Mini Project 2 - Sentiment Analysis

Group 9

December 2018

1 Introduction

According to an assignment we have tried and to some extend successfully built machine learning algorithm that learns to recognize sentiment of an amazon review.

2 Reviews & GloVe

To test and later see the accuracy (results) of the model we have used 10,261 amazon reviews on musical instruments that have 5-core rating. In order to gain additional information about certain words in review message we have used 50 dimensional GLoVe file (embeddings) that simply in the vectores provides additional information about the word that we can use.

3 Data processing

As to start off data processing we have loaded reviews and embeddings into data sets and tokenized reviews in a way where we have splitted sentences into single words and altered the 5-core rating into 3 categories - 0: negative, 1: neutral, 2: positive. After the tokenization the data set looked like the sample below. After that we have flattened the data set to single words in column word and finally assigned proper embedding vectors to specific words and structured the data set into features (embeddings) and labels (categories)

```
tokenized                            flatten
+---+--------------------+-----+     +---+--------+---+
| id|               words|label|     | _1|      _2| _3|
+---+--------------------+-----+     +---+--------+---+
|  0|[good, not, much,...|  2.0|     |  0|    good|2.0|
|  2|[jake, the, produ...|  2.0|     |  0|     not|2.0|
|  4|[it, does, the, j...|  2.0|     |  0|    much|2.0|
|  6|[good, windscreen...|  2.0|     |  0|      to|2.0|
|  8|[no, more, pops, ...|  2.0|     |  0|   write|2.0|
| 10|[the, best, cable...|  2.0|     |  0|   about|2.0|
| 12|[monster, standar...|  2.0|     |  0|   here,|2.0|
| 14|[didn't, fit, my,...|  1.0|     |  0|     but|2.0|
| 16|[great, cable, pe...|  2.0|     |  0|      it|2.0|
| 18|[best, instrument...|  2.0|     |  0|    does|2.0|
| 20|[one, of, the, be...|  2.0|     |  0| exactly|2.0|
| 22|[it, works, great...|  2.0|     |  0|    what|2.0|
| 24|[has, to, get, us...|  1.0|     |  0|    it's|2.0|
| 26|[awesome, i, love...|  2.0|     |  0|supposed|2.0|
| 28|[it, works!, i, b...|  2.0|     |  0|     to.|2.0|
| 30|[definitely, not,...|  0.0|     |  0| filters|2.0|
| 32|[durable, instrum...|  2.0|     |  0|     out|2.0|
| 34|[fender, 18, ft.,...|  2.0|     |  0|     the|2.0|
| 36|[so, far, so, goo...|  2.0|     |  0|     pop|2.0|
| 38|[add, california,...|  2.0|     |  0| sounds.|2.0|
+---+--------------------+-----+     +---+--------+---+
```

```
ready for training
+--------------------+-----+
|            features|label|
+--------------------+-----+
|[0.32940111923076...|  1.0|
|[0.29777697180616...|  2.0|
|[0.24063043809523...|  2.0|
|[0.02920190000000...|  2.0|
|[0.25275608301886...|  2.0|
|[0.28202170885167...|  0.0|
|[0.23154156111111...|  2.0|
|[0.17441562857142...|  2.0|
|[0.25496338085106...|  2.0|
|[0.32703426315789...|  2.0|
|[0.25917704918032...|  2.0|
|[0.13906502,-0.02...|  2.0|
|[0.05009434347826...|  1.0|
|[0.35522498688524...|  2.0|
|[0.37010584202898...|  2.0|
|[0.35257505625000...|  0.0|
|[0.15081176359223...|  0.0|
|[0.11517424210526...|  2.0|
|[0.29338983840000...|  2.0|
|[0.20986450833333...|  2.0|
+--------------------+-----+
```

4 Training & evaluation

Firstly, we have split data set into 80:20 sets where 80% of the data were used to train the model and 20% to evaluate the accuracy. We have defined a neural network layers for our model where first layer consists of 50 neurons as we use 50 dimensional embeddings file for input, 100 middle layer and 3 output layer as we want to output into 3 categories. We have set model to train on max 100 iterations and after that provide us with accuracy:

```
Test set accuracy = 0.867748279252704
```