Shirin Mohebbi

Neural Networks and Deep Learning (Spring 2021)

Homework #2: Kohonen Self-Organizing Map (SOM)

May 14, 2021

Phase 1: Document Preprocessing

1. Remove all non-letter characters from the documents.
2. Extract all words of the document and remove the short words (length ≤ 2).
3. Remove all stop words (e.g., ‘a’, ‘and’, ‘what’, ...), given in file ‘stopwords.txt’.
4. Compute the feature vector for each document, using TF-IDF weighting scheme.

$$v_{ij} = \log(1 + tf_{ij}) \times \log(\frac{N}{df_j})$$

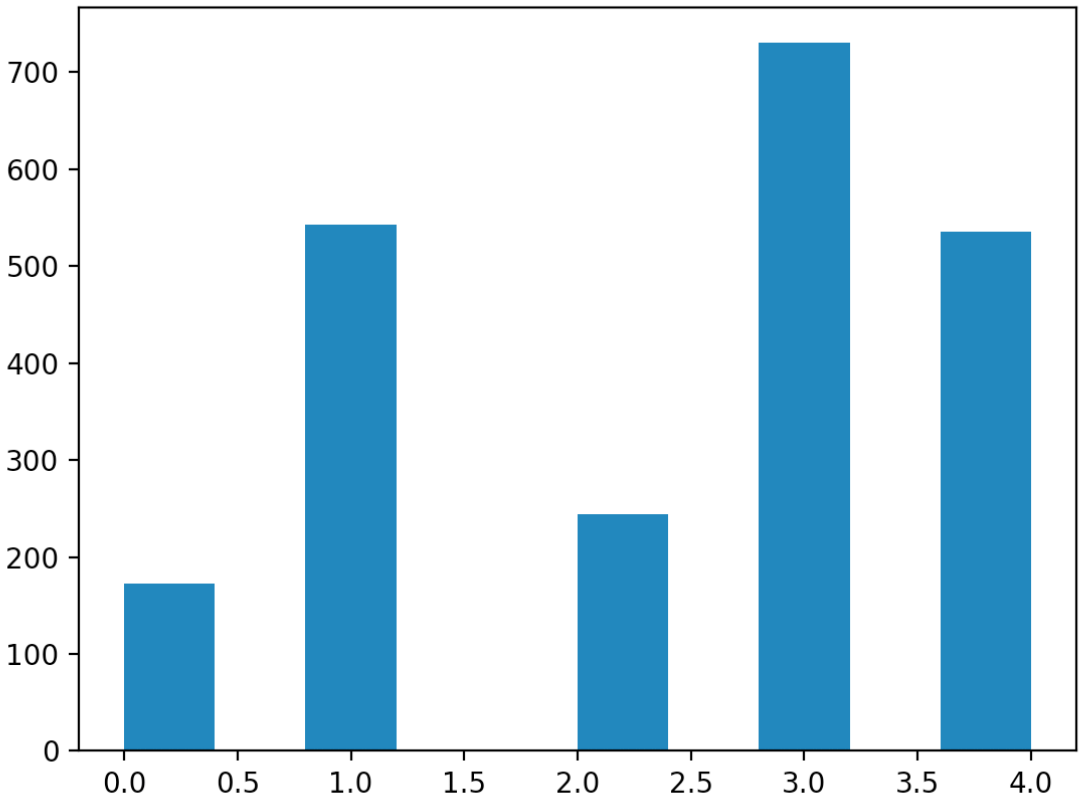
fisrt i read data from ‘bbc-text.csv’ and ‘stopwords.txt’. then in the pre-process(self, content, stopWord) function i preprocess the data base on 4 steps above. and save vsms list in ‘vsm-np.txt’ file.

Phase 2: SOM Clustering

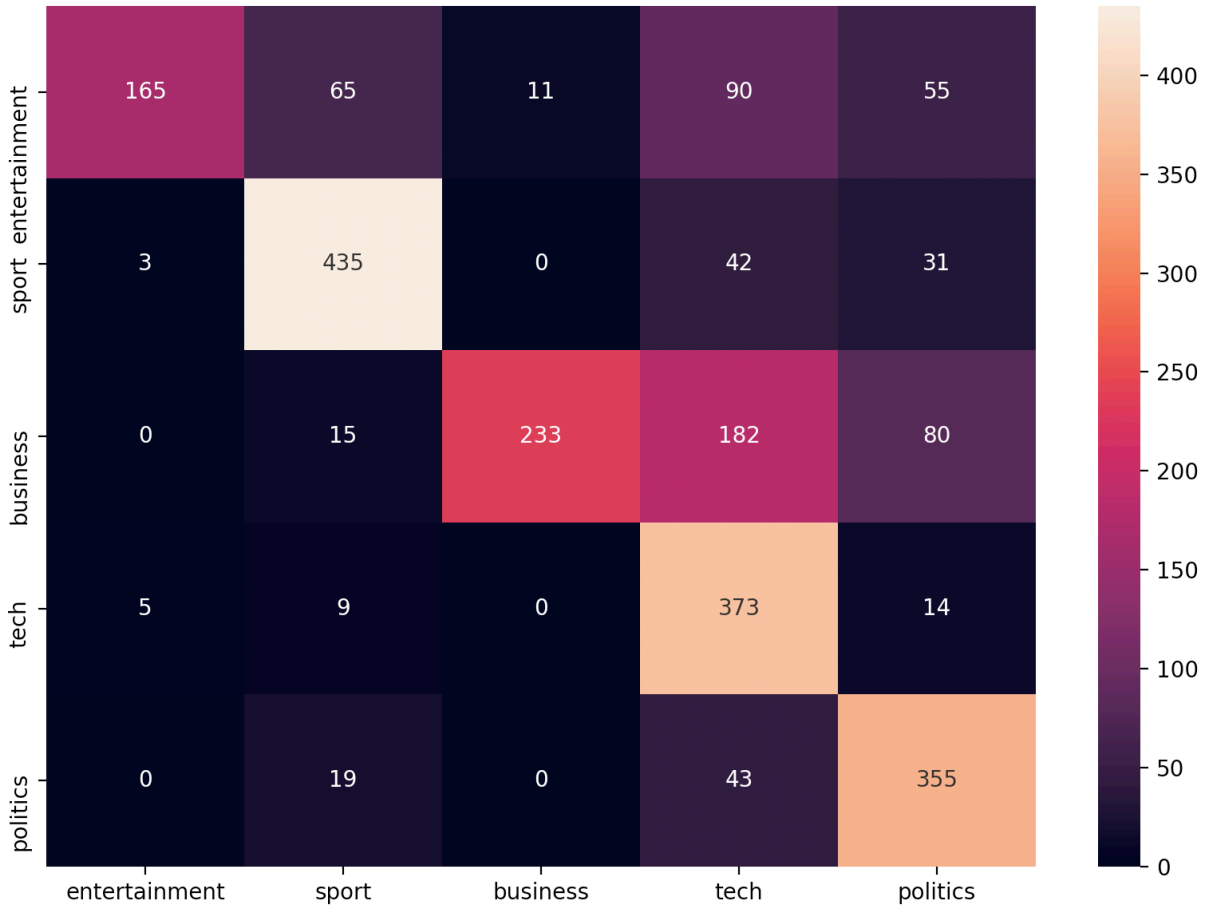
a) Winner-takes-all approach

for this part first i randomly initialize weight vector which is size = (5,28965) because we have 5 clusters. then i set the learning parameters. lin each epoc all datas are chosen to train the model. but the order of choosing data is random. for each data we find distance between data and output neurons. and find closest to data as k. then update winner neuron(k) based on its distance to data. the learning iterations goes till 500 epoc reached or when largest change(max norm change in w in previous epoc) is less than 0.02. for prediction we find distance of each sample to each 5 output neuron and set label of closest to data.

hit plot:



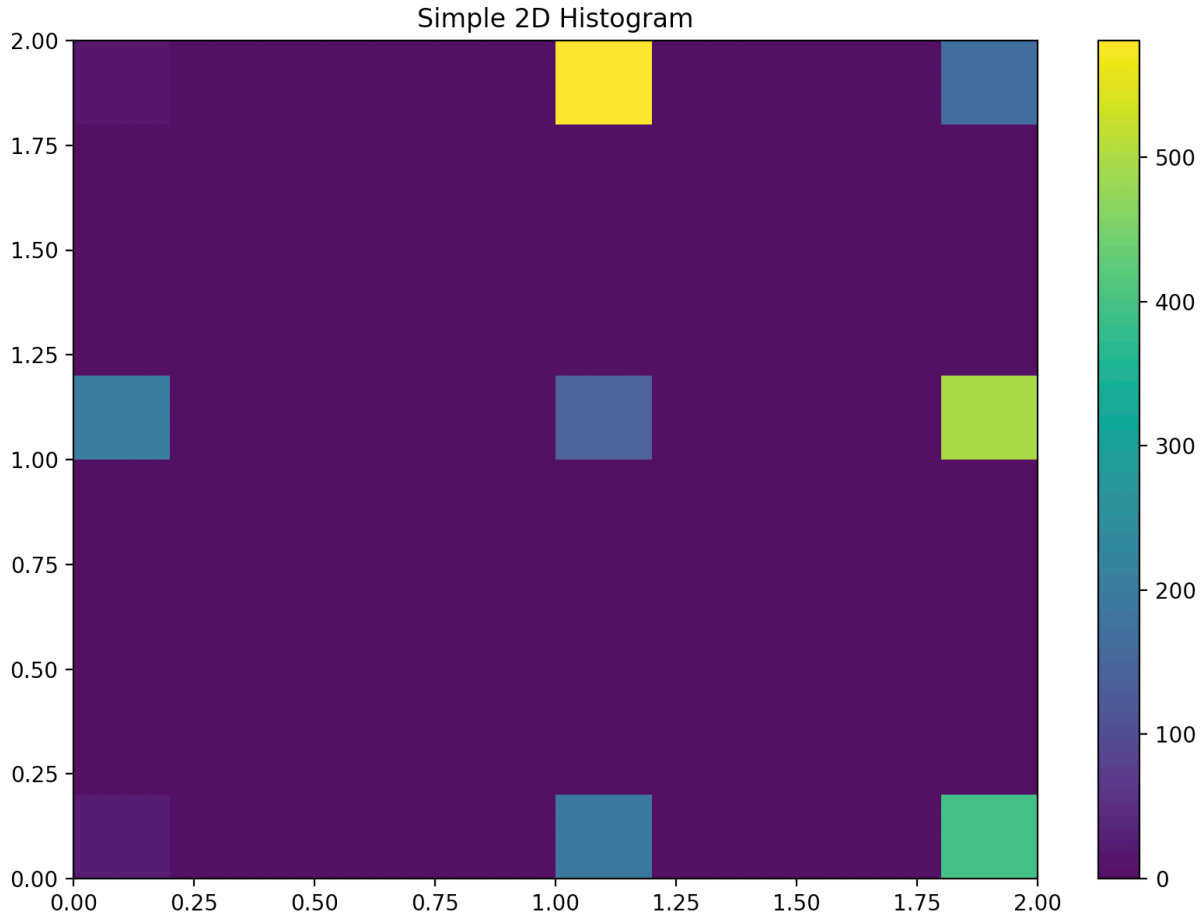
confusion matrix:



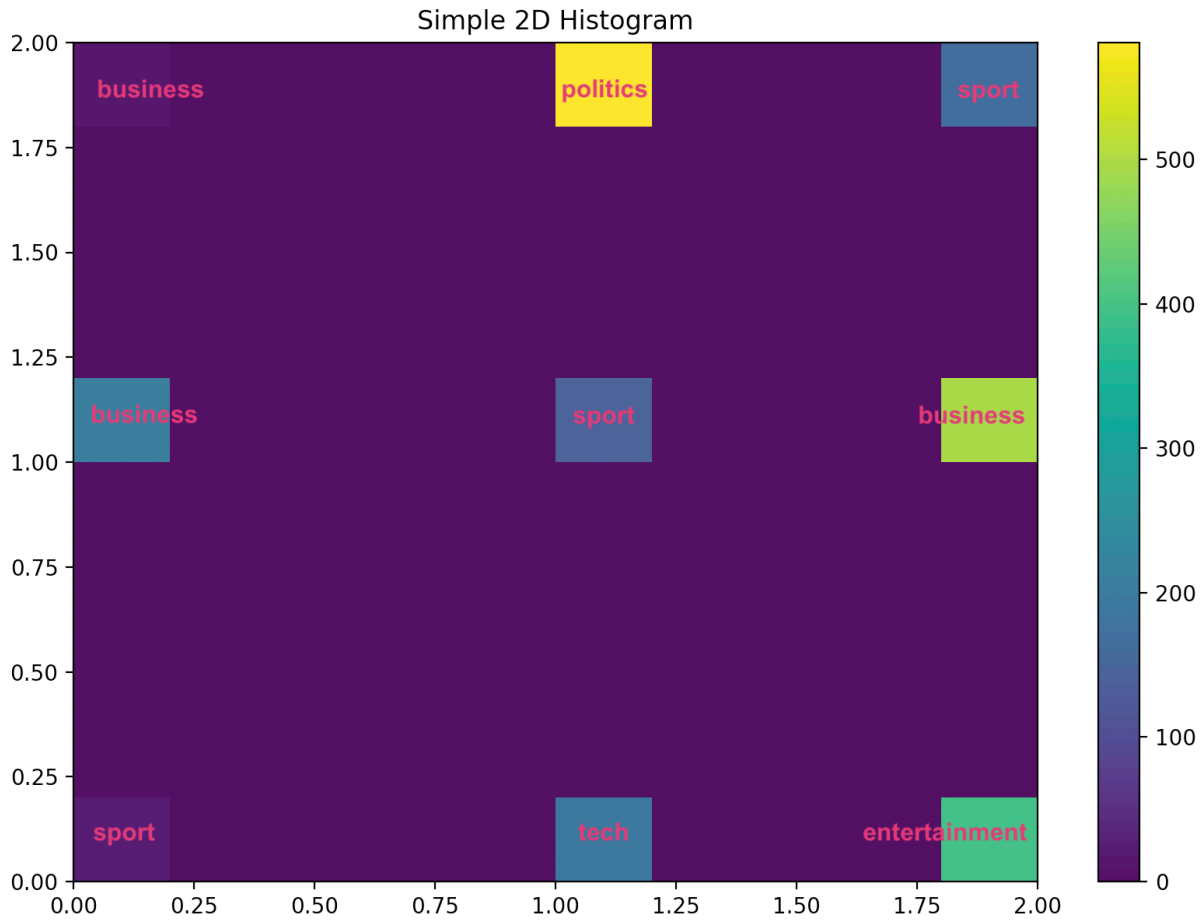
b) On-center, off-surround approach

this part is much like the previous one, the difference is in output neurons, here we have 3*3 output neurons. and in every iteration furthermore winner's weights, also neighbour's weights will update base on their distance to winner.

hit plot:




```
(2225, 28965)
matrix [['business', 'politics', 'sport'], ['business', 'sport', 'business'],
 ['sport', 'tech', 'entertainment']]
all distances 426.89186572298325
```


Simple 2D Histogram
business
politics
sport
business
sport
business
sport
tech
entertainment
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
500
400
300
200
100
0