computational social media

assignment 2

daniel gatica-perez
skanda muralidhar

09.03.2018
assignment 2: basic 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 the trends emerging from the data and some potential biases; and (c) learn basic ideas behind self-reported personality traits.
introductory material (1): big-five personality traits
what is a personality trait?

"stable individual differences in the reactivity of mental mechanisms designed to respond to particular classes of situations »

source: oxford university press
the big-five personality traits

“the Big-Five traits have been broadly accepted as a way of presenting all the major traits of a person at the highest level of abstraction” (Gosling, 2003)

<table>
<thead>
<tr>
<th>Extraversion</th>
<th>outgoing, enthusiastic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>aloof, quiet</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>prone to stress &amp; worry</td>
</tr>
<tr>
<td></td>
<td>emotionally stable</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>organized, self-directed</td>
</tr>
<tr>
<td></td>
<td>spontaneous, careless</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>trusting, empathetic</td>
</tr>
<tr>
<td></td>
<td>uncooperative, hostile</td>
</tr>
<tr>
<td>Openness</td>
<td>creative, imaginative</td>
</tr>
<tr>
<td></td>
<td>practical, conventional</td>
</tr>
</tbody>
</table>

(Tupes & Christal, 1961; Norman, 1963; Goldberg, 1981; Costa & McCrae 1985; Digman, 1990)
the lens model: a basic model for interpersonal perception

Figure 2. The interpersonal circle. Adapted from “A Psychological Taxonomy of Trait-Descriptive Terms: The Interpersonal Domain” by J. S. Wiggins, Journal of Personality and Social Psychology, 1979, Vol. 37, p. 400. Copyright © 1979 by the American Psychological Association. Adapted by permission.

a big-five trait questionnaire
each item below is responded on a likert scale (5 or 7 points): from “strongly disagree” to “strongly agree”

**Extraversion items**
- I am the life of the party
- I don't mind being the center of attention
- I feel comfortable around people
- I start conversations
- I talk to a lot of people at parties

**Agreeableness items**
- I am interested in people
- I sympathize with others' feelings
- I take time out for others
- I feel others' emotions
- I make people feel at ease

**Openness items**
- I have a rich vocabulary
- I have a vivid imagination
- I have excellent ideas
- I am quick to understand things
- I use difficult words

**Neuroticism items**
- I am easily disturbed
- I change my mood a lot
- I get upset easily
- I have frequent mood swings
- I worry about things

**Conscientiousness items**
- I am always prepared
- I pay attention to details
- I get chores done right away
- I like order
- I follow a schedule

**TIPI - Ten-Item Personality Instrument**
(Gosling, 2003)
10 questions, 2 per trait

**NEO FFI - Five-Factor Inventory**
(Costa & McCrae, 1992)
60 questions, 12 per trait

http://en.wikipedia.org/wiki/Big_Five_personality_traits
introductory material (2):
basic network features
Network Features

- **Network size**: Total number of connections
- **Betweenness centrality**: The proportion of shortest paths that the given node lies on
- **Density**: Number of ties, expressed as a percentage of the number of possible ties
- **Brokerage**: Role played by a node which 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.

[Introducing Social Networks Degenne, Forse 1999]
Dataset Details

- Consists of 9900 Facebook status updates from 250 users
- Collected using the myPersonality application [Stillwell & Kosinski]
- Contains information about big-5 self-reported personality traits and social network structure
  - Personality traits obtained from 100-item International Personality Item Pool (IPIP) questionnaire [http://ipip.ori.org/]
Assignment #2

1. Download the dataset from the following site:
   http://mypersonality.org/wiki/doku.php?id=download_databases

2. The downloaded dataset should contain 3 files: (a) csv file with data (b) corresponding paper (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
   - Status updates per user (Mean, Median, Min, Max, Std) – status level stats

4. Compute the following descriptive temporal statistics
   - Time difference (deltas) (in hours) between consecutive status updates (Mean, Median, Min, Max, Std). Discard data from users that only have one status update. Compute this in 2 ways:
     (a) At the status level: using all status updates as if they were independent (don’t mix data from different users to compute deltas!)
     (b) At the user level: first compute the mean time difference for each user separately, and then use the mean value as attribute of each user to compute mean, median, etc.
     (c) How different are the various statistics from the 2 ways of doing it?
   - Plot the histogram of timestamps of all status updates on a month-by-month basis to cover all the time period. Do you see any pattern?
Assignment #2

5. Plot the cumulative distribution (like Slide 23 of Week-2 lecture) for
   • number of status updates per user
   • Number of characters-per-message (at the status level)
   • time difference between consecutive status updates (at the status level)
   • Depending on the specific variable to plot, use either a linear X axis
     or a logarithmic one.

6. Using the Network features
   • Generate descriptive statistics for each network feature per user (Mean, Median, Min, Max, Std) – *user level stats*
   • Plot the cumulative distribution for each network feature per user.

7. Using the Personality trait features
   • Generate descriptive statistics for each trait (Mean, Median, Min, Max, Std) – *user level stats*
   • Plot the cumulative distribution for each big-5 trait.

8. Please submit the following
   • Write a short report including the results (as Tables) and the plots,
     and discussing any findings you find worth commenting.
   • Include your code as well.
assignment #2: logistics and deadlines

1. Use any programming language of your choice. A list of functions in Python is provided for your reference in the next slide

2. HARD DEADLINE TO SUBMIT REPORT: Wed 21.03.2018, midnight
YOU HAVE 10 DAYS TO COMPLETE THE ASSIGNMENT
   - 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
Python Functions

- Using Pandas `[from pandas import DataFrame as df]
  - data.describe()

- Using NumPy[import numpy as np]
  - np.mean(data)
  - np.median(data)
  - np.min(data)
  - np.max(data)
  - np.std(data)