Assignment 2

Team Helsinki: Sergey Shin, Ansat Abirov, Azat Sultanov, Ruslan Kim

Project description

Performing this task we were processing the Twitter stream provided by the Queryfeed in the following way:

1. Each tweet was preprocessed (converted to the appropriate format from the initial HTML)
2. Each preprocessed tweet was then normalized and standardized
3. After normalization the standard Spark Machine Learning Pipeline was applied in order to determine the emotional meaning of the tweet

Tasks split:

Sergey

- Preprocessing
- Building prediction model

Ruslan

- Preprocessing
- Prediction model
- Report

Azat

- Fitting the model
- Normalization

Ansat

- Deploying
- Normalization
- Fitting the model
- Queryfeed API integration
- Core structure creation

Preprocessing

The preprocessing actually consists of transforming the html-like tweet fields to the RDD and Datasets afterwards.

Custom normalization

We found it very useful to describe the custom normalization we implemented more precisely. We used the following techniques:

1. Lower case
2. Delete all links
3. Delete all new lines
4. Remove all non-usable symbols like (‘.’, ‘;’ and etc)

ML pipeline

Our pipeline is built of the following stages:

1. Tokenizer - to split strings into words (tokens)
2. HashingTF - to convert input tokens into Feature Vectors
3. Applying Model (Linear Regression) - making predictions

ML Model choice

We found Logistic Regression to be the best model for our choice due to its high accuracy and simultaneously high performance in terms of the training time.
We selected the model across the following ones: Support Vector Classifier, Decision Tree Pure Classifier, Decision Tree Classifier with Random forests, Decision Tree Classifier with AdaBoosting.

Problems

While implementing the project we faced a lot of problems with fetching Queryfeed stream in a proper way: without fetch error and in a correct time intervals.

Moreover, selecting the best ML model among the described ones required a lot of time.

Conclusion

To sum up, while performing this assignment, we processed the stream and analyze whether each tweet had positive or negative mood inside. Finally we got 92% accuracy on 50 samples checked by hands.