**Harvard University Extension School**

**"Principles of Big Data Processing"**

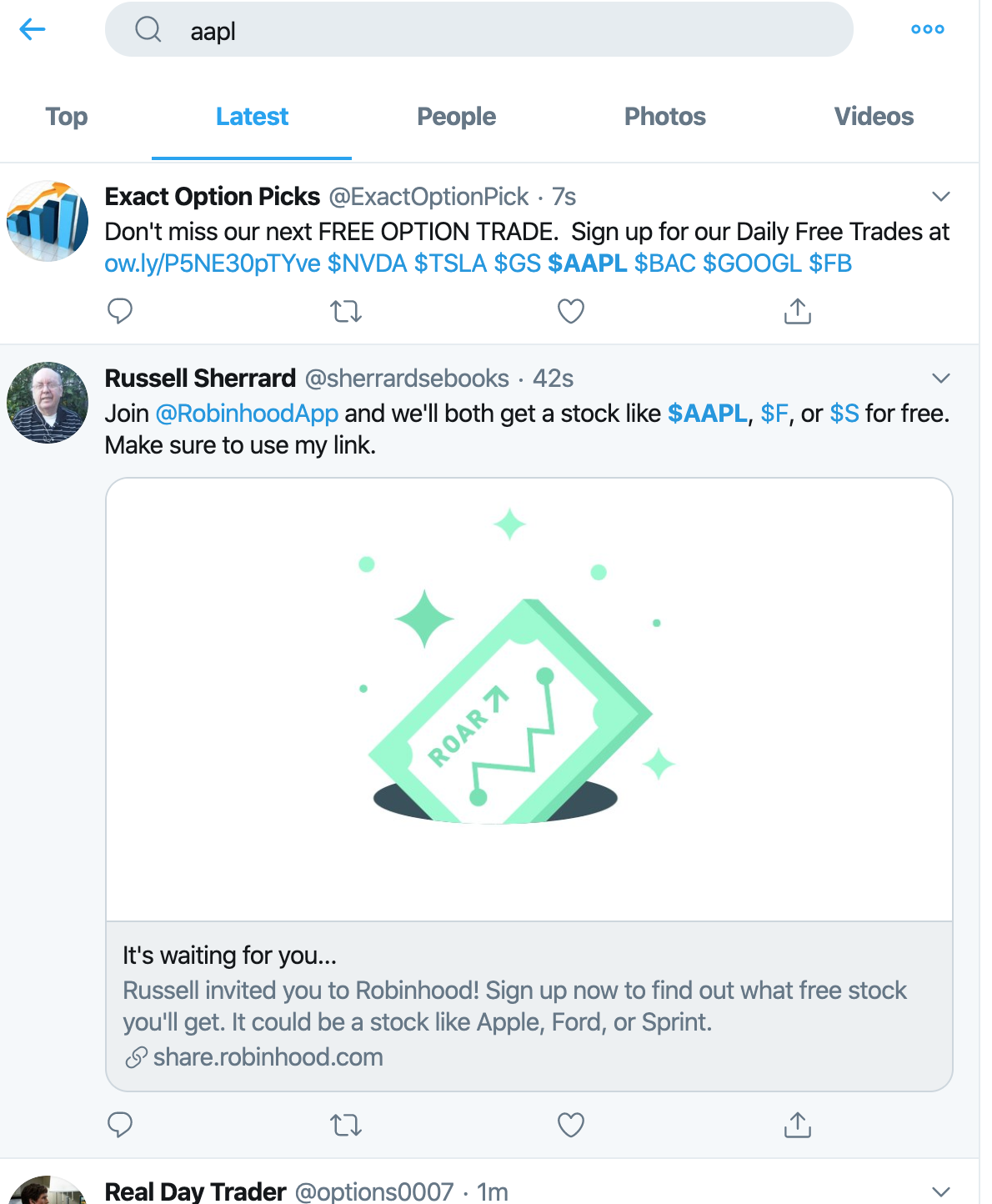
**CSCI E-88, Fall 2019**

**Final Project Proposal**

**by Loi Cheng**

## Project Goal and Problem Statement

This project’s goal is to study real-time tweets of the 30 stock symbols in the Dow Jones Industrial Average. We will demonstrate how to build a system that collects only twitter data that mentions these stock symbols and indexes them into ElasticSearch for further analytics. For example, below are some mentions for AAPL, which is Apple Inc.



## Big Data Source

Twitter streaming data

## Expected Results

As a result of this processing pipeline, I expect to be able find which stocks have higher mentions, and also if any of the stock mentions are concentrated at any regions in the world.

## Processing Pipeline

Flume

Kafka

Kafka ES Connector

ES



twitter

## Pipeline Overview and Technologies used

* Collection tier: Flume with experimental Twitter streaming source
  + Flume will be ingesting data from Twitter, filtered by stock symbol tags
* Messaging Tier: Kafka
  + Flume will push events into Kafka for further processing
* Stream Processing Tier: Kafka ElasticSearch Connector
  + Connector will index data into indexes
* Visualization Tier: we will use Kibana with ElasticSearch to visualize received data and discover which stocks are popular and in which area

## New Technology/Framework used

* Flume Twitter Source
* Kafka ES Connector