Fractal AI Triathlon Stage 2

Problem statement and data sets

It is important for companies to understand the sentiments of their customers or clients in order to remain competitive and grow their businesses. Traditionally this depended on feedback surveys and informal, small-scale and often inefficient methods. Today, however, there is a great deal of information on social media and e-commerce platforms related to customers' experiences with products and services. This is available for companies to analyze and better understand their customers' experience and satisfaction.

Accurately identifying the sentiment in a body of text requires accuracy (to be useful) and automation (to deal with the scale and nature of data available today). Machine learning models provide one solution to this problem.

Here we have collected reviews for various products and services from different sources written between December 2016 and March 2017.

Problem Statement:

Build a machine learning model using reviews about companies' products and services that will predict whether reviews are positive or negative.

Training data set:
The training data set is divided into positive and negative reviews:

Positive: 122470 reviews

Negative: 7256 reviews

Note the imbalance between the number of observations in each class.

Data will be available in the **shared** directory on the AWS instance provided.

Data dictionary

Each review is in JSON format and contains the following fields:

'author': author of this review

'crawled': site from where it is crawled

'entities': if any

'external_links': if any links

'highlightText': if any

'highlightTitle': if any

'language': language of the review

'locations': if tracked of the author

'ord_in_thread': if any

'organizations': organization of author

'persons': if any

'published': date of publish

'text': review written by author

'thread': thread of review

'title': title of review

'url': link of review

'uuid': id of review

Test Dataset:

The test dataset has 3645 reviews in the same JSON format.

Evaluation:

Your model will be scored using **Matthews Correlation Coefficient (**MCC).

MCC is defined as follows:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC ranges from -1 (worst, complete disagreement between prediction and observation) to +1 (complete concordance between prediction and observation) with a value of zero indicating that the model is no better than random prediction. Read more about Mathews Correlation Coefficient here.

Why MCC?

1. MCC is considered one of the best single number representations of the confusion matrix - the formula for MCC considers all four cells in the confusion matrix.
2. MCC is particularly well-suited to unbalanced data sets.
3. MCC does not depend on which class is defined as positive and which one as negative.

What to submit:

1. Output: You should create a csv called “output.csv” in the format given in the example below:

Filename	Prediction
fdeufd53d7ed.json	negative
fdedmbkd537.json	positive
ceded37ddded.json	negative
cedeffd53ddd7.json	positive
qweedeufd537.json	negative

2. Code: please submit the code used to generate your predictions. Accepted formats are .ipynb, .py and .R. Please read the triathlon rules for more information.

How to submit:

1. Please name the output csv in the format : TeamMember1EmpID_TeamMember2EmpID.csv

Ex : If the Employee ID of Team Member 1 is F001 and 2 is F002 then the file is to be named as F001_F002.csv. In case the team has only a single member name as F001.csv. Download the file to your local machine

2. Code: please submit the code used to generate your predictions. Accepted formats are .ipynb, .py and .R. Please read the triathlon rules for more information. Please name the code file in the same format as mentioned above. Ex. F001_F002.ipynb (or .py or .R). Download the file to your local machine.

3. Put both files into a folder and create the zip in the same format as shown above. Ex. : F001_F002.zip and upload .

4. Enroll in the hive course http://hive.fractalanalytics.com/blocks/manage/courseinfo.php?id=433 if already enrolled during stage I you can directly submit your submission file , go to view content , go to Click here for submissions and upload your zip file .

Please note:

We will be providing access to an AWS instance to each competing team, as we did in the previous stage. These instances have been extensively tested by the FAA team with the following findings:

- Solving this problem using an ML approach took about one second to train the model
- Solving this problem using a deep learning approach took approximately 150 seconds per epoch

Please use this information to allocate the resources made available to you. Also, **when not using your instance please shut it down, and restart it when you wish to continue working on the problem.**

If you have questions or issues, please contact us:

1. Email faateam@fractalanalytics.com
2. Use the Slack group for the triathlon to communicate with us and your fellow participants (join the Slack group)