Regression model comparison

December 4, 2021

0.1 Import Packages

```
In [95]: #spark sql imports
         from pyspark.sql import *
         from pyspark.sql.functions import *
         from pyspark.sql.types import *

         import matplotlib.pyplot as plt
         %matplotlib inline
```

```
In [2]: #create spark
        spark = SparkSession.builder.appName('RedditData').config("spark.jars.packages").enabl
```

```
In [3]: #connect to dataset

        df = spark.read \
            .option("quote", "\"")  \
            .option("escape", "\"") \
            .option("ignoreLeadingWhiteSpace",True) \
            .csv("hdfs:///user/yizhou/data/group_project/data_cleaned1.csv",inferSchema=True,
```

```
In [4]: df_nlp = df.select('subreddit', 'clean_comment', 'ups', 'downs', 'score').dropna().dro
```

```
In [19]: rc_t = df_nlp.sample(False, .1)
         rc_t.write.format('json').save('hdfs:///user/yizhou/data/group_project/rc_t')
```

```
In [20]: rc_t = spark.read.format('json').load('hdfs:///user/yizhou/data/group_project/rc_t/*'
         rc_t.cache()
         print(rc_t.count())
```

```
3740
```

```
In [6]: df_nlp.show(10)
```

```
+-------------+--------------------+---+-----+-----+
|    subreddit|       clean_comment|ups|downs|score|
+-------------+--------------------+---+-----+-----+
|    AskReddit|definitely not sa...|  2|    0|    2|
|          nba|most pathetic soc...|217|    0|  217|
|         news|  ill take that bet | 46|    0|   46|
|          nfl|domo arigoto mari...|  3|    0|    3|
|          nba|    lol tony parker |  1|    0|    1|
|    AskReddit|knew group people...|  1|    0|    1|
|    AskReddit|              latex |  2|    0|    2|
|todayilearned|issue people beli...|  3|    0|    3|
|    AskReddit|then how come pos...|  1|    0|    1|
|    AskReddit|sitting the floor...|  2|    0|    2|
+-------------+--------------------+---+-----+-----+
only showing top 10 rows
```

```
In [96]: ## Exploring TF-IDF features
```

```
In [101]: from collections import Counter
          import string
          import nltk

          def term_freq_mapper(comment):
              body = comment['clean_comment']
              #tokens = nltk.tokenize.word_tokenize(body.lower())
              tokens = [word.strip(string.punctuation) for word in body.lower().split()]
              counter = Counter(tokens)
              return (comment['subreddit'], counter)

          term_freq = rc_t.rdd \
              .map(term_freq_mapper) \
              .reduceByKey(lambda a,b: a+b)
          term_freq.cache()
```

```
Out[101]: PythonRDD[1718] at RDD at PythonRDD.scala:53
```

```
In [102]: sub_term_freq_res_0 = term_freq.take(1)[0]
          sub_0 = sub_term_freq_res_0[0]
          term_freq_res_0 = sub_term_freq_res_0[1]
          print(sub_0)
          print(sorted(list(term_freq_res_0.items()), key=lambda t_f:t_f[1], reverse=True)[0:5
```

```
todayilearned
[('the', 151), ('that', 75), ('you', 65), ('and', 56), ('they', 33), ('not', 31), ('was', 30),
```

```
In [103]: # document frequency
          doc_freq = term_freq \
              .flatMap(lambda sub_counter: list(sub_counter[1])) \
              .map(lambda word: (word, 1)) \
              .reduceByKey(lambda a, b: a + b) \
              .sortBy(lambda word_docfreq: word_docfreq[1], False)
          doc_freq.cache()
```

```
Out[103]: PythonRDD[1730] at RDD at PythonRDD.scala:53
```

```
In [104]: print(doc_freq.collect()[0:50])
```

```
[('and', 21), ('the', 19), ('you', 18), ('that', 17), ('way', 17), ('are', 17), ('for', 17), (
```

```
In [105]: #number of document
          num_docs = term_freq.count()
          print(num_docs)
```

```
48
```

```
In [106]: #Inverse Document Frequency
          import math

          inv_doc_freq = doc_freq \
              .map(lambda t_df: (t_df[0], math.log(num_docs / t_df[1]))) \
              .sortBy(lambda t_idf: t_idf[1], True)
          inv_doc_freq.cache()
```

```
Out[106]: PythonRDD[1738] at RDD at PythonRDD.scala:53
```

```
In [107]: inv_doc_freq_res = inv_doc_freq.collect();
          print(inv_doc_freq_res[0:50])
```

```
[('and', 0.8266785731844679), ('the', 0.9267620317414504), ('you', 0.9808292530117262), ('that
```

```
In [108]: sub_1 = 'nba'
          term_freq_res_1 = term_freq.sortByKey().lookup(sub_1)[0]
          print(sub_1)
          print(sorted(list(term_freq_res_1.items()), key=lambda t_f:t_f[1], reverse=True)[0:5
```

```
nba
[('the', 216), ('and', 80), ('that', 74), ('you', 48), ('this', 46), ('game', 40), ('was', 39)
```

```
In [120]: sub_2 = 'funny'
          term_freq_res_2 = term_freq.sortByKey().lookup(sub_2)[0]
          print(sub_2)
          print(sorted(list(term_freq_res_2.items()), key=lambda t_f:t_f[1], reverse=True)[0:5
```

```
funny
[('the', 150), ('you', 76), ('and', 69), ('that', 67), ('this', 38), ('but', 33), ('are', 32),
In [114]: inv_doc_freq_map_res = inv_doc_freq.collectAsMap()
          inv_doc_freq_map_res
```

```
Out[114]: {'and': 0.8266785731844679,
           'the': 0.9267620317414504,
           'you': 0.9808292530117262,
           'that': 1.037987666851675,
           'way': 1.037987666851675,
           'are': 1.037987666851675,
           'for': 1.037987666851675,
           'not': 1.037987666851675,
           'but': 1.037987666851675,
           'really': 1.0986122886681098,
           'was': 1.1631508098056809,
           'they': 1.1631508098056809,
           'have': 1.1631508098056809,
           'from': 1.1631508098056809,
           'good': 1.1631508098056809,
           'with': 1.1631508098056809,
           'when': 1.1631508098056809,
           'there': 1.2321436812926323,
           'what': 1.2321436812926323,
           'like': 1.2321436812926323,
           'all': 1.2321436812926323,
           'don': 1.2321436812926323,
           'people': 1.2321436812926323,
           'this': 1.2321436812926323,
           'being': 1.2321436812926323,
           'even': 1.3062516534463542,
           'over': 1.3062516534463542,
           'well': 1.3062516534463542,
           'right': 1.3062516534463542,
           'very': 1.3062516534463542,
           'just': 1.3062516534463542,
           'who': 1.3062516534463542,
           'then': 1.3062516534463542,
           'other': 1.3062516534463542,
           'because': 1.3062516534463542,
           'said': 1.3062516534463542,
           'never': 1.3062516534463542,
           'about': 1.3062516534463542,
           'more': 1.3062516534463542,
           'your': 1.3062516534463542,
           'against': 1.3062516534463542,
           'best': 1.3062516534463542,
           'first': 1.3862943611198906,
           'into': 1.3862943611198906,
           'which': 1.3862943611198906,
           'any': 1.3862943611198906,
           'their': 1.3862943611198906,
           'those': 1.3862943611198906,
           'know': 1.3862943611198906,
           'can': 1.3862943611198906,
           'off': 1.3862943611198906,
           'say': 1.3862943611198906,
           'them': 1.3862943611198906,
           'some': 1.3862943611198906,
           'get': 1.3862943611198906,
           'see': 1.3862943611198906,
           'too': 1.3862943611198906,
           'his': 1.3862943611198906,
           'most': 1.3862943611198906,
           'here': 1.3862943611198906,
           'out': 1.3862943611198906,
           'been': 1.3862943611198906,
           'how': 1.3862943611198906,
           'would': 1.3862943611198906,
           'only': 1.3862943611198906,
           'doesn': 1.3862943611198906,
           'much': 1.3862943611198906,
           'idea': 1.3862943611198906,
           'great': 1.3862943611198906,
           'many': 1.3862943611198906,
           'could': 1.3862943611198906,
           'lot': 1.3862943611198906,
           'does': 1.3862943611198906,
           'why': 1.3862943611198906,
           'make': 1.3862943611198906,
           'need': 1.3862943611198906,
           'work': 1.4733057381095203,
           'second': 1.4733057381095203,
           'old': 1.4733057381095203,
           'yes': 1.4733057381095203,
           'buy': 1.4733057381095203,
           'while': 1.4733057381095203,
           'thought': 1.4733057381095203,
           'one': 1.4733057381095203,
           'two': 1.4733057381095203,
           'yeah': 1.4733057381095203,
           'before': 1.4733057381095203,
           'mean': 1.4733057381095203,
           'remember': 1.4733057381095203,
           'reddit': 1.4733057381095203,
           'man': 1.4733057381095203,
           'still': 1.4733057381095203,
           'fun': 1.4733057381095203,
           'pretty': 1.4733057381095203,
           'now': 1.4733057381095203,
           'give': 1.4733057381095203,
           'after': 1.4733057381095203,
           'down': 1.4733057381095203,
           'did': 1.4733057381095203,
           'makes': 1.4733057381095203,
           'than': 1.4733057381095203,
           'think': 1.4733057381095203,
           'shit': 1.4733057381095203,
           'usually': 1.4733057381095203,
           'top': 1.4733057381095203,
           'fact': 1.4733057381095203,
           'got': 1.4733057381095203,
           'live': 1.4733057381095203,
           'love': 1.4733057381095203,
           'different': 1.4733057381095203,
           'long': 1.4733057381095203,
           'had': 1.4733057381095203,
           'went': 1.4733057381095203,
           'thing': 1.4733057381095203,
           'has': 1.4733057381095203,
           'hope': 1.4733057381095203,
           'also': 1.4733057381095203,
           'getting': 1.4733057381095203,
           'least': 1.4733057381095203,
           'guess': 1.4733057381095203,
           'isn': 1.4733057381095203,
           'always': 1.4733057381095203,
           'better': 1.4733057381095203,
           'ever': 1.4733057381095203,
           'big': 1.4733057381095203,
           'again': 1.4733057381095203,
           'day': 1.4733057381095203,
           'something': 1.4733057381095203,
           'look': 1.4733057381095203,
           'him': 1.4733057381095203,
           'having': 1.4733057381095203,
           'person': 1.4733057381095203,
           'didn': 1.4733057381095203,
           'want': 1.4733057381095203,
           'going': 1.4733057381095203,
           'will': 1.4733057381095203,
           'anything': 1.4733057381095203,
           'actually': 1.4733057381095203,
           'probably': 1.4733057381095203,
           'working': 1.4733057381095203,
           'looking': 1.5686159179138452,
           'used': 1.5686159179138452,
           'guys': 1.5686159179138452,
           'called': 1.5686159179138452,
           'lol': 1.5686159179138452,
           'saying': 1.5686159179138452,
           'someone': 1.5686159179138452,
           'nothing': 1.5686159179138452,
           'same': 1.5686159179138452,
           'kind': 1.5686159179138452,
           'wouldn': 1.5686159179138452,
           'little': 1.5686159179138452,
           'believe': 1.5686159179138452,
           'things': 1.5686159179138452,
           'place': 1.5686159179138452,
           'everyone': 1.5686159179138452,
           'our': 1.5686159179138452,
           'around': 1.5686159179138452,
           'gets': 1.5686159179138452,
           'close': 1.5686159179138452,
           'else': 1.5686159179138452,
           'try': 1.5686159179138452,
           'lost': 1.5686159179138452,
           'back': 1.5686159179138452,
           'seen': 1.5686159179138452,
           'guy': 1.5686159179138452,
           'reason': 1.5686159179138452,
           'these': 1.5686159179138452,
           'fucking': 1.5686159179138452,
           'won': 1.5686159179138452,
           'making': 1.5686159179138452,
           'time': 1.5686159179138452,
           'years': 1.5686159179138452,
           'fuck': 1.5686159179138452,
           'full': 1.5686159179138452,
           'instead': 1.5686159179138452,
           'put': 1.5686159179138452,
           'part': 1.5686159179138452,
           'anyone': 1.5686159179138452,
           'less': 1.5686159179138452,
           'sorry': 1.5686159179138452,
           'stupid': 1.5686159179138452,
           'were': 1.5686159179138452,
           'its': 1.5686159179138452,
           'take': 1.5686159179138452,
           'should': 1.5686159179138452,
           'understand': 1.5686159179138452,
           'real': 1.5686159179138452,
           'point': 1.5686159179138452,
           'trying': 1.5686159179138452,
           'without': 1.5686159179138452,
           'sure': 1.5686159179138452,
           'wasn': 1.5686159179138452,
           'week': 1.5686159179138452,
           'through': 1.5686159179138452,
           'exactly': 1.5686159179138452,
           'bad': 1.5686159179138452,
           'damn': 1.5686159179138452,
           'every': 1.5686159179138452,
           'she': 1.5686159179138452,
           'tell': 1.6739764335716716,
           'video': 1.6739764335716716,
           'star': 1.6739764335716716,
           'find': 1.6739764335716716,
           'literally': 1.6739764335716716,
           'talking': 1.6739764335716716,
           'means': 1.6739764335716716,
           'chance': 1.6739764335716716,
           'unless': 1.6739764335716716,
           'high': 1.6739764335716716,
           'start': 1.6739764335716716,
           'sound': 1.6739764335716716,
           'needs': 1.6739764335716716,
           'set': 1.6739764335716716,
           'win': 1.6739764335716716,
           'wrong': 1.6739764335716716,
           'stuff': 1.6739764335716716,
           'each': 1.6739764335716716,
           'either': 1.6739764335716716,
           'enough': 1.6739764335716716,
           'both': 1.6739764335716716,
           'taking': 1.6739764335716716,
           'money': 1.6739764335716716,
           'read': 1.6739764335716716,
           'heard': 1.6739764335716716,
           'dude': 1.6739764335716716,
           'last': 1.6739764335716716,
           'took': 1.6739764335716716,
           'under': 1.6739764335716716,
           'wait': 1.6739764335716716,
           'fan': 1.6739764335716716,
           'where': 1.6739764335716716,
           'might': 1.6739764335716716,
           'thanks': 1.6739764335716716,
           'story': 1.6739764335716716,
           'using': 1.6739764335716716,
           'pay': 1.6739764335716716,
           'hate': 1.6739764335716716,
           'cause': 1.6739764335716716,
           'looks': 1.6739764335716716,
           'few': 1.6739764335716716,
           'though': 1.6739764335716716,
           'care': 1.6739764335716716,
           'year': 1.6739764335716716,
           'ass': 1.6739764335716716,
           'started': 1.6739764335716716,
           'haha': 1.6739764335716716,
           'watch': 1.6739764335716716,
           'comment': 1.6739764335716716,
           'yourself': 1.6739764335716716,
           'keep': 1.6739764335716716,
           'hard': 1.6739764335716716,
           'maybe': 1.6739764335716716,
           'until': 1.6739764335716716,
           'name': 1.6739764335716716,
           'face': 1.6739764335716716,
           'game': 1.6739764335716716,
           'made': 1.6739764335716716,
           'nobody': 1.6739764335716716,
           'post': 1.6739764335716716,
           'knew': 1.6739764335716716,
           'let': 1.6739764335716716,
           'show': 1.6739764335716716,
           'another': 1.6739764335716716,
           'whole': 1.6739764335716716,
           'end': 1.6739764335716716,
           'matter': 1.6739764335716716,
           'came': 1.6739764335716716,
           'new': 1.6739764335716716,
           'says': 1.6739764335716716,
           'joke': 1.6739764335716716,
           'world': 1.6739764335716716,
           'god': 1.6739764335716716,
           'life': 1.6739764335716716,
           'may': 1.791759469228055,
           'quite': 1.791759469228055,
           'stop': 1.791759469228055,
           'play': 1.791759469228055,
           'comes': 1.791759469228055,
           'away': 1.791759469228055,
           'almost': 1.791759469228055,
           'sad': 1.791759469228055,
           'help': 1.791759469228055,
           'once': 1.791759469228055,
           'come': 1.791759469228055,
           'seems': 1.791759469228055,
           'saw': 1.791759469228055,
           'screen': 1.791759469228055,
           'problem': 1.791759469228055,
           'possible': 1.791759469228055,
           'left': 1.791759469228055,
           'super': 1.791759469228055,
           'friend': 1.791759469228055,
           'true': 1.791759469228055,
           'size': 1.791759469228055,
           'feel': 1.791759469228055,
           'such': 1.791759469228055,
           'ago': 1.791759469228055,
           'hand': 1.791759469228055,
           'happens': 1.791759469228055,
           'shouldn': 1.791759469228055,
           '100': 1.791759469228055,
           'car': 1.791759469228055,
           'shut': 1.791759469228055,
           'head': 1.791759469228055,
           'hey': 1.791759469228055,
           'already': 1.791759469228055,
           'far': 1.791759469228055,
           'open': 1.791759469228055,
           'her': 1.791759469228055,
           'agree': 1.791759469228055,
           'aren': 1.791759469228055,
           'state': 1.791759469228055,
           'jpg': 1.791759469228055,
           'days': 1.791759469228055,
           'today': 1.791759469228055,
           'goes': 1.791759469228055,
           'wanted': 1.791759469228055,
           'own': 1.791759469228055,
           'done': 1.791759469228055,
           'works': 1.791759469228055,
           'watching': 1.791759469228055,
           'worse': 1.791759469228055,
           'team': 1.791759469228055,
           'level': 1.791759469228055,
           'check': 1.791759469228055,
           'definitely': 1.791759469228055,
           'giving': 1.791759469228055,
           'truly': 1.791759469228055,
           'step': 1.791759469228055,
           'use': 1.791759469228055,
           'kinda': 1.791759469228055,
           'change': 1.791759469228055,
           'white': 1.791759469228055,
           'able': 1.791759469228055,
           'likely': 1.791759469228055,
           'together': 1.791759469228055,
           'next': 1.791759469228055,
           'job': 1.791759469228055,
           'nice': 1.791759469228055,
           'happen': 1.791759469228055,
           'hell': 1.791759469228055,
           'news': 1.791759469228055,
           'assuming': 1.791759469228055,
           'terrible': 1.791759469228055,
           'entire': 1.791759469228055,
           'black': 1.791759469228055,
           'non': 1.791759469228055,
           'player': 1.791759469228055,
           'mind': 1.791759469228055,
           'surprised': 1.791759469228055,
           'gonna': 1.9252908618525775,
           'women': 1.9252908618525775,
           'children': 1.9252908618525775,
           'bunch': 1.9252908618525775,
           'everything': 1.9252908618525775,
           'meant': 1.9252908618525775,
           'per': 1.9252908618525775,
           'given': 1.9252908618525775,
           'whatever': 1.9252908618525775,
           'beat': 1.9252908618525775,
           'supposed': 1.9252908618525775,
           'poor': 1.9252908618525775,
           'fucked': 1.9252908618525775,
           'myself': 1.9252908618525775,
           'middle': 1.9252908618525775,
           'three': 1.9252908618525775,
           'fine': 1.9252908618525775,
           'a': 1.9252908618525775,
           'doing': 1.9252908618525775,
           'wow': 1.9252908618525775,
           'anyway': 1.9252908618525775,
           'weird': 1.9252908618525775,
           'cool': 1.9252908618525775,
           'line': 1.9252908618525775,
           'plus': 1.9252908618525775,
           'side': 1.9252908618525775,
           'interested': 1.9252908618525775,
           'course': 1.9252908618525775,
           'sell': 1.9252908618525775,
           'issues': 1.9252908618525775,
           'food': 1.9252908618525775,
           'somewhere': 1.9252908618525775,
           'free': 1.9252908618525775,
           'interesting': 1.9252908618525775,
           'playing': 1.9252908618525775,
           'ones': 1.9252908618525775,
           'seeing': 1.9252908618525775,
           'run': 1.9252908618525775,
           'bit': 1.9252908618525775,
           'dick': 1.9252908618525775,
           'college': 1.9252908618525775,
           'minutes': 1.9252908618525775,
           'rather': 1.9252908618525775,
           'shitty': 1.9252908618525775,
           'series': 1.9252908618525775,
           'worry': 1.9252908618525775,
           'played': 1.9252908618525775,
           'half': 1.9252908618525775,
           'thinking': 1.9252908618525775,
           'case': 1.9252908618525775,
           'lack': 1.9252908618525775,
           'family': 1.9252908618525775,
           'knows': 1.9252908618525775,
           'entirely': 1.9252908618525775,
           'regardless': 1.9252908618525775,
           'truth': 1.9252908618525775,
           'absolutely': 1.9252908618525775,
           'school': 1.9252908618525775,
           'dumb': 1.9252908618525775,
           'special': 1.9252908618525775,
           'thinks': 1.9252908618525775,
           'times': 1.9252908618525775,
           'small': 1.9252908618525775,
           'favorite': 1.9252908618525775,
           'breaking': 1.9252908618525775,
           'consider': 1.9252908618525775,
           'number': 1.9252908618525775,
           'sense': 1.9252908618525775,
           'sweet': 1.9252908618525775,
           'wish': 1.9252908618525775,
           'worth': 1.9252908618525775,
           'obviously': 1.9252908618525775,
           'wtf': 1.9252908618525775,
           'difference': 1.9252908618525775,
           'massive': 1.9252908618525775,
           'similar': 1.9252908618525775,
           'mom': 1.9252908618525775,
           'thread': 1.9252908618525775,
           'store': 1.9252908618525775,
           'honestly': 1.9252908618525775,
           'history': 1.9252908618525775,
           'career': 1.9252908618525775,
           'games': 1.9252908618525775,
           'forgot': 1.9252908618525775,
           'whether': 1.9252908618525775,
           'subreddit': 1.9252908618525775,
           'past': 1.9252908618525775,
           'dont': 1.9252908618525775,
           'sort': 1.9252908618525775,
           'situation': 1.9252908618525775,
           'considered': 1.9252908618525775,
           'later': 1.9252908618525775,
           'higher': 1.9252908618525775,
           'company': 1.9252908618525775,
           'move': 1.9252908618525775,
           'seriously': 2.0794415416798357,
           'room': 2.0794415416798357,
           'couldn': 2.0794415416798357,
           'cases': 2.0794415416798357,
           'often': 2.0794415416798357,
           'key': 2.0794415416798357,
           'enjoy': 2.0794415416798357,
           'gives': 2.0794415416798357,
           'red': 2.0794415416798357,
           'wonder': 2.0794415416798357,
           'basically': 2.0794415416798357,
           'rule': 2.0794415416798357,
           'happened': 2.0794415416798357,
           'drive': 2.0794415416798357,
           'online': 2.0794415416798357,
           'lives': 2.0794415416798357,
           'imagine': 2.0794415416798357,
           'city': 2.0794415416798357,
           'front': 2.0794415416798357,
           'bed': 2.0794415416798357,
           'skin': 2.0794415416798357,
           'group': 2.0794415416798357,
           'round': 2.0794415416798357,
           'throw': 2.0794415416798357,
           'attempt': 2.0794415416798357,
           'wall': 2.0794415416798357,
           'sign': 2.0794415416798357,
           'background': 2.0794415416798357,
           'pass': 2.0794415416798357,
           'home': 2.0794415416798357,
           'feels': 2.0794415416798357,
           'worst': 2.0794415416798357,
           'points': 2.0794415416798357,
           'thank': 2.0794415416798357,
           'must': 2.0794415416798357,
           'opinion': 2.0794415416798357,
           'shoot': 2.0794415416798357,
           'seemed': 2.0794415416798357,
           'steal': 2.0794415416798357,
           'found': 2.0794415416798357,
           'speed': 2.0794415416798357,
           'seconds': 2.0794415416798357,
           'haven': 2.0794415416798357,
           'voice': 2.0794415416798357,
           'coming': 2.0794415416798357,
           'summer': 2.0794415416798357,
           'source': 2.0794415416798357,
           'word': 2.0794415416798357,
           'pro': 2.0794415416798357,
           'wants': 2.0794415416798357,
           'becomes': 2.0794415416798357,
           'ridiculous': 2.0794415416798357,
           'posted': 2.0794415416798357,
           'apparently': 2.0794415416798357,
           'neither': 2.0794415416798357,
           'picture': 2.0794415416798357,
           'missed': 2.0794415416798357,
           'dad': 2.0794415416798357,
           'between': 2.0794415416798357,
           'self': 2.0794415416798357,
           'ask': 2.0794415416798357,
           'music': 2.0794415416798357,
           'lose': 2.0794415416798357,
           'yet': 2.0794415416798357,
           'death': 2.0794415416798357,
           'baby': 2.0794415416798357,
           'hold': 2.0794415416798357,
           'during': 2.0794415416798357,
           'mention': 2.0794415416798357,
           'league': 2.0794415416798357,
           'downvoted': 2.0794415416798357,
           'broken': 2.0794415416798357,
           'football': 2.0794415416798357,
           'exists': 2.0794415416798357,
           'words': 2.0794415416798357,
           'house': 2.0794415416798357,
           'completely': 2.0794415416798357,
           'power': 2.0794415416798357,
           'heart': 2.0794415416798357,
           'please': 2.0794415416798357,
           'mad': 2.0794415416798357,
           'system': 2.0794415416798357,
           'died': 2.0794415416798357,
           'assume': 2.0794415416798357,
           'rules': 2.0794415416798357,
           'extremely': 2.0794415416798357,
           'piece': 2.0794415416798357,
           'bullshit': 2.0794415416798357,
           'gave': 2.0794415416798357,
           'somebody': 2.0794415416798357,
           'perfect': 2.0794415416798357,
           'throughout': 2.0794415416798357,
           'behind': 2.0794415416798357,
           'huge': 2.0794415416798357,
           'happening': 2.0794415416798357,
           'technically': 2.0794415416798357,
           'future': 2.0794415416798357,
           'information': 2.0794415416798357,
           'dead': 2.0794415416798357,
           'since': 2.0794415416798357,
           'totally': 2.0794415416798357,
           'men': 2.0794415416798357,
           'hit': 2.0794415416798357,
           'proof': 2.0794415416798357,
           'type': 2.0794415416798357,
           'players': 2.0794415416798357,
           'kid': 2.0794415416798357,
           'straight': 2.0794415416798357,
           'definition': 2.0794415416798357,
           'form': 2.0794415416798357,
           'til': 2.0794415416798357,
           'woman': 2.0794415416798357,
           '2nd': 2.0794415416798357,
           'certain': 2.0794415416798357,
           'afford': 2.0794415416798357,
           'leave': 2.0794415416798357,
           'sounds': 2.0794415416798357,
           'seem': 2.0794415416798357,
           'evidence': 2.0794415416798357,
           'except': 2.0794415416798357,
           'annoying': 2.0794415416798357,
           'outside': 2.0794415416798357,
           'everywhere': 2.0794415416798357,
           'fall': 2.0794415416798357,
           'issue': 2.0794415416798357,
           '1': 2.0794415416798357,
           'average': 2.0794415416798357,
           'turn': 2.0794415416798357,
           'amount': 2.0794415416798357,
           'running': 2.0794415416798357,
           'etc': 2.0794415416798357,
           'become': 2.0794415416798357,
           'call': 2.0794415416798357,
           'yea': 2.0794415416798357,
           'fault': 2.0794415416798357,
           'ruined': 2.0794415416798357,
           'weeks': 2.0794415416798357,
           'learn': 2.0794415416798357,
           'drink': 2.0794415416798357,
           'quality': 2.0794415416798357,
           'fast': 2.0794415416798357,
           '000': 2.0794415416798357,
           'happy': 2.0794415416798357,
           'simply': 2.0794415416798357,
           'police': 2.0794415416798357,
           'style': 2.0794415416798357,
           'okay': 2.0794415416798357,
           'book': 2.0794415416798357,
           'office': 2.2617630984737906,
           'low': 2.2617630984737906,
           'night': 2.2617630984737906,
           'simple': 2.2617630984737906,
           'child': 2.2617630984737906,
           'heavy': 2.2617630984737906,
           'recent': 2.2617630984737906,
           'starts': 2.2617630984737906,
           'legal': 2.2617630984737906,
           'brother': 2.2617630984737906,
           'biggest': 2.2617630984737906,
           'marketing': 2.2617630984737906,
           'vagina': 2.2617630984737906,
           'closer': 2.2617630984737906,
           'impact': 2.2617630984737906,
           'crap': 2.2617630984737906,
           'concept': 2.2617630984737906,
           'defensive': 2.2617630984737906,
           'easily': 2.2617630984737906,
           'burn': 2.2617630984737906,
           'eat': 2.2617630984737906,
           'watched': 2.2617630984737906,
           'killed': 2.2617630984737906,
           'kills': 2.2617630984737906,
           'notice': 2.2617630984737906,
           'sex': 2.2617630984737906,
           'rights': 2.2617630984737906,
           'depends': 2.2617630984737906,
           'prove': 2.2617630984737906,
           'damage': 2.2617630984737906,
           'eyes': 2.2617630984737906,
           'owner': 2.2617630984737906,
           'conversation': 2.2617630984737906,
           'sleep': 2.2617630984737906,
           'boy': 2.2617630984737906,
           'allowed': 2.2617630984737906,
           'bring': 2.2617630984737906,
           'role': 2.2617630984737906,
           'season': 2.2617630984737906,
           'deserve': 2.2617630984737906,
           'local': 2.2617630984737906,
           'looked': 2.2617630984737906,
           'ball': 2.2617630984737906,
           'personal': 2.2617630984737906,
           'possibly': 2.2617630984737906,
           'loss': 2.2617630984737906,
           'i': 2.2617630984737906,
           'teams': 2.2617630984737906,
           'gotten': 2.2617630984737906,
           'personally': 2.2617630984737906,
           'dollars': 2.2617630984737906,
           'holy': 2.2617630984737906,
           'yep': 2.2617630984737906,
           'slightly': 2.2617630984737906,
           'random': 2.2617630984737906,
           'important': 2.2617630984737906,
           'american': 2.2617630984737906,
           'asking': 2.2617630984737906,
           'bottom': 2.2617630984737906,
           'worked': 2.2617630984737906,
           'bought': 2.2617630984737906,
           'available': 2.2617630984737906,
           'common': 2.2617630984737906,
           'paid': 2.2617630984737906,
           'tried': 2.2617630984737906,
           'rest': 2.2617630984737906,
           'lower': 2.2617630984737906,
           'young': 2.2617630984737906,
           'however': 2.2617630984737906,
           'idiot': 2.2617630984737906,
           'wars': 2.2617630984737906,
           'vote': 2.2617630984737906,
           'luck': 2.2617630984737906,
           'air': 2.2617630984737906,
           'talk': 2.2617630984737906,
           'cheaper': 2.2617630984737906,
           'door': 2.2617630984737906,
           'stay': 2.2617630984737906,
           'lucky': 2.2617630984737906,
           'pants': 2.2617630984737906,
           'month': 2.2617630984737906,
           'negative': 2.2617630984737906,
           'problems': 2.2617630984737906,
           'knowledge': 2.2617630984737906,
           'actual': 2.2617630984737906,
           'pick': 2.2617630984737906,
           'civil': 2.2617630984737906,
           'third': 2.2617630984737906,
           'karma': 2.2617630984737906,
           'decent': 2.2617630984737906,
           'weight': 2.2617630984737906,
           'bet': 2.2617630984737906,
           'based': 2.2617630984737906,
           'wearing': 2.2617630984737906,
           'road': 2.2617630984737906,
           'legit': 2.2617630984737906,
           'south': 2.2617630984737906,
           'wins': 2.2617630984737906,
           'dogs': 2.2617630984737906,
           'america': 2.2617630984737906,
           'hear': 2.2617630984737906,
           'aid': 2.2617630984737906,
           'along': 2.2617630984737906,
           'song': 2.2617630984737906,
           'shot': 2.2617630984737906,
           'met': 2.2617630984737906,
           'longer': 2.2617630984737906,
           'support': 2.2617630984737906,
           'soon': 2.2617630984737906,
           'above': 2.2617630984737906,
           'taste': 2.2617630984737906,
           'opposite': 2.2617630984737906,
           'stand': 2.2617630984737906,
           'area': 2.2617630984737906,
           'fans': 2.2617630984737906,
           'hurt': 2.2617630984737906,
           'classic': 2.2617630984737906,
           'hours': 2.2617630984737906,
           'wife': 2.2617630984737906,
           'sources': 2.2617630984737906,
           'large': 2.2617630984737906,
           'beer': 2.2617630984737906,
           'parents': 2.2617630984737906,
           'question': 2.2617630984737906,
           'country': 2.2617630984737906,
           'light': 2.2617630984737906,
           'street': 2.2617630984737906,
           'funny': 2.2617630984737906,
           'bias': 2.2617630984737906,
           'ads': 2.2617630984737906,
           'catch': 2.2617630984737906,
           'thoughts': 2.2617630984737906,
           'clear': 2.2617630984737906,
           'doubt': 2.2617630984737906,
           'link': 2.2617630984737906,
           'trust': 2.2617630984737906,
           'scared': 2.2617630984737906,
           'wonderful': 2.2617630984737906,
           'recently': 2.2617630984737906,
           'older': 2.2617630984737906,
           'fight': 2.2617630984737906,
           'spread': 2.2617630984737906,
           'himself': 2.2617630984737906,
           'attention': 2.2617630984737906,
           'turns': 2.2617630984737906,
           'fit': 2.2617630984737906,
           'cops': 2.2617630984737906,
           'pre': 2.2617630984737906,
           'cut': 2.2617630984737906,
           'pull': 2.2617630984737906,
           'mother': 2.2617630984737906,
           'ring': 2.2617630984737906,
           'couple': 2.2617630984737906,
           'healthy': 2.2617630984737906,
           'major': 2.2617630984737906,
           'multiple': 2.2617630984737906,
           'shits': 2.2617630984737906,
           'accept': 2.2617630984737906,
           'penalty': 2.2617630984737906,
           'pulled': 2.2617630984737906,
           'deal': 2.2617630984737906,
           'fuckin': 2.2617630984737906,
           'water': 2.2617630984737906,
           'friends': 2.2617630984737906,
           'article': 2.2617630984737906,
           'population': 2.2617630984737906,
           'lots': 2.2617630984737906,
           'law': 2.2617630984737906,
           'sub': 2.2617630984737906,
           'angry': 2.2617630984737906,
           'rice': 2.2617630984737906,
           'excuse': 2.2617630984737906,
           'tree': 2.2617630984737906,
           'places': 2.2617630984737906,
           'report': 2.2617630984737906,
           'winning': 2.2617630984737906,
           'fair': 2.2617630984737906,
           'phone': 2.2617630984737906,
           'attack': 2.2617630984737906,
           'quickly': 2.2617630984737906,
           'account': 2.2617630984737906,
           'killing': 2.2617630984737906,
           'girl': 2.2617630984737906,
           'hero': 2.2617630984737906,
           'finally': 2.2617630984737906,
           'walk': 2.2617630984737906,
           'answer': 2.2617630984737906,
           'especially': 2.2617630984737906,
           'choice': 2.2617630984737906,
           'themselves': 2.2617630984737906,
           'others': 2.2617630984737906,
           'pool': 2.2617630984737906,
           'table': 2.2617630984737906,
           'shape': 2.2617630984737906,
           'needed': 2.2617630984737906,
           'everybody': 2.2617630984737906,
           'surprise': 2.2617630984737906,
           'international': 2.2617630984737906,
           'expensive': 2.2617630984737906,
           'individual': 2.2617630984737906,
           'serious': 2.2617630984737906,
           'reference': 2.2617630984737906,
           'statement': 2.2617630984737906,
           'kids': 2.2617630984737906,
           'abuse': 2.2617630984737906,
           'lines': 2.2617630984737906,
           'internet': 2.2617630984737906,
           'smart': 2.2617630984737906,
           'mouth': 2.2617630984737906,
           'shows': 2.2617630984737906,
           'liked': 2.2617630984737906,
           'easy': 2.2617630984737906,
           'crazy': 2.2617630984737906,
           'community': 2.2617630984737906,
           'kill': 2.2617630984737906,
           'nearly': 2.2617630984737906,
           'ways': 2.4849066497880004,
           'alone': 2.4849066497880004,
           'didnt': 2.4849066497880004,
           'miss': 2.4849066497880004,
           'late': 2.4849066497880004,
           'body': 2.4849066497880004,
           'bro': 2.4849066497880004,
           'extra': 2.4849066497880004,
           'tip': 2.4849066497880004,
           'properly': 2.4849066497880004,
           'ended': 2.4849066497880004,
           'example': 2.4849066497880004,
           'unique': 2.4849066497880004,
           'generation': 2.4849066497880004,
           'expected': 2.4849066497880004,
           'wear': 2.4849066497880004,
           'five': 2.4849066497880004,
           'gas': 2.4849066497880004,
           'steve': 2.4849066497880004,
           'list': 2.4849066497880004,
           'majority': 2.4849066497880004,
           'receiving': 2.4849066497880004,
           'comments': 2.4849066497880004,
           'sadly': 2.4849066497880004,
           'google': 2.4849066497880004,
           'brain': 2.4849066497880004,
           'original': 2.4849066497880004,
           'hitting': 2.4849066497880004,
           'moving': 2.4849066497880004,
           'order': 2.4849066497880004,
           'iirc': 2.4849066497880004,
           'drunk': 2.4849066497880004,
           'safety': 2.4849066497880004,
           'roll': 2.4849066497880004,
           'acts': 2.4849066497880004,
           'among': 2.4849066497880004,
           'keeps': 2.4849066497880004,
           'valid': 2.4849066497880004,
           'referred': 2.4849066497880004,
           'cost': 2.4849066497880004,
           'french': 2.4849066497880004,
           'asked': 2.4849066497880004,
           '2000': 2.4849066497880004,
           'near': 2.4849066497880004,
           'father': 2.4849066497880004,
           'smith': 2.4849066497880004,
           'gay': 2.4849066497880004,
           'prefer': 2.4849066497880004,
           'human': 2.4849066497880004,
           'swear': 2.4849066497880004,
           'ran': 2.4849066497880004,
           'across': 2.4849066497880004,
           'tech': 2.4849066497880004,
           'logic': 2.4849066497880004,
           'count': 2.4849066497880004,
           'court': 2.4849066497880004,
           'hardest': 2.4849066497880004,
           'shown': 2.4849066497880004,
           'sports': 2.4849066497880004,
           'due': 2.4849066497880004,
           'twist': 2.4849066497880004,
           'odd': 2.4849066497880004,
           'finals': 2.4849066497880004,
           'awful': 2.4849066497880004,
           'van': 2.4849066497880004,
           'known': 2.4849066497880004,
           'released': 2.4849066497880004,
           'effect': 2.4849066497880004,
           'bonus': 2.4849066497880004,
           'pressure': 2.4849066497880004,
           'experience': 2.4849066497880004,
           'nor': 2.4849066497880004,
           'cats': 2.4849066497880004,
           'obvious': 2.4849066497880004,
           'tie': 2.4849066497880004,
           'war': 2.4849066497880004,
           'theory': 2.4849066497880004,
           'strong': 2.4849066497880004,
           'fly': 2.4849066497880004,
           'loved': 2.4849066497880004,
           'spend': 2.4849066497880004,
           'energy': 2.4849066497880004,
           'slap': 2.4849066497880004,
           'reasons': 2.4849066497880004,
           'thrown': 2.4849066497880004,
           'posts': 2.4849066497880004,
           'prices': 2.4849066497880004,
           'meanwhile': 2.4849066497880004,
           'match': 2.4849066497880004,
           'sport': 2.4849066497880004,
           'land': 2.4849066497880004,
           'sucks': 2.4849066497880004,
           'boring': 2.4849066497880004,
           'alright': 2.4849066497880004,
           'missing': 2.4849066497880004,
           'moment': 2.4849066497880004,
           'harder': 2.4849066497880004,
           'balls': 2.4849066497880004,
           'english': 2.4849066497880004,
           'scene': 2.4849066497880004,
           'mate': 2.4849066497880004,
           'trash': 2.4849066497880004,
           'afraid': 2.4849066497880004,
           'rare': 2.4849066497880004,
           'upvote': 2.4849066497880004,
           'immediately': 2.4849066497880004,
           'beach': 2.4849066497880004,
           'expect': 2.4849066497880004,
           'health': 2.4849066497880004,
           'ice': 2.4849066497880004,
           'quote': 2.4849066497880004,
           'windows': 2.4849066497880004,
           'page': 2.4849066497880004,
           'force': 2.4849066497880004,
           'charge': 2.4849066497880004,
           'groups': 2.4849066497880004,
           'photo': 2.4849066497880004,
           'walking': 2.4849066497880004,
           'current': 2.4849066497880004,
           'impressive': 2.4849066497880004,
           'design': 2.4849066497880004,
           'driving': 2.4849066497880004,
           'reminds': 2.4849066497880004,
           'annoyed': 2.4849066497880004,
           'mods': 2.4849066497880004,
           'cop': 2.4849066497880004,
           'technology': 2.4849066497880004,
           'regular': 2.4849066497880004,
           'rich': 2.4849066497880004,
           'guessing': 2.4849066497880004,
           'learned': 2.4849066497880004,
           'realize': 2.4849066497880004,
           'twice': 2.4849066497880004,
           'build': 2.4849066497880004,
           'trouble': 2.4849066497880004,
           'caught': 2.4849066497880004,
           'bag': 2.4849066497880004,
           'useless': 2.4849066497880004,
           'faster': 2.4849066497880004,
           'students': 2.4849066497880004,
           'blue': 2.4849066497880004,
           'suck': 2.4849066497880004,
           'argument': 2.4849066497880004,
           'break': 2.4849066497880004,
           'terms': 2.4849066497880004,
           'fuckers': 2.4849066497880004,
           'incredible': 2.4849066497880004,
           'journey': 2.4849066497880004,
           'kick': 2.4849066497880004,
           'share': 2.4849066497880004,
           'anyways': 2.4849066497880004,
           'green': 2.4849066497880004,
           'items': 2.4849066497880004,
           'cup': 2.4849066497880004,
           'otherwise': 2.4849066497880004,
           'posting': 2.4849066497880004,
           'pack': 2.4849066497880004,
           'son': 2.4849066497880004,
           'filled': 2.4849066497880004,
           'practice': 2.4849066497880004,
           'benefit': 2.4849066497880004,
           'defense': 2.4849066497880004,
           'chill': 2.4849066497880004,
           'mine': 2.4849066497880004,
           'weak': 2.4849066497880004,
           'tumblr': 2.4849066497880004,
           'meaning': 2.4849066497880004,
           'difficult': 2.4849066497880004,
           'saved': 2.4849066497880004,
           'although': 2.4849066497880004,
           'base': 2.4849066497880004,
           'thats': 2.4849066497880004,
           'everyday': 2.4849066497880004,
           'personality': 2.4849066497880004,
           'network': 2.4849066497880004,
           'complain': 2.4849066497880004,
           'systems': 2.4849066497880004,
           'cannot': 2.4849066497880004,
           'recommend': 2.4849066497880004,
           'fired': 2.4849066497880004,
           'talent': 2.4849066497880004,
           'release': 2.4849066497880004,
           'protect': 2.4849066497880004,
           'math': 2.4849066497880004,
           'correct': 2.4849066497880004,
           'science': 2.4849066497880004,
           'card': 2.4849066497880004,
           'tons': 2.4849066497880004,
           'compare': 2.4849066497880004,
           'bank': 2.4849066497880004,
           ...}
```

0.2 NLP Pipeline

```
In [7]: from pyspark.ml import Pipeline
        from pyspark.ml.feature import HashingTF, IDF, RegexTokenizer, Tokenizer, CountVectori
        from pyspark.ml.regression import RandomForestRegressor, LinearRegression, DecisionTre
        from pyspark.ml.evaluation import RegressionEvaluator
```

```
In [44]: #tokenization
         regexTokenizer = RegexTokenizer().setInputCol("clean_comment").setOutputCol("comment_

         #remove stop words
         #StopWordsRemover.loadDefaultStopWords("english")
         remover = StopWordsRemover().setInputCol("comment_tokenized").setOutputCol("filtered"

         #TF-IDF
         countVector = CountVectorizer(inputCol="filtered", outputCol="features")
         #hashingTF = HashingTF().setInputCol("filtered").setOutputCol("features").setNumFeatu

         #label indexer
         indexer = StringIndexer(inputCol = "score", outputCol = "label").setHandleInvalid("ke

         #build pipeline
         pipeline = Pipeline(stages=[regexTokenizer, remover, countVector, indexer])
         df_new = pipeline.fit(df_nlp).transform(df_nlp)
```

```
In [52]: df_sub = pipeline.fit(rc_t).transform(rc_t)
         df_sub = df_sub.drop('comment_tokenized','filtered')
```

```
In [ ]: # document frequency
        doc_freq = term_freq \
            .flatMap(lambda sub_counter: list(sub_counter[1])) \
            .map(lambda word: (word, 1)) \
            .reduceByKey(lambda a, b: a + b) \
            .sortBy(lambda word_docfreq: word_docfreq[1], False)
        doc_freq.cache()
```

```
In [ ]: print(doc_freq.collect()[0:50])
```

```
In [ ]: #number of document
        num_docs = term_freq.count()
        print(num_docs)
```

```
In [ ]: #Inverse Document Frequency
        import math

        inv_doc_freq = doc_freq \
            .map(lambda t_df: (t_df[0], math.log(num_docs / t_df[1]))) \
            .sortBy(lambda t_idf: t_idf[1], True)
        inv_doc_freq.cache()
```

```
In [ ]: inv_doc_freq_res = inv_doc_freq.collect();
        print(inv_doc_freq_res[0:50])
```

```
In [ ]: sub_1 = 'politics'
        term_freq_res_1 = term_freq.sortByKey().lookup(sub_1)[0]
        print(sub_1)
        print(sorted(list(term_freq_res_1.items()), key=lambda t_f:t_f[1], reverse=True)[0:50]
```

```
In [ ]: sub_2 = 'programming'
        term_freq_res_2 = term_freq.sortByKey().lookup(sub_2)[0]
        print(sub_2)
        print(sorted(list(term_freq_res_2.items()), key=lambda t_f:t_f[1], reverse=True)[0:50]
```

```
In [ ]: inv_doc_freq_map_res = inv_doc_freq.collectAsMap()
        inv_doc_freq_map_res
```

```
In [ ]: tfidf_list_1 = list(map(lambda t_f: (t_f[0], t_f[1] * inv_doc_freq_map_res[t_f[0]]),
                          term_freq_res_1.items()))

        print(sub_1)
        print(sorted(tfidf_list_1, key = lambda t_fidf: t_fidf[1], reverse = True)[0:50])
```

```
In [53]: train_sub, test_sub = df_sub.randomSplit([0.8, 0.2])
```

```
In [47]: train_sub.count(), test_sub.count()
```

```
Out[47]: (2965, 775)
```

```
In [48]: df_new = df_new.drop('comment_tokenized','filtered')
         df_new.show(10)
```

```
+---------------+--------------------+---+-----+-----+--------------------+-----+
|      subreddit|       clean_comment|ups|downs|score|            features|label|
+---------------+--------------------+---+-----+-----+--------------------+-----+
|          funny|kids don think li...|  1|    0|    1|(32868,[0,4,145],...|  0.0|
|         videos|most our politici...|  1|    0|    1|(32868,[0,78,3587...|  0.0|
|      AskReddit|yknow friends pla...|  2|    0|    2|(32868,[11,12,39,...|  1.0|
|      AskReddit|read that many ti...|  6|    0|    6|(32868,[72,121,12...|  7.0|
|           news|those are only tw...|  3|    0|    3|(32868,[73,411,19...|  2.0|
|      AskReddit|silly person seed...|  5|    0|    5|(32868,[64,90,141...|  5.0|
|leagueoflegends|foxdrop great but...|  1|    0|    1|(32868,[0,2,6,10,...|  0.0|
|           pics|that because terr...|  3|    0|    3|(32868,[417,1418]...|  2.0|
|      AskReddit|last time had one...|  3|    0|    3|(32868,[3,6,70,64...|  2.0|
|            nba|more than anyone ...|  1|    0|    1|(32868,[109,422],...|  0.0|
+---------------+--------------------+---+-----+-----+--------------------+-----+
only showing top 10 rows
```

```
In [49]: df_new.dtypes
```

```
Out[49]: [('subreddit', 'string'),
          ('clean_comment', 'string'),
          ('ups', 'string'),
          ('downs', 'string'),
          ('score', 'string'),
          ('features', 'vector'),
          ('label', 'double')]
```

```
In [50]: train_df, test_df = df_new.randomSplit([0.8, 0.2])
```

```
In [51]: train_df.count(), test_df.count()
```

```
Out[51]: (30685, 7689)
```

0.3 Logistic Regression

```
In [55]: from pyspark.ml.classification import LogisticRegression

         lr = LogisticRegression(maxIter=10, regParam=0.3, elasticNetParam=0.8, featuresCol="fe

         # Fit the model
         lrModel = lr.fit(train_sub)

         # Print the coefficients and intercept for logistic regression
         #print("Coefficients: " + str(lrModel.coefficients))
         #print("Intercept: " + str(lrModel.intercept))
```

```
In [35]: # trainingSummary = lrModel.summary
         # trainingSummary.roc.show()
         # print("areaUnderROC: " + str(trainingSummary.areaUnderROC))
```

```
In [ ]: # Obtain the objective per iteration
        objectiveHistory = trainingSummary.objectiveHistory
        print("objectiveHistory:")
        for objective in objectiveHistory:
            print(objective)
```

```
In [69]: lr_pred = lrModel.transform(test_sub)
         # print('accuracy %s' % accuracy_score(y_pred, y_test))
         # print(classification_report(y_test, y_pred,target_names=my_tags))
```

```
In [57]: lr_pred.printSchema()
```

```
root
 |-- clean_comment: string (nullable = true)
 |-- downs: string (nullable = true)
 |-- score: string (nullable = true)
 |-- subreddit: string (nullable = true)
 |-- ups: string (nullable = true)
 |-- features: vector (nullable = true)
 |-- label: double (nullable = false)
 |-- rawPrediction: vector (nullable = true)
 |-- probability: vector (nullable = true)
 |-- prediction: double (nullable = false)
```

```
In [70]: from pyspark.ml.evaluation import MulticlassClassificationEvaluator

         evaluator = MulticlassClassificationEvaluator(labelCol="label", predictionCol="predic

         print("accuracy:",evaluator.evaluate(lr_pred, {evaluator.metricName: "accuracy"}))
         print("f1:",evaluator.evaluate(lr_pred, {evaluator.metricName: "f1"}))
```

```
accuracy: 0.398110661268556
f1: 0.2267232530390425
```

```
In [75]: evaluator = RegressionEvaluator(labelCol="label", predictionCol="prediction", metricNa
         rmse_label = evaluator.evaluate(lr_pred)
         print("Root Mean Squared Error (RMSE) on test data for 'score' = %g" % (rmse_label))
```

```
Root Mean Squared Error (RMSE) on test data for 'score' = 28.3704
```

0.4 Linear Regression

```
In [79]: from pyspark.ml.regression import LinearRegression

         lnr = LinearRegression(featuresCol="features", labelCol="label").setMaxIter(10).setReg

         lnrModel = lnr.fit(train_sub)
```

```
In [72]: lnr_pred = lnrModel.transform(test_sub)
```

```
In [77]: evaluator = RegressionEvaluator(labelCol="label", predictionCol="prediction", metricNa
         rmse_label = evaluator.evaluate(lnr_pred)
         print("Root Mean Squared Error (RMSE) on test data for 'score' = %g" % (rmse_label))
```

```
Root Mean Squared Error (RMSE) on test data for 'score' = 27.87
```

0.5 Decision Tree

```
In [81]: from pyspark.ml.regression import DecisionTreeRegressor

         tree = DecisionTreeRegressor().setLabelCol("label").setFeaturesCol("features")

         treeModel = tree.fit(train_sub)
```

```
In [82]: tree_pred = treeModel.transform(test_sub)
```

```
In [83]: evaluator = RegressionEvaluator(labelCol="label", predictionCol="prediction", metricNa
         rmse_label = evaluator.evaluate(tree_pred)
         print("Root Mean Squared Error (RMSE) on test data for 'score' = %g" % (rmse_label))
```

```
Root Mean Squared Error (RMSE) on test data for 'score' = 26.0653
```

0.6 Gradient-boosted Regression Tree

```
In [84]: from pyspark.ml.regression import GBTRegressor

         gbt = GBTRegressor().setLabelCol("label").setFeaturesCol("features").setMaxIter(10)

         gbtModel = gbt.fit(train_sub)
```

```
In [88]: gbt_pred = gbtModel.transform(test_sub)
```

```
In [89]: evaluator = RegressionEvaluator(labelCol="label", predictionCol="prediction", metricNa
         rmse_label = evaluator.evaluate(gbt_pred)
         print("Root Mean Squared Error (RMSE) on test data for 'score' = %g" % (rmse_label))
```

```
Root Mean Squared Error (RMSE) on test data for 'score' = 26.0882
```

0.7 Random Forest

```
In [87]: from pyspark.ml.regression import RandomForestRegressor

         forest = RandomForestRegressor().setLabelCol("label").setFeaturesCol("features")

         forestModel = forest.fit(train_sub)
```

```
In [90]: forest_pred = forestModel.transform(test_sub)
```

```
In [91]: evaluator = RegressionEvaluator(labelCol="label", predictionCol="prediction", metricNa
         rmse_label = evaluator.evaluate(forest_pred)
         print("Root Mean Squared Error (RMSE) on test data for 'score' = %g" % (rmse_label))
```

```
Root Mean Squared Error (RMSE) on test data for 'score' = 26.7029
```