Presentation for 25.02. <https://docs.google.com/presentation/d/15XnZCc_s5TuuTBM4zfSw3qlruCSoBhl5eU_hhCGgKJs/edit?usp=drivesdk>

Translation API:

<https://stackabuse.com/text-translation-with-google-translate-api-in-python/>

speech violation classification:

<https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data>

BERLIN

DATA we have:

Data from Versammlungsbehörde - State

1. <https://github.com/demodiff/berlin/tree/main/data%2Fjsonl>

Json file: versammlungsbehörde ->28.06-21-12-21 every day protest in berlin

more data we could have from here: 2022 - now -> not scraped jet !

> no participants information

> but: location(start, end),route, date, time, topic

1. <https://zenodo.org/records/10094245>

2018 -2022 -> sorted version here:

> no exact location, route or time

> but: berlin, date, topic and expected + actual participants

(<https://docs.google.com/spreadsheets/d/1c12PRI_Bv1z44Nktvj9jFU2WRw5AxvxHbQfqOWl7aFk/edit?usp=drivesdk> - exel updated version from zenodo !

[<https://github.com/davidpomerenke/german-protest-registrations/tree/v1.0.0?tab=readme-ov-file>Data to download would be probably double ? Data from Berlin 2020-2022 possible ) -> read.me is important to check out for more data to newspapers (maybe helpful for participants , …..)

<https://fragdenstaat.de/anfrage/demonstrationen-in-den-jahren-2020-2022/#nachricht-762832> only year 2022 probably that data is double but its filtered, sorted and with participants, so thats a plus]

Newspaper Data :

1. ACLED got an account(Paula) only allowed to get data from past 3 years :

newspaper data:

> event type: riot or protest; subtype: violent or non-violent; organization names; newspaper: names, national or city news, time reported; location: quarters, latitude, longitude, population of quarters, …; crowd size;

<https://docs.google.com/spreadsheets/d/13tzNXDDcp9nwD2PYd7647NBt3WWRQree9i_sqi1ERUU/edit?usp=drivesdk> - exel sheet here you can download as csv yourself

some data in json, xsxl, … to download here: <https://we.tl/t-m7sG5hK246>(you don't need the german data anymore just the report (for 2.), versammlungsbehörde (for 1. ) &(ifg(frag den staat data for 2.)))

other data i linked up ^ updated zenodo + newspaper data from ACLED

Our Main inspo for what problems we could solve with that data is:

<https://interaktiv.tagesspiegel.de/lab/demo-atlas-berlin-wofuer-wird-in-meiner-nachbarschaft-demonstriert/>

<https://protestdata.eu/bundesweite-analyse>

international inspo:

<https://carnegieendowment.org/publications/interactive/protest-tracker#>

<https://acleddata.com/early-warning/>

Poblem models:

1. ecological protests sort them into different organizations find out if they are right wing or left

multi-class problem for Data Science

or even

NER(named-entity-recognition)

**Problem:**

* 1. **Predict from what is happening right now on social media that will happen in future.** We have information from Social Media (TWITTER/X), and there are #s, topics, catchy phrases, videos
  + Assumption: Social media gives signal about probable protests
* 2. We have social media information (posts and invites for protests)
* 3. We have social media information (abouts protests that happened) 1990 - 2020 and how did ~~they develop~~

**Mentoring session:**

We pre-select #s. Then we filter by the #s

Protest-type -> [‘Protection of the forest’, ‘war’,...]

Protection of the forest: [#saveamazonia, #savetherainforest, …] -> “Protection of the forest

(1.000) data points

Flu: [#covid, #flu, …] -> Flu

(1.000) data points

5.000 data points randomly from twitter (Taylor Swift, US election, Soccer, Dota2, …) → [“

**Problem v2 - proposal:**

Given some texts from social media (twitter/x, facebook, tiktok) infer the protests and the “\*power” that they will may happen.

*\*Power is the strength that the protest may have, we could measure it through related posts*

by Tugba

By using the [**Versammlungen im Land Berlin**](https://daten.berlin.de/datensaetze/versammlungen-im-land-berlin) data we can have regression and classification tasks. Regression is a type of supervised learning task where the goal is to predict a continuous target variable.

In regression, the output variable takes continuous values. For example duration of protest in the future.

We can evaluate the regression model with some metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), etc.

In classification, the output variable takes categorical values. For example, next protest area can predict depends on the zip and street.

We can evaluate the classification model with some metrics like accuracy, precision, recall, F1-score, etc.