Topic 1: Implement a Tweet Clustering System

Clustering in the context of text essentially means grouping a set of related texts into logical groups (topics). For example, if you have a set of sentences from news articles on the Web that talk about President Obama and some that talk about President Bush, you can group those sentences into 2 groups. One group about Obama, and the other about Bush.

In this project, you will be working on clustering a set of **Tweets**. Tweets as you know covers a broad range of topics and the goal is to cluster these Tweets by topics. For example, if you have 300 Tweets about the iPhone 5s, the topics within these tweets may be *"complaint"*, *"poor customer service"*, *"display"* and *"user-friendly"*. *So,* you would need to write a program to discover such topics using a Web API. In this project you will do the following:

1. **Collect Tweets:** Write Java code to collect a set of Tweets (>500) using the Twitter Streaming API for a few brands or products (e.g. CitiBank, Bank of America, iPhone 5s, Samsung Note 4, Microsoft Surface Pro, etc). You need to collect Tweets for at least 2 brands/products and each should have at least 500 Tweets.
2. **Clean Tweets:** Write Java code to clean the Tweets (e.g. remove duplicate tweets, lowercase text, remove unwanted text like @,!,hyperlinks, etc). You can use regular expressions for this.
3. **Cluster Tweets:** Write Java code to cluster these Tweets using the RxNLP's Sentence Clustering API. You can subscribe to the free version on mashape or contact me for an API Key. You will send an API request in JSON to the **Sentence Clustering** endpoint and you will get back a response in JSON. You will have to parse the JSON response and make it usable by your code. The results should be stored in XML for visualization purposes.

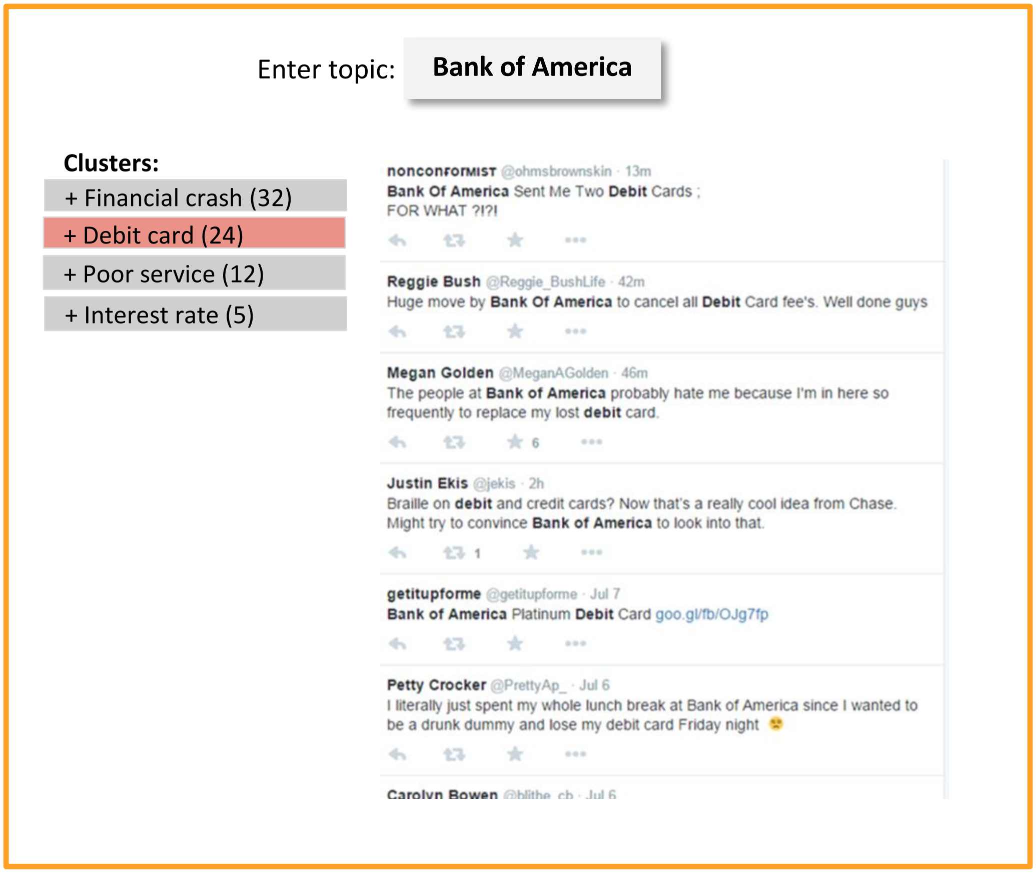
   Note:

   - This code should run on-demand. Meaning, given a new set of "cleaned Tweets" the code should be able to access the clustering API and output results in XML.

- Your XML file should be comprehensive. Meaning for each topic, there needs to be the list of tweets, and all other information provided by the clustering API.
- For your Demo, you can use the browser to show the XML output or create a user interface (see below) for extra points.

4. **User Interface for visualization (10 extra points):** Provide a user interface piece to visualize the results. This can be written in any language of your choice using any technology as long as the previous 3 code pieces are written **ONLY in Java**. This is left to your creativity. **Some Ideas:**
    a. C# FrontEnd with Java Backend
    b. JSP + HTML/CSS

Here is an example **mock-up of the System.** You are free to use different visualization approaches as long as the core code specified above is written in Java.



Enter topic: **Bank of America**

Clusters:

+ Financial crash (32)
+ Debit card (24)
+ Poor service (12)
+ Interest rate (5)

nonconformist @ohmsbrownskin · 13m
Bank Of America Sent Me Two **Debit** Cards ;
FOR WHAT ?!?!

Reggie Bush @Reggie_BushLife · 42m
Huge move by **Bank Of America** to cancel all **Debit** Card fee's. Well done guys

Megan Golden @MeganAGolden · 46m
The people at **Bank of America** probably hate me because I'm in here so frequently to replace my lost **debit** card.

Justin Ekis @jekis · 2h
Braille on **debit** and credit cards? Now that's a really cool idea from Chase. Might try to convince **Bank of America** to look into that.

getitupforme @getitupforme · Jul 7
Bank of America Platinum **Debit** Card goo.gl/fb/OJg7fp

Petty Crocker @PrettyAp_ · Jul 6
I literally just spent my whole lunch break at Bank of America since I wanted to be a drunk dummy and lose my debit card Friday night 😟

Carolyn Bowen @blithe_cb · Jul 6

What you will learn with this project:

- Clustering text at a high-level
- Accessing Web-APIs
- Data cleaning using regular expressions and other types of pre-processing
- I/O on static data (collected Tweets)
- JSON and Parsing JSON into a usable format
- UI interaction with your Java back-end (extra points)
- Simple visualization techniques (extra points)