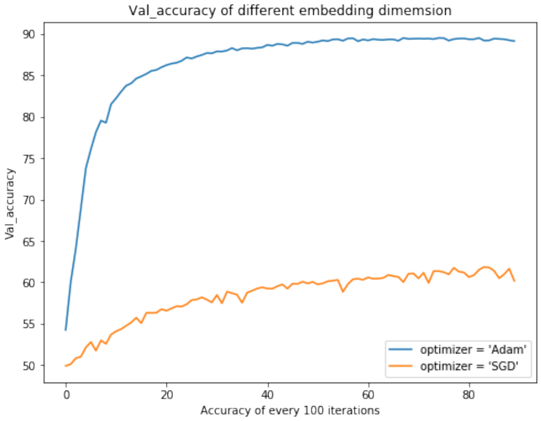
	1 thounsand iterations	2 thounsand iterations	3 thounsand iterations	4 thounsand iterations	5 thounsand iterations	6 thounsand iterations
Embeding dimension = 50	59.68	85.14	87.86	88.90	89.34	89.14
Embeding dimension = 100	57.06	86.40	88.82	89.44	89.42	89.22
Embeding dimension = 200	69.98	87.78	89.20	89.32	88.92	88.66
Embeding dimension = 300	65.02	88.08	89.38	89.28	89.12	88.52

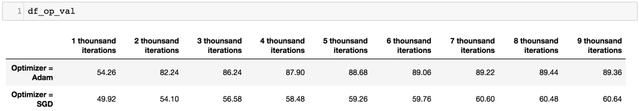


3 Optimization hyperparameters

3.1 Optimizer itself

Adam optimizer gets the best validation accuracy 89.36 after 10 epochs



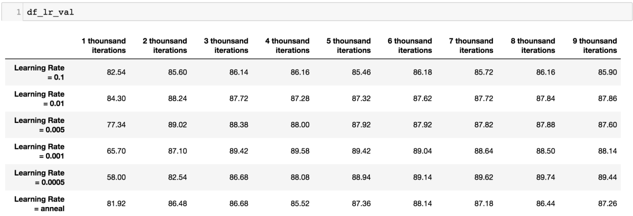


```
df_op_val
```

	1 thounsand iterations	2 thounsand iterations	3 thounsand iterations	4 thounsand iterations	5 thounsand iterations	6 thounsand iterations	7 thounsand iterations	8 thounsand iterations	9 thounsand iterations
Optimizer = Adam	54.26	82.24	86.24	87.90	88.68	89.06	89.22	89.44	89.36
Optimizer = SGD	49.92	54.10	56.58	58.48	59.26	59.76	60.60	60.48	60.64

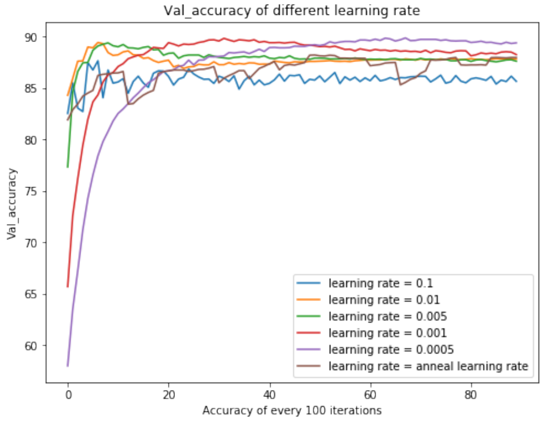
3.2 Learning rate

Learning rate = 0.0005 gets the best validation accuracy 89.44 after 15 epochs, even better than learning rate anneal.



```
df_lr_val
```

	1 thousand iterations	2 thousand iterations	3 thousand iterations	4 thousand iterations	5 thousand iterations	6 thousand iterations	7 thousand iterations	8 thousand iterations	9 thousand iterations
Learning Rate = 0.1	82.54	85.60	86.14	86.16	85.46	86.18	85.72	86.16	85.90
Learning Rate = 0.01	84.30	88.24	87.72	87.28	87.32	87.62	87.72	87.84	87.86
Learning Rate = 0.005	77.34	89.02	88.38	88.00	87.92	87.92	87.82	87.88	87.60
Learning Rate = 0.001	65.70	87.10	89.42	89.58	89.42	89.04	88.64	88.50	88.14
Learning Rate = 0.0005	58.00	82.54	86.68	88.08	88.94	89.14	89.62	89.74	89.44
Learning Rate = anneal	81.92	86.48	86.68	85.52	87.36	88.14	87.18	86.44	87.26



3.3 Final validating and testing accuracy

I choose n=4, lr=0.0005, embedding dimension =100, vocabulary size = 20000, adam optimizer for the model. Get val Acc 89.10 and test Acc 88.27

3.4 3 right and wrong predictions in validation dataset

Right prediction index: [1, 3, 4], label: ['Negative', 'Negative', 'Negative']
1,PROM NIGHT (2008)directed by: Nelson McCormickstarring...Add to all that predictable plot turns, a terrible soundtrack and a big lack of respect to the original material, and you have quite a stinker.
3,I really hate this retarded show, it SUCKS!...
4,Not a knock on Korman as he was very funny on the Carol Burnett show.
Wrong prediction index: [0, 2, 15], label: ['Negative', 'Positive', 'Positive']
0,It's unlikely that anyone except those who adore silent films will appreciate any of the lyrical camera-work and busy (but scratchy) background score that accompanies this 1933 release...
2,What another reviewer called lack of character development, I call understatement...
15,I have to admit to enjoying bad movies. I love them I watch all of them...