Reddit Comment Text Analysis

Chuyu Chen

Agenda

1. Business Problem
2. Exploratory Data Analysis & Preparation
3. Text Classification
4. Sentiment Analysis
5. Path Forward

Business Problem

Executive Summary

To help Reddit understand their topics, categorize comments attitude, and predict comments' likes and dislike scores.

Text analysis can be used to better understand the current Reddit community through the subreddits.

1. **Target popular topics** using word clouds.
2. **Categorize the attitude** based on the comments.
3. **Predict scores** for each comment accordingly.

Business Problem

Text Classification w/ Spark NLP

- Based on the body of the post, predict a post’s success before it’s submitted (good v.s. bad ratings)
- Potentially help Redditors gain upvotes & improve user engagement

Sentiment Analysis w/ Pretrained Model

- Understand people's opinions from a post
- Potentially help Reddit gain an overview of the wider public opinion behind certain topics

Link to Sentiment Analysis Pretrained Model:
https://www.johnsnowlabs.com/detect-sentiment-emotion/

EDA & Data Preparations

Data Overview

A small portion of publicly available reddit comment datasets
-- comments posted in **December 2019**

Breaking an enormous dataset to explore

- Original Dataset released by Reddit: 1+ TB with 1.7 billion comments
- Practical subset: 35.2G with ~146 million comments
- Sampled dataset for analysis: 460k comments

An aggregated table with 20 variables and 5.8M records.

Columns include:

- Body
- subreddit
- score
- ups
- downs

Link to datasource in GCP BigQuery:
https://console.cloud.google.com/bigquery?referrer=search&authuser=3&project=nimble-net-337716&d=reddit_comments&p=fh-bigquery&t=2019_12&page=table&ws=!1m5!1m4!4m3!1sfh-bigquery!2sreddit_comments!3s2019_12

Tools & Platforms

Google Cloud Platform

Google BigQuery

Google Cloud Storage

Cloud Dataproc

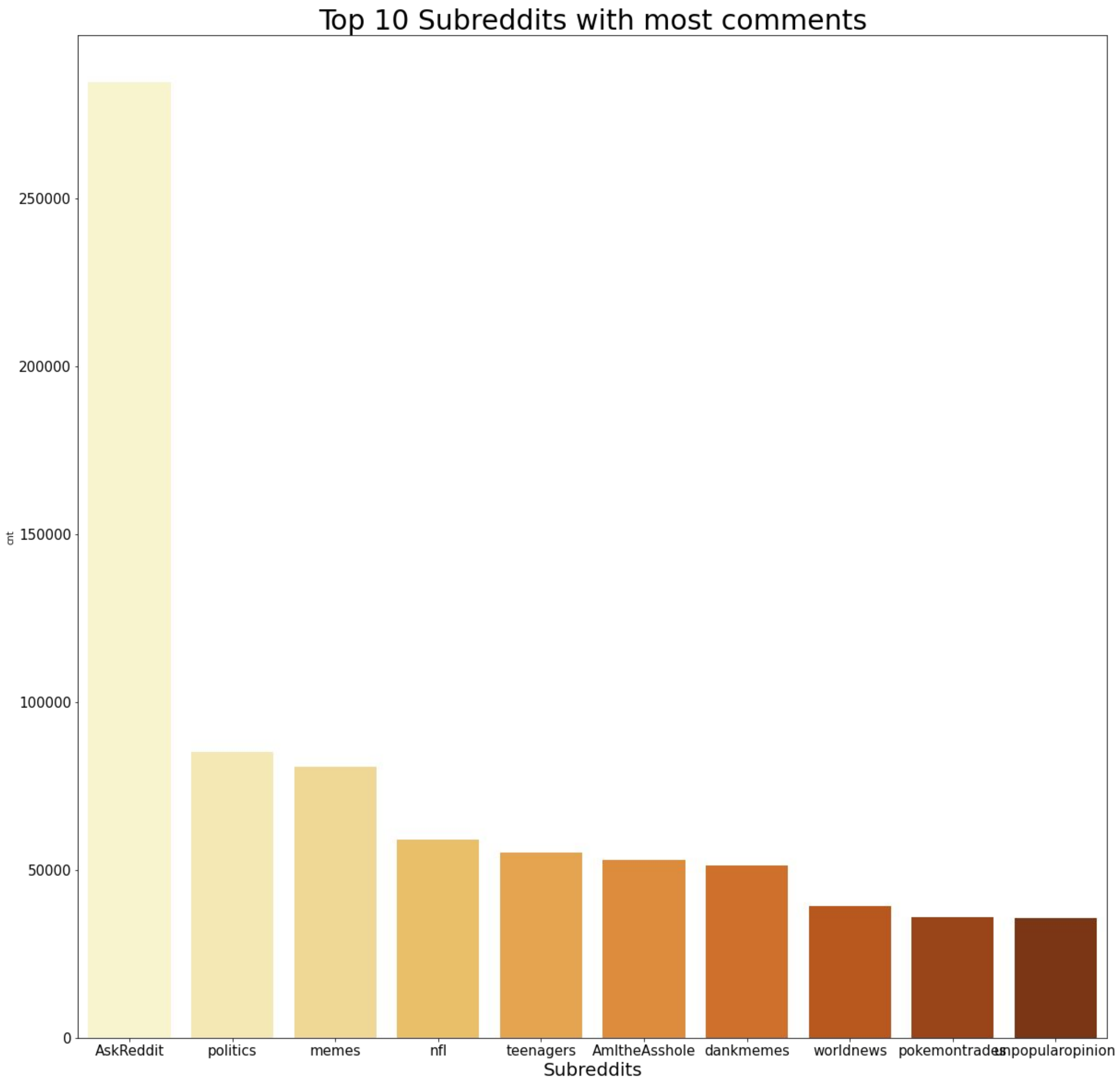
jupyter

CO

John Snow LABS

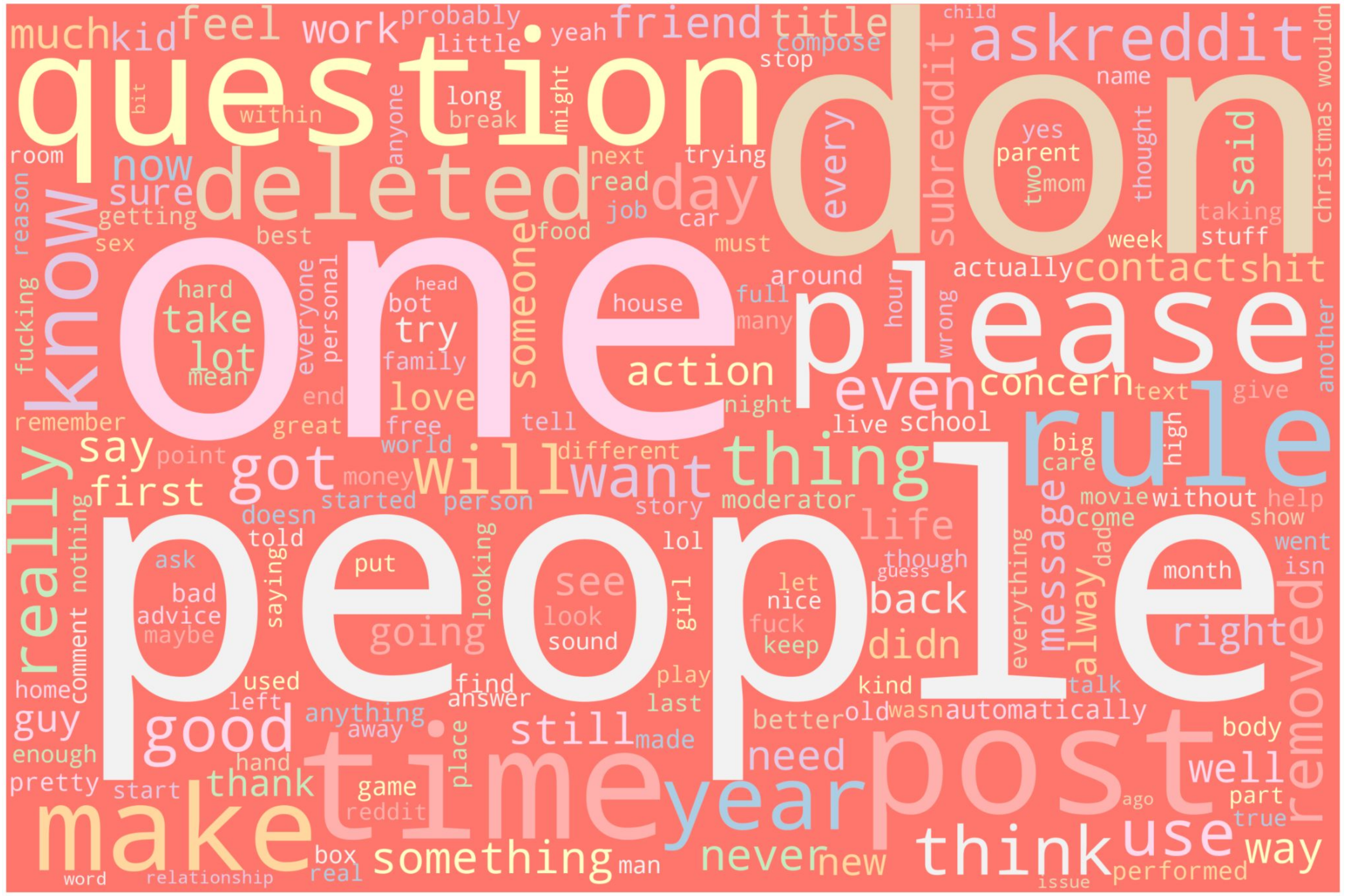
APACHE Spark™

Exploratory Data Analysis

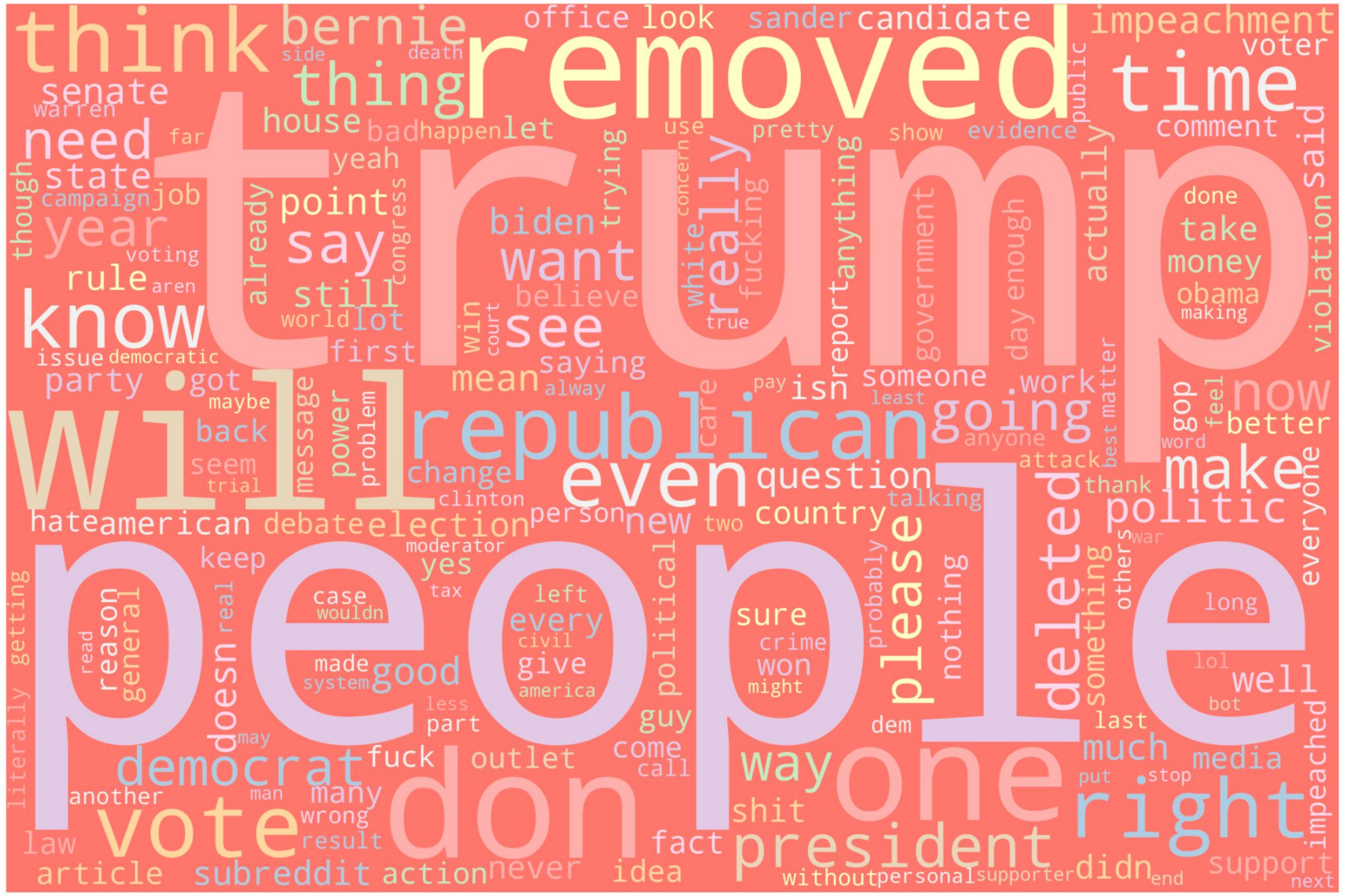


- Calculated Top 10 subreddits based on number of comments
- AskReddit was by far the most popular subreddit

Exploratory Data Analysis - Word Cloud



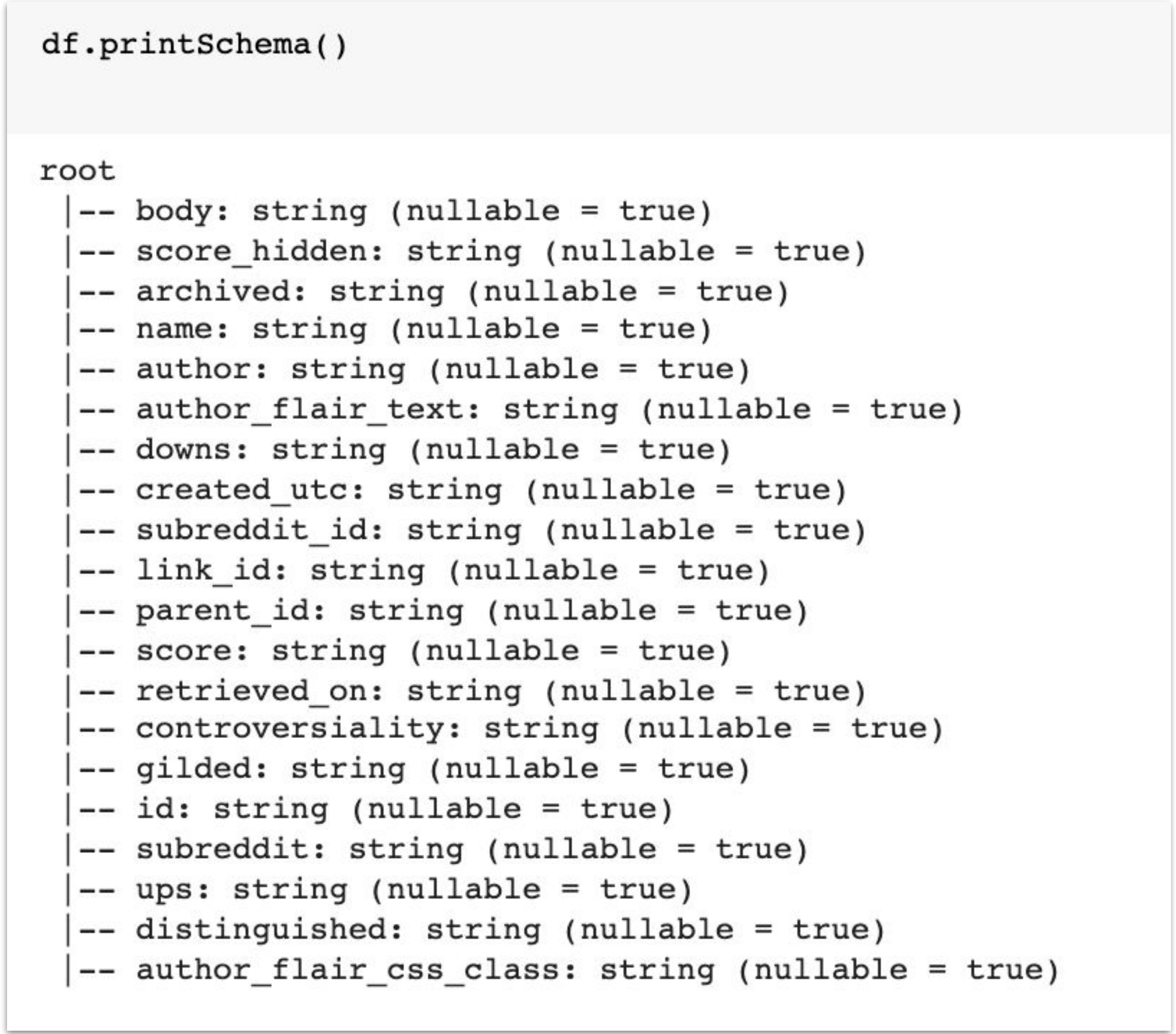
Subreddit: AskReddit



Subreddit: Politics

Exploratory Data Analysis - Data Snapshot & Variables

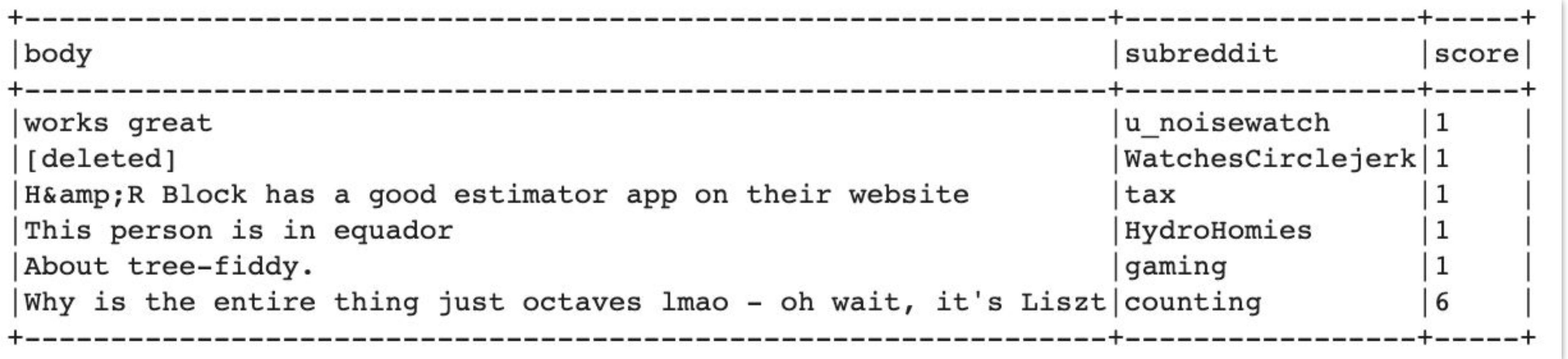
Initially 20 variables; convert data types & subset data



```
df.printSchema()
```

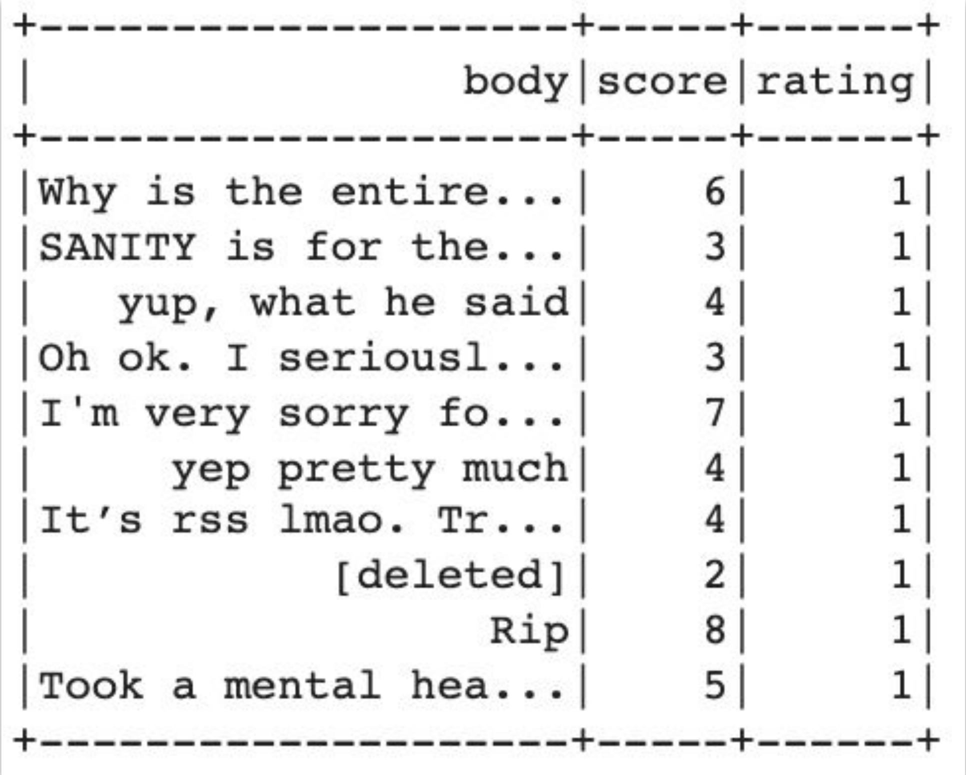
```
root
 |-- body: string (nullable = true)
 |-- score_hidden: string (nullable = true)
 |-- archived: string (nullable = true)
 |-- name: string (nullable = true)
 |-- author: string (nullable = true)
 |-- author_flair_text: string (nullable = true)
 |-- downs: string (nullable = true)
 |-- created_utc: string (nullable = true)
 |-- subreddit_id: string (nullable = true)
 |-- link_id: string (nullable = true)
 |-- parent_id: string (nullable = true)
 |-- score: string (nullable = true)
 |-- retrieved_on: string (nullable = true)
 |-- controversiality: string (nullable = true)
 |-- gilded: string (nullable = true)
 |-- id: string (nullable = true)
 |-- subreddit: string (nullable = true)
 |-- ups: string (nullable = true)
 |-- distinguished: string (nullable = true)
 |-- author_flair_css_class: string (nullable = true)
```

Only use columns ‘body’, ‘score’, ‘subreddit’

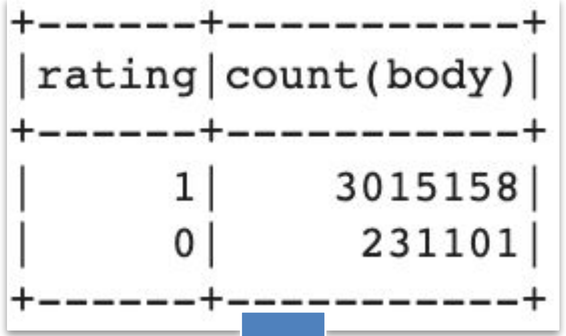


```
+---------------------------------------------------------------+-----------------+-----+
|body                                                           |subreddit        |score|
+---------------------------------------------------------------+-----------------+-----+
|works great                                                    |u_noisewatch     |1    |
|[deleted]                                                      |WatchesCirclejerk|1    |
|H&amp;R Block has a good estimator app on their website        |tax              |1    |
|This person is in equador                                      |HydroHomies      |1    |
|About tree-fiddy.                                              |gaming           |1    |
|Why is the entire thing just octaves lmao - oh wait, it's Liszt|counting         |6    |
+---------------------------------------------------------------+-----------------+-----+
```

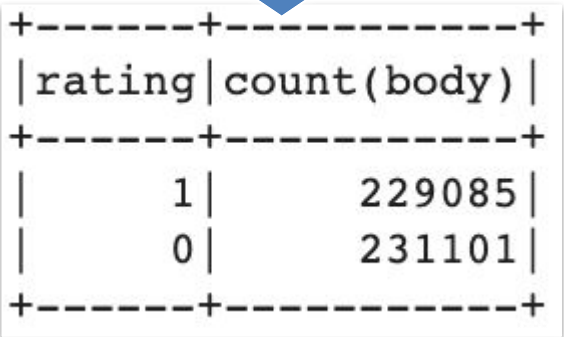
Assign labels to positive / negative ratings; subset & create balanced data



```
+--------------------+-----+------+
|                body|score|rating|
+--------------------+-----+------+
|Why is the entire...|    6|     1|
|SANITY is for the...|    3|     1|
|   yup, what he said|    4|     1|
|Oh ok. I seriousl...|    3|     1|
|I'm very sorry fo...|    7|     1|
|     yep pretty much|    4|     1|
|It’s rss lmao. Tr...|    4|     1|
|           [deleted]|    2|     1|
|                 Rip|    8|     1|
|Took a mental hea...|    5|     1|
+--------------------+-----+------+
```



```
+------+-----------+
|rating|count(body)|
+------+-----------+
|     1|    3015158|
|     0|     231101|
+------+-----------+
```



```
+------+-----------+
|rating|count(body)|
+------+-----------+
|     1|     229085|
|     0|     231101|
+------+-----------+
```

reddit

Text Classification

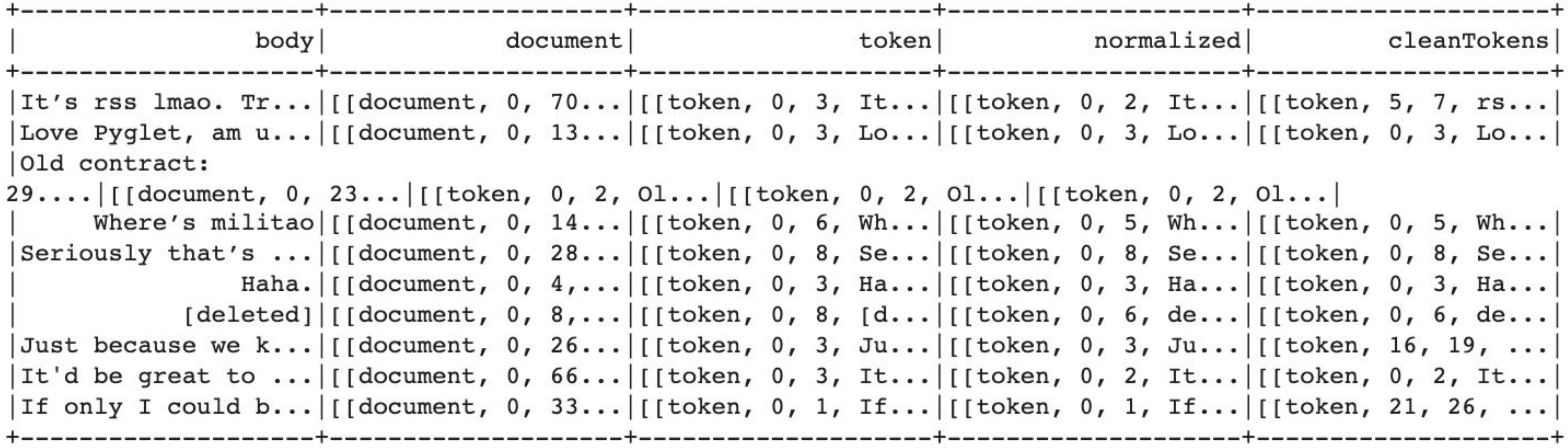
Classification Pipeline

DocumentAssembler

Tokenizer

Normalizer

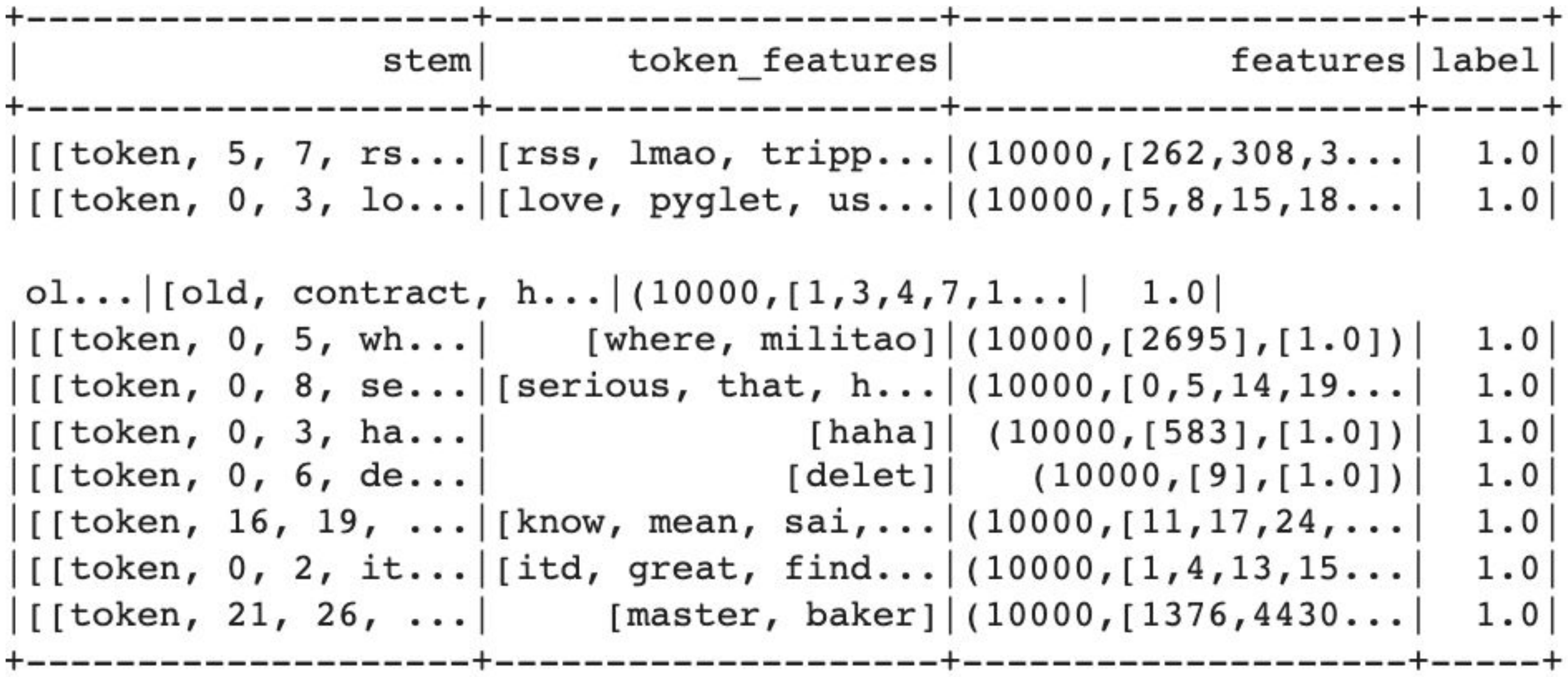
StopwordsCleaner



```
+--------------------+--------------------+--------------------+--------------------+--------------------+
|                body|            document|               token|          normalized|         cleanTokens|
+--------------------+--------------------+--------------------+--------------------+--------------------+
|It’s rss lmao. Tr...|[[document, 0, 70...|[[token, 0, 3, It...|[[token, 0, 2, It...|[[token, 5, 7, rs...|
|Love Pyglet, am u...|[[document, 0, 13...|[[token, 0, 3, Lo...|[[token, 0, 3, Lo...|[[token, 0, 3, Lo...|
|Old contract:
29....|[[document, 0, 23...|[[token, 0, 2, Ol...|[[token, 0, 2, Ol...|[[token, 0, 2, Ol...|
|     Where’s militao|[[document, 0, 14...|[[token, 0, 6, Wh...|[[token, 0, 5, Wh...|[[token, 0, 5, Wh...|
|Seriously that’s ...|[[document, 0, 28...|[[token, 0, 8, Se...|[[token, 0, 8, Se...|[[token, 0, 8, Se...|
|               Haha.|[[document, 0, 4,...|[[token, 0, 3, Ha...|[[token, 0, 3, Ha...|[[token, 0, 3, Ha...|
|           [deleted]|[[document, 0, 8,...|[[token, 0, 8, [d...|[[token, 0, 6, de...|[[token, 0, 6, de...|
|Just because we k...|[[document, 0, 26...|[[token, 0, 3, Ju...|[[token, 0, 3, Ju...|[[token, 16, 19, ...|
|It'd be great to ...|[[document, 0, 66...|[[token, 0, 3, It...|[[token, 0, 2, It...|[[token, 0, 2, It...|
|If only I could b...|[[document, 0, 33...|[[token, 0, 1, If...|[[token, 0, 1, If...|[[token, 21, 26, ...|
+--------------------+--------------------+--------------------+--------------------+--------------------+
```

Classification Pipeline

Stemmer | Finisher | CountVectorizer | StringIndexer



```
+--------------------+--------------------+--------------------+-----+
|                stem|      token_features|            features|label|
+--------------------+--------------------+--------------------+-----+
|[[token, 5, 7, rs...|[rss, lmao, tripp...|(10000,[262,308,3...|  1.0|
|[[token, 0, 3, lo...|[love, pyglet, us...|(10000,[5,8,15,18...|  1.0|

 ol...|[old, contract, h...|(10000,[1,3,4,7,1...|  1.0|
|[[token, 0, 5, wh...|    [where, militao]|(10000,[2695],[1.0])|  1.0|
|[[token, 0, 8, se...|[serious, that, h...|(10000,[0,5,14,19...|  1.0|
|[[token, 0, 3, ha...|              [haha]| (10000,[583],[1.0])|  1.0|
|[[token, 0, 6, de...|             [delet]|   (10000,[9],[1.0])|  1.0|
|[[token, 16, 19, ...|[know, mean, sai,...|(10000,[11,17,24,...|  1.0|
|[[token, 0, 2, it...|[itd, great, find...|(10000,[1,4,13,15...|  1.0|
|[[token, 21, 26, ...|     [master, baker]|(10000,[1376,4430...|  1.0|
+--------------------+--------------------+--------------------+-----+
```

→

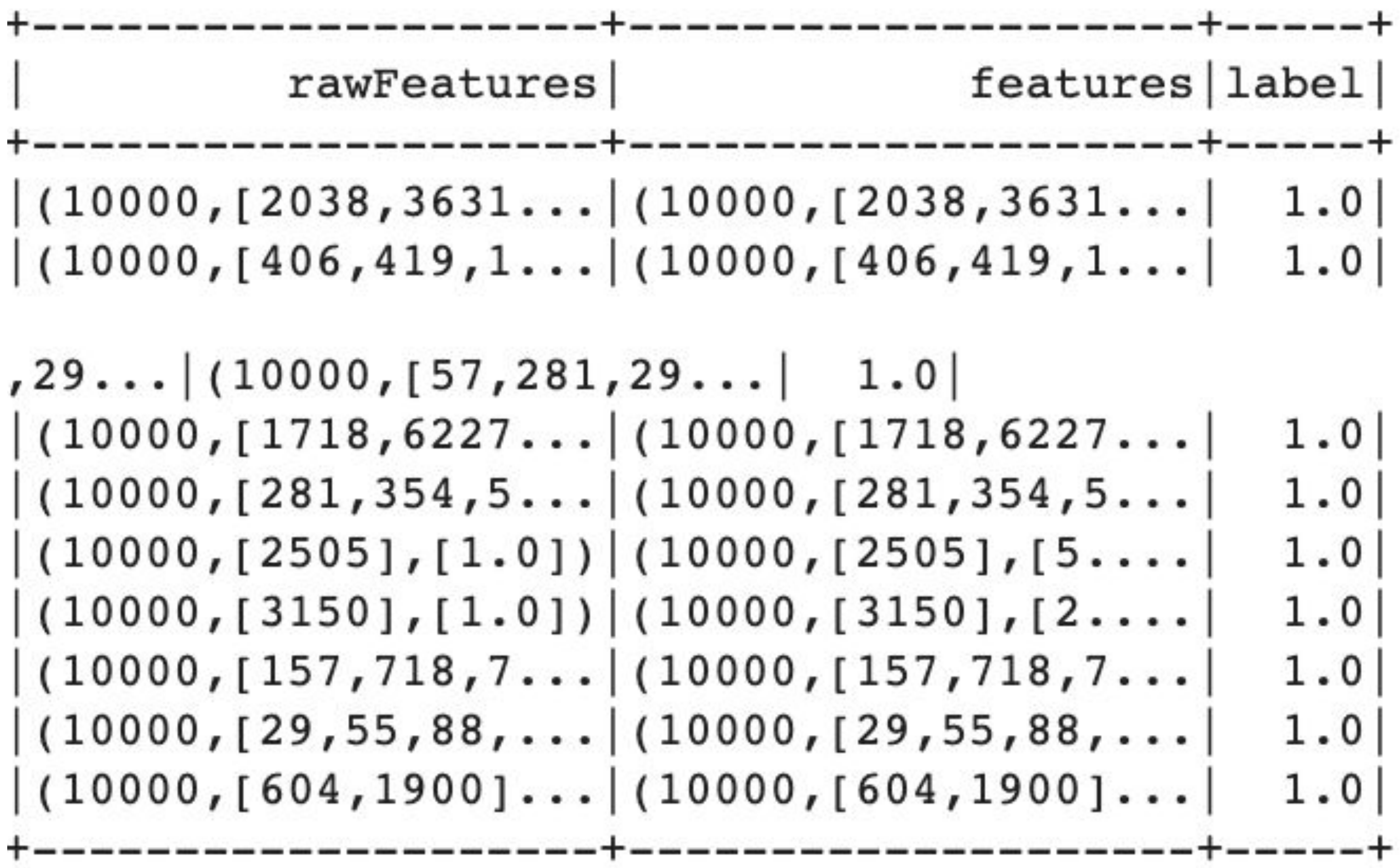
Logistic Regression w/ CountVectorizer

Classification Pipeline

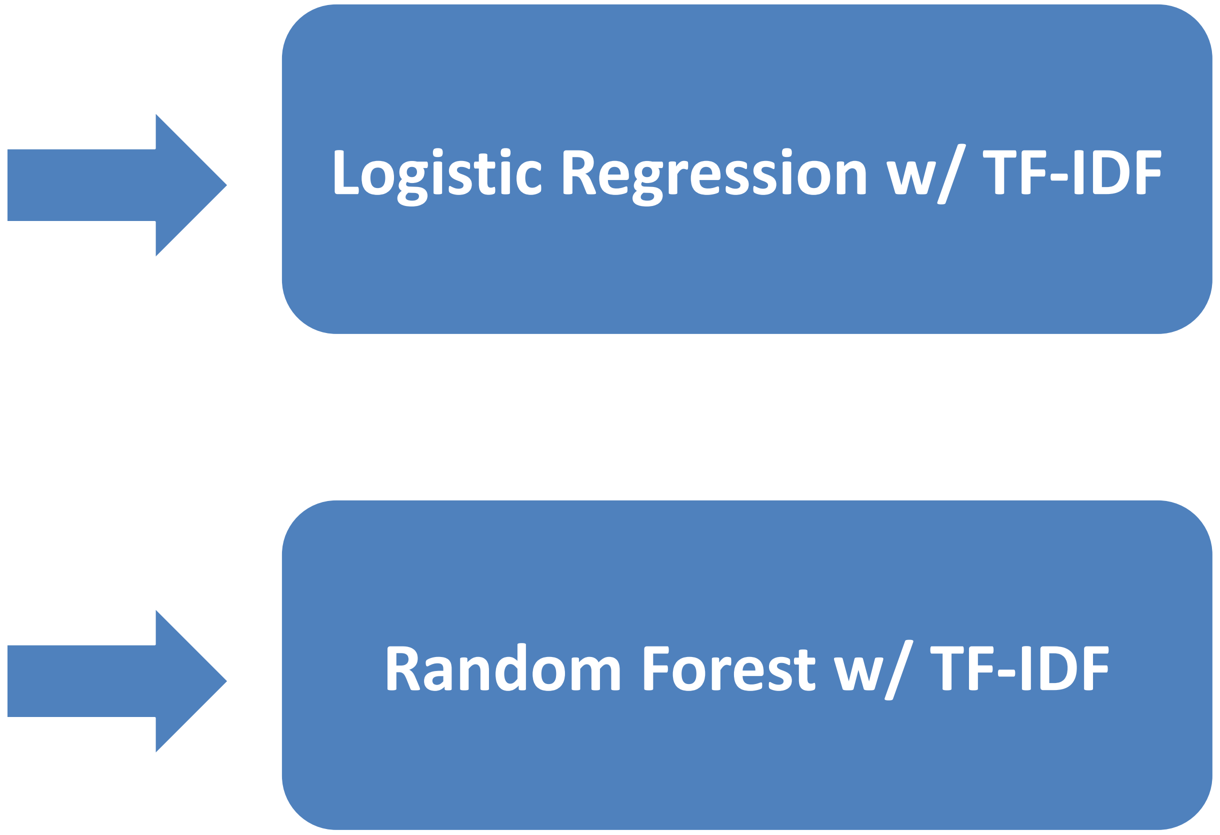
HashingTF

IDF

StringIndexer



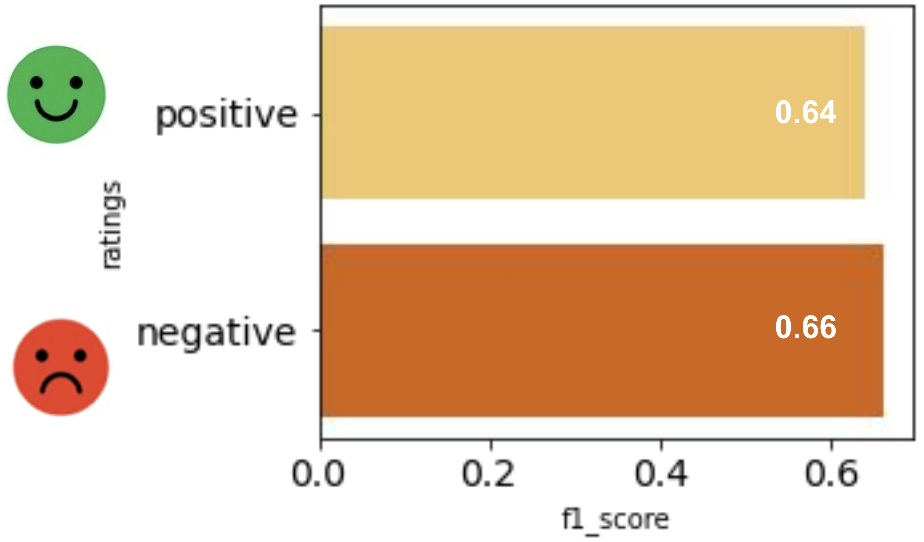
```
+--------------------+--------------------+-----+
|         rawFeatures|            features|label|
+--------------------+--------------------+-----+
|(10000,[2038,3631...|(10000,[2038,3631...|  1.0|
|(10000,[406,419,1...|(10000,[406,419,1...|  1.0|

,29...|(10000,[57,281,29...|  1.0|
|(10000,[1718,6227...|(10000,[1718,6227...|  1.0|
|(10000,[281,354,5...|(10000,[281,354,5...|  1.0|
|(10000,[2505],[1.0])|(10000,[2505],[5....|  1.0|
|(10000,[3150],[1.0])|(10000,[3150],[2....|  1.0|
|(10000,[157,718,7...|(10000,[157,718,7...|  1.0|
|(10000,[29,55,88,...|(10000,[29,55,88,...|  1.0|
|(10000,[604,1900]...|(10000,[604,1900]...|  1.0|
+--------------------+--------------------+-----+
```

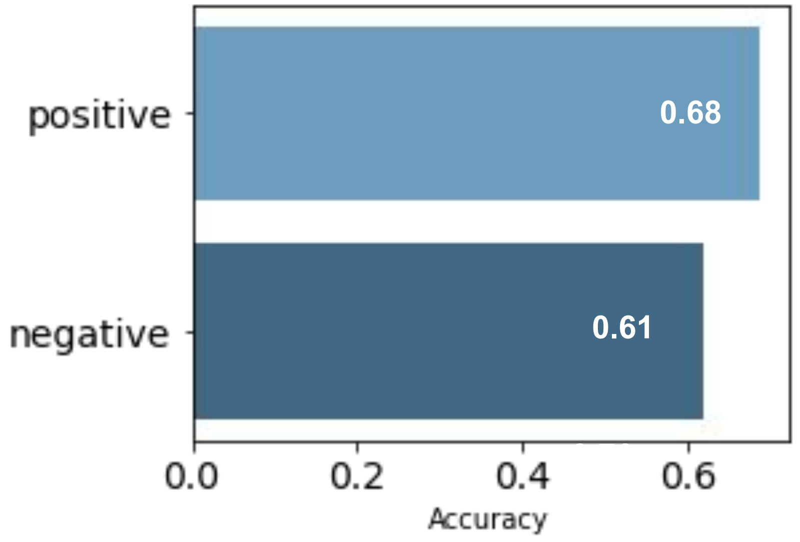


Logistic Regression w/ TF-IDF

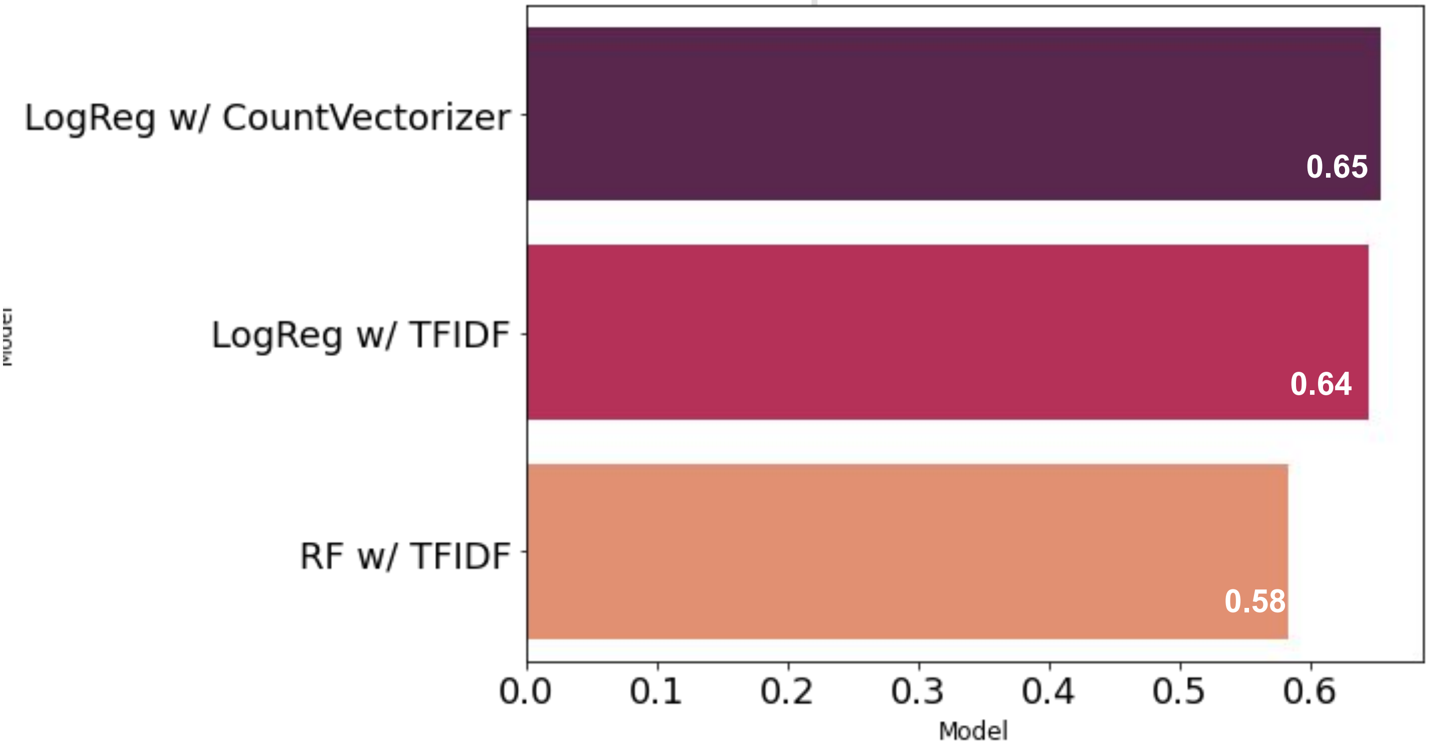
Random Forest w/ TF-IDF

Validation : Text Classification



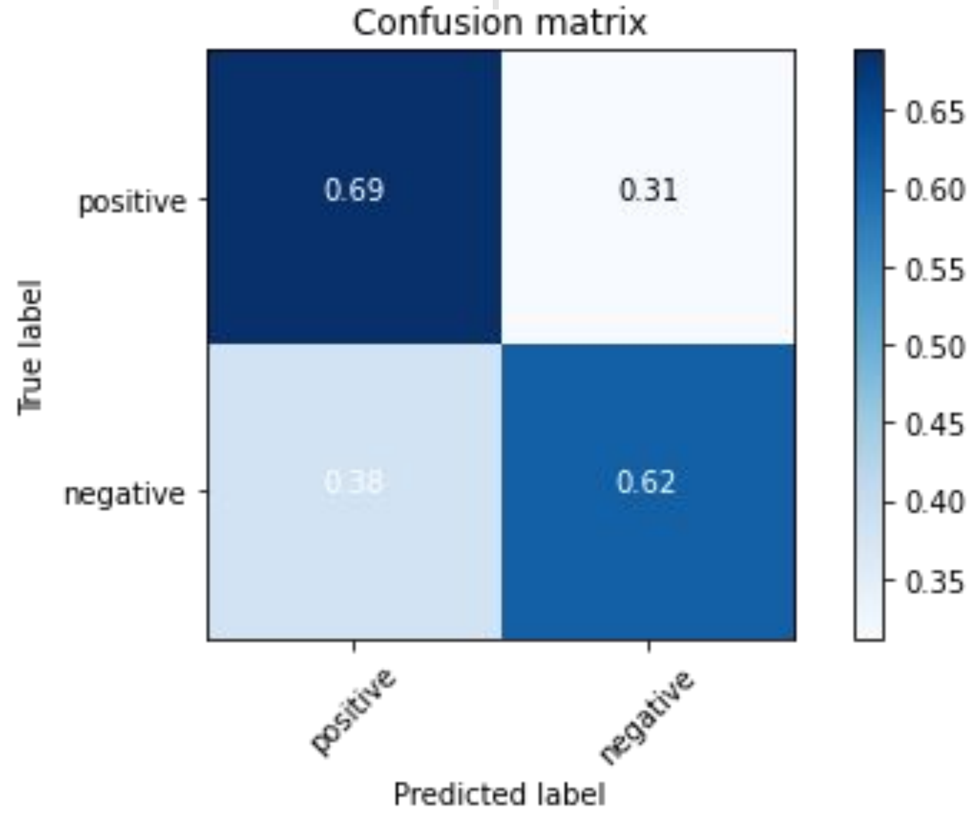


1. Model Validation - Accuracy



2. Logistic Regression F1 & Accuracy Scores

3. Logistic Regression Confusion Matrix

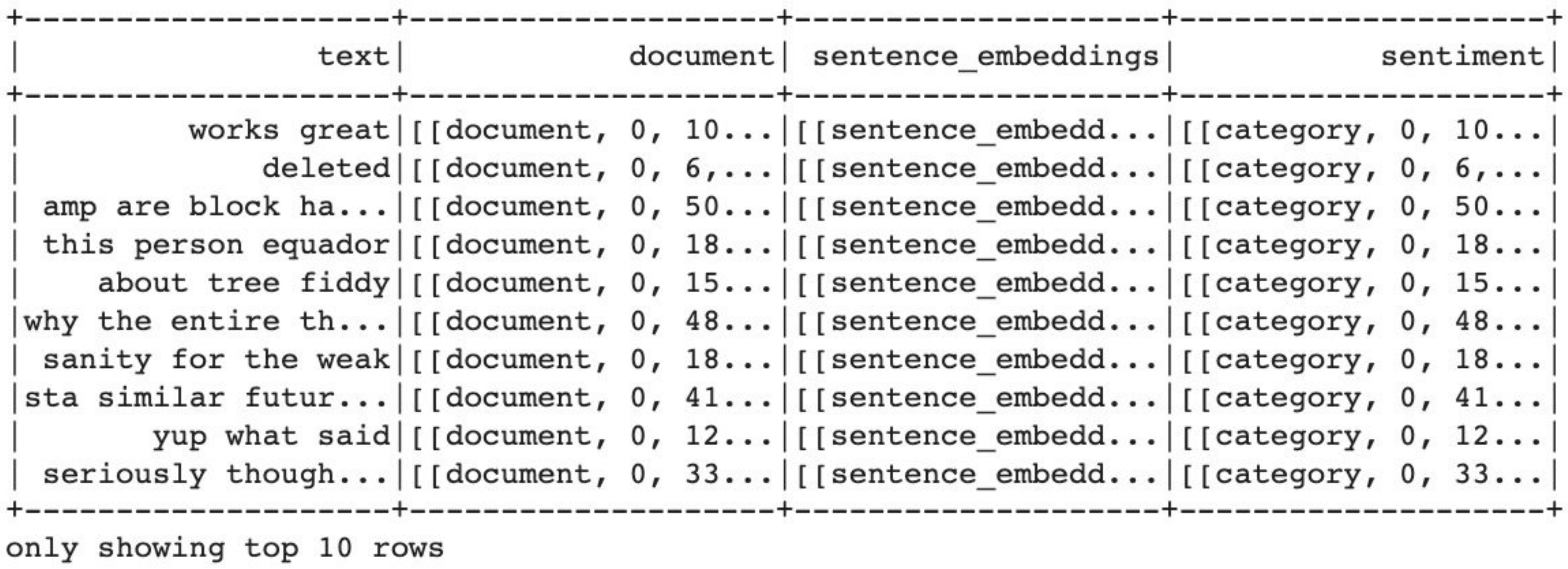


reddit

Sentiment Analysis

Model Pipeline

DocumentAssembler UniversalSentenceEncoder SentimentDLModel

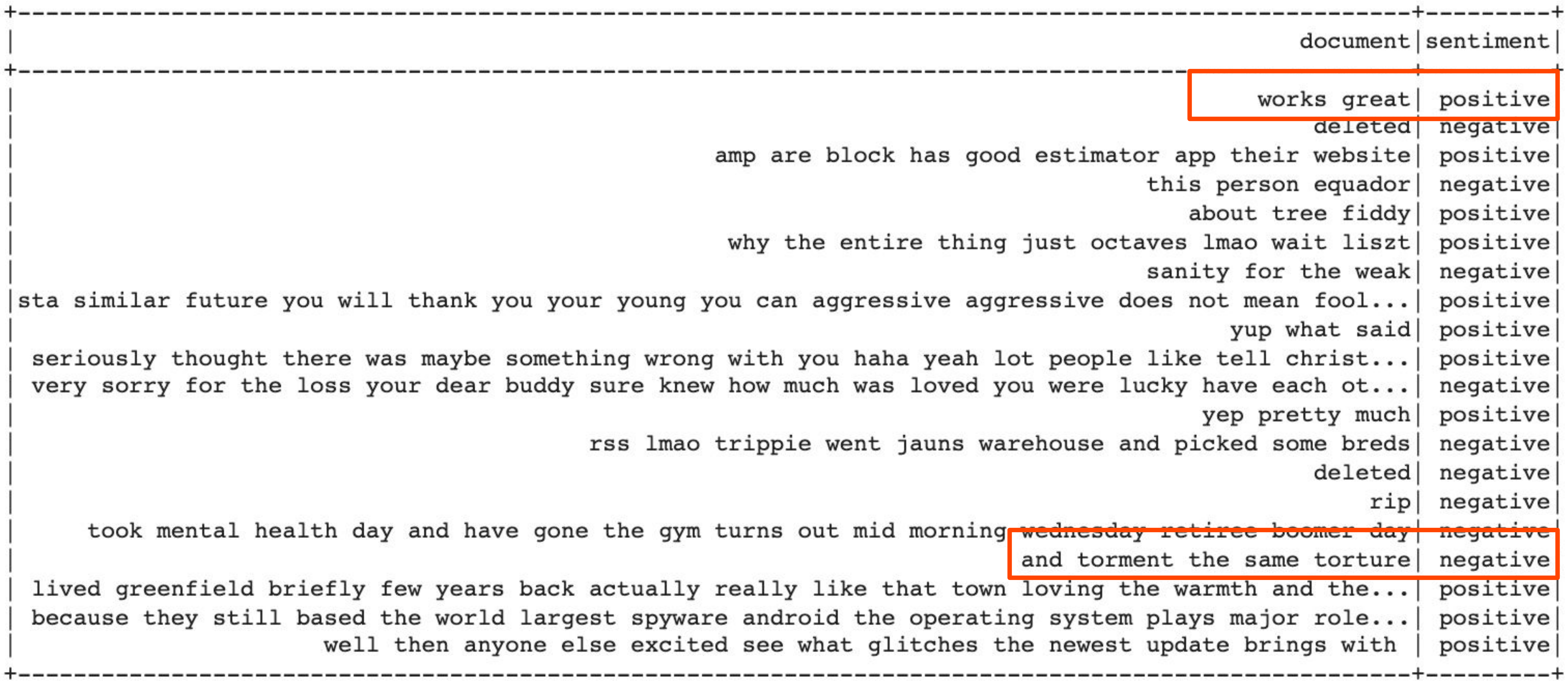


```
+--------------------+--------------------+--------------------+--------------------+
|                text|            document| sentence_embeddings|           sentiment|
+--------------------+--------------------+--------------------+--------------------+
|         works great|[[document, 0, 10...|[[sentence_embedd...|[[category, 0, 10...|
|             deleted|[[document, 0, 6,...|[[sentence_embedd...|[[category, 0, 6,...|
| amp are block ha...|[[document, 0, 50...|[[sentence_embedd...|[[category, 0, 50...|
| this person equador|[[document, 0, 18...|[[sentence_embedd...|[[category, 0, 18...|
|    about tree fiddy|[[document, 0, 15...|[[sentence_embedd...|[[category, 0, 15...|
|why the entire th...|[[document, 0, 48...|[[sentence_embedd...|[[category, 0, 48...|
| sanity for the weak|[[document, 0, 18...|[[sentence_embedd...|[[category, 0, 18...|
|sta similar futur...|[[document, 0, 41...|[[sentence_embedd...|[[category, 0, 41...|
|       yup what said|[[document, 0, 12...|[[sentence_embedd...|[[category, 0, 12...|
| seriously though...|[[document, 0, 33...|[[sentence_embedd...|[[category, 0, 33...|
+--------------------+--------------------+--------------------+--------------------+
only showing top 10 rows
```

Reference: John SnowLabs Sentiment Analysis Pretrained Model

https://www.johnsnowlabs.com/detect-sentiment-emotion/

Model Pipeline



```
+----------------------------------------------------------------------------------------------------+---------+
|                                                                                            document|sentiment|
+----------------------------------------------------------------------------------------------------+---------+
|                                                                                         works great| positive|
|                                                                                             deleted| negative|
|                                                  amp are block has good estimator app their website| positive|
|                                                                                 this person equador| negative|
|                                                                                    about tree fiddy| positive|
|                                                   why the entire thing just octaves lmao wait liszt| positive|
|                                                                                 sanity for the weak| negative|
|sta similar future you will thank you your young you can aggressive aggressive does not mean fool...| positive|
|                                                                                       yup what said| positive|
| seriously thought there was maybe something wrong with you haha yeah lot people like tell christ...| positive|
| very sorry for the loss your dear buddy sure knew how much was loved you were lucky have each ot...| negative|
|                                                                                     yep pretty much| positive|
|                                         rss lmao trippie went jauns warehouse and picked some breds| negative|
|                                                                                             deleted| negative|
|                                                                                                 rip| negative|
|     took mental health day and have gone the gym turns out mid morning wednesday retiree boomer day| negative|
|                                                                        and torment the same torture| negative|
| lived greenfield briefly few years back actually really like that town loving the warmth and the...| positive|
| because they still based the world largest spyware android the operating system plays major role...| positive|
|                      well then anyone else excited see what glitches the newest update brings with | positive|
+----------------------------------------------------------------------------------------------------+---------+
```

Summary

- Promote comments which have potentials to get positive score (upvotes) and positive sentiments to the top, so that we can improve user participation / user acquisition
- Understand the value of created content
- Understand opinions concerning current problems like their political and social views
- Target product & service feedback through comment text analysis

Comparison of Model Performance

Algorithm	Accuracy Score
Logistic Regression w/ CountVectorizer	0.65
Logistic Regression w/ TF-IDF	0.64
Random Forest w/ TF-IDF	0.58

Path Forward

- Most comments have score as 0 and 1, which brings us an super imbalanced dataset.
- Ways to improve model accuracy:
  - Model Tuning
  - Upsample / downsample data; train on entire dataset
  - Other modeling techniques:
    - Spark NLP Bert Embeddings
    - ELMO Embeddings
    - Deep learning models
- More computing power
  - Currently 4vCPU, 15GB memory

References

- Google Cloud Platform BigQuery public dataset (Reference for Exporting data)
  - https://cloud.google.com/bigquery/docs/exporting-data

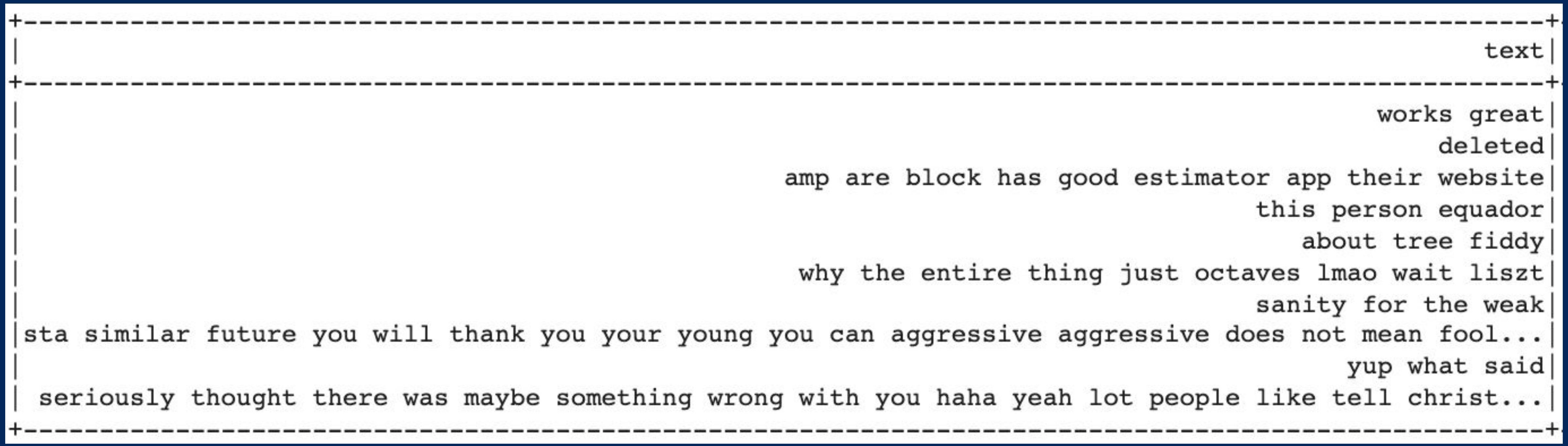
- https://www.johnsnowlabs.com/detect-sentiment-emotion/

“

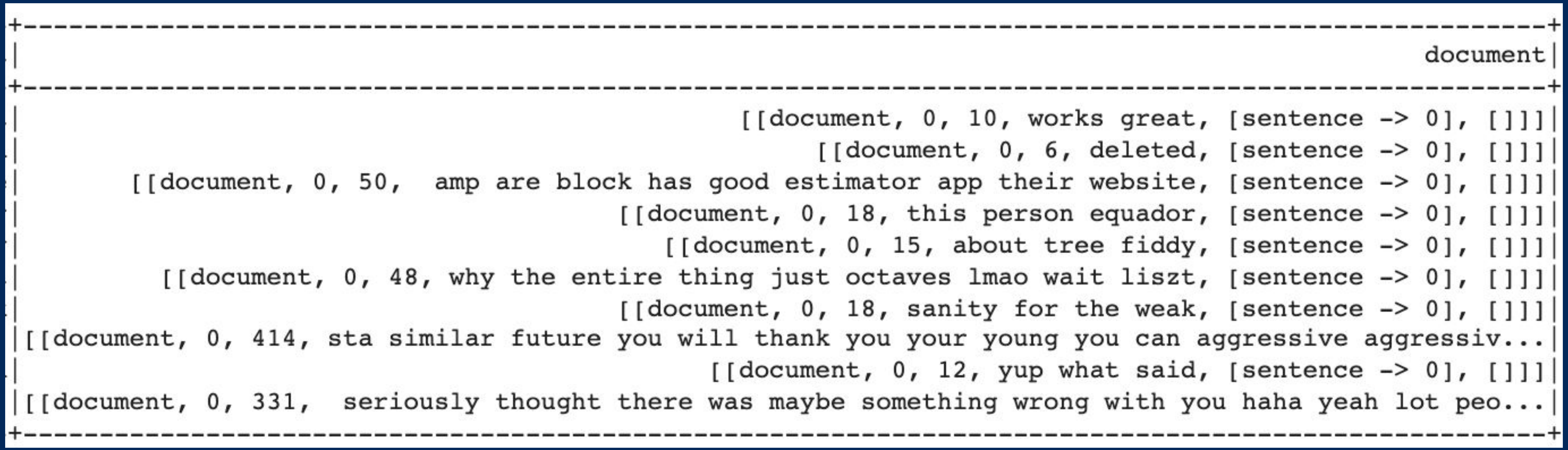
Thank you!

Appendix

Sentiment Analysis Pipeline (Full Output)

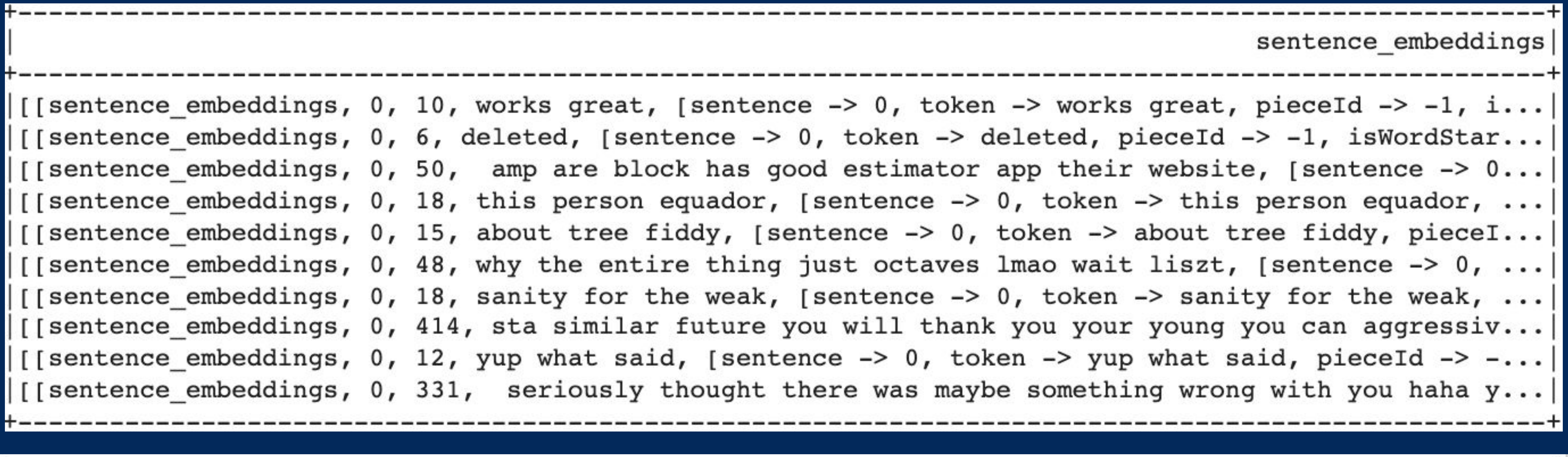


```
+----------------------------------------------------------------------------------------------------+
|                                                                                                text|
+----------------------------------------------------------------------------------------------------+
|                                                                                         works great|
|                                                                                             deleted|
|                                                 amp are block has good estimator app their website|
|                                                                                 this person equador|
|                                                                                    about tree fiddy|
|                                                   why the entire thing just octaves lmao wait liszt|
|                                                                                 sanity for the weak|
|sta similar future you will thank you your young you can aggressive aggressive does not mean fool...|
|                                                                                       yup what said|
| seriously thought there was maybe something wrong with you haha yeah lot people like tell christ...|
+----------------------------------------------------------------------------------------------------+
```

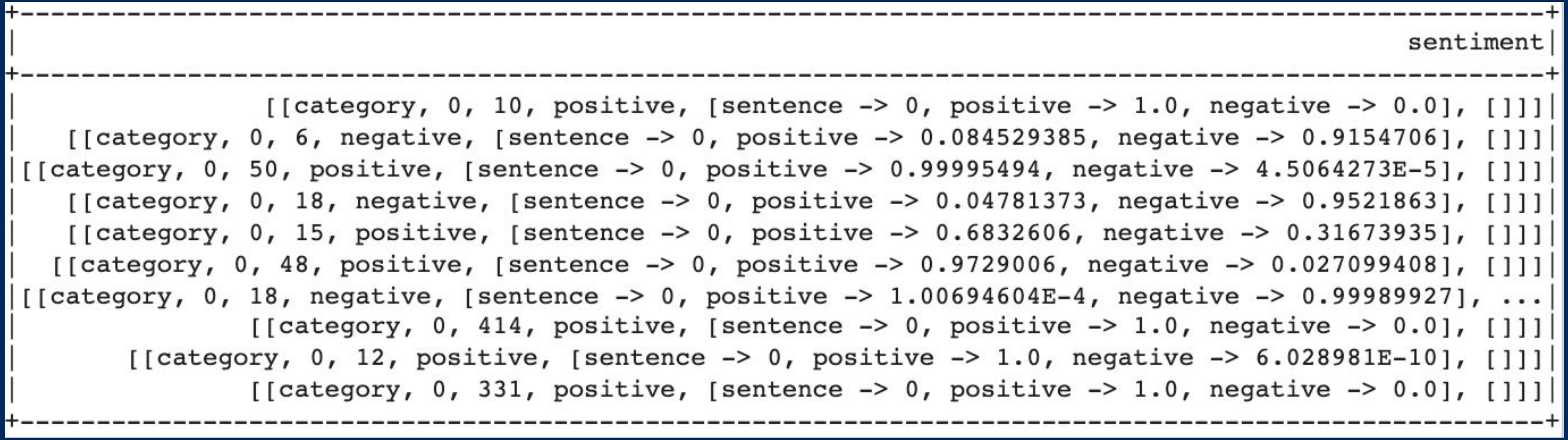


```
+----------------------------------------------------------------------------------------------------+
|                                                                                            document|
+----------------------------------------------------------------------------------------------------+
|                                               [[document, 0, 10, works great, [sentence -> 0], []]]|
|                                                    [[document, 0, 6, deleted, [sentence -> 0], []]]|
|       [[document, 0, 50,  amp are block has good estimator app their website, [sentence -> 0], []]]|
|                                       [[document, 0, 18, this person equador, [sentence -> 0], []]]|
|                                          [[document, 0, 15, about tree fiddy, [sentence -> 0], []]]|
|         [[document, 0, 48, why the entire thing just octaves lmao wait liszt, [sentence -> 0], []]]|
|                                       [[document, 0, 18, sanity for the weak, [sentence -> 0], []]]|
|[[document, 0, 414, sta similar future you will thank you your young you can aggressive aggressiv...|
|                                             [[document, 0, 12, yup what said, [sentence -> 0], []]]|
|[[document, 0, 331,  seriously thought there was maybe something wrong with you haha yeah lot peo...|
+----------------------------------------------------------------------------------------------------+
```

Sentiment Analysis Pipeline (Full Output)

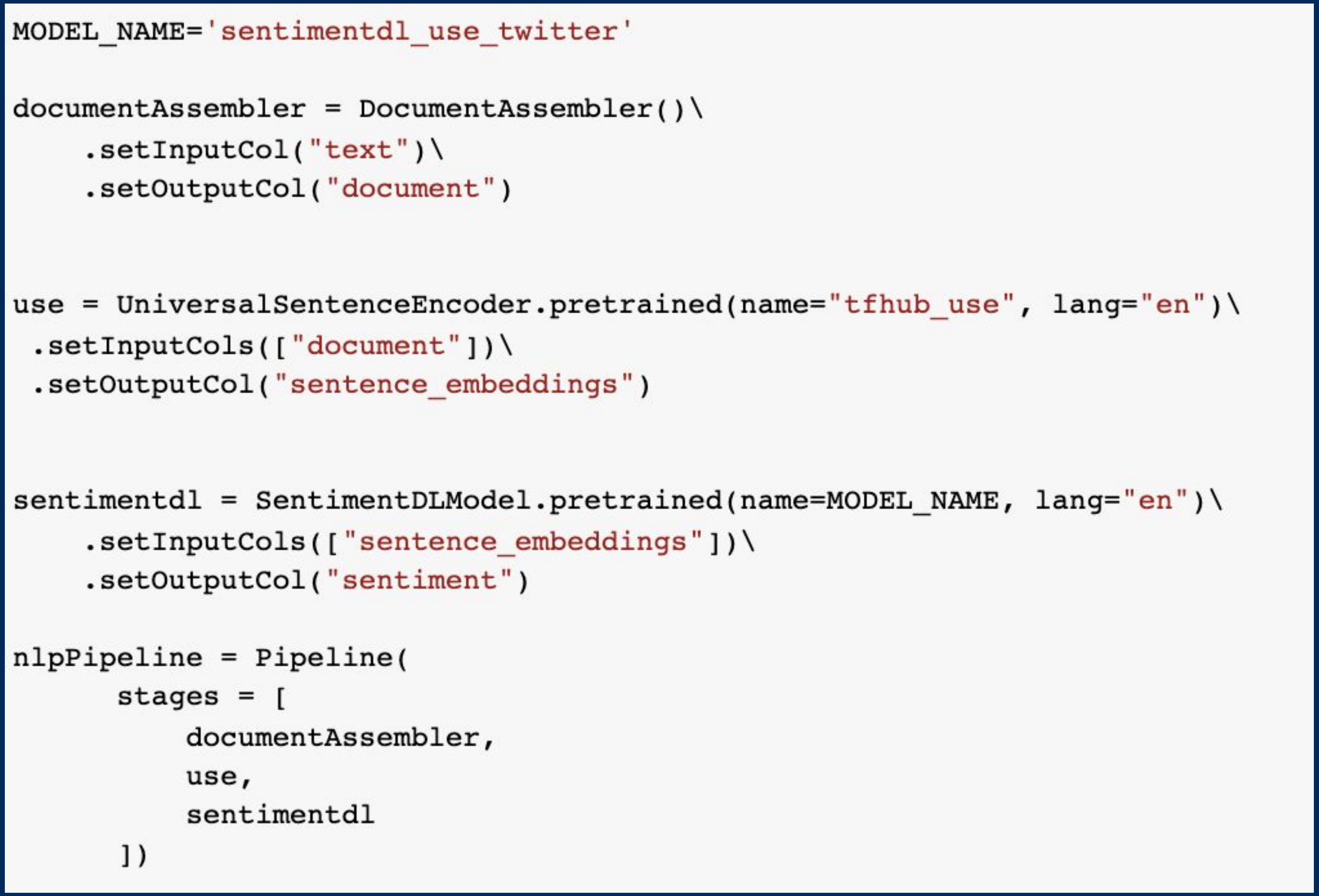


```
+----------------------------------------------------------------------------------------------------+
|                                                                                 sentence_embeddings|
+----------------------------------------------------------------------------------------------------+
|[[sentence_embeddings, 0, 10, works great, [sentence -> 0, token -> works great, pieceId -> -1, i...|
|[[sentence_embeddings, 0, 6, deleted, [sentence -> 0, token -> deleted, pieceId -> -1, isWordStar...|
|[[sentence_embeddings, 0, 50,  amp are block has good estimator app their website, [sentence -> 0...|
|[[sentence_embeddings, 0, 18, this person equador, [sentence -> 0, token -> this person equador, ...|
|[[sentence_embeddings, 0, 15, about tree fiddy, [sentence -> 0, token -> about tree fiddy, pieceI...|
|[[sentence_embeddings, 0, 48, why the entire thing just octaves lmao wait liszt, [sentence -> 0, ...|
|[[sentence_embeddings, 0, 18, sanity for the weak, [sentence -> 0, token -> sanity for the weak, ...|
|[[sentence_embeddings, 0, 414, sta similar future you will thank you your young you can aggressiv...|
|[[sentence_embeddings, 0, 12, yup what said, [sentence -> 0, token -> yup what said, pieceId -> -...|
|[[sentence_embeddings, 0, 331,  seriously thought there was maybe something wrong with you haha y...|
+----------------------------------------------------------------------------------------------------+
```



```
+----------------------------------------------------------------------------------------------------+
|                                                                                           sentiment|
+----------------------------------------------------------------------------------------------------+
|                [[category, 0, 10, positive, [sentence -> 0, positive -> 1.0, negative -> 0.0], []]]|
|   [[category, 0, 6, negative, [sentence -> 0, positive -> 0.084529385, negative -> 0.9154706], []]]|
|[[category, 0, 50, positive, [sentence -> 0, positive -> 0.99995494, negative -> 4.5064273E-5], []]]|
|   [[category, 0, 18, negative, [sentence -> 0, positive -> 0.04781373, negative -> 0.9521863], []]]|
|   [[category, 0, 15, positive, [sentence -> 0, positive -> 0.6832606, negative -> 0.31673935], []]]|
|  [[category, 0, 48, positive, [sentence -> 0, positive -> 0.9729006, negative -> 0.027099408], []]]|
|[[category, 0, 18, negative, [sentence -> 0, positive -> 1.00694604E-4, negative -> 0.99989927], ...|
|               [[category, 0, 414, positive, [sentence -> 0, positive -> 1.0, negative -> 0.0], []]]|
|        [[category, 0, 12, positive, [sentence -> 0, positive -> 1.0, negative -> 6.028981E-10], []]]|
|               [[category, 0, 331, positive, [sentence -> 0, positive -> 1.0, negative -> 0.0], []]]|
+----------------------------------------------------------------------------------------------------+
```

Sentiment Analysis Pipeline



```
MODEL_NAME='sentimentdl_use_twitter'

documentAssembler = DocumentAssembler()\
    .setInputCol("text")\
    .setOutputCol("document")


use = UniversalSentenceEncoder.pretrained(name="tfhub_use", lang="en")\
 .setInputCols(["document"])\
 .setOutputCol("sentence_embeddings")


sentimentdl = SentimentDLModel.pretrained(name=MODEL_NAME, lang="en")\
    .setInputCols(["sentence_embeddings"])\
    .setOutputCol("sentiment")

nlpPipeline = Pipeline(
      stages = [
          documentAssembler,
          use,
          sentimentdl
      ])
```