Feature-Based Tweets Sentiment Analysis on Movies

Partly Cloudy

Chaoran Fu, Kai Kang

Dec 8, 2015

Subject and Problem

- IMDb, Rotten Tomatoes: One score per User
- Movie has several features

Creed (2015)

PG-13 | 133 min | Drama, Sport |
25 November 2015 (USA)

8.6

Your rating: ★★★★★★★★★★ -/10

Ratings: **8.6**/10 from 13,918 users Metascore: 82/100

Subject and Problem

- There are much more reviews on Twitter with different features(actors, director, music…)

Feature-based method is valuable for tweets

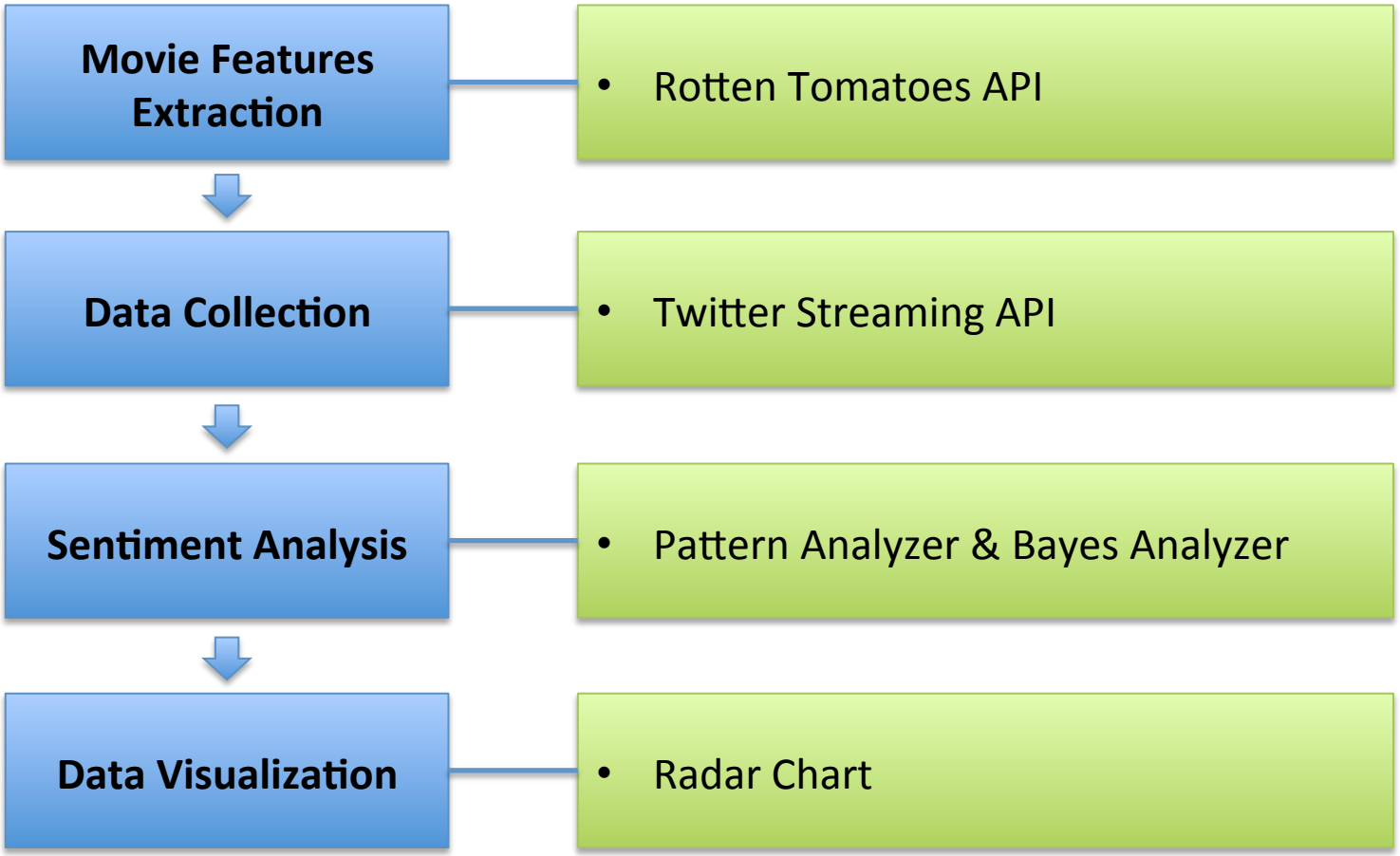
Previous Work

- Sentiment Analysis rarely used on tweets
- Feature-Based Focused on some other topics
- Rough sifting, only eliminating the tweets with URL or images(possibly ads)

Our Approach

- Implemented sentiment analysis
  - Pattern Analyzer for sifting
  - Naïve Bayes Analyzer for rating
- Tested on Spark
- Evaluated the performance
  - Comparing with Single thread
  - Comparing different data partitions

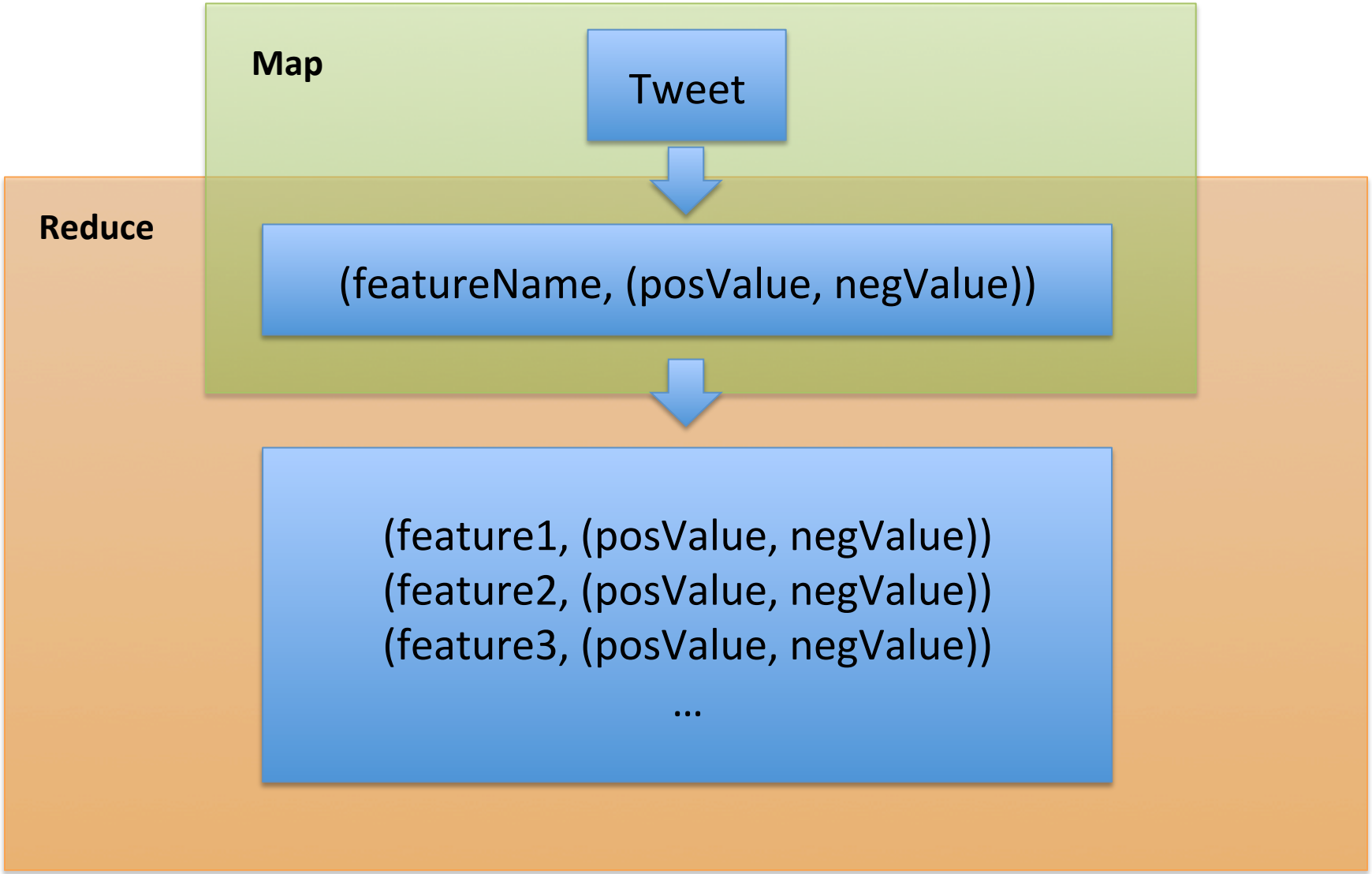
System Design



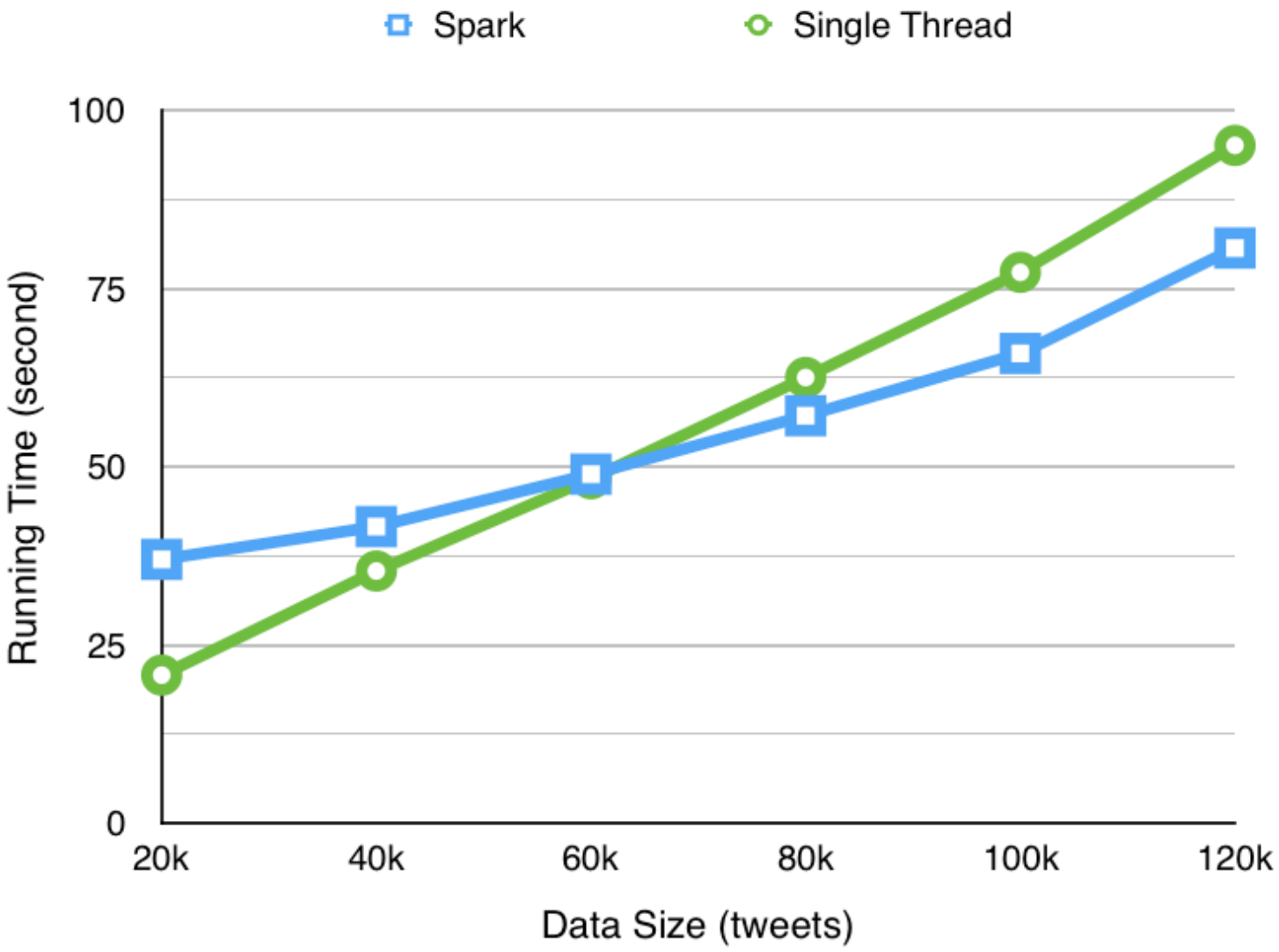
Algorithm

- Naïve Bayes Analyzer
  - Training: Movie Review dataset
  - Tweet → (pos/neg, posValue, negValue)
- Pattern Analyzer
  - Tweet → Subjectivity
  - “#Creed #Stallone” → 0
- Rating
  - Score = posValue / (posValue + negValue) * 5.0

Spark - MapReduce



Results – Spark VS Single thread

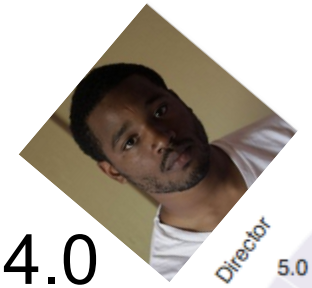


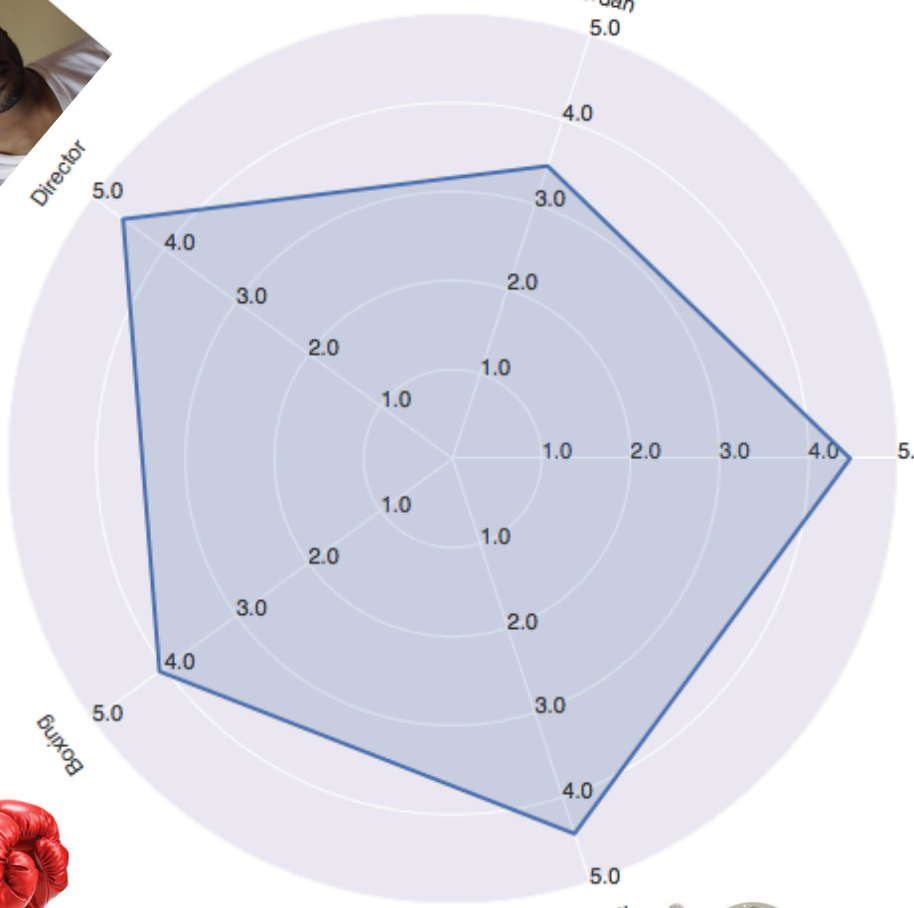
Results – Spark configurations

- 20K tweets data (1 movie, 2 days)
  - Different cores, different partitions → 40s
- 2000K tweets data (10 movies, 20 days)
  - 133 MB → 4 partitions
  - 2 cores / 4 cores : 14min / 11min

Results – Movie Rating

- 28,752 tweets
- General Score: 4.0





Future Prospects

- More movies, Larger data, Higher speed
- Better performance on Spark
- Maybe a real time service

Questions?