computational social media

assignment #2

08.03.2019
The assignment: hands-on exercise with Facebook data
Goals

1. You are asked to compute basic statistics from a publicly available dataset generated by a set of Facebook users.

2. Through the exercise, you will (a) practice basic concepts about descriptive statistics at the level of messages and at the level of users; (b) reflect about trends (and some potential biases) emerging from the data; and (c) apply concepts about personality traits discussed in class.
Dataset details

- 9900 Facebook status updates from 250 users
- The data was collected using the myPersonality application (Stillwell & Kosinski, discussed in class)
- It contains three types of features: (1) text status reports; (2) big-5 self-reported personality traits; (3) social network structure features
- Personality traits obtained from 100-item International Personality Item Pool (IPIP) questionnaire (http://ipip.ori.org/)

Dataset details (2)

- **AUTHID** is the list of all users (anonymized); there are 250 unique AUTHIDs
- **Status**: Text status updates by the users (9900 in total)
- **Columns (C-G)** are 5 self-rated scores, one score for each personality trait (extraversion, neuroticism, agreeableness, conscientiousness, openness)
- **Columns (N-T)** are 7 network features; see next slide
- **Date**: Date and time when status update was posted.
Network features in the dataset

- **Network size**: Total number of connections a node has
- **Betweenness centrality**: The proportion of shortest paths that a given node lies on
- **Density**: Number of ties, expressed as a percentage of the number of possible ties
- **Brokerage**: role played by a node that mediates contact between two other nodes.
- **Transitivity**: Node A is connected to Node B and Node B is connected to Node C, then Node A is connected to Node C. i.e., “friend of a friend is a friend”.

Instructions

1. Download the dataset from:

2. The downloaded dataset should contain 3 files: (a) csv file with data; (b) paper that describes the dataset; (c) readme.txt

3. Using the data, compute the following descriptive statistics
   • Total number of users – user-level stats
   • Total number of status updates – status level stats
   • Statistics of status updates per user (Mean, Median, Min, Max, Std)

4. Plot the cumulative distribution (like in Slide 20 of Lecture 2 part 1) for
   • Number of status updates per user (i.e., at the user level)
   • Number of characters per status message (i.e., at the status update level)
Instructions (2)

5. For 2 of the 7 network features (Network Size, and a second feature of your own choice), do the following:
   - Compute descriptive statistics for each feature per user (Mean, Median, Min, Max, Std)
   - Plot the cumulative distribution for each feature per user.

6. For the 5 personality trait features, do the following:
   - Compute descriptive statistics for each trait per user (Mean, Median, Min, Max, Std)
   - Plot the cumulative distribution for each big-5 trait per user.

7. Compute a correlation matrix (at the user level), indicating correlation values and p-values for
   - Big-5 traits (5x5 matrix)
   - Big-5 traits vs. network features (5x7 matrix)

8. Submit the following
   - A short report including results, tables, and plots, and discussing any findings you find worth commenting.
   - Include your code as well.
Logistics and deadline

1. Python as programming language

2. In case of questions, contact Skanda by email.

3. Deadline to submit assignment (report+code): **Tue 19.03.2019, 5pm**
   - send by email to:
     - daniel.gatica-perez@epfl.ch
     - skanda.muralidhar@epfl.ch
   - pdf format
   - submit your assignment even if it is not complete
   - late assignments will not be given any credit