Assignment #4: Social Networks and Text Processing
Spring 2016
CS3943
Prof. Rumi Chunara
Total: 30 points

All questions must be completed in R. Implement and comment your code so that anyone reading the file can reproduce the code easily (e.g. set the file path once at the beginning of the script where it can be easily changed). Save the code as an R markdown file, and upload it to NYU classes.

1. Intro Network Analyses (10 points).
The purpose of this question is to get you started with network analysis. The data we will use derived from Internet-posted recipes. There are two files both in the Resources section of NYU Classes: subs.txt, which contains the edges (with directions) and keys.txt, which contains the ingredients (node labels). An edge exists between ingredient *i* and ingredient *j*, if *j* was recommended as a substitute for *i* in at least 5% of the comments recommending substitutions, e.g. "I liked this recipe, but I used Cinnamon instead of sugar..." Therefore the edges are *directed*.

a) Read in the files and visualize the network (try using ggplot2 or networkD3 libraries).
b) Calculate the degree centrality of each node.
d) Which are the most "connected" node(s).
c) Visually determine what are the furthest ingredients from cocoa powder.

2. Crawling Twitter (10 points).
In class we learned about using the Twitter API in R. This question repeats what we did, asking you to implement it on your own with a small change to increase the size of the network. Use the hashtag #SXSW2016 which we selected in class (we selected it because has enough coverage and is generated in a somewhat closed community).

a) Download 100 users ids that have tweeted about this, and their friends/followers. *Note that due to rate limits you may need to include a pause in order to be able to download data on this many users.*
b) Assess and plot the degree distribution of your network (choose either in-degree or out-degree and motivate why you chose the metric).
c) Visualize the network (try using ggplot2 or networkD3 libraries).

3. Developing a Language Model (for *n-gram* word sequences) (10 points).

a) Download Tweets from each user above that mention the hashtag you selected (over an appropriate time period).
b) For n = 1, n = 2 and n = 3, submit the list of the 10 most frequent sequences.
c) For n = 1, n = 2 and n = 3, submit the sum of all frequencies of all sequences for that n.
d) Using these frequencies, generate a distance measure for individuals (e.g. they share the X most common frequency 3-gram terms, or 2-gram terms, or 1-gram term). How does this network look compared to the network generated in question 2?