Using Apache Storm to track location-based sentiments
Apache Storm - Overview

• Fault-tolerant, distributed stream processing
• Peak load 1 million tuples/second per node
• Used by many major companies
• Similar to MapReduce but runs infinitely
• More mature than other models (e.g. Samza)
Apache Storm – Spouts and Bolts

Data

Spout

Data

Spout

Shuffle

Bolt

Emit tuple

Bolt

Track lineage
Project

- Twitter as an example for unbounded data streams
- Limited filters in streaming API => process raw tweets in Storm
- Storm does not scale well on large states
- Employ stream approximation techniques to capture statistics
Architecture

Twitter stream filtered by bounding box

TwitterSpout

Filter Bolt

Tracker Bolt

Location Service

- Cluster to predefined locations

- Filter matching tweets
- Top k words
- ...
Approximating stream frequencies

Count-Min Sketch: trading off precision/space

Source: https://highlyscalable.wordpress.com/2012/05/01/probabilistic-structures-web-analytics-data-mining/
#StarWars, #TheForceUnleashed

Location = New York, tweets = 48568, topic related = 14
Location = Ferguson, tweets = 70362, topic related = 12
Location = Los Angeles, tweets = 14723, topic related = 3
Location = Montreal, tweets = 3847, topic related = 0

Most popular words in relevant tweets:
Word = #StarWars, count = 29
Word = #nerdlife, count = 29
Word = #lego, count = 29
Word = http://t.co/nsGsufTsq3, count = 29
Word = good!!, count = 29
Word = Lego, count = 28
Word = Star, count = 28
Word = Aaaand, count = 28
Word = so, count = 28
Word = version, count = 28

Total tweets this session = 95000
#nerdlife, #lego

Lego Star Wars: Episode VII - The Force Awakens Teaser Trailer
Future work

• Evaluate sketch data structures against different data rates
• Comparison to Naiad
• Simple positive/negative filter
• Hierarchical clustering of locations (AGNES)