Data Description

The dataset you will be working with comes from Venmo, the most popular US-based Peer-to-Peer (P2P) payment application (https://en.wikipedia.org/wiki/Venmo). You have been provided with a sample of approximately 7 million transactions dated between 2009 and 2016. The dataset contains the following fields (in csv format):

- **User1**: Unique user ID – user1 is the person who initiates the transaction.
- **User2**: Unique user ID – user2 is on the receiving end of the transaction.
- **Transaction_type**: Charge (requesting money) or payment (paying money) – can ignore
- **Description**: user specified description of what the money was used for.
- **Is_business**: 0/1 flag indicating whether user1 or user2 is a business entity – can ignore
- **Transaction_id**: unique ID for every transaction.

Social Network Analysis (Problem 1)

Data: /home/public/course/venmoSample.csv

The transaction file represents Venmo's social network: the nodes of the network correspond to users, and the edges represent the transactions between users. One of the fundamental properties of every network is its degree distribution (https://en.wikipedia.org/wiki/Degree_distribution), i.e. the number of edges/connections an individual has to other nodes in the network. There exist two types of network: directed (e.g. Facebook) and undirected (e.g. Twitter: user A following user B does not imply that user B follows user A); depending on the network type, the number of degrees of each node can vary. Venmo is a directed network, as the act of user A sending money to user B has a direction associated with it. In directed networks the exist two types of degrees: in-degree (i.e. number of incoming connections) and out-degree (i.e. number of outgoing connections).

For this problem you have to perform the following three tasks:

1. Plot Venmo's degree distribution by treating the network as undirected.
2. Plot Venmo's in-degree and out-degree distributions.
3. What is the percentage of reciprocal transactions in the network (this should be a single number)? Reciprocity is defined as follows: both user A and user B have sent money to each other. Create a plot that shows the percentage of reciprocal transactions over time, with six month increments between January 1st 2010 (start date) and June 1st 2016 (end date).

Notes:

- You can assume that all transactions are payments, not charges. In other words, you do not have to create separate degree distributions for payments and charges.
- You are free to use either dataframes or RDD API, or a mix of the two. (You are encouraged to use dataframes.)

Emoji & Text Analysis (Problem 2)

Data: /home/public/course/venmo/venmoSample.csv

One of the most fun aspects of using Venmo is picking emojis to describe your spending habits. Various analyses have tried to identify the most popular emojis in Venmo: https://bankinnovation.net/2016/09/what-do-people-use-venmo-for-check-the-emojis/.

In this exercise, you have to perform the following three tasks:

1. Find the top 10 most popular emojis on Venmo.
2. Find the top 5 most popular emojis on Venmo by weekday (similar to the link above).
3. Analyzing the content of all transaction messages in a consistent manner is a hard task; some descriptions only contain text, others only emojis and so on. In order to facilitate the classification of transactions into different categories (e.g. food, drinks, utilities), you should cluster transaction messages using text-based attributes. Examples of attributes include, but are not limited to, the number of characters in a message, presence of emojis in the message, etc. In your write-up explain your reasoning behind the attributes you selected (you need at least 5 attributes). When deciding what attributes to use, keep in mind that the end goal is to improve the classification of messages by applying a separate text classification algorithm to each cluster, as messages within the same cluster share some basic structural similarities. You DO NOT have to perform any text classification for this problem. Include a brief overview of your results in the write-up – some sort of visualization is expected.

Notes:

- Working with emojis can be challenging. While there exist various libraries that can parse emojis, all of the above tasks must be completed WITHOUT the use of any external python libraries.
- For parts 1 & 2 note that it is possible for a transaction message to contain more than one emoji, or multiple copies of the same emoji. In your write-up explain how you dealt with these issues.
- You are free to use either dataframes or RDD API, or a mix of the two. (You are encouraged to use dataframes.)