Scalable Document Classification with Naive Bayes in Spark

Team Pavo

Overview

- Data Cleaning
- Modeling
- Google Cloud
- Prediction

Data Clearning

- Remove all punctuations and stopping words.
- For special characteristics,
  - INT, 2LINT, 3LINT, 5LINT, and 6LINT
  - DFLOAT, LFLOAT, 2LFLOAT, 3LFLOAT, 5LFLOAT, and 6LFLOAT
  - DATE
  - CURRENCY
  - PUREFRAC
  - MIXEDFRAC

Modeling

- Naive Bayes Modeling
  - Calulate word counts
  - Add Laplace smoothing
  - Take log of the probabilities
  - Classical Naive Bayes Modeling
- Logistic Regression Modeling

Google Cloud

- Data Storage
  - Main python file is saved along with sample data
- DataProc
  - Before creating a cluster make sure a billing account is added to that project. Open Google Cloud consloe--Billing--add billing details
  - Create a cluster-gcloud dataproc clusters create cluster-name Manually set master and worker configuration by using GCP console.
  - Setting up a Job: gcloud dataproc jobs submit spark --cluster cluster-name -mainpthonfile.py-arguments/ use GCP console

Prediction

More grams, better results?