One of you asked me how to proceed with Phase 2. Here are some tips:

1. Think how you will store tweets into Apache Spark. Look at what features of Spark can be useful. You can store all the attributes in tweets or a subset of the attributes depending on what analytical queries you are planning to pose. You could explore stream processing over tweets too. Choose the one that you are most comfortable with.

2. Design interesting analytical queries. For example, I may want to know the number of tweets per language during a particular time duration on a particular topic. I could visualize this using a pie chart. If I am operating over streaming tweets, I can refresh this pie chart periodically or have moving lines in a plot.

3. Have a good plan and complete what is required for phase 2. Add extra features if you have time.