TEXT-AS-DATA RESEARCH PROJECT

PITCH, PROPOSAL, FINAL PRESENTATION and REPLICABLE CODE PACKAGE

(MSc in Economics, WiSo Universität Hamburg)

Winter Semester 2021 – 2022

Instructor : Huyen Nguyen
Website: http://huyenttnguyen.com
E-mail: huyen.nguyen-1@uni-hamburg.de

Task & percentage of final grade	Due date
1) pitch recording (max 5'/team) (15%) 2) research proposal (max 1 page) (15%)	Wednesday 24.11.2021 @ 23.59 CEST (upload BOTH components to Open Olat)
Final oral presentation (max 25'/team) & slides (50%)	To be confirmed (early 02.2022) Slides to be uploaded 3 days before your presentation (to Open Olat)
Replicable code package (20%)	3 days before your presentation (to Open Olat)

What is a research project?

You research project can be anything from a replication exercise of latest research papers, an improvement of the empirical techniques the papers use, or answering a new research question using existing/self-sourced datasets. The main requirement is that the project should *use at least two tools* covered in class.

Some practical tips for you on how to read a research paper productively in social sciences:
https://www.icpsr.umich.edu/files/instructors/How_to_Read_a_Journal_Article.pdf

Inspiration for interesting and relevant research projects in social sciences:
https://github.com/causaltext/causal-text-papers

Data sources?

You can choose one of the following options (Whichever your option is, do check and decide EARLY ENOUGH IN ADVANCE)

1) Use your own procured data sources (if it is raw data scraped from the Web, make sure that you account properly for the necessary time for data preprocessing)

2) Check one of the databases below:
   - List of interesting database
   - U.S. Congressional Record, https://data.stanford.edu/congress_text
   - CourtListener, https://github.com/idc9/law-net
   - Chris Bail's list of digital data sets
   - Data is plural spreadsheet
   - Google toolbox: earch engine for data
   - Liste: https://github.com/awesomedata/awesome-public-datasets
   - Dataverse: data repository from Harvard
   - APIs list: https://github.com/public-apis/public-apis or https://apilist.fun/ or https://www.programmableweb.com/apis/directory
   - Datasets on Kaggle
   - Project ideas

3) Ask the instructor for sample data sources

Midterm research pitch & research design report (30%)

Preliminary topic choice & consultation session registration (for Week 8) should be done at the latest by **Week 5, Wednesday 10.11.2021.** This sign-up list will be available in class to sign up during Week 3 to Week 5.

By **23.59 on Wednesday 24.11**, in groups of two to three students (group matching coordination via Slack channel must be finished by Wednesday 27.10), you are required to submit the following items:

1) A 5-min pitch video recording of your research idea in .mp4 format
2) A 1-page detailed research design of your group (motivation, related literature, data, and approach)

Each task contributes equally to the 30% grade component you receive for the course i.e. 15% for the 5-min video recording, 15% for the research design.

1) Research pitch

The 5-min pitch video recording should serve as an inspiring pitch of ALL group members about your research project. Pitch it as if your audience is a to a non-technical decision-maker, such that he/she/they can understand the key ideas of and get excited about your research project.

NOTE: 5-minute means **exactly** 5 minutes, not one or a few seconds more.

Should you upload anything more than exactly 5-minute, your team will be disqualified and notified, once the instructor has checked if your video meets the time limit after the deadline. You will then be

required to reupload, which means that your team will incur a grade loss because of late submission in this task.

Regarding the technology you can use, some familiar options include Powerpoint (instructions here) Zoom (instructions). Alternatively, you can play around with the professional pitch narrative and slide decks that are used to pitch to company investors.
https://www.pitchtape.com/
https://www.loom.com/blog/remote-pitch-deck-best-practices

Whichever technology you use, make sure your group names, individual names are visible on the first slide, while the rest of the slides, voices, faces are clear and visible throughout the entire 5 minutes.

2) Research design report

The 1-page research design report should explain the motivation, research question, related literature, data, hypotheses, expected results and desired contribution to a non-technical decision-maker.

Below is a detailed sample guideline:

Research Question

- *What is your research question?*
- *Discuss the motivation for the project -- why is it interesting or necessary?*
- *Briefly summarize at least 3 previous related papers.*

Dataset/Corpus

- *Describe the corpus or dataset you will use for your project -- where you obtained it, or where you will obtain it from.*
- *How will the corpus help you answer your research question?*
- *Describe how you will organize and label the documents, and outline how you will merge them with associated metadata (if applicable).*
- *Perform a preliminary inspection of your corpus and explain what preprocessing/cleaning steps may be needed.*
- *If applicable, report some summary statistics about the data.*

Methods

- *Outline your approach. What tools will you use to solve your research problem?*
- *E.g., for machine learning: What model will you use? How will you evaluate and validate your model? How will you interpret or explain your model?*
- *For an econometric analysis, write out your regression equation. What are the empirical assumptions and how will you assess robustness of your results?*

Timeline

Provide a project timeline for finishing by mid January 2022.

Final presentation (25') (50%)

The presentation should be using slides and *last MAX 25 minutes*. Please submitted your presentation in the Open Olat folder THREE DAYS before your assigned exam date. One submission per group only, but do not forget to put the names of all group members in the first slide.

Grading components:

- Individual:
  - Presentation skill: 10%
- Group
  - Quality of the material (slides): 10%
  - Summary: 10%
  - Critical assessment: 15% (i.e. uncertainty, implications, and limitations of your work)
  - Suggestion: 5% (e.g. way(s) to improve the methodology, extend data limitations, test another mechanism, address a complementary research question)

Reproducible Code Package (20%)

Additionally, you need to give a **replication code package (20% of final grade)**. Similar to the presentation slides, *this .zip file* also needs to be uploaded to the dedicated Open Olat folder THREE DAYS before your assigned exam date. This .zip file must include:

(1) the data you use
(2) the codes you wrote to generate the results in your oral presentation (in .py/.ipynb preferably)
(3) a 1-page README file (in .txt) with any further step-by-step miscellaneous explanations of how to load the data, Python version and what packages to use to run your files.

Full credit will be given only if the instructor can reproduce all results in the oral presentation in one click. Inspiration can be found here.