CENG313 Introduction to Data Science

Fall 2022 – Gazi University, Computer Engineering Department

Instructor: Dr. Duygu Sarıkaya email: duygusarikaya@gazi.edu.tr

TA: Dr. Berrin İşlek Bilge email: islekberrin@hotmail.com

FINAL PROJECT INSTRUCTIONS

Final Project Goals:

- **Explore** key data science and machine learning concepts and novel approaches / apply data science and machine learning algorithms to real-world tasks.
- **Do research** in data science. You are expected to do independent research for your project in addition to following the course material.
- **Learn to work together for a common goal** even though your technical background might vary.
- **Learn soft skills** (such as leadership, working in a group, managing time, and communication skills).
- **Create a portfolio,** you can use this work as part of your portfolio which might be useful in the future. You might even want to submit a paper (to credible venues only, and please note that all papers to be submitted share the authorship between teammates and the instructor of the class as the supervisor, so if you have plans for this you should contact me beforehand)
- **Contribute to your team, learn and have fun in the process!**

What are the deliverables of the final project?

- Proposal
- Milestone
- Final Deliverables: Poster with Infographics, Final Technical Report, Presentation (in person "and" video), Project Codes, Peer Review

You are a journalist / data scientist and your team is assigned a data journalism / data story telling project or case study to be published. Some examples can be seen here (please follow the links for each article mentioned):

https://gijn.org/2021/10/22/data-journalism-top-10-climate-change-fake-google-reviews/

https://gijn.org/2021/10/29/data-journalism-top-10-disappearing-bumblebees-methane-leaks-tax-hot-siberia/

https://gijn.org/2021/10/14/data-journalism-top-10-nobel-prizes-hong-kong-fuel-shortages/

More from gijn: https://gijn.org/stories/

https://sites.lib.uh.edu/datajournalism/overview/examples-of-award-winning-data-journalism/

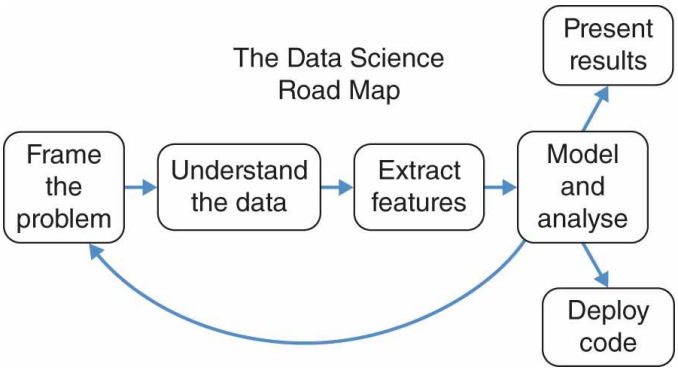
https://datajournalism.com/read/longreads/covid-19-data-journalism

https://www.import.io/post/8-fantastic-examples-of-data-journalism/

https://www.theguardian.com/membership/datablog/2021/sep/13/numbers-you-can-tell-stories-with-a-decade-of-guardian-data-journalism

Basically you will take a topic, learn many aspects about it / do research about the topic, find **multiple** datasets and data sources you can use, follow the **Data Science Road Map** as we have seen in class (except for data deployment: you will submit a jupyter notebook instead). **Exploratory Data Analysis and Visualization** will be key components of your project, as well as utilizing **machine learning algorithms** (for classification, clustering or regression).

For an interesting article you should have strong and wherever possible dynamic **visualizations**, **multiple sources and types of data** (**curating/scraping your data, using APIs to get data from famous platforms, using data from social media such as Twitter, event based data (e.g. Twitter hashtags) and user interactions (e.g. followers, retweets)**, **insights driven from the data**, **different aspects**, how / for which aspect you use **machine learning**, **creativity,** and **good presentation/communication methods**. Adding **key/historical information**, an **introduction**, **a background story**, **infographics, or animations with this information** for each component will make your article much stronger.



Your first task as a team will be to pick a topic. **The topics this year are:**

- **k-pop**
- **Arts, Literature and Media**
- **Global Sustainable Development Goals with a focus on:**

  Global Inflation and Cost of Living around the World

  Global Energy Crisis

  Climate Change

We will have a **tournament** among groups in each category, details will be announced later.

Just to give you an idea about what you can do on the topics:

- **k-pop:**

For example, you can use Spotify's API to get data about k-pop in general or a k-pop band (e.g. BTS), you can compare k-pop/this band to other bands, other music genres and trends. Among the data you can get are: 11 acoustic qualities for each song, tracks, artists, popularity etc. For example, you can explore what acoustic characteristics a band has and how does it differ from other bands / genres / all time favorite (top) songs? Can you predict whether a song belongs to that band by just looking at these characteristics? How do the characteristics for each album change?

You can create a small dataset of song lyrics / their English translations and visualize which words are used most frequently / cluster words based on topics / sentiment. How do these keywords / most frequently used words/ sentiments change for each album? How do they change over time?

You can find more data sources (for example this survey: https://figshare.com/articles/dataset/KPOP_DATA_xlsx/12093648/2 , Google trends / analytics, Twitter: for example, the hashtags fort he band in general e.g. #bts , their fans #btsarmy or the ones used during significant events: #LightItUpBTS during Grammys, #dynamitebts on the song Dynamite's release day or #btsinbusan Their recent concert in Busan: what are the main sentiments of the tweets with the hashtag? What other hashtags are used? A World map that visualizes the density of tweets posted according to the countries, a World map that shows how the number of tweets per country changes in time, the community analysis e.g. who posts the initial tweets that are retweeted the most, a graph that shows the most frequent posters and their interaction with other users, are these users the same for multiple related hashtags? Which accounts are the most influential? What are some interesting findings you have come across?)

For example, for the background story, what is k-pop's history? What are the most popular bands for each year? A visualization of how the most popular bands change? (an example of a chart / time series plot changing over time: https://www.youtube.com/watch?v=DpgsPbMXwn8) What events / albums / songs etc. has been turning points in the success of a specific band? Which companies are more successful? Which bands drive their success? Collaborations with other artists, bands etc.

These are only some questions to give you an idea but you should not be limited to these suggestions. Ideally you should have different components as above but you can come up with different data sources and questions.

Some other sources for Video Games and Sustainable Development Goals topics:

- **Arts, Literature and Media**

  Create projects focusing on data visualization or data analysis on art history data, artistic collaborations, gallery attendances, data from movie/tv show/music industries, scripts of theater plays, data about audience interest such as library circulation or loan information etc. You can ask specific questions such as "How does the proportion of fiction written by British authors or by women change across time?", "How do the themes/most frequently used words of books written by a particular author change over time?" You can analyze correlations between frequently used words and emotions using basic sentiment analysis or nature language processing (NLP) techniques. You can analyze differences in a genre across different regions of the world, discover influences and compare timelines. Predict how many copies of a certain book has the potential to be sold in each region of the world depending on the genre, most frequent words used, gender of characters, etc.

  Sample datasets:
  https://www.nga.gov/open-access-images/open-data.html
  https://paperswithcode.com/dataset/wikiart (you can use artist, genre, objects information in paintings and possibly simple features (first and second-order statistics, color distributions, co-occurrence matrixes etc) that you can extract from images)
  https://www.artnome.com/art-data
  https://pro.europeana.eu/pages/datasets/data/itemtype/paintings
  https://artsdatathon.org/data/datasets/

https://www.kaggle.com/datasets/raynardj/classic-english-literature-corpus
https://hcommons.org/deposits/item/hc:26955/ (data Access: https://github.com/tedunderwood/noveltmmeta)
https://www.kaggle.com/datasets/kingburrito666/shakespeare-plays
https://www.kaggle.com/datasets/kewagbln/shakespeareonline
https://analytics.hathitrust.org/datasets
https://www.data-in-brief.com/article/S2352-3409(22)00117-2/fulltext#relatedArticles
https://www.opensourceshakespeare.org/
https://dataspace.princeton.edu/handle/88435/dsp019306t2441

You can also curate your own data using publicly available and free (non copyrighted) texts. For example as the copyright protection lasts for the life of the author plus an additional 70 years, many classic novels do not have copyrights and their digitalized versions are often shared freely.

- What are **Sustainable Development Goals**: https://sdgs.un.org/goals

Some example articles / visualizations:

https://www.wsj.com/articles/the-bumblebees-plight-why-they-are-disappearing-in-the-u-s-11634992056?mod=e2tw

https://www.theguardian.com/environment/ng-interactive/2021/oct/13/uk-us-china-how-the-worlds-carbon-centre-of-gravity-moved-over-200-years

https://www.theguardian.com/environment/ng-interactive/2021/oct/14/climate-change-happening-now-stats-graphs-maps-cop26?

https://www.washingtonpost.com/climate-environment/interactive/2021/russia-greenhouse-gas-emissions/

Some data sources:

World Bank Data

National Climatic Data Center NOAA

ClimateData.us

UNICEF Data

undata

SocioEconomic Data and Applications Center SEDAC

Deliverable 1: Proposal

Deadline: 16th November 2022, Wed , 23:59 (**It is a good idea to start working on your project as soon as possible, so don't wait until the last moment for this step**)

On the 23rd of November 2022, 9:30, each team will be presenting a 3 minute teaser of their projects in class.

You should do some research, explore ideas and datasets, and brainstorm together. Then you will write a project proposal (PDF format, 300-500 words – not counting the cover page or logo) to be uploaded

on lms/guzem. You should indicate the topic you choose (one of the three main topics), the specific topics you plan to work on, the main questions that will be driving your project, exploratory questions you are planning to search answers for, title of the project, name and logo of the team, the full names of all of your team members.

Your proposal should include:

- Motivation: Which specific topic will you work on? Why did you pick this problem? What is your motivation to pick this problem? How is it interesting? What is the main questions you are looking to answer? What are some exploratory / additional questions you will be asking?
- Datasets: Which datasets are available in this domain? Which datasets do you plan to use? What are possible data sources you can use to collect data? Give some information on the data. Please indicate where you can download the dataset, what APIs you can use or how you can collect the data.
- Method: What machine learning techniques are you planning to apply? What will you be using these techniques for? (it is okay to refine this point or make changes at later stages but you should ideally have a rough plan)

Deliverable 2: Milestone

Deadline: 7th of December 2022, Wed, 23:59

The milestone will help you make sure you're on track, and should describe what you've accomplished so far, and very briefly say what else you plan to do. You should write it as if it's an “early draft" of what will turn into your final project (please check the final technical report template indicated in the final deliverables). You can write it as if you're writing the first few pages of your final project report, so that you can re-use most of the milestone text in your final report.

Your milestone should be at most 3 pages, excluding references. Similar to to the proposal, it should include

- Motivation: Which specific topic will you work on? Why did you pick this problem? What is your motivation to pick this problem? What is the background story? How is it interesting? What is the main questions you are looking to answer? What are some exploratory / additional questions you will be asking?
- Datasets: Which datasets / data sources have you used? Give some information on the data. Please indicate where you can download the dataset, what APIs you used or how you collected the data. Please explain the data processing / cleaning techniques you used and why?
- Method: What machine learning techniques have you applied? What are you be using these techniques for? Why did you choose a specific algorithm for the problem?

Final deliverables: Poster with Infographics , Final Technical Report, Presentation (in person and video), Peer Review, Project Codes

(Deadline : 28th of December 2022, 9:30 – in class presentations,

in class presentation files and poster infographics are to be sent on the day before via Slack team channels(27th of December 12:00), all deliverables are due midnight 28th of December 2022, 23:59 to be sent on guzem and Slack team channels)

Presentation (In person "and" video): The presentation will be around 7 minutes. We will have in class presentations "and" video presentations, which means, **we will have video presentations that are prerecorded but you will need to be present in person for the Q&A session**.

Consider this the **story** that will be published for a newspaper. Please adopt the good practices in data storytelling, data journalism, case studies and enrich your presentation with infographics, animations, visualizations. The background story, key/important information and storytelling with data, creativity, communication will be important aspects. **The presentation will be graded on how well you tell a story with data (data journalism / data storytelling)**, e.g. the insights driven from data, questions asked/answered, data sources used, visualization, communication, background story, presentation skills etc.

The video presentation will be uploaded to Youtube. You may use Zoom or some other recording software to record your presentations, and add graphics, videos to your presentations. You can present in person or just do voiceover, you may all take turns to present but it is okay if one person presents as well, please make sure to not include sensitive information to your videos (such as phone numbers) as they will be made public. You will share the link to the video among other deliverables to be submitted on lms/guzem (please do not try to upload the video itself, it won't work).We will make a playlist of all video presentations. Do not forget to make your video public. Consider the deliverables and your video presentation as portfolio you can include in your CV when you graduate!

Poster with Infographics: This will basically be a **poster with infographics intended for press medium which summarizes your presentation**. It should be informative and visually appealing / communicative as much as possible. The poster will be graded on similar points as the presentation. You will be uploading the poster in PDF format.

Some Examples:

https://www.zencos.com/blog/infographics-basics-example-design-guide/

https://blink-designs.co.uk/infographics2

https://visual.ly/community/Infographics/other/data-visualization-data-storytelling

Final Technical Report: Up to 5 pages at most excluding References. The **technical** report which will include: Project Name, Team Name/Logo, Title, Author(s), Abstract, Introduction, Methodology, Dataset, Experiments, Results, Conclusions, References. Please download the file at https://www.acm.org/binaries/content/assets/publications/article-templates/pubform.docx (ACM Template for SIG Site). **The reports will be graded on the quality, clarity, and the technical content**. Please make sure to discuss the motivation you pick certain techniques/algorithms and to discuss the outcomes as well. You will be uploading the report in PDF format.

Project Codes:

The programming language is Python with no exceptions and you will be submitting a Jupyter notebook. Please submit a zip file with the code for your final project as Jupyter notebook extension is not allowed (you might include a link to a Github repository in addition). You do not have to include the data or additional libraries, instead write up a README file giving information about which libraries / APIs you have used, where we can download the data etc).

Peer Review (collected individually)

Using the Slack's "direct message" to me (please do not send emails, or messages on guzem/lms: it is already hard to keep track of the emails) , please write your group's name, and the name and surname of the members, their contributions (describe what each team member worked on and contributed to the project. This is to make sure team members are carrying a fair share of the work for projects) and for each member please evaluate their performance based on this criteria (just write the number which you think best represents the members' performance):

0: The member did not contribute at all / was not present

1: The member contributed only marginally (e.g. the member helped write a proposal but then was not present for the rest of the project)

2: The member was present through the whole process, did contribute some however overall, they didn't carry out the tasks they were assigned.

3: The member was present through the whole process, they did contribute to a limited extent (e.g. they did carry out tasks to a large extend but communication was difficult etc.)

4: The member was present through the whole process and did their best to contribute to the project.

5: The member was present through the whole process and they did extraordinarily well to complete the project.