Computational Social Science

Research project proposition

Students:

Nađa Jeličić

Dhurim Sylejmani

Mentor:

JProf. Dr. Claudia Wagner

Project overview

- Analyzing factors leading to successful or bad results that students achived on Harvard and MIT open online courses.
- Extracting conclusions based on **gender**, **interaction** with courses, **origin** of students, **level of education** and other attributes.
- Creating descriptive statistics and conducting research on data.

Questions of interest

- **Why** did certain students achieve better results? Does for example level of education have a correlation with success?
- Comparing the success of students based on gender? Based on education? Based on amount of interaction with materials? Based on the popularity of courses taken? Dropout rates?
- **How** central in the network are the more successful students?
- Can we run an experiment by matching similar students?

Data set proposition

- Open dataset named “HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0, created on May 14, 2014”.
- 641 139 rows, detailed history of student’s data and participation.
- This dataset is at the level of one row per--person, per--course.

For acces follow link:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2F26147&studyListingIndex=1_66ddd8428ef019414859146e978e

Methods to analyse data

- **Descriptive statistics with Python**
  - Transferring data to Pandas data structures (mainly Data Frame) will allow better manipulation and clearer representation of data
  - Following, the comparison of the data can be conducted based on the **features**
  - Presenting results visually in numerous different styles like bar or scatter plots using Matplotlib library
  - Using measures of descriptive statistics like mean, median, variance, standard deviation etc.

Methods to analyse data

- **Buillding a network - Network analysis**
  - Using students and courses as nodes in the network would allow further analysis of their connections
  - Calculating centrality measures
  - Managing data: not all data needs to be included in the network aim is at least 70% but extreme cases can be eliminated
  - Possibility of folding the network to create a network of students and access their connections

Methods to analyse data

- **Matching**
  - Using of matching methods to run an experiment. Analysing the effects of different treatments (level of education, number of courses started…) on final success.
  - Using data about students to group similar people based on attributes other than the one used as treatment.

Additional idea

- Possibility for further analysis can be comparison with the data on Harvard and MIT traditional courses.
- For example there are statistics showing graduation and dropout rates
- Are the factors for success different online and offline

Overview of data attributes

- course_id
- user_id
- registered: 0/1
- certified: 0/1 anyone who earned a certificate
- origin - country or region
- LoE: Level of education
- YoB: Year of birth
- gender
- date of enrolment
- nevents: number of interactions ….

Thank you for the attention!