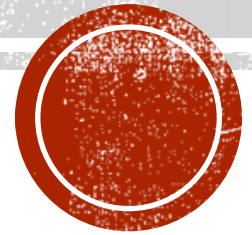


# TRENDING AND SENTIMENT ANALYSIS ON TWEETS

Team 3

Kuan Hu [hu.kua@husky.neu.edu](mailto:hu.kua@husky.neu.edu)

Ningtong Bai [bai.ni@husky.neu.edu](mailto:bai.ni@husky.neu.edu)



# PROJECT GOAL

- Utilize Twitter Restful API to acquire real tweets
- Parse twitter sources to JSON Format
- Build Topic Model (LDA model)
- Process Stream with Spark Streaming
- Perform Sentiment Analysis



# USE CASE

- INPUT:
  - ◆ Keyword (optional)
- OUTPUT:
  - ◆ Top 10 trending topics (including hashtags) from upcoming tweets containing that keyword in 1 minute
  - ◆ The sentiment score for each trending topic

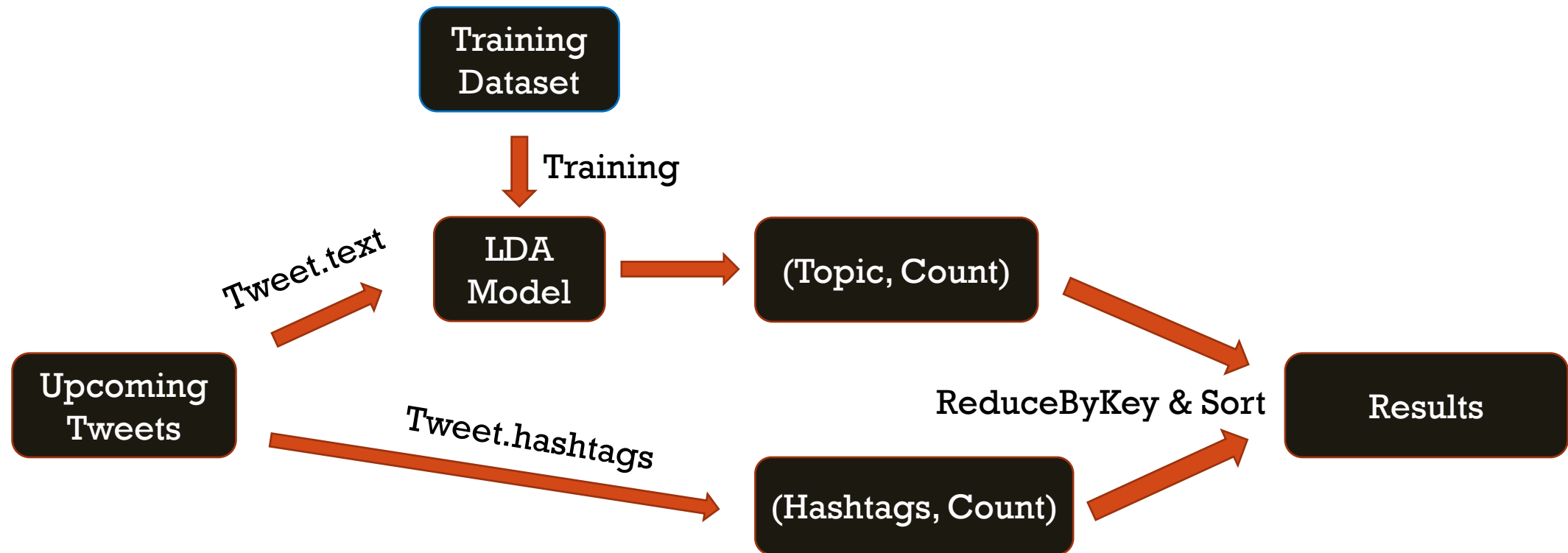


# METHODOLOGY

- |                              |    |   |
|------------------------------|----|---|
| ▪ Search & Streaming API     | => | Acquiring Tweets  |
| ▪ JSON Format Transformation | => | Parsing Twitter Sources                                   |
| ▪ Tweets Text Filter         | => | Filtering Language, Emoji, Stop Words                     |
| ▪ Topic Model Training       | => | Training LDA Model  |
| ▪ Spark Streaming            | => | Reading Tweets and ranking topics<br>(Mapping & Reducing) |
| ▪ Stanford NLP               | => | Calculating Sentiment                                     |



# TRENDING TOPICS & HASHTAGS



# SENTIMENT ANALYSIS

- Utilize Stanford Natural Language Process library to process, split, and parse input tweets
- Determine sentiment attitude (positive, neutral, negative) on tweets
- Calculate sentiment scores and assign it to tweets



# PROGRAMING & TEST

- ▼ scala
  - ▼ Ingest
    - Ingest
    - Tweet
  - ▼ Retrival
    - DownloadTwitterText
    - GenerateTopTrends
    - InferenceTopics
    - TwitterClient
    - TwitterLADModel
  - ▼ SentimentAnalysis
    - SentimentAnalysis

- ▼ scala
  - ▼ Ingest
    - TweetSpec
  - ▼ Rtrivial
    - TwitterLDAModelSpec
  - ▼ SentimentAnalysis
    - SentimentAnalysisSpec



# RESULTS

---

Time: 1513031450000 ms




---

(Movie,0.5263)  
(#bluecoveday,2.0000)  
(#dcsuperherogirls,0.0000)  
(#mlscup!,1.0000)  
(#traffic,0.0000)  
(#climatechange,1.0000)  
(#flashfrost...,0.0000)





# ACCEPTANCE CRITERIA

- Ensure topic generated reasonably from tweets
  - The LDA model will generate correct topic with a tweet
  - Acceptance Criteria: 60% accuracy 
- Rank Top 10 trending topics
  - The top 5 topics should related to the keywords
  - Acceptance Criteria: 60% accuracy 
- Sentiment Analysis
  - Characteristics of the tweet are correctly determined
  - Acceptance Criteria: 90% accuracy 



# CODE REPOSITORY

- <https://github.com/KuanHu/TopHitsAnalysisOnTwitter>

# REFERENCE

- <https://developer.twitter.com/en/docs/tweets/search/api-reference>
- <http://ampcamp.berkeley.edu/big-data-mini-course/realtime-processing-with-spark-streaming.html>
- <https://fangjian0423.github.io/2016/02/10/sparkstreaming-programming-guide/>

