Capstone Proposal

Lára Hrönn Pétursdóttir

January 11th 2017

All the people of earth have this in common: One day is one earth's rotation.
None has more or less time in a day - just more or less sunlight.

The project's domain background

We have probably all heard: "I´m an "A" type or "B" type". But are we really?

There are many forms used to divide people into personalities or types. Some as simple as categorising people in to A, B and even C types and other more complicated. Known within the field of psychology is for example the Psychometric Evaluation of the Revised Temperament and Character Inventory (TCI).

> *"TCI is a set of tests designed to identify the intensity of and relationships between the seven basic personality dimensions of Temperament and Character, which interact to create the unique personality of an individual. Basic personality factors are important predictors of risky health behaviours where they often result in a lifestyle more open to certain health issues" (http://psychobiology.wustl.edu/what-is-the-tci/).*

Simplified labels that represent how people spend their days would be helpful in a vide variety of fields. That would help include the influence of lifestyle in wide array of research and evaluations.

For example when evaluating a sample, a part of that evaluation could be labelling people depending on how they spend their day where that can affect their buying behaviour, health, family life, mood, level of happiness and so on.

A problem statement

Can we use unsupervised learning and clustering to divide people into groups / types / personalities based on how they spend their days? The first part of the problem is to find if it is possible in general to categorise people in this way. The second part is to find how many groups there would be and how they interact with each other.

The datasets and inputs

The American Time Use Survey (ATUS) is the Nation's first federally administered, continuous survey on time use in the United States. The goal of the survey is to measure how people divide their time among life's activities.

In ATUS, individuals are randomly selected from a subset of households that have completed their eighth and final month of interviews for the Current Population Survey (CPS). One individual age 15 or over is randomly chosen from each sampled household. This "designated person" is interviewed by telephone once about his or her activities on the day before the interview--the "diary day."

ATUS designated persons are preassigned a day of the week about which to report. Interviews occur on the day following the assigned day. For example, a person assigned to report about a Monday would be contacted on the following Tuesday.

In the time diary portion of the ATUS interview (the portion that will be used in this project), survey respondents sequentially report activities they did between 4 a.m. on the day before the interview ("yesterday") until 4 a.m. on the day of the interview. For each activity, respondents are asked how long the activity lasted and responds are in hours.

The ATUS describes the amount of time Americans spend doing unpaid, non market work, which could include unpaid childcare, eldercare, housework, and volunteering. The survey also provides information on the amount of time people spend in many other activities, such as religious activities, socialising, exercising, and relaxing. In addition to collecting data about what people did on the day before the interview, ATUS collects information about where and with whom each activity occurred, and whether the activities were done for one's job or business. Demographic information—including sex, race, age, educational attainment, occupation, income, marital status, and the presence of children in the household—also is available for each respondent. Although some of these variables are updated during the ATUS interview, most of this information comes from earlier CPS interviews, as the ATUS sample is drawn from a subset of households that have completed month 8 of the CPS.

In 2003, 3,375 households leaving the CPS sample were selected each month (about 40,500 over the whole year) and divided in 12 categories:

Household type	Race/ethnicity of household reference person in CPS			Total
	Hispanic	Non-Hispanic, black	Non-Hispanic, nonblack	
With at least one child under 6	1,500	1,000	5,400	7,900
With at least one child between 6 and 17	1,400	1,400	7,800	10,600
Single adult, no children under 18	800	1,800	5,900	8,500
Two or more adults, no children under 18	1,500	1,600	10,400	13,500
Total	5,200	5,800	29,500	40,500

In December 2003 and later it was reduced to 2,190 pr month divided by table 2 divided in the same 12 categories:

Household type	Race/ethnicity of household reference person in CPS			Total
	Hispanic	Non-Hispanic, black	Non-Hispanic, nonblack	
With at least one child under 6	1,200	600	3,400	5,200
With at least one child between 6 and 17	1,200	900	4,900	7,000
Single adult, no children under 18	700	1,600	4,300	6,600
Two or more adults, no children under 18	1,200	1,400	5,000	7,600
Total	4,300	4,500	17,600	26,400

The response rates for each year range from 48.5% to 57.8%:

Year	Response rate (percent)
2003	57.8
2004	57.3
2005	56.6
2006	55.1
2007	52.5
2008	54.6
2009	56.6
2010	56.9
2011	54.6
2012	53.2
2013	49.9
2014	51.0
2015	48.5

The activity classification system is a 3-tiered system with 17 major (or first-tire) categories, each having 2 additional levels of detail. Each third-tier activity category also contains a list of examples of activities that fall into that category. Codes are periodically updated prior to the start of each year's data collection.

All data is represented with a numerical value and can be translated using the ATUS Data Dictionary. In the ATUS Activity file each respondent is represented with a 14 number ID (TUCASEID) and answers with activity he or she did (TUACTIVITY_N), when it started and when it finished (HH:MM:SS). The ATUS Activity Summary file uses the respondent ID, his age and shows how many minutes (MMM) the respondent spends doing a certain activity

A solution statement

Use a series of unsupervised learning and clustering techniques to read cleaned data from the American Time Use Survey to see if it is possible to create simplified labels for people depending on how they spend their days.

A benchmark model

There is no obvious methodology agains which I can benchmark and there for I will compare it the statistical description of the ATUS summary and perhaps a more simple model K-Means clustering model. I will also select 5 sample respondents from the ATUS and see how they fit within the clusters.

A set of evaluation metrics

Silhouette score - to determine how many clusters are optimal
t-Distributed Stochastic Neighbour Embedding - dimensionality reduction for visual evaluation (see if there are clusters or not).

An outline of the project design

- Data Exploration
  - What information does it contain
  - Statical description of the dataset
  - Distribution and relevance and correlation
  - Deal with NaN
  - Remove or merge features if necessary
  - Merging with other dataset if necessary

- Data Preprocessing
  - Feature and scaling
  - Identify outliers and decide if they should be removed or not

- Feature Transformation
  - Load the dataset into a big sparse matrix so we only look at the values that are non zero
  - TruncatedSVD to make it into a smaller informative matrixesKMeans to group into clusters
  - TSNE to get coordinates for a scatterplot
  - bokeh to make interactive graphic

www.bls.gov/tus/lexicons.htm
https://www.bls.gov/tus/lexiconnoex0315.pdf
https://www.bls.gov/tus/atusintcodebk0315.pdf
https://www.kaggle.com/bls/american-time-use-survey
http://www.simplypsychology.org/personality-a.html
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2810834/
http://psychobiology.wustl.edu/what-is-the-tci/
http://www.sciencedirect.com/science/article/pii/S0191886901000800

https://www.youtube.com/watch?v=2lpS6gUwiJQ
http://lvdmaaten.github.io/tsne/