Social science dataset (Swedish version)

Overview

The dataset contains about 11 200 tweets in Swedish that mention terms related to diabetes (diabetes, diabetik, insulin, blodsocker, hypogly, diabulimi, hba1c, flashmätar, flash-mätar, and långtidsvärde). The tweets were collected over 267 days and the twitter language API was used to determine that the tweet was in Swedish.

Sample questions

The following are some of the questions that researcher would use a dataset such as this to investigate. Feel free to be inspired by or start from either of these.

- What is the size of the (active) Swedish diabetes network on twitter.
- Who are the central actors and which are their roles.
- Are there any smaller networks within the larger networks, and do these smaller networks have distinct profiles?
- Are there any overlapping networks, are the members in this network part of other networks, and if so, what are their profiles (e.g., other illnesses, activites, etc.)
- Can we trace users between different social media (e.g., Twitter and Instagram), and can we use this information to extend the network?

Detailed information

The dataset consists of a single file that contains the tweets in JSON format. Each row in the file is a JSON object. A JSON object is a set of key-value pairs, where the key is a string and the value can be of different types. For example, the key-value pair `screen_name":sofiabremsjo"` has key `screen_name` and the value is a text string that represents the author's screen name (`sofiabremsjo`). You can find a description of all the keys in a tweet JSON object at https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object.html.

References and additional data

You can find more information about the Twitter API at their developer site https://developer.twitter.com/en.html.

Example

The following Python example finds the number of distinct screen names in the data set.

```python
import json

# Set of screen names
sn = set([])

# Read the tweets file line by line and convert each line to a JSON object.
# Extract the 'user' key and then the 'screen_name' key. Add the name
# to the set of names (sn)
with open('diabetes_tweets.json') as f:
        for row in f:
                jo = json.loads(row)
                sn.add(jo["user"]["screen_name"])
```

```python

print('There are', len(sn), 'unique screen names in the dataset.')
```

You can also use jq to parse the data. The following set of commands achieves the same as the previous Python program.

```
jq < diabetes_April4.json .user.screen_name | sort -u | wc -l
```