UNIVERSITEIT TWENTE.

Data Science [201400174]

Course year 2018/2019, Quarter 1B

DATE
November 22, 2018

TEACHERS
Maurice van Keulen
Christin Seifert
Mannes Poel
Karin Groothuis-Oudshoorn
Faiza Bukhsh
Miha Lavric

PROJECT OWNERS
Faiza Bukhsh
Karin Groothuis-Oudshoorn
Maurice van Keulen
Mannes Poel
Michel van Putten
Luc Wismans

COURSE COORDINATOR
Christin Seifert
Maurice van Keulen

Part II

Projects

Project 1

Project 1: Decision support for University timetables [TIMETABLES]

1.1 Introduction

Project owner: Maurice van Keulen (Original source of the data: Rudy Oude Vrielink)
Primary topic: DPV

Every year students in higher education grade the institution they are studying at. The National Students Enquiry (NSE) shows that many higher education institutions have work to do concerning their timetabling. The combined results give us information about the opinion on many topics. Unfortunately, the timetabling is one of the processes that are hard to improve but is considered below standard in Twente, being at both the UT and Saxion. You may consult the Elsevier's survey, (in Dutch), at: http://onderzoek.elsevier.nl/onderzoek/beste-studies-2015/17/universiteit-twente-enschede/1152. Although the results are not unsatisfactory, we think that we should spend effort to obtain improvement. You will dive into the results of timetabling and advise us how to do that.

The project has the following deliverables to be submitted on Blackboard.

1. Slides of your presentation
2. Report
3. Any source files or intermediary data files

1.2 Description of data set

You have the data, consisting of all timetables from all educational programmes of several years, from both organisations.

Note that the data may be considered incomplete in some ways. This is not something that we 'fix', because it is a real life situation. In daily life you almost never find complete datasets, all cleaned and ready to be explored. Nevertheless, should you need to find more data or other views on this data, please consult the

following websites: roosters.saxion.nl or rooster.utwente.nl, where you can find the source data of the given timetables.

Please note that a lot of this data contains names, mostly of teachers. Since the data is freely available on the internet, there is no issue with privacy. Nevertheless, it is considered wise and decent to not spread this data outside our institutions. Please do not share, forward, mail, copy or in any way distribute this data, because it is meant solely for the purpose of learning and advising.

Blackboard contains a lot of files concerning the timetables of Saxion and UT. The data files are called "ActivitiesUT_yyyy1-yyyy2_v2.xlsx", with yyyy1 being the first, and yyyy2 the last year of the file.

1.3 Description of challenge

The UT and Saxion are both planning to start a programme in the field of education logistics. This programme is set up as a series of interrelated projects in an action design research method. This means: small steps, starting with scientific research with pilots, giving the results as advice to the organisation before they make decisions, then setting up the project based on the advice followed by the implementation. Results after implementation go back to being researched, and the next step can be taken. The research you are about to be doing, is one of those steps.

You have the data, consisting of all timetables from all educational programmes of several years, from both organisations. These data can be researched in order to advise the organisation. Each group has 2 tasks in this project. One is to check the compliance of the timetables to the given performance indicators (KPIs). The second is to explore the timetables and find patterns or other facts standing out.

Please remember that the goal of this project is to give your advise to the management. What is your interpretation and what should UT and/or Saxion do, after your study of the facts and figures?

1.3.1 Strategy 1: Compliance

Given the set of Key Performance Indicators (KPIs) per institution, check to what extend the timetables comply with their own set. In case you interpret the given sets of KPIs as a 'wish list' more than as a list of measurable indicators, you may feel free to translate the set of KPIs to more measurable ones. See to what extend the timetables comply with the KPIs, but also look at the other rules and preferences. To what extend do the timetables follow those? Do this for every rule, every preference and every KPI for both institutions. What can you tell? (Congratulations! You have just done what almost no higher education institution has ever done before!)

1.3.2 Strategy 2: Exploration

This task is about exploring the data and looking for pattern that nobody has seen before. Examples of questions are:

- How far do students and teachers have to walk on campus each day?
- How much waste (free hours in between classes) do teachers have?
- How many free hours do teachers and students have, on average? Who has the most?
- How many contact hours do teachers and students have? Which course has the most?
- What is the maximum used capacity for each building? For each course/programme? For each faculty? What is the minimum?
- How many inconsistencies do you see?

- How much do teachers or students have to walk around campus to follow courses? Who is leading with the most kilometres per day/week/module?
- How many hours does each course schedule, on average, at minimum, at maximum?
- How many students are in every course? Per week/module?
- Can you compare the contact hours of the 2nd year in 2013/14 (non-TOM education) to 2014/15 (TOM education)? How much is the increase in contact hours due to the introduction of TOM?
- Do the same for 1st years.
- What teacher teaches with most other teachers?

Write down your findings in a report directed to the management.

1.3.3 Strategy 3: Trend analysis

Both Strategy 1: Compliance, and Strategy 2: Exploration, do not indicate which persons, courses, are important *now*. In this strategy, we will analyse the data as time series data. Divide the data in sections of a module, a quarter, a semester, or a year, ranging from September 2013 to July 2016. What can you tell about the trends that you see? Also compare your data between both institutions, UT and Saxion. Examples of analyses that you might do are:

- Plot how many courses are given in each time span;
- Determine the correlation using regression analysis or Pearson's correlation;
- What lecture halls gained importance over the years? (i.e. what are trending halls?)
- What lecture halls show decreased importance over the years? (i.e. what are nostalgic halls?)
- What teachers / faculties / gained importance over the years?
- etc.

1.4 Tips and suggestions

1.4.1 Creating new geographic maps for use in Tableau

If you'd want to show data using a geographic map and the available maps in Tableau do not suffice, then you can also import your own from *shapefiles* using a Geographic Information System (GIS) such as QGIS or ArcGIS. A shapefile is a special file that stores the polygones that make up a map.

See http://onlinehelp.tableau.com/v10.1/pro/desktop/en-us/help.htm#maps_shapefiles.html%3FTocPath%3DDesign%2520Views%2520and%2520Analyze%2520Data%7CBuild%2520and%2520Use%2520Maps%7C_____1 for details.

Project 2

Project 2: Business Intelligence [BI]

2.1 Introduction

Project owner: Chintan Amrit
Primary topic: DPV

There are several data sets available the contain data about sales and other business aspects of several businesses. The challenge of this project is to come up with 4 interesting business questions based on the Balanced Scorecard perspectives that can be answered with the particular dataset. The balanced scorecard perspectives are:

1. Financial: How do we look to our Shareholders?
2. Customer: How do our Customers See Us?
3. Internal Business Process: What should we do that is Excellent?
4. Employee and Organization Innovation and Learning: Can we continue to Improve and Add Value?

The project then follows the same steps as done in the DPV topic: design a star or snowflake schema using the multidimensional modeling approach, prepare the data (extract, transform and clean) using the ETL approachand store it in a DBMS, and design and realize a visualization.

The deliverables are:

- The report explaining the design decisions of each step. Pictures of the design of the star or snowflake schema, the design of the ETL flow, and the design of the visualization / dashboard should be included in the report (screenshots are also allowed).
- Source files of (1) database schema, (2) ETL flow, and (3) dashboard.

2.2 Description of data set

2.3 Description of challenge

2.3.1 Task 1: Business Questions and Multidimensional Modeling

Please choose **one** of the datasets available on Blackboard. You need to come up with 4 interesting business questions based on the Balanced Scorecard perspectives that can be answered with the particular dataset. *Note: Please come up with non-trivial business questions that requires a bit of thought in the datawarehouse modeling. For understanding what makes a good business problem(s), see the paper on Balanced Scorecard on Blackboard.* The balanced scorecard perspectives are:

1. Financial: How do we look to our Shareholders?
2. Customer: How do our Customers See Us?
3. Internal Business Process: What should we do that is Excellent?
4. Employee and Organization Innovation and Learning: Can we continue to Improve and Add Value?

Based on these business questions, think of Key Performance Indicators (KPIs) as well as metrics that you think best represent a solution to the business problem. Then design a Star or Snowflake schema (OLAP model) to answer the business questions.

2.3.2 Task 2: ETL

1. In this task you need to first create the data warehouse in MySQL/PostgreSQL for the chosen dataset using the OLAP model you made in the previous task.
2. Prepare and store the data in this database.

2.3.3 Task 3: Visualisation

In this assignment you address the four business problems raised in Task 1 by creating a dashboard with a visualization based on the metrics for the required KPIs. The steps for this task are as follows:

1. Connect to the data in the MySQL datawarehouse from Tableau.
2. Visualize the metrics using Tableau by creating a dashboard with appropriate visualisations to measure the KPIs.

2.4 Tips and suggestions

2.4.1 Hint 1

Hint for parts 1 and 2 if using Sakila dataset:
`http://www.percona.com/live/mysql-conference-2012/sessions/starring-sakila-building-data-warehouses-and-bi-solutions-using-mysql-and-pentaho`

2.4.2 Hint 2

It may happen in your project that there is not enough data in the available data set. For example, you want to show a trend over several years, but there is only one years worth of data. In that case you are allowed to artificially add more data to your data set by generating it randomly. See
`http://kedar.nitty-witty.com/blog/generate-random-test-data-for-mysql-using-routines`

If you experience issues running the `populate_fk` function provided on the page referred above, please use the file `generate_random_data.sql` attached. You can use it by running the sql statement and then calling for example: `call populate_fk('sakila','film',10,'Y');`

Project 3

Project 3: Transport domain [TRANSPORT]

3.1 Introduction

Project owner: Luc Wismans
Primary topic: DPV or SEMI
Possibly combinable with TS as the data can be viewed as a time series.

Transportation is about moving of people and goods from A to B. Being able to transport people and goods is a prerequisite for economic growth and the consequence of the separation of production and consumption. There are various means of transportation (e.g. car, train or bicycle) being serviced by private companies or public authorities. Although transportation brings utility it also comes with a cost. Unwanted side effects (i.e. externalities) are caused by transportation affecting for instance the air quality, climate, safety and noise. Furthermore, there is a difference between the user needs and resulting behaviour and the societal needs and desired behaviour. Simple examples show that individuals pursuing their own objectives (e.g. shortest travel time from A to B) does not result in the optimal situation for society as a whole (e.g. minimal total delay in the system).

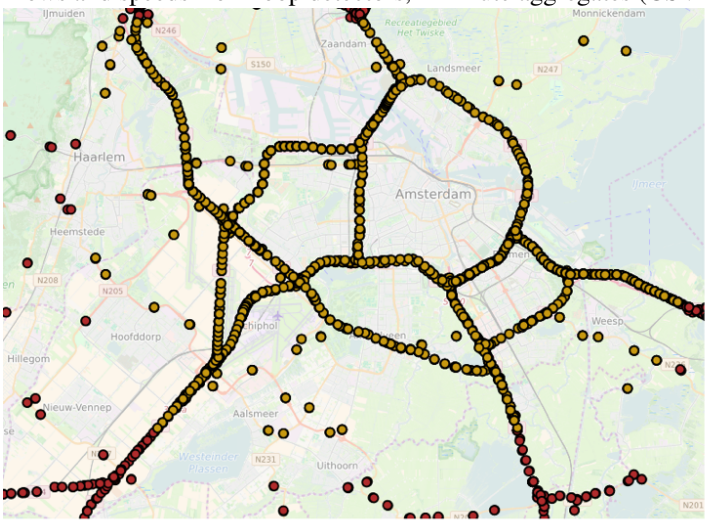
Road authorities are always working on improving the transport system balancing the societal objectives related to economic growth, minimizing externalities and user needs (i.e. sustainability). For this purpose they can adapt the system taking hard measures (infrastructural changes including deployment of intelligent transport systems) or influence the system providing services like traveller information. In most cases the infrastructure is owned by and a responsibility for governmental authorities as were the services provided on these networks. In the case of services these were at least controlled by the government (e.g. public transportation and provision of information). However, the past few years there is a shift of services provided by private parties not only because governmental authorities allow them to do so (providing data to such parties), but also as a result of an increase in data availability as well as ICT technologies not necessarily for transportation purposes deployed by private companies. Loop detector data, GPS, GSM, Bluetooth, WiFi, camera, smart card data, AVL and dedicated smartphone apps are examples of the sources capable of providing data of interest for transportation. Accurate maps/ topology of networks are needed to be able to map these data sources, offering the opportunity to connect and interpret this data for transportation purposes. Other spatial and temporal types of data like, socio economic data, points of interests, weather, deployment of measures, time tables and lines of PT might be of interest because of correlations with traffic conditions.

This data is obviously of interest for governmental authorities and private parties, because it allows to improve the decision support information for themselves or their customers. The enormous increase of data availability opens opportunities to better understand the current transport system (e.g. what are the traffic conditions, where are problems and when do they occur, and how do traffic conditions change as a result of construction works), to monitor the transport system (e.g. route choice effects of measures taken), as well as to improve predictions of the future (e.g. what will be the traffic conditions in the coming hour, what will be the travel time from A to B tomorrow during rush hour, etc.). Furthermore, it is useful to have some knowledge on GIS software packages like the open software QGis.

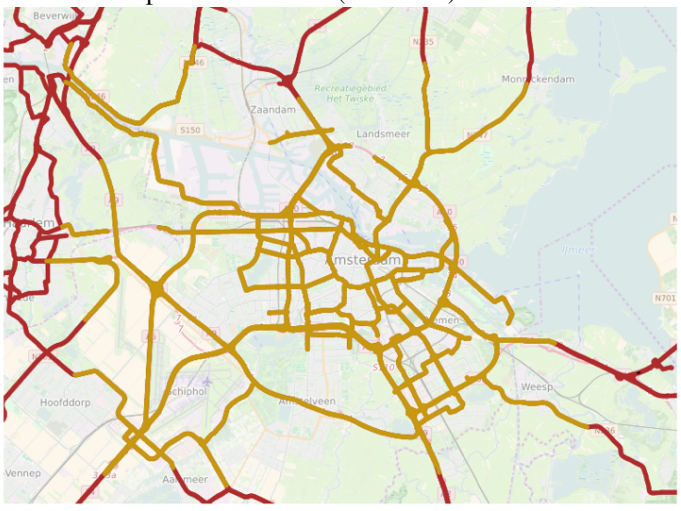
3.2 Description of data set

For this domain there are several data sources available. All data from NDW and BE-mobile is for the same time period June 1st 2016 till July 18th 2016 (for (part of the) Amsterdam region.

1. Data delivered by NDW containing
   - Flows and speeds from loop detectors, 1 minute aggregates (CSV-files)



   - Travel times of predefined routes (CSV-files)



   - Status information on occurrence, whole of Netherlands (XML),
     – traffic measures,

  - opening bridges,
  - road works,
  - traffic jams
  - traffic reports
  - status opening rush hour lanes

2. Data delivered by BE-mobile

   - Tripdata within bounding box part of A10 west
   - Important: Use of this data requires you to sign an agreement regarding the use.

3. CDR data Estonia

   - Outbound Call Detail Records of Estonians in 7 other countries (roaming data). Already connected to mast locations using OpenCellid.

3.2.1 Description NDW data speed, flow and traveltime

Extensive descriptions (in Dutch) of data can be found in

- NDWInterfacebeschrijvingversie2.2 (1).pdf
- handleiding.pdf

Datasets:

1. Datasets containing all measurements:
    - Filename speeds: utwente snelheden groot amsterdam.zip
    - Filename flows: utwente intensiteiten groot amsterdam.zip
    - Filename travel times: utwente reistijden groot amsterdam _20160916T115957_197.zip
2. Datasets containing metadata:
    - Filename speeds: utwente snelheden groot amsterdam 1 dag met metadata_20160916T105028_197.zip
    - Filename flows: utwente intensiteiten groot amsterdam 1 dag met metadata (2)_20160916T104708_197.zip
    - Filename travel times: utwente reistijden groot amsterdam 1 dag met metadata_20160916T103803_197.zip

Files within zip-files: Complete dataset is separated in several files Csv-files: "," separates fields

The raw data used by NDW are 1-minute aggregates. This means that the available datasets contain the raw data of NDW. As a result several fields are not filled, because these are used when higher aggregates would be delivered by NDW (e.g. numberOfInputValuesused, standerdDeviation and dataError).

The datasets containing all measurements need to be combined with the datasets containing metadata to add locational data (i.e. to be able to connect the measurements to the location of measurement). The metadata contains the data of 1 day containing all measurement locations (and additionally more information of possible interest, like lane number and lat lon position of measurement) and the measurements for one day, the dataset containing all measurements contain the measurements of all days between June 1st 2016 till July 18th 2016. Furthermore, if there is data available per lane and/or data per vehicle class (e.g. cars and trucks), these measurements for the same location and same minute will be presented in a separate line. For this purpose you also need to combine the complete dataset with the metadata file.

Important attribute fields are:

- MeasurementSiteReference and index: to connect the metadata information to all measurements
- PeriodStart and periodEnd indicating the timeinterval of measurement
- avgVehicleFlow: measurement of flow
- avgVehicleSpeed: measurement of speed
- avgTravelTime: measurement of traveltime
- measurementSide: side of road i.e. eastbound or westbound
- specificLane: lane nr, number starts left
- specificVehicleCharacteristics: vehicle class
- startLocatieForDisplayLat and startLocatieForDisplayLong: lat-lon location of measurement (or for travel time starting point) based on ETRS89 system, which is the same as WSG84
- generatedSiteName: description of location (unfortunately in Dutch)
- Specific for traveltime
- computationMethod: method used to compute the average measurement
- measurementEquipmentUsed: sensor used to measure, Dutch description (e.g. loop detector in Dutch "Lus")
- startLocatieForDisplayLat and startLocatieForDisplayLong: lat-lon location of starting point section of measurement
- eindLocatieForDisplayLat and eindLocatieForDisplayLong: lat-lon location of end point section of measurement
- lengthAffected: length of section

3.2.2 Description Status information

Available information on data description can be found in DATEX-II Dutch Profile 2015-2a (NP2015-2a).pdf, which is in English.

Datasets

- ActualTraffic.zip
- Bridge.zip
- ONDA.zip
- RoadMaintenance.zip
- SRTI.zip
- TrafficInfo.zip

Datasets contain data for days between June 1st 2016 till July 18th 2016. However, in this case the available data for the Dutch network.

3.2.3 Description of BE-mobile data

The Be-mobile data contains trip data of individual vehicles on the Amsterdam region network for all days between June 1st 2016 till July 18th 2016. Note that these are floating car data of a sample of vehicles equipped with devices providing these information. Further note that there is a possible bias in this sample (i.e. it is not a random sample of all vehicles driving within the Amsterdam region network).

Two datasets:

1. Trip data:
    - Archive 1.zip
    - Archive 2.zip
    - Archive 3.zip
2. SegmentId information
    - staticDataFiltered.xlsx

Files within the zipfiles are numbered, these numbers do not have any meaning. The csv files contain the following information:

- anonimized vehicle ID
- time stamp: YYYYMMDDHHMMSS
- segmentID
- traveltime (ms)
- covered distance (mm), not neceserally the same as segment length. Gps positions were mapped on segments and between two positions the route is determined. End of route can be placed at certain position on a segment.

staticDataFiltered.xlsx contains background information on segments and contains following information:

- SegmentID
- BeginLongitude and BeginLatidude: lat lon position starting point segment
- EndLongitude and EndLatitude: lat lon position end point segment
- OptimalTTMs: freeflow travel time segment in ms
- Lengthmm: length of segment in mm

3.2.4 Description of CDR data

Filename: 1week_outbound_data_extended.csv

Attributes:

- pos_time: date and time measurement
- usr_id: unique user identification number,
- mcc: country code
- lac: locational area code

- cell_id: identification number of mast (based on openCellid.org),
- lon, lat: location of measurement based on location mast according to openCellid.org
- type: reason of connection with mast:
  - HDR (Header Record, all types)
  - MOC (Mobile Originating Call, outgoing mobile call)
  - MTC (Mobile Terminating Call, incoming mobile call)
  - SMMO (Mobile Originating SMS Event, outgoing SMS)
  - SMMT (Mobile Terminating SMS Event, incoming SMS)
  - Data (GPRS/UMTS event, outgoing data session)
  - TRL (Trailer Record, all types)

3.3 Description of challenge

We provide you with four example challenges you could work on in your project.

1. State estimation The data provided for Amsterdam for the same time period are different sources which can be used to estimate the traffic conditions (speeds and flows) on the entire network. NDW data provides speed and flow measurements on locations based on all passing vehicles. NDW also provides travel times on routes and the Be-mobile floating car data provides spatio-temporal information for all locations based on a proxy of all vehicles. The challenge is to use this data to provide the spatio-temporal state estimates (e.g. speeds and flows for segments for every minute) as complete as possible. Also other state variables like routechoice/turnfractions or demand can be of interest.

2. Prediction The data provided for Amsterdam can be used to build a prediction module which predicts future (can be the next minutes or the next day) traffic states or travel times

3. Influence of roadworks on traffic conditions The provided NDW data also contains information on for example traffic measures, road works and bridge openings. This challenge is about analysing the impact of these aspects on the traffic conditions. Are there correlations and can we derive knowledge which we can use for future decision.

4. Reconstructing routes The outbound CDR records of Estonians can be used to analyse their trips. The challenge is to reconstruct the trips and routes of people. Is it possible to derive the mode of transport?

For all challenges it is required to connect the available data with a network and provide visualizations. First steps could be:

- Analysing and understanding of data set, e.g. by making figures of measurements for a specific location or specific user for a day or multiple days, computing averages, checking plausibility, determine whether there is data missing, etc
- Visualize locations of measurements
- Select suitable part of network for case study and determine what data is available for this part and which is not
- Connect data with a road network, visualizing locations
- Determine desired outcome and possible ways to compute this
- ...

3.4 Tips and suggestions

3.4.1 Creating new geographic maps for use in Tableau

If you'd want to show data using a geographic map and the available maps in Tableau do not suffice, then you can also import your own from *shapefiles* using a Geographic Information System (GIS) such as QGIS or ArcGIS. A shapefile is a special file that stores the polygones that make up a map.

See http://onlinehelp.tableau.com/v10.1/pro/desktop/en-us/help.htm#maps_shapefiles.html%3FTocPath%3DDesign%2520Views%2520and%2520Analyze%2520Data%7CBuild%2520and%2520Use%2520Maps%7C_____1 for details.

Project 4

Project 4: Predicting neurological outcome in patients with a severe postanoxic encephalopathy [EEG]

4.1 Introduction

Project owner: Michel J.A.M. van Putten
Primary topic: DM

Each year, about 7000 patients with a postanoxic coma after a cardiac arrest are admitted to the Intensive Care Unit. Early prediction of neurological outcome is highly relevant, not only for the treating physicians, but also for family members. This can prevent futile treatment, but will also assist in providing care for those with a high probability of good recovery.

We and others have shown that early recording of the electroencephalogram (EEG) allows reliable prediction of both poor and good outcome in a significant percentage of patients (about 50-60%). While these recordings are typically assessed by visual analysis, machine learning may assist or even outperform human visual assessment.

We provide you with a dataset with various quantitative EEG features and the neurological outcome, the Cerebral Performance Category Score (CPC).

Deliverables

- key result: ROC curves for poor and good outcome, including confidence intervals
- report
- presentation

References

The description of the data set contains references to several papers where more details are given on the specific features. Furthermore, these two papers are recommended: [**?**, **?**]. As a general text on machine

learning, the project owner recommended

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. http://doi.org/10.1023/A:1010933404324

4.2 Description of data set

The data is contained in two excel files, featuresNEW_12hrs.xls (corresponding to 12 hours after CA) and featuresNEa_24hrs.xls (corresponding to 24 hours after CA) from several patients. Some patients can have both 12hrs and 24hrs EEG.

In excel files, column 1 corresponds to Neurological outcome (or `$Patient Outcome`) which is the CPC score grouped into two categories (binary): good (CPC scores [1-2] denoted as "1") and poor (CPC scores [3-5] denoted as "0").

Columns 2 to 45 corresponds to different features extracted from the EEG. In addition to standard features, we also extract amplitude and frequency modulation features from EEG. The HT is a popularly used tool in the field of neuroscience that provides an automatic method for separating the signal spectrum into amplitude modulation (AM) and frequency modulation (FM) components [**?**]. Given an input EEG signal $x[n]$, the AM and FM can be obtained as,

$$\begin{aligned} AM &= w[n,n]\,|z[n]| \\ FM &= \frac{1}{2\pi} w[n,n] \frac{d\angle z[n]}{dn} \end{aligned}$$

where $z[n]$ is the analytic associate of the signal $x[n]$, and $w[n,n]$ is 2D Hamming window of duration H_t (4s) and bandwidth H_f seconds (4s).

Below is the description of features:

- **Time domain features** [**?**]:
  'Nonlinear energy', 'Activity' ,'Mobility', 'Complexity', 'RMS Amplitude', 'kurtosis', 'skewness'
  AM - 'meanAM' (mean AM), 'stdAM' (standard deviation of AM), 'SkewAM' (Skewness of AM), 'KurtAM' (Kurtosis of AM)
  'BSR' - Burst suppression ratio defined as

$$BSR = \frac{\text{duration of EEG in suppression state (amplitude} \leq 5\mu V)}{\text{total duration of EEG}} \times 100$$

- **Frequency domain features** (obtained using standard Fourier transform):
  Power in subband: 'delta'(0.5-4 Hz), 'theta'(4-8 Hz), 'alpha'(8-12 Hz), 'spindle' (12-16 Hz), 'beta'(16-32 Hz), 'total' (0.5-32Hz)
  Corresponding normalized power (normalized with total spectral power): 'delta_tot' (delta/total), 'theta_tot' (theta/total), 'alpha_tot' (alpha/total), 'spindle_tot' (spindle/total), 'beta_tot' (beta/total)
  Corresponding normalized power (normalized with delta spectral power): 'alpha_delta' (alpha/delta), 'theta_delta' (theta/delta), 'spindle_delta' (spindle/delta), 'beta_delta' (bata/delta)
  Corresponding normalized power (normalized with theta spectral power): 'alpha_theta' (alpha/theta), 'spindle_theta' (spindle/theta), 'beta_theta' (beta/theta)
  FM: 'fhtife1' (mean FM), 'fhtife2' (standard deviation of FM), 'fhtife3' (skewness of FM), 'fhtife4' (Kurtosis of FM)
  'sef' (spectral edge frequency) [**?**, **?**], 'df' (peak frequency)

- **Entropy domain features** [**?**, **?**]:
  'svd_ent' (Singular value decomposition entropy), 'H_spec' (spectral entropy) [**?**], 'SE' (State entropy) [**?**], 'saen' (sample entropy) [**?**], 'abs(renyi)' (Renyi entropy) [**?**], 'abs(shan)' (Shannon entropy) [**?**], 'perm_entr' (permutation entropy) [**?**], 'FD' (fractal dimension) [**?**]

4.3 Description of challenge

We would like to improve on our earlier estimates, about 50%, in prediction accuracy, both for good outcome (CPC=1 or 2) and poor outcome (CPC =3,4 or 5). For poor outcome, this should be reached at a specificity of 100% and for a good outcome for specificity of 95% or better.

4.4 Tips and suggestions

You can use various classifiers, e.g. SVM, decision trees, random forests. Results must be shown as ROC curves, including confidence intervals. Apply techniques that provide information about which features are most relevant for the prediction.

Note that for most patients, data from various hours after arrest is available. Start with using data for the same hours after arrest, e.g 12h or 24h.

How does the prediction change if you use different hours after arrest? try to explain from a neurophysiological and biological perspective.

Project 5

Project 5: Text classification or Named Entity Recognition [TCNER]

5.1 Introduction

Project owner: Maurice van Keulen
Primary topic: IENLP
Well combinable with DM or SEMI.

In the realm of information extraction and natural language processing, there are two suggested projects to select from

- **Text Classification**
  Recommend a conference to a researcher given the title of his new article
- **Named Entity Recognition**
  Extract and classify named entities from tweets

In case that a group wants to suggest a different project, the group should submit an initial project proposal to be reviewed. In this case the group is encouraged to meet with the project owner / topic teacher to discuss project ideas.

The suggested projects represent two challenges. For each challenge, you will be given a training set to train and tune your system on. The test set will be given to you one week before the presentation date. You should apply your system on the test set and send the results to the instructor a week before the presentation date. Final results of all groups on the test set will be presented by the instructor on the day of the presentations.

Note that the project grading is not related to the achieved results. However, the top performing system in each of the suggested projects will be rewarded with a symbolic gift.

5.2 Description of data set

The data for *Text Classification* is Conference Proceedings training data from the paper below.

Reference: Machine Learning in Automated Text Categorization http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf.

The data for *Named Entity Recognition* is NER Twitter training data from the paper below.

References: (1) A survey of named entity recognition and classification. http://brown.cl.uni-heidelberg.de/~sourjiko/NER_Literatur/survey.pdf. (2) Named Entity Recognition: A Literature Survey http://www.cfilt.iitb.ac.in/resources/surveys/rahul-ner-survey.pdf.

5.3 Description of challenge

5.3.1 Text Classification

- Design and implement a system to recommend a conference to a researcher given the title of his new article.
- The system should use the provided Conference Proceedings training data. You should implement the sub tasks (feature extraction, dimensionality reduction, and classifier) by yourself.
- You are free to select the algorithms you prefer for each sub task. However it is recommended that you test and compare multiple methods.
- Evaluate you system on the training set by using the cross-validation approach. Provide the confusion matrix of your system output.
- Evaluation should be done in terms of Micro-average precision, recall and F1 measures.
- Prior to the final presentation, you will be given a test set. Use your system to classify the given article titles into the set of the pre-specified conferences.
- Your submitted results should be a text file where each line is formatted as follows:
  <article id>Tab Separation <the recommended class textgreater

5.3.2 Named Entity Recognition

- Design and implement a system to extract and classify named entities in tweets.
- The system should use the provided NER Twitter training data. You should implement the sub tasks (feature extraction, and classifier) by yourself.
- You are free to select the algorithms you prefer for each sub task. However it is recommended that you test and compare multiple methods.
- Evaluate you system on the training set by using the cross-validation approach.
- Evaluation should be done in terms of micro-average of precision, recall and F1 measures.
- Prior to the final presentation, you will be given a test set. Use your system to extract and classify named entities.
- Your submitted results should be a text file where each line is formatted as follows:
  <tweet id>Tab Separation <list of entity-type/mention pairs separated by semicolons ordered by their appearance in the tweet text>

5.4 Tips and suggestions

Project 6

Project 6: Automatic detection of Atrial fibrillation (AF) episodes [AF]

6.1 Introduction

Project owner: Mannes Poel
Primary topic: DM or TS

Atrial fibrillation (AF) occurs as a complication postoperatively form cardiac surgery. AF results in stasis of the blood. In the postoperative period AF can induce delirium and neurocognitive decline, thereby prolonging the hospital stay. [1] On the long term serious complications like thromboembolic diseases, stroke and heart failure can be induced by AF. These complications result in increased morbidity and mortality and prolonged hospital stays. [2-7] Precise ECG monitoring is important to detect AF as soon as possible. Then complications caused by AF can be obviated due to a fast intervention. The challenge of this project it is to develop an algorithm/method that can detect automatically episodes of AF (minimum of 30 seconds) from (preprocessed) ECG data. Framing it differently, the research questions is: "To what extent can one automatically detect episodes of AF?"

6.1.1 Background

Manual detection of AF in ECG record is time-consuming, especially in the case of large datasets consisting of 24-hour ECGs. When automating the detection, the physician can be deprived of work and research can be accelerated. Also, such an algorithm may result in the direct detection of AF during ECG monitoring, thereby creating the possibility for a fast treatment of AF. This underlines the need for an algorithm to automatically detect AF for analysis purposes. Automatic AF detection provides a faster analysis of long term ECGs. Hereby opportunities arise for better diagnostics and for gaining more insight into postoperative AF on a larger scale. Automatic quantification of AF may help to get insight in the yet unsolved underlying problem of AF.

AF is defined as a period of at least 30 seconds in which an irregular ventricular rate and P peaks are absent. [8] These two ECG characteristics indicate the rapid abnormal atrial activity seen in AF. An AF detection algorithm based on R-R interval irregularity is preferred, due to the prominence of QRS complexes, making it more robust to noise. [9] Therefore this algorithm is also based on the R-R interval irregularity.

6.2 Description of data set

In this project you will work with preprocessed ECG data from the Erasmus Medical Centre in Rotterdam of the department electrophysiology. Data was obtained within 10 days post-operatively of CABG surgery. Atrial fibrillation (an arrhythmia) occurs frequently after cardiac surgery. Atrial fibrillation is an arrhythmia that does not have to sustain constantly, it comes and goes, and therefore often passes by unnoticed in the patient's hospital stay.

ECG

From the patient a 12 lead ECG is recorded. A semi-automatic program analyzed the ECGs for annotation of the R peaks (see green dots in Figure 6.1). R peak detection was manually audited by a physician and atrial fibrillation or other arrhythmias were labeled.

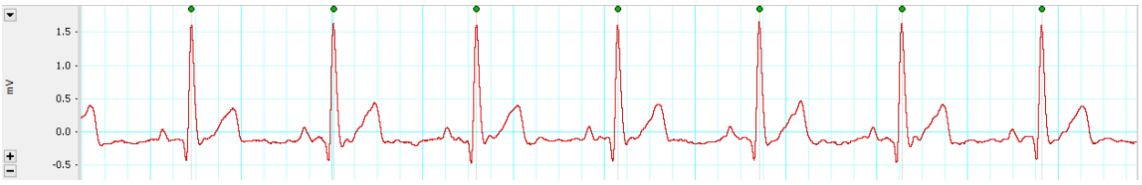


Figure 6.1: Example ECG. The green dots annotate the R peaks

From this a text file was formed (Figure 6.2) with the time, R-R intervals in milliseconds, and more details on the observed rhythm.

Preprocessing

The algorithm you will build will be based on the R-R intervals, and more specifically on the irregularity of these intervals. This data has been preprocessed for you to make it easier to work with.

1. First of all the array of R-R intervals was split in arrays covering periods of 30 seconds. Each of this we will call a sample.
2. All samples for which no control could be made (due to artifacts, loss of server contact, limited data, etc.) were excluded.
3. All samples that included unphysiological high R-R intervals were excluded.
4. 30 bins of 50 milliseconds were created covering R-R intervals of 200 ms up to 1700 ms. For each sample the frequency of an R-R interval occurring in a certain bin was counted. See Figure 6.3.
5. These frequencies of occurring were then normalized.
6. All samples used were shuffled.

Excel file

In the excel file placed on blackboard is the data of 150.000 periods of half a minute of different patients. Around 36.000 of these periods are labeled as AF. Each row represents a sample. The first 30 columns show the values as explained above (normalized frequency of R-R interval in 30 bins). The 31st column is the label. Zero (0) means no AF was reported in this sample by the physician. One (1) means AF was reported by the physician. For students who want to investigate the raw data, this data can be made available on request.

[1] Alqahtani AA. Atrial fibrillation post cardiac surgery trends toward management. Heart Views. 2010 Apr 1;11(2):57.

```
23:56:21        845     N
23:56:22        845     N
23:56:23        855     N
23:56:24        845     N
23:56:25        840     N
23:56:26        840     N
23:56:26        845     N
23:56:27        830     N
23:56:28        825     N
23:56:29        825     N
23:56:30        835     N
23:56:31        815     N
23:56:31        815     N
23:56:32        820     N
23:56:33        830     N
23:56:34        830     N
23:56:35        835     N
23:56:35        835     N
23:56:36        840     N
23:56:37        850     N
23:56:38        850     N
23:56:39        835     N
23:56:40        830     N
```

Figure 6.2: Example of text file that is the result of the processed ECG. In the first column record time, second column R-R interval in milliseconds and in the last columns annotations regarding the rhythm have been made.

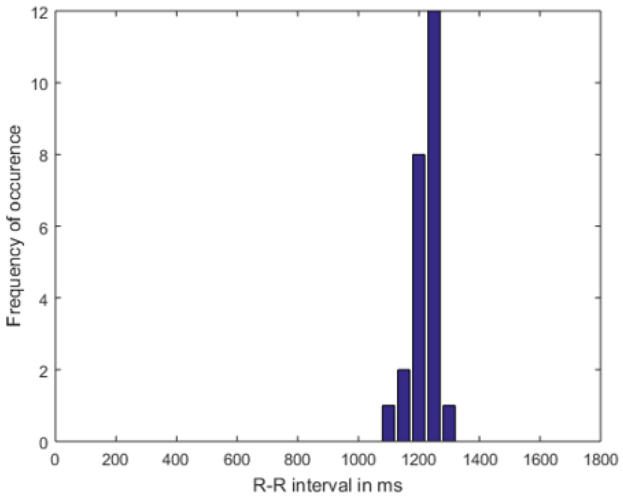


Figure 6.3: This is an example of how the data would look like visualized after step 4. Here you see that within this 30 second episode 24 R-peaks were detected mostly with an R-R interval around 1200 ms

[2] W.H. Maisel, J.D. Rawn, and W.G.Stevenson. Atrial Fibrillation after Cardiac Surgery; Review. Annals of Internal Medicine, 135(12):1061–1073, 2001.

[3] N. Echahidi, P. Pibarot, G. O'Hara, and P. Mathieu. Mechanisms, prevention, and treatment of atrial fibrillation after cardiac surgery. Journal of the American College of Cardiology, 51(8):793–801, February 2008.

[4] N.S. Peters, R.J. Schilling, P. Kanagaratnam, and V. Markides. Atrial fibrillation: strategies to control, combat, and cure. Lancet, 359(9306):593–603, February 2002.

[5] S.M. Narayan, M.E. Cain, and J.M. Smith. Atrial fibrillation. Lancet, 350(9082):943–50, September 1997.

[6] S.S. Chugh, R. Havmoeller, K. Narayanan, D. Singh, M. Rienstra, E.J. Benjamin, R.F. Gillum, Y.H. Kim, J.H. McAnulty, Z.J. Zheng, M.H. Forouzanfar, M. Naghavi, G. Mensah, M. Ezzati, and C.J.L. Murray. Worldwide epidemiology of atrial fibrillation: A global burden of disease 2010 study. Circulation, 129(8):837–847, 2014.

[7] M.P. Turakhia, M.D. Solomon, M. Jhaveri, P. Davis, M.R. Eber, R. Conrad, N. Summers, and D. Lakdawalla. Burden, timing, and relationship of cardiovascular hospitalization to mortality among Medicare beneficiaries with newly diagnosed atrial fibrillation. American Heart Journal, 166(3):573–580, 2013.

[8] V. Fuster, L.E. Ryden, R.W. Asinger, D.S. Cannom, H.J. Crijns, R.L. Frye, J.L. Halperin, and et al. ACC/AHA/ESC Guidelines for the Management of Patients With Atrial Fibrillation: Executive Summary A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the European Society of Cardiology Committee, volume 104. 2001, ISBN: 8006116083.

[9] N. Larburu, T. Lopetegi, and I. Romero. Comparative study of algorithms for atrial fibrillation detection. In 2011 Computing in Cardiology, pages 265–268. IEEE, 2011

6.3 Description of challenge

The goal of the project is to answer the research question "To what extent can one automatically detect episodes of AF?"

Method.

Of course one should apply DM techniques to answer this question based on the given data. This implies that you should:

1. Design/select different DM models, such as Decision Trees, which one you would like to validate. The selection of promising models can be based on related work. Which models did others use and what was their performance. Select 3 or 4 models which you want to compare on the given data set.
2. Validate (measure the performance) of the selected models on the given data set. Of course one is interested in the performance of the models on new, unseen, data. How do you estimate such a performance, and which performance measure to use (accuracy, precision, recall). Are the results sound, i.e. what you expect?
3. One problem with the data set is that it is imbalanced, much more non AF then AF samples. Look up in the literature how to deal which such imbalanced data sets.
4. Compare the different models and select the best one. Explain the selection procedure.
5. Finally answer the research question and discuss the practical applicability of an automatic AF detector.

6.3.1 Requirements report

The final report of your case study should least cover the following topics:

- Short description of the case.
- Short summary of a literature study on the case under consideration. This can give you insight in useful methods for the DM approach.
- Remember that results must be reproducible, otherwise they are of no use. Hence give a clear conceptual description of the approach taken (methodology):
  - Description of the attributes (user profile)used and if useful attribute selection.
  - Which classification models did you use and why.
  - Design and evaluation of the selected models:
    - Assessment of the performance of your models (split training testing, number of repetitions,confusion matrix, etc.).
    - Give an estimate, including confidence intervals, of how the selected models will perform in practice.
  - Determine which model is the best; give a statistical underpinning.
  - Comparison with approaches (models) found in the literature. Explain if this comparison is possible or not. This is part of the discussion.
- Global outline of your report could be:
  - Introduction in which you explain the DM challenge and your research questions.
  - Related work (optional, can lead to bonus)
  - Methodology. A clear description of your approach.
  - Results.
  - Discussion.
  - Conclusions. In the conclusions you answer your research questions based on the results and discussion.

6.4 Tips and suggestions

Here you can find some practical hints for dealing with WEKA and the data.

- Inspect your results, for instance the decision tree. Does it contain sound features, does it make sense?
- It training takes too long you can reduce the number of folds or resample the data (Preprocessing $\rightarrow$ Filter $\rightarrow$ Resample) and afterwards validate the most promising models on the whole dataset. Be aware that you do not filter out all AF instances.
- Determine if feasible which features are most predictive and does this make sense.
- If time allows determine new features (based on related work) and validate if this increases the performance.
- If Weka runs out of memory you can increase the heapsize in the Weka configuration file called "RunWeka".

Project 7

Project 7: Linked Open Data [LOD]

7.1 Introduction

Project owner: Maurice van Keulen
Primary topic: SEMI

The project is about enriching data about cultural events:[1]

> Pick a theatre of your own choosing and demonstrate how you can enrich the data of that theater.

7.2 Description of data set

Data is not available directly, but based on information from the theater's website and established data sets and ontologies:

- Construct an RDF data set containing the information from the theatre's website
- Linking the data to established Linked Open Data sets, such as DBpedia.

7.3 Description of challenge

The added value of Semantic Web technology can be *demonstrated* by, among other things,

- Running example SPAQRL queries finding data in a way not possible with the original website's or Google's search facilities.
- Using the remote querying facilities of SPARQL to query Linked Open Data on the web.

[1] You may choose to do something else than cultural events. This has been chosen because it is a semantically rich domain. If you want to focus on another domain, please consult the topic teacher to determine if the domain is suitable enough.

Note that these are possibilities and suggestions. There actually is much freedom in this project to go your own way and to focus your attention. You need to, however, stay within the domain of cultural performances and the global goal of semantic enrichment. Minimal requirements for the project are

- It should enrich the original data with Linked Open Data on the web.
- It should demonstrate that now something non-trivial can be done.

Deliverables for the project are

- Presentation slides
- Report (PDF) explaining
  1. the specific goal of your project,
  2. the applied Semantic Web technogies,
  3. the resulting RDF data set and how it was constructed,
  4. the design of the software / system / website you developed, and
  5. examples that demonstrate non-trivial queries.
- Source files of all artifacts used and produced in the project such as data sets, web pages, software code, example queries, etc.

Please combine the above files in a ZIP-archive and upload that to Blackboard.

The grading for the project is based on the following principles

- How much *understanding* is shown
- The depth and quality of the developed components
- The power and potential of the enrichment

7.4 Tips and suggestions

Project 8

Project 8: Probabilistic Data Integration [PDI]

8.1 Introduction

Project owner: Maurice van Keulen
Primary topic: PDBDQ

This project is about probabilistically integrating semantic duplicates in a given data set about music albums.

A semantic duplicate is a set of different data items that actually refer to the same entity/object in the real world. For example, one can have two data items describing a music album with slightly different strings for the album, artist and song names, or there may be songs missing, typos, language differences, and many more possible problems.

The approach suggested here is:

- Match the data from the discs to determine a *matching score* for each pair of discs.
- Set *two thresholds* τ_b and τ_t that define three classes of matches:
  - **Non-match** any match below τ_b is for sure **not a match**
  - **Match** any match above τ_t is for sure **a match**
  - **Uncertain** for any match between τ_b and τ_t it is uncertain whether or not it refers to the same disc or not.
- Merge the matching and uncertain matching discs with the probabilistic approach taught in Probabilistic Databases and Data Quality [PDBDQ] .
- Evaluate the result using the "ground truth" data provided.
- Experiment with and evaluate any other aspects that interests you.

8.2 Description of data set

The data set about music albums is onelinercddb_discs.xml. The data is a list of discs representing music albums. The ground truth is `cddb_9763_dups.xml` where the duplicates are already indi-

cated in the data as it is a list of pairs of music albums. Source of this data can be found at http://hpi.de/naumann/projects/repeatability/datasets/cd_datasets.html

It is also possible to bring you own data. In that case, you need to pay attention to the following

- You need approval from the project owner / topic teacher.
- It needs to contain an element of probabilistic integration in it (typically there are two data source that do not perfectly match).
- It needs to contain imperfections and aspects of untrustworhtiness in it (possibly artificially created).

8.3 Description of challenge

The goal of the project is

> Write a program in a language of your own choosing that reads the discs file, probabilistically integrates the duplicates, faithfully represents any uncertainty that exists, and demonstrate the strength of doing so for a certain analytical goal.

The file with the duplicates is provided for your convenience. You can use it to first start with the merging part of the project, developing software for merging pairs of discs and representing the uncertainty resulting from merging in the resulting data. You can then focus on matching discs, developing a matching technique based on string matching of certain attributes to derive possible duplicates (these need not be precisely the same as the ones in cddb_9763_dups.xml but should be 'close'). The file with the duplicates can also be used to check and assess how well your matching strategy works.

Deliverables for the project are

- Presentation slides
- Report (PDF) explaining
  - The specific analytical goal for the integration.
  - The identified data quality problems.
  - How these data quality problems are represented in the data.
  - Which data processing actions you have developed to achieve this.
  - How you demonstrate the strength of probabilistic representation of data quality problems in your project.
- Source files of all artifacts used and produced in the project such as datalog programs, raw data files, data conversion scripts, example queries, etc.

The grading for the project is based on the following principles

- How much *understanding* is shown
- The depth and quality of the developed data processing actions
- The depth and quality of the representation of data quality problems in the resulting data.
- The depth and convincingness of the demonstration of the strengths of the method.

It doesn't matter what your main focus is or which other tools and programming languages you use; it does matter how well well you have accomplished the task you have set out to do and how much understanding you show.

8.4 Tips and suggestions

Project 9

Project 9: Referral Advice [RA]

9.1 Introduction

Project owner: Mannes Poel
Primary topic: DM

Low back pain (LBP) is the most common cause for activity limitation and has a tremendous socioeconomic impact in Western society. In primary care, LBP is commonly treated by general practitioners (GPs) and physiotherapists. In the Netherlands, patients can opt to see a physiotherapist without referral from their GP (so called 'self-referral'). Although self-referral has improved the choice of care for patients, this also requires that a patient knows exactly how to select the best next step in care for his or her situation (GP, physiotherapist or self-care), which is not always evident. We would like to automatize the referral advice (no human made decision) and want to know which features are relevant in this referral advice, since this could shorten the questionnaire.

9.2 Description of data set

The dataset contains 1288 fictive patient cases on low back pain that were judged by healthcare professionals on referral advices on a 5 points scale. The cases were constructed around 15 features (see Figure 9.1), but the expectation is that not all features are equally relevant to determine the referral advice.

9.3 Description of challenge

The research questions of this assignment are:

1. To what extend can one, based on the outcome of the questionnaire, automatically determine the 5 point score for each dimension (GP, physiotherapist, self-care)? The final referral advice is a deterministic algorithm based on the three scores and hence of no importance for the Data Mining case.

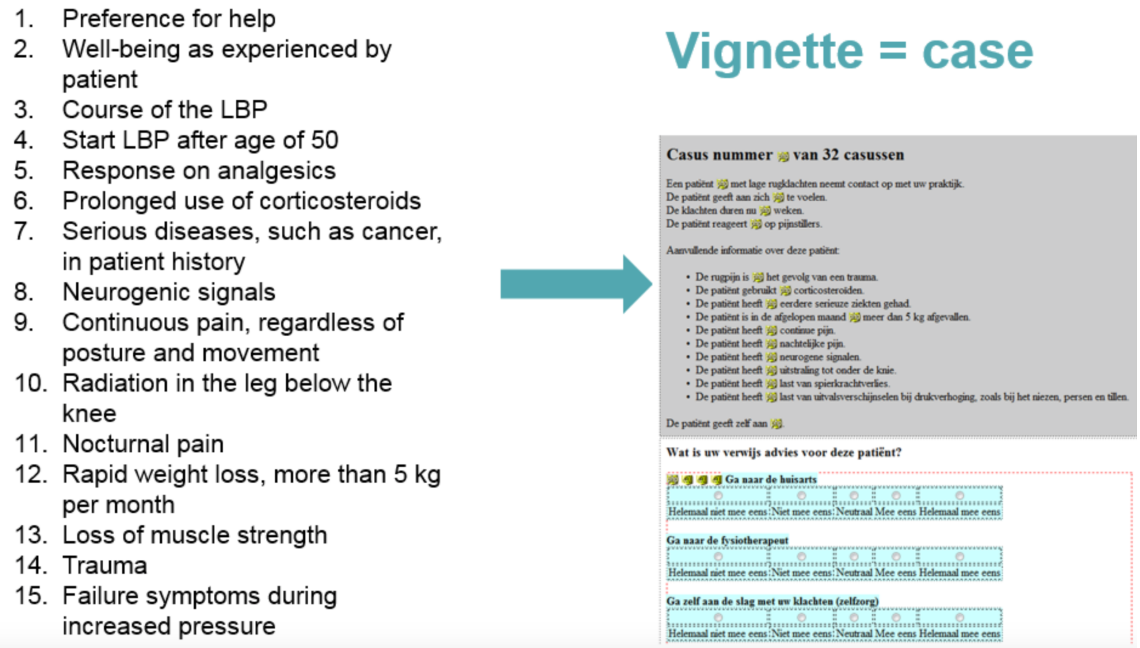


Figure 9.1: Features and sketch of case

2. To what extend can one reduce the number of questions, i.e. which questions can be left out without major decrease in performance?

9.4 Tips and suggestions

Project 10

Project 10: Displacement prediction in South Sudan [DISPRED]

10.1 Introduction

Project owner: Mannes Poel
Primary topic: DM

International migration is a complex phenomenon, and in the recent years there has been detected an increase in migration and displacement occurring due to conflict, persecution, environmental degradation and change, and a profound lack of human security and opportunity. Migration is increasingly seen as a high-priority policy issue by many governments, politicians and the broader public throughout the world. The current global estimate is that there were around 244 million international migrants in the world in 2015. The great majority of people in the world do not migrate across borders; much larger numbers migrate within countries. There are more than 65.6 million people who are forcibly displaced around the world. Out of the 65.6 million, 40.3 million people are internally displaced within the borders of their own country and 22.5 million seek safety crossing international borders, as refugees. With the increase of violent conflict and other conditions that exacerbate forced displacement, this figure is estimated to rise in the upcoming years.

This project is motivated by a request of the UNHCR, the United Nations Refugee Agency, to research the possibilities of creating a predictive engine of internal population displacement within South Sudan and its neighboring countries. The UNHCR is most interested in the most influential factors that affect the displacement of the People of Concern (POCs).

10.2 Description of data set

The starting point in the data folder is the reach report, a pdf with a long name but startine with reach_ssd. It is a report about the factors that exacerbate refugee movement - besides conflict. In page 3, there are the results of the survey they conducted, with their main findings.

It describes the pulling and pushing factors of the South Sudanese Crisis and the available relevant data sources have been placed in this folder. For further insights, or extra research you can refer to the data source at https://data.humdata.org/dataset and search for the countries of interest. We also used the same source to extract the data relevant to the South Sudan Crisis.

ACLED south sudan.csv All the conflict (both violent and non-violent) quantified per incident as collected by ACLEDdata.org in South Sudan

Refugees from South Sudan all.json UNHCR-collected. All aggregated refugees per month since the conflict started whose origin is from South Sudan. CoO (country of origin)

Given South Sudan location, there are 5 potential countries of arrivals/movement where refugees preferred to move: Uganda, Sudan, Democratic Republic of Congo (DRC), Kenya and Ethiopia.

Each country has a .json file, summarizing the arrivals to the country per month called *refugees to [country].json*

Plus, there is another .json file with the preferred state/region/location inside the country of arrival (CoA). This is called *refugees to [country] per [region/district/location].json*.

We are also adding the Kenya sex disaggregation to see patterns of movement between men and women and locations. We do not have that file for the other countries, only Kenya. It is called *refugees to Kenya per sex.json*

WFP prices food.csv
Adding the prices of food, as according to the REACH report (PDF enclosed, page 3) it is also another factor for movement, in addition to conflict.

More documents describing the situation and for further analysis can be found here (link in digital pdf).

10.3 Description of challenge

The main research questions to be addressed in this project are:

1. How can machine learning assist in predicting human displacement in the country of South Sudan?
2. What are the most influential factors that affect the displacement of the People of Concern (POCs)?

10.4 Tips and suggestions

In order to make the project more manageable one could look first at one region within South Sudan and try to model or predict the human displacement and determine the most relevant factors. And, if feasible, extend the research to more regions. The best models could differ per region and also the most relevant factors.

Project 11

Project 11: Predicting surgical case durations for a Thorax centre [PSCD]

11.1 Introduction

Project owner: Karin Groothuis-Oudshoorn

Primary topic: DPV and/or DM

In modern healthcare, organizations face the challenge of delivering more and better quality care with less human and financial resources. This is mainly due to rising demand for healthcare and increasing expenditures. Efficiency is directly linked with quality, as inefficient care processes use up valuable resources and displace more useful care. Efficiency improvements are therefore very valuable for hospitals. MST is a top-clinical medical center located in the region of Twente and is one of the biggest non-academic hospitals in the Netherlands. The medical center compromises of two inpatient clinics in Enschede and Oldenzaal and two outpatient clinics in Haaksbergen and Losser. The inpatient clinic in Enschede has moved her patients as of 2016 to the newly built location Koningsplein. This new location provides a capacity of 739 beds. Thorax Centrum Twente (TCT) is a center within MST, specializing in diagnosis and treatment of cardiothoracic diseases. Multidisciplinary medical care is delivered through several cardiothoracic-related specialties such as cardiology and cardiac surgery. TCT is one of the 16 thorax centers in the Netherlands and has grown rapidly after its establishment in September 2004. One reason for this is their short waiting list for open heart surgery, making TCT an interesting medical center for patients. TCT performs approximately 1,100 to 1,200 open-heart surgeries per year, mainly coronary and heart valve surgeries. TCT experiences a high rate of operating rooms working beyond regular operating time. High amounts of overtime result in unnecessary costs and low staff satisfaction. A recent study among Dutch hospitals suggests that more accurate predictions of surgical case duration and altering the sequencing of surgical cases on an OR-schedule can improve efficiency[1].

References

[1] van Veen-Berkx E, Elkhuizen SG, van Logten S, et al. Enhancement opportunities in operating room utilization; with a statistical appendix. J. Surg Res. 2015;194(1):43-51.

Deliverables

- Report
- Presentation

11.2 Description of data set

The dataset comprised of 4087 surgical cases performed from January 2013 to January 2016 at TCT. The surgical case duration is given in minutes. The hospital stay time and IC stay time is given in days. Unknown data is indicated with 'NULL' or 'onbekend'. The data is in the file 'surgical_case_duration.csv'. In tables 11.1-11.4 you can find descriptions of the variables in the dataset. The different levels are given. In case of type of surgery not all labels are translated (we left out the less frequent ones).

11.3 Description of challenge

The challenge of this project is to identify patterns in surgical case durations and to derive prediction models and / or classification models for surgical case duration to support OR-planners at TCT in making the most efficient OR-schedules with the available patient level data in order to decrease the overtime at TCT, while maintaining the current OR-utilisation rate.

11.4 Tips and suggestions

- Some variables (features) have a lot of categories and some of those categories have few observations. In that case it can be wise to recode these categories to e.g. 'other types' or leave those observations out. But please explain and give arguments if you do so!

Table 11.1: Description of Surgery-related variables

Variablename	Variable (English)	Categories (definition)
Operatietype	Surgery type	Aortic Valve Replacement (AVR) AVR + MVP Bentall Procedure Coronary Artery Bypass Graft (CABG) CABG + AVR CABG + MVP Epicardial LV-lead (Epicardiale LV-lead) Lobectomy or segment resection (Lobectomie of segmentresectie) Mediastinoscopy Mitral Valve Plasty (MVP) MVP + Tricuspid Valve Plasty (TVP) Mitral Valve Replacement (MVR) Nuss bar removal Nuss-procedure Refixation of the sternum (Refixatie sternum) Rethoracotomy (Rethoractomie) Removal of steel wires (Staaldraden verwijderen) VATS Boxlaesie (video assisted thoracic surgery) Wound debridement (wondtoilet) Other types
Benadering	Surgical approach	Full sternotomy (Volledige sternotomie) Left antero lateral (Antero lateraal links) Right antero lateral (Antero lateraal rechts) Left postero lateral (Postero lateraal links) Right postero lateral (Postero lateraal rechts) Partial sternotomy (Partiele sternotomie) Other approaches: Parasternaal links, Parasternaal rechts, Dwarse sternotomie, Xiphoidaal, NULL)
Chirurg	Surgeon	Surgeon 1 – 15 Other specialism (Ander specialisme)
Anesthesioloog	Anesthesiologist	3 – 19, Unknown (onbekend)
OK	Operation room	HCK1, HCK3, HCK4, OK 1, OK 10, OK 11, OK 3, OK 4, OK 5, OK 9, TOK1, TOK2, TOK3, TOK4, else (Elders)
Casustype	Urgency	Elective (planned on the elective program) (Electief) Emergency ($<$ 24 hours) (Spoed $<$ 24 uur) Acute ($<$ 30 minutes) (Acuut $<$ 30 minuten) Acute (Spoed) Acute ($<$ 5 hours) (Spoed $<$ 5 uur) Unknown (NULL)
Dagdeel	Time of day	Morning (7:00 – 12:00) Afternoon (12:00 – 18:00) Evening and night (18:00 – 7:00)
Aantal anastomosen	Amount of bypasses	Continuous variable
HLM	Cardiopulmonary bypass use	Yes (heart-lungmachine usage planned for surgery) No

Table 11.2: Description of Surgery-related variables

Variablename	Variable (English)	Categories (definition)
Leeftijd	Patient age	Continuous
Geslacht	Patient gender	Male Female
AF	Presence of atrial fibrillation	Yes (AF rhythm present) No
Chronische longziekte	Presence of chronic lung disease	Yes (long term use of bronchodilators or steriods for lung disease) No
Extracardiale vaatpathie	Presence of extracardial arteriopathy	Yes (claudication, carotid occlusion or 50% stenosis, amputation for arterial disease or previous or planned intervention on the abdominal aorta, limb arteries or carotids) No
Active endocarditis	Presence of active endocarditis	Yes (patient still on antibiotic treatment for endocarditis) No
Hypertensie	Presence of hypertension	Yes No
Pulmonale hypertensie	Presence of pulmonary hypertension	Normal (no increased pulmonary artery pressure) Moderate (pulmonary artery systolic pressure 31-55 mmHg) Severe (pulmonary artery systolic pressure $>$ 60mmHg)
Slechte mobiliteit	Presence of poor mobility	Yes (severe impairment of mobility secondary to musculoskeletal or neurological dysfunction) No
Hypercholesterolemie	Presence of hypercholesterolemia	Yes No
Perifeer vaatlijden	Presence of peripherial vascular disease	Yes No
Linker ventrikel	Left ventricle	Good (LV ejection fraction $>$50%) (Goed) Moderate (LV ejection fraction 31-50% (Matig) Poor (LV ejection fraction $\leq$ 30% (Slecht) Very poor (Heel slecht)
Nierfunctie	Renal function	Normal (creatine clearance $>$ 85 ml/min Moderate (creatine clearance 50-85 ml/min) (Matig) Poor (creatine clearance $<$ 50 ml/min or dialysis) (Slecht) Dialyse
DM	Presence of diabetes mellitus requiring insulin	Yes (diagnosis DM requiring insulin) No
Eerdere hartchirurgie	Previous heart surgery	Yes (heart surgery in the patient's history) No
Kritische preoperatieve status	Critical pre-OR state	Yes (ventricular tachycardia or ventricular fibrillation or aborted sudden death, preoperative cardiac massage, preoperative ventilation before anesthetic room, preoperative inotropes or IABP, preoperative acute renal failure (anuria or oliguria $<$10ml/hr)) No
Myocard infact $<$90 dagen	Mycordial infarction before surgery	Yes (MI within 90 days before surgery) No
Aorta chirurgie	Aortic surgery	Yes (planned surgery on the aorta) No
Euroscore1	Euroscore1	Continuous variable
Euroscore2	Euroscore II	Continuous variable

Table 11.3: Description of Surgery-related variables (Continuation)

Variablename	Variable (English)	Categories (definition)
CCS	Canadian Cardiovascular Society (CCS) score for angina	0 (no symptoms) 1 (angina only during strenuous or prolonged physical activity) 2 (slight limitation, with angina only during vigorous physical activity) 3 (symptoms with everyday living activities, i.e. moderate limitation) 4 (inability to perform any activity without angina or angina at rest, i.e. severe limitation)
NYHA	New York Heart Association (NYHA) score - dyspnea	1 (cardiac disease, but no symptoms and no limitation in ordinary physical activity, e.g. no shortness of breath when walking, climbing stairs etc.) 2 (mild symptoms (mild shortness of breath and/or angina) and slight limitation during ordinary activity) 3 (marked limitation in activity due to symptoms, even during less-than-ordinary activity, e.g., walking short distances (20-100 meters). Comfortable only at rest 4 (severe limitations, experiences symptoms even while at rest, Mostly bedbound patients).

Table 11.4: Description of Outcome variables

Variablename	Variable (English)	Categories (definition)
Geplande operatieduur	Planned surgery duration	Continuous outcome
Operatieduur	Surgery duration	Continuous outcome
Ziekenhuis ligduur	Hospital days	Continuous outcome
IC ligduur	Intensive care days	Continuous outcome

Project 12

Project 12: Web Harvesting for Smart Applications [SDSI]

12.1 Introduction

Project owner: Maurice van Keulen
Primary topic: SEMI
Good to combine with PDBDQ, DPV, DM, or IENLP

The High Tech Systems Park of Thales in Hengelo[1] is also used as a kind of laboratory, called "Fieldlab The Garden".[2] One of the projects making use of this lab is "Secure Data Sharing Innovation" (SDSI). Part of this project focuses on the development of and experimenting with *Smart Applications*.

This project is related to this activity. The idea is that with technology that can autonomously and robustly harvest data from the web, one can develop smart applications. For example, finding indications of possible unknown side effects of medicines. One could harvest all messages from a web forum for a certain disease, extract information about (a) medicines people report using and (b) which side effects they report having (topic IENLP!), and compare that with the leaflets of these medicines to determine if some reported side effects are unknown (i.e., not mentioned in the leaflet).

12.2 Description of data set

There is no given data set, but you are expected to choose a website yourself and harvest data from it. There is a requirement to use data from at least two sources, where at least one is harvested from the web. The other could, for example, also be a linked open data set (see topic SEMI). The data of web pages is stored in `XHTML` format, which is an XML-format. In other words, you can use the technology of the SEMI topic to manipulate it and to extract specific bits of data from it.

There are web harvesters / crawlers available on the web that you can use as a service. You can also write your own program that fetches pages (websites that use JavaScript to dynamically load data and construct

[1] http://hightechsystemspark.com/about-high-tech-systems-park/
[2] http://hightechsystemspark.com/smart-industry/the-garden/

a web page in the client, can be harveste by using the FireFox browser in *headless* mode, see https://developer.mozilla.org/en-US/Firefox/Headless_mode. In certain circumstances, you could resort to manually saving the individual web pages, but that obviously has its limitations.

12.3 Description of challenge

The challege is to

> Demonstrate potential of data harvested from the web for developing smart applications by (a) integrating data from at least 2 sources (at least one is harvested from the web), and (b) designing and implementing a proof of concept of a smart web/mobile app

12.4 Tips and suggestions

Project 13

Project 13: Web Harvesting of Online Healthcare Communities [WHCC]

13.1 Introduction

Project owner: Mohsen Jafari Songhori
Primary topic: SEMI
Good to combine with PDBDQ, DPV, DM, or IENLP

Nowadays, ubiquity of online communities gives us invaluable dataset in the form of electronic peer-to-peer communication, which may help us understand social influence and collective behavior dynamics. Especially, to some extent, messages exchanged in health-related online communities can reflect the intricacies of human health behavior as experienced in real time at individual, community, and societal levels [3]Relevantly, in some online platforms, the online community is not just a place for the public to share physician reviews or medical knowledge, but also a physician-patient communication platform [4].Those platforms can be seen as a solution for lack of medical resources in developing countries (ibid). That is, the online health care community is a potential solution to alleviate the phenomenon of long hospital queues and the lack of medical resources in rural areas. However, except a few papers [1,2], online healthcare communities have received little attention and investigation efforts.

This project is related to one of the online healthcare communities, called *CancerConnect* (http://cancerconnect.com/). The idea is that with technology that can autonomously and robustly harvest data from this website, one can develop smart applications and also data analytics-based implications. For example, one could harvest all posts and comments of active users in one of the communities (like Breast Cancer community on this platform), and then, finding indications of possible unknown side effects of medicines. In more detail, one could harvest all messages from a web forum for a certain disease, extract information about (a) medicines people report using and (b) which side effects they report having (topic IENLP!), and compare that with the leaflets of these medicines to determine if some reported side effects are unknown (i.e., not mentioned in the leaflet).

13.2 Description of data set

There is no given data set, but part of your task is to harvest information from the forums of the healthcare community website *CancerConnect* (http://cancerconnect.com/). Overall, we expect that you can harvest the website, and store it as structured data. There are web harvesters / crawlers available on the web that you can use as a service. You can also write your own program that fetches pages (websites that use JavaScript to dynamically load data and construct a web page in the client, can be harveste by using the FireFox browser in headless mode, see https://developer.mozilla.org/en-US/Firefox/Headless_mode or using the `wget` command on linux/mac. In certain circumstances, you could resort to manually saving the individual web pages, but that obviously has its limitations.

About the data items, you should at least collect the following items: (1) User ID, (2) User Name, (3) each post by User, (4) each comment on each post by user, (5) timestamp for each posts/comment.

13.3 Description of challenge

The challenge is to analyse the data according to interesting questions you come up with, potentially leading to new insights. You could use different approaches including data preparation and visualization, natural language processing, semantic analysis or data mining. Additionally, it is possible to build network of interactions among users (who has commented/liked/disliked the posts of whom) across times and gain some insights about the community.

13.4 Tips and suggestions

Network analysis tools are not part of a topic taught. If you want to research in this direction helpful tools are gephi and the networkx library for python.

13.5 References

1 Daren C Brabham, Kurt M Ribisl, Thomas R Kirchner, and Jay M Bernhardt. Crowdsourcing applications for public health. American journal of preventive medicine, 46(2):179–187, 2014.

2 Jie Mein Goh, Guodong Gao, and Ritu Agarwal. The creation of social value: Can an online health community reduce rural–urban health disparities? Mis Quarterly, 40(1), 2016.

3 Sahiti Myneni, Nathan Cobb, and Trevor Cohen. In pursuit of theoretical ground in behavior change support systems: analysis of peer-to-peer communication in a health-related online community. Journal of Medical Internet Research, 18(2), 2016. 2

4 Jying-Nan Wang, Ya-Ling Chiu, Haiyan Yu, and Yuan-Teng Hsu. Understanding a nonlinear causal relationship between rewards and physicians' contributions in online health care communities: Longitudinal study. Journal of Medical Internet Research, 19(12), 2017.

Project 14

Project 14: Mining Crowd-based Inventions [CBIP]

14.1 Introduction

Project owner: Mohsen Jafari Songhori
Primary topic: SEMI
Good to combine with PDBDQ, DPV, DM, or IENLP

Nowadays, ubiquity of online platforms has provided opportunities for generating and selecting innovative ideas, and even developing new products! One of such platforms is Quirky (www.quirky.com). It is a crowdsourcing-based invention platform for household gadgets (e.g., kitchen tools, electronic accessories) founded in 2009. Quirky has more than 405,000 registered members and enthusiastic inventors, ranging from stay-at-home parents, professors, and artists, to engineers. Quirky users submit idea briefs and vote on briefs they want to assist. Of ideas submitted each week, those that pass the initial idea screening phase move rapidly into New Product Development (NPD), during which other users, (i.e., cocreators) self-select to participate in various development subtasks [1].

Product development subtasks start from Research, in which cocreators complete online surveys and indicate consumer preferences, potential uses, and design directions for the new product. Ideas then move to Design, in which cocreators translate and validate the inventor's vision by sketching, modeling, and refining the product design. Near the product design completion are multiple subtasks, including Name, where cocreators submit product names; Tagline, where cocreators propose keywords summarizing product characteristics; and Style, where cocreators provide finishing touches for the new product (e.g., colors and materials). Enhance Design occurs only for products that require special attention, such as those that experience engineering or concept consolidation problems. Finally, cocreators recommend an appropriate price level for the product in Price. Quirky also hires industrial designers, mechanical engineers, and product managers to shepherd ideas through the NPD process.

Quirky's professionals meet weekly with external experts to pick the best developed projects, which sell in both online and offline channels including Quirky.com, Bed Bath & Beyond, Best Buy, and Target. The Quirky community develops products in four main consumer categories: electronics & power, home & garden, kitchen, and travel & adventure. The platform acknowledges significant and legitimate contributions of each user in the ideation and development phases by rewarding them with influence scores that earn them

revenue-sharing rights if the product eventually sells in the marketplace. The influence score is a system Quirky used to record user contributions in NPD. Each percentage of influence represents the contributor's share of revenue of the products sold (i.e., influence score earned multiplied by 10% of the product's revenue).

In crowdsourcing-based NPD processes, information related to each product ideation and development is visible to the crowd. Sales figures for each launched product through all channels are updated immediately.

Overall, this project is related to Quirky (`www.quirky.com`), see figure 14.1 for screenshots. The idea is that with technology that can autonomously and robustly harvest data from this website, one can develop smart applications and also data analytics-based implications. For example, one could harvest all ideas and comments of all users, from ideation and design, till launch stages, and then, finding hidden patterns among successful inventions and users. In more detail, one could harvest all tags/pictures/text from Quirky, and store all information in a structured way, and for example, tracked comprehensive information on all ideas that successfully passed through the ideation and development phases and were launched in the marketplace. This includes the sub-task inputs the ideas received (e.g., number of contributors, contents) from co-creators during all phase.

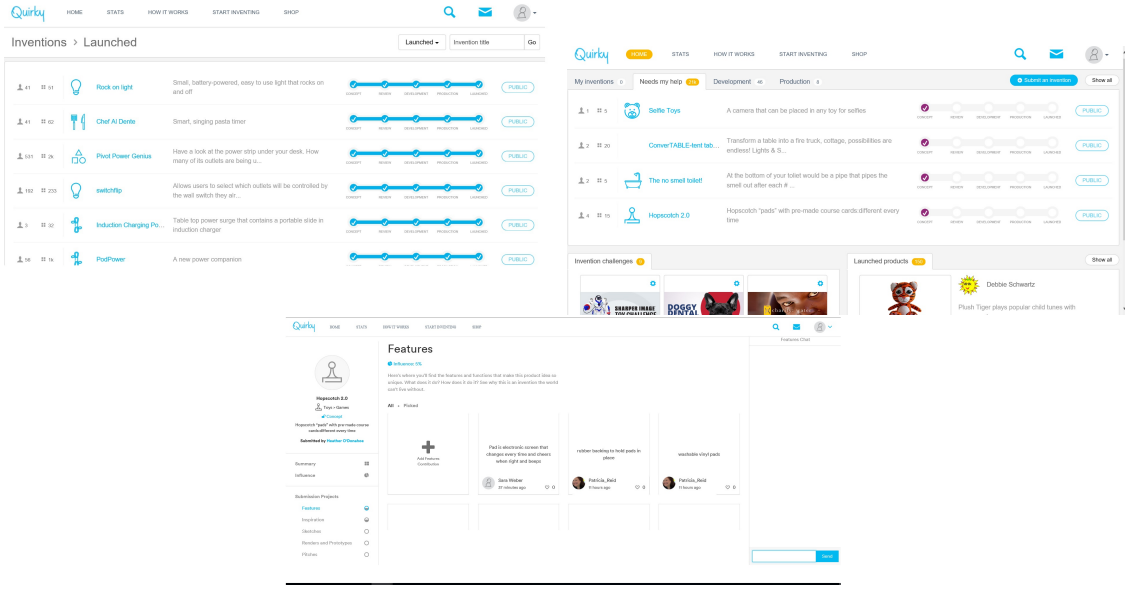


Figure 14.1: Quirky Platform.

14.2 Description of data set

We collected data publicly available on the Quirky website between February and August 2013, and captured detailed information on 20,702 ideas initiated by 1877 users from the website's incipience until August 2013. Of these ideas, we tracked comprehensive information on 89 that successfully passed through the ideation and development phases and were launched in the marketplace. This includes the subtask inputs the ideas received (e.g., number of contributors, contents) from 4509 cocreators during the development phase.

There is no given data set, but part of your task is to harvest information from Quirky. Overall, we expect that you can harvest the website, and store it as structured data. There are web harvesters / crawlers available on the web that you can use as a service. You can also write your own program that fetches pages (websites that use JavaScript to dynamically load data and construct a web page in the client, can be harveste by using the FireFox browser in headless mode, see `https://developer.mozilla.org/en-US/Firefox/`

`Headless_mode` or using the `wget` command on linux/mac. In certain circumstances, you could resort to manually saving the individual web pages, but that obviously has its limitations.

About the data items, you should at least collect the following items: (1) User ID, (2) User Name, (3) ideas/commented post by User, (4) each comment on each post by user, (5) timestamp for each activity of each user, (6) photos/tags by each user, (7) all previous information should include stage of development.

14.3 Description of challenge

The challenge is to analyse the data according to interesting questions you come up with, potentially leading to new insights. You could use different approaches including data preparation and visualization, natural language processing, semantic analysis or data mining. Additionally, it is possible to build network of interactions among users (who has commented/liked/disliked the posts of whom) across times and gain some insights about the community.

14.4 Tips and suggestions

Network analysis tools are not part of a topic taught. If you want to research in this direction helpful tools are gephi and the networkx library for python.

14.5 References

1 John Jianjun Zhu, Stella Yiyan Li, and Michelle Andrews. Ideator expertise and cocreator inputs in crowdsourcing-based new product development. Journal of Product Innovation Management, 34(5):598–616, 2017.

Project 15

Project 15: Process discovery and analysis [PDA]

15.1 Introduction

Project owner: Faiza Bukhsh
Primary topic: PM

15.1.1 Deliverables

The answers for this assignment should be given in a report and a presentation. The reason for doing so is that these documents are always produced when showing managers the results of process mining analysis in real-life situations. This way they can communicate the results for other levels of administration in the hierarchy. Therefore, to guide you in producing this report, we have defined an outline with the points your report should include. This outline is as follows:

(a) Cover page —Size: 1 page Includes the title of your report, your names, student numbers quartile of the course and name of the project owner and date

(b) Introduction —Size: At most 2 pages. The introduction should clearly state the aim of this report. As the target audience are managers of organizations, the introduction should indicate what these managers can find in the report in an appealing way (so that the managers are motivated to read the rest of the document.

(c) Analytical Results—Size: At most 1 page of text per question answered (Note that this page limit does not include figures!) This section should present the results of your analysis. The idea is that you create subsections for every point, you should include: (i) The questions addressed; (ii) Your answer for these questions; (iii) Screen shots of mined models/results that support your answer

(d) Conclusion—Size: At most 2 pages. This section should summarize your main findings and suggestions on how to improve the process that you have analyzed

15.2 Description of data set

There are two different log files available namely Municipality_1.xes, Municipality_2.xes (available on blackboard). Events are labeled with both a code and a Dutch and English label. Each activity code consists of three parts: For instance '01_HOOFD_xxx' indicates the main process and '01_BB_xxx' indicates the 'objections and complaints' ('Beroep en Bezwaar' in Dutch) subprocess. The last three digits hint on the order in which activities are executed, where the first digit often indicates a phase within a process.

15.3 Description of challenge

In the project, you will use the skills gathered in the previous exercises to explore the data set. Although in general we do not want to prescribe tools for the projects, we advise to use ProM in the project. In the topic assignments, we only saw a small portion of its possibilities. Note that some of the plug ins may be unstable, so you may encounter one that does't work as expected.

The project is about performance and conformance checking. ProM has the possibility to replay an event log on the discovered process. Data set of project is from Dutch municipalities. You will explore the data set and discover the process. The data contains building permit applications over a period of approximately four years. The cases in the log contain information on the main application as well as objection procedures in various stages. Furthermore, information is available about the resource that carried out the task and on the cost of the application (attribute SUMleges).The municipalities want to find possible points for improvement on the organizational structure. Moreover if some of the processes will be outsourced then they should be removed from the process and the applicant needs to have these activities performed by an external party before submitting the application. Management wants to know will outsourcing effect the organizational structures of municipalities? The organization would like to streamline their business process and has asked you for advice about perform exploration (e.g., about occurrence and start/end of events, used resource, relations between resources, performance, etc.), process discovery and performance/conformance analysis. Write an advisory report with performance statistics, bottlenecks, etc. and present recommendations for the organization on how to improve and enhance their business process.

15.4 Tips and suggestions