**Springboard –Foundations of Data Science**

**Yoganand Mandali - Capstone project**

**Topic:**

1. Social Media Text Analysis using Twitter
2. Social Media Text Analysis using Facebook
3. Entity Resolution in R

**Proposal for topics 1 and 2 -**

**Problem:**

Customer focused businesses need to understand real-time customer choices, intentions and sentiments. Social media analytics provide an understanding of what the customer experience is and helps to achieve better consumer service and increase the effectiveness of marketing.

**Audience:**

The client for this analysis will be business development and product managers who develop customer centric products or services.

With social media analytics, insights could be gained about which business objectives can gain an advantage. Also, obtaining feedback on services and products, enhancing public opinion about a business division or a specific product can be achieved.

**Data:**

The countless number of tweets, blogs, comments and complaints regarding products and services form the input data for the analysis. This huge volume of information contains consumer sentiments that can be used to evaluate users' experience with a particular product or service.

The aim is to gather data from social networking engines using Streaming APIs and Search APIs developed for R

**Approach:**

Search social media engine (Twitter, FB) for product or service related posts

Summarize for each product or service

Assign a score for each sentiments on each post

Create sentiment word lists to differentiate positive, negative and neutral sentiments

**Deliverables:**

* R-code
* Slide deck

**Proposal for topic 3 –**

**Problem:**

Large amounts of data using by data driven companies don’t usually come from single or similar sources. Sources and the quality as well as quantity of data related to the same entity or variable can be different. Streamlining the data from various different sources by removing redundancy, creating unique keys for each variable is important before analyzing the data for trends or for decision making.

**Audience:**

The clients for this problem can be any firm which collects data from different sources where they have different sources providing the same data points but the contents of data are different.

For example- One source provides only phone and email of a person whereas the other sources provide address and email.

**Data:**

Data is still being researched.

**Approach:**

One of the following 3 methods -

* Record linkage
* Deduplication
* Clustering based ER

**Deliverables:**

* R-code
* Slide deck