Introduction

- Social media becomes a platform of user generated text streams to express opinion.

Big Data

- Those user generated unstructured text data is Big Data with the following properties: Volume, Variety and Velocity.

- Two ways to process Big Data:
  - Batch mode analytics
  - Streaming analytics

Sentiment Analysis (Opinion mining)

- Natural Language Processing Techniques and
- Text Analysis Methods
- Extract, Quantify Subjective information in a text span.

Research goal:

- Investigates if stream-processing big data social media sentiment service with analytics can offer:
  - Scalability (enormous volume)
  - Efficient near-real time data processing and data analytics (sentiment analysis) compared to batch mode counterparts.

Contribution:

- Developed real time (streaming) Framework for complex multiphase Sentiment Analysis:
  - Social Media Data Stream Sentiment Analysis Service (SMDSSAS)
- Two Topic Sentiment Models for Social Media Sentiment Analysis.
  - Deterministic Topic Model – Instances of positive and negative sentiment in context of user provided topic word(s).
  - Probabilistic Model – Identify instance of positive and negative sentiment by log of the ratio of sentiment count per topic correlated messages.

System Architecture

- Created DStream to request Twitter data stream from Twitter Streaming API

Twitter Streaming API

- Created DStream to request Twitter data stream from Twitter Streaming API

Apache Spark Streaming

- Internally Spark streaming receives:
  - live input data streams
  - divides streams into batches (resilient distributed RDDs) using Map and reduce

- Processed results pushed out in DStream: Continuous abstracted stream of data (RDDs)

Deterministic Topic Model

- Given the presumption that topic and sentiment can be jointly inferred:
  - Counted instances of positive and negative sentiment in context of user provided topic word(s).
  - Likelihood estimated as relative frequencies
  - Tweets categorized by polarity and subjectivity

Probabilistic Model

- Per-word, log-based calculation.
- Calculates log of the ratio of individual positive and neutral subjective word count per topic correlated tweet.

SMDSSAS Data Storage/Stream Layers

- SMDSSAS Feature Extraction/Prediction Layers
- Real-Time Cross Validation

Conclusions

- Our topic model in SMDSSAS on the pre-election Twitter data stream successfully predicted the candidate Donald Trump as the winner of the 2016 Presidential Election. Positive Polarity of Trump significantly higher compared to Clinton’s 0.60 vs 0.069 (Shown below for Deterministic Model)

References