KU LEUVEN
LEUVEN STATISTICS RESEARCH CENTRE

Modern Data Analytics

Exam Project GOZ39a

Summary

There is no exam for this course, instead a project has to be handed in. The project involves a data analytics project. The projects are assigned (randomly) to the different teams but you are allowed to make your preferences known :

- Water Security
- Heat waves impacts
- President Obama
- What makes a great speech?
- Eurovision
- Covid 19 in the USA

Open Research Questions

The research questions are really "open". In other words: **these are just a start**. The same holds for the initial dataset. By no means are you restricted to use the suggested data. On the contrary, a good data-scientist is capable to "think out of the box" and to retrieve data from other sources and blend these together.

It would be a shame, if the only thing you achieved in this course, is the production of a vanilla Jupyter Notebook where you delivered some standard charts using a technique picked up during the lectures. You should have advanced much further on the learning curve than the topics I handed out during the lectures.

And Neo4j? Graph databases are everywhere, even if your project at first sight does not deal with nodes and relationships, think twice. The following url shows how Word2vec was put at work in Neo4j (https://bit.ly/3j54swq)

What is expected from each Team?

- In a nutshell, **we want you to think as a data-scientist throughout the whole production pipeline:** retrieving & pre-processing data, exception handling, building a model, hyperparameter optimization, etc...
- We expect that you bring the topics explained during the course into practice. Your team should be able to bring value to the data. You can use techniques that were not covered during the course and can bring other python packages into the project. This is considered a plus.
- Make sure you start from the same python environment, used in the course. Of course you can update packages, install new ones, ...

- Make sure that you understand the underlying mathematics in the approach that you use (supervised, unsupervised, nlp, AI,).

What do you need to hand in ?

The deliverables shall consist of:

1. A program (python and/or cypher)
   Code can be cypher scripts, a jupyter notebook or both.
   If you managed to deploy your solution as an app, this is considered a big plus.
2. A report (pdf)
3. A presentation (slides)

Deliverable

The project is a group-effort which has to result in the **Three** deliverables mentioned above

1. Your Python Code / Cypher scripts have to be shared on a GitHub account.
2. Report (pdf) of maximum 2 pages
3. Each team will be invited for a presentation on-campus or on-line (teams are free to choose).

Exam / Presentation

There are 3 dates scheduled for the exam / presentation:

- June 10th
- June 23rd
- June 30th

You should already have been assigned one of these dates. If some of the team-members are not scheduled on the same date, feel free to propose an alternative date.

During this presentation the team will present their findings (15 min) and will answer questions (10 minutes). Each team member will receive a question on the project and a theoretical question on some items covered during the lectures.

Assigned Project

You will receive a mail **before April 11th** with the project assigned to you.

Delivery Date

Before June 8th, 2021
Failure to deliver the report in time, results in a no-pass grade for this course

Grading

- 15% on the presentation and Q&A
- 20% on the written report
- 60% on the work done (project)
- 5% peer evaluation

Grading Criteria

Below is in bullet-point format a **non-exhaustive** list of the criteria that we will take into account when we evaluate your work. Each of the projects are described in detail in the appendices of this document. Some of the projects were already covered in the 2021 edition of this course. Do not cut/paste code-snippets, reports, etc… from the previous edition.

Modeling

- Are you able to reach out to different data-sets outside the assigned dataset?
- Visualisation
- Code: Style & Organisation of your Python/Cypher Code.
- Does the code actually work?
  We should be able to clone your code on github and run it on our computer. Make sure that you use a **requirements.txt** file to specify the python packages you require.
- Delivery App
  If you deliver an App, the code should be on Github. The app should be deployed.
  Note that after the Easter break, we plan a course on app-building in Python.

Content of the report

- Your pipeline: from retrieving data to the actual model
- Introduction and problem statement
- Research method & scientific character of the work done
- Argumentation
- Results: discussion/ interpretation
- General conclusion
- Coherence/ logical composition
- Originality & creativity
- References

Presentation

- Presentation: used language
- Presentation: content/ accessibility
- Presentation: form/ composition/ timing
- Understanding underlying mathematics
- Answering questions on Python, Machine Learning, Cypher Code

Failure

There are three ways to obtain a "no-pass" result:

- Your project did not receive a pass grade
- Your team did not hand in a report in time.
- Your team did receive a pass grade put you failed to answer the questions you were asked during the Q&A

In case of a "no-pass" result, students can participate in the August exam. Here new projects will be made available. The August session will be individual, not in team work.

Q&A

There will be Q&A session every Friday afternoon (17.30-18.30) where you can ask questions. The schedules of these Q&A's is available on Toledo. If you think that your question is too specific to be covered on the Q&A session, you can drop me an email. I'll respond as soon as I can.

Appendix A

Water Security

Introduction

The current climate change scenario predicts that almost half of the world’s population will live in areas of high water stress by 2050 with limited access to fresh clean water. Governments, national, and international institutions, as well as water management companies, are looking for solutions that can address this growing global water demand. Cities are encouraged to take action on water security, to build resilience to water scarcity and manage this finite resource for the future.

Proposed Research Question

Are all cities reporting consistent data ? Are there data gaps in some regions ? Can you predict the likelihood of a water shortage ? What about the population densities in those areas ?...

Initial Data Set

The place to **start** your enquiry is (data.cdp.net and www.cdp.net). The climate disclosure panel (CDP) is a not-for-profit charity that runs the global disclosure system for investors, companies, cities, states and regions to manage their environmental impacts. The data for cities and regions is free for access & downloading.

Appendix B

Heat Wave Impacts

Introduction

In the 1960s, Major cities experienced, on average, about two heat waves per year. In the 2010s, that number rose to more than six heat waves per year. These heat waves are also lasting longer, on average 47 days longer than in 1960. Even under different climate models and emission scenarios, results indicate that extreme heat events worsen.

Heatwaves, or heat and hot weather that can last for several days, can have a significant impact on society, including a rise in heat-related deaths. More than 70 000 people died during the 2003 heatwave in Europe. Workers who are exposed to extreme heat or work in hot environments may be at risk of heat stress. Exposure to extreme heat can result in occupational illnesses and injuries. Heat stress can result in heat stroke, heat exhaustion, heat cramps, or heat rashes. Humidity is an important factor in heat index assessment. When the humidity is high, water does not evaporate as easily and so it becomes difficult for the body to cool off through sweating.

More than 10 nuclear power plants were shut down in France because of cooling-water related issues ! Heatwaves also have an impact on productivity of workers and impact the food supply.

Proposed Research Questions

- Can you predict the impact on mortality as a function of different parameters (Location, Age, …) caused by heath waves?
- Can you spot trends in heat waves?
- Can you predict the impact on the economy (GDP) resulting from heatwaves
- ...

Initial Data Sets

- International disaster database https://public.emdat.be
- Heat data https://earthdata.nasa.gov/learn/pathfinders/disasters/extreme-heat
- Temperature in different countries and cities https://sedac.ciesin.columbia.edu/data/set/sdei-global-uhi-2013/data-download
- Population Data https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count-rev11/data-download
- Heat Stress https://www.cdc.gov/niosh/topics/heatstress
- Poverty Related Data https://sedac.ciesin.columbia.edu/data/sets/browse?facets=theme%3Apoverty
- Heathwaver & Mortality https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3324776/

Appendix C

President Obama

Introduction

Barack Obama has held two consecutive terms as US President. He has given multiple speeches. Much of those have been transcribed and analysed. In this thesis you must analyse the written text of his speeches in detail using various NLP techniques. The analysis of the text should be done using extra variables: economic growth, level of the stock market, approval ratings…

Proposed Research Question

What variables have an impact on the speech of the US President? Be as broad as possible in your approach. Look at economic variables, approval ratings, international political situation, elections,…

Dataset

https://www.americanrhetoric.com/barackobamaspeeches.htm
Each of the speeches has a pdf.

Appendix D

What makes a great speech?

Introduction

The influence of a speech on a group of human beings cannot be underestimated. The impact of a couple of words, well aligned in a sentence can have a tremendeous impact on the behaviour of a group of people. Think about the inaugural speech of President Kennedy "*…do not ask what your country can do for you, but what you can do for your country…*". But what makes a good speech?

Proposed Research Question

Analyse using a wide range of NLP techniques a series of important speeches (English) and bring insight to why these speeches were so important.

Data Set

Full text, audio, and video database of the 100 most significant American political speeches of the 20th century, according to 137 leading scholars of American public address, as compiled by Stephen E. Lucas (University of Wisconsin-Madison) and Martin J. Medhurst (Baylor University): https://www.americanrhetoric.com/top100speechesall.html

Each speech is provided in pdf-format.

Appendix E

Eurovision

Introduction

Eurovision is a song contest amongs European countries. The particularity is that countries must vote to each other to determine the winner. There is a semi-final and a final. The total score is a combination of what the (professional) jury attributes to a song and the public. The maximum score a song can obtain is 12. Each year we observe how some countries consistently vote for each other. Far from an honest competition.

Proposed Research Question

Examine from a data-science perspective the scoring system in the Eurovision competition. Could you make a prediction how countries will vote ? What are the underlying features ? (neigbouring countries, language, culture, ….)

Data Set

- The scores for the 2013,2014 and 2015 contests are availble on the following S3 bucket (https://goz39a.s3.eu-central-1.amazonaws.com/eurovision/eurovision.txt). These datapoints are defined as nodes in a Neo4j graph. There is a popular Neo4j graph gist where the issue of mutual voting was examined: https://neo4j.com/graphgists/eurovision-votes/
  (Of course I expect you to do better than this)
- The 2016 data set is available at https://eurovision.wetransfer.com/downloads/a1da4b5eb0395e58b71016dce076564a20170409152448/6754ae

- The 2017 dataset is available at https://eurovision.wetransfer.com/downloads/79b392c2f33731977e70a9ce555f1d1620170614172858/fcced0

Appendix F

Covid 19 in the USA

Introduction

The NY times has published (and keeps doing so) Covid 19 data for the USA’s territory.
The dataset incorporates:

- covid 19 cases at county-level
- covid 19 cases in prison
- covid 19 cases on college and university campuses.

Proposed Research Question

Starting from the dataset below, provide insight on the evolution of covid 19 in the USA. Note that providing insight is much more that producing a couple of nice charts. We expect to see a data scientist at work.

Data Set

https://github.com/nytimes/covid-19-data