Foundations of Computational Social Systems: Projects Guide

The course assessment is based on a final course project. Projects are done in groups that can have at most 4 students and must mix students with different backgrounds (at least two different backgrounds per group). The purpose of the project is to show what you have learned in the course in terms of data management, analysis, statistics, and interpretation. Projects have to start from a research question and have an empirical focus, but at the same time critically reflect on methods, conclusions, and limitations.

The timeline of the project deadlines can be found below. The deadlines are up to the end of the day marked by each date.

- **19.12.2023:** Registration of projects.
- **03.02.2024:** 11:00: Project presentations session day 1.
- **17.02.2024:** Deadline for submission of final report.

We offer a second date for project presentations and the submission of the final report for project teams that cannot make it on the first date:

- **03.03.2024:** 11:00 Second date for project presentations session day 2.
- **17.03.2024:** Second date for submission of final report.

You can get guidance for your project on two special sessions: at 11:00 on 09.01.2024 and 16.01.2024.

The steps to do the course project are the following:

1. **Form a group and choose a topic**

Get in touch with fellow students over the course discord server or around lectures to form a group. You can use the same group composition as in Introduction to Computational Social Systems, but you don't have to.

2. **Project groups registration**

In the last week before the Christmas break (22.12), your group has to register for the project on TC group assignment.

3. **Project presentations**

During project presentation sessions, students give 7-minute presentations about the state of results of their projects. Students get questions from lecturers regarding their understanding of the topic and feedback on how to improve the project for the final report.

Each presentation at least has to contain four slides:

1) **Project title, names of group members**
2) **Research question and motivation**
3) **Data and methods**
4) **Results**
5) **Conclusion**

All members of the group have to participate in their presentation (for example presenting one slide each) and be present to reply to questions by the lecturers. The time for the presentation is restricted to 7 minutes. It is important that you learn to synthesize and summarize the project, please practice the presentation substantially because taking longer will affect the project grade.

You can get guidance for your project on the designated sessions (09.01.2024 and 16.01.2024) or on discord up until the first project presentations date (23.01.2024). After that, no guidance will be provided via discord or email.

4. **Submit the final report**

Send a final report as a PDF document (max. 6 pages, min. font size 11pt) via TC. References do not count towards this 6-page limit. Projects can contain a link to a Github repository including the code to produce results, datasets if they can be shared, and additional figures or tables that can be referenced from the project report.

Project reports should follow this structure:

- Motivation: What question(s) do you seek to answer and why?
- Data retrieval: Explain the interfaces or resources you used to collect all data necessary for the project.
- Data processing: Explain how you filtered data, normalized values, computed additional variables, etc.
- Analysis: Perform statistical analyses and visualizations that assess the question(s).
- Conclusion: Evaluate answers to the question and their reliability.
- Critique: Identify limitations and alternative explanations for your results.

Plots should be correctly shown (named axes, visible scales) and writing has to be understandable. The motivation part is very important. Argue why your project is relevant and what we can learn about human behavior with it.

Be careful with statistics in the analysis part. Use the methods covered in the course to assess the uncertainty of your answers and comment on these results in the conclusion part. The critique part is essential. An important objective of this course is to develop a critical understanding of the opportunities and limitations of the data and methods we will use.

After your project presentations, we might suggest feedback on your current project analysis and results. We will keep track of this feedback and expect that you incorporate it in your final project report. If something could not be done then we expect that you explain why. No more feedback or support will be given after the first project presentations day (03.02.2023). Students who aim to the second date should get early input on their project.

Project grading

The grade for the project is composed of 50% for the presentation and 50% for the final report. Extra points are given when projects are based on open science principles (e.g. data and codes are available in a Github repository) and when data sources or methods go beyond what is covered in the course. Projects do not need to report "positive results", what is important is that you show how your have addressed your research questions, document any issues or deviations, and critically reflect on methods and results. Remember that on top of the project presentation and report you can individually achieve an additional 4x5% (up to 20%) by handing in the exercise solutions.

Examples of project topics

Replicating a previous paper

You can select a previous paper and take the same question and methods used in the paper. You don't need to do the exact same thing, but the question can be the same. For example, you can replicate a study about sentiment on Twitter but with Reddit data. In this case, it is important that you explain in your report the question and methods of the paper you replicate, how you have deviated from those, and how your results agree, disagree, or are inconclusive. Many papers share tweet ids or other kinds of ids for you to generate your own dataset in a similar way, we welcome that kind of replication but not replications based solely on the data shared by the authors of the original paper.

Building on your Introduction to Computational Social Systems project

You can seize the readings you do for Introduction to Computational Social Systems and propose a question and a plan based on that. This has the upside that you could merge both projects afterwards in one report and use it to showcase what you learned or towards a scientific publication. Some examples of how this can be done:

- **Bias and fairness in recommender systems.** Use a dataset of book reviews and a recommender system method (for example collaborative filtering) to evaluate if books written by women are recommended less often. You can compare various recommendation algorithms and assess their gender biases.

- **Integrating Survey Data and Digital Trails Data.** Take international survey data, for example from the World Values Survey, and compare values across countries with a language-independent measure of online behavior. You can use this to test if long-term orientation in the World Values Survey is correlated with the Future Orientation Index, or if the volume of accesses to Wikipedia pages about climate change in all languages is correlated with concerns about the climate or average altitude of a country.

- **Fighting online misinformation.** Use a set of links to websites with misinformation (e.g. promoting ivermecting as COVID-19 treatment) and gather tweets or Facebook public posts through Crowdtangle. You can start from what previous research has done, for example testing if emotional content on posts helps to get more retweets or shares on Facebook or testing if there is substantial inequality in sharing links such that 90% of shares come from 10% of users.

- **Algorithm aversion.** Search on Twitter or Reddit for posts containing the word "algorithm" and apply sentiment analysis. Do your own annotations to identify if sentiment analysis can find expressions of aversion. Run text analysis to identify if there are different reasons for the aversion and the different meanings of the word.

- **Social media polarization.** Take a recent Twitter controversial trending topic and retrieve tweets about it from the last week. Take the most active users and get network data between them in terms of replies, retweets, quotes, and followings. Quantify modularity in these networks and argue what aspect of polarization this measures and what other aspects are missing with this analysis.

- **Technologies for self-tracking.** Download your own Twitter, Facebook, or Instagram datasets and process your own information as if it was a self-tracker of relevant information such as location, activity, or sentiment. You can invite colleagues to donate data to you if they are actively consenting to your analysis. Then you can assess a question about social media use, for example whether passive use is associated with negative sentiment.

Propose your own idea

Here are some examples of further ideas for you to see how a project plan can look like. Feel free to choose among these but research by yourself if this kind of data and methods are still available.

- Popularity assortativity of musicians in Spotify. Retrieving data about the latest popular musicians on a genre through the Spotify API, using the measurement of popularity given by Spotify. The project validates the popularity measurement by analyzing correlations with the number of followers of musicians. Then builds the collaboration network between musicians based on co-authorship links of appearing in the same song. The project presents descriptive statistics of the network and visualizations. The final question is tested by measuring assortativity with respect to popularity, assessing it with a permutation test.

- Testing Social Impact Theory in the case of COVID19. Measuring impact through Google trends volumes and number of tweets with terms related to the disease. Immediacy measured as the distance between countries and early sources of spreading. Number of sources measured as deaths in those countries. The project focuses on a series of countries where Twitter and Google are widely used, testing the theory through the relationship between impact, strength, and immediacy.

- Sentiment about Donald Trump and income at the regional level. Testing a hypothesis about a correlation between sentiment in tweets mentioning #Trump in a region correlated and its income and education levels. Used syuzhet and VADER. Manual annotation of some tweets showed limitations, for example some dictionaries have "trump" as a positive word, thus the sentiment analysis method was changed accordingly. Geographic location information based on text processing of user location. Test based on a regression model including two education levels from census data too. Robustness with two sentiment analysis methods but not "shopping around methods" including too many.

Frequently asked questions

- **Do I have to work in a group?**
  Yes, individual projects are not allowed. You can use collaborative tools like Github and videoconference technologies to coordinate your project. You can divide tasks as long as you all contribute to the project. Some exceptions about the group composition can be considered for groups composed only of doctoral researchers at TU Graz, contact us if that is your case.

- **Do we need to start with a question?**
  Yes, and this question has to be relevant about human behavior. Remember that this course is not data-driven, so please do not start with a cool dataset and then see what to do with it. In general, good research questions can be answered with a yes/no answer or with a number. For example: "are emotional tweets retweeted more often?" or "is there a correlation between Twitter misspelling frequency and unemployment?". Questions that start with "how" or "what" are usually not good, like "how do people use TikTok?" or "what is Austria's favorite Pokemon?".

- **Can we change the question or project topic?**
  Only before you present your project. Your project presentation has to include your question and you cannot change it afterwards. If you have problems regarding methods or data, you should document that in your project report, but not change the question because of that.

- **What if some dataset or method turns out to work differently than we planned?**
  This can always happen and there is no problem with changing some plans about the data retrieval or analysis, as long as you document these well in your report and argue about the changes. What you cannot change is the question and motivation of your project or do arbitrary, unjustified changes to the plan.

- **What if we don't get "nice" results or don't fit our expectations?**
  As long as you followed your plan, the particular results that you get will not factor into the grade. If you get a negative result or a surprising answer, comment on that in the conclusion section and reflect on it in your critique section. Follow our recommendations and feedback during the course and comment on them in your report if some reason you couldn't follow them.

- **Can our question be about methods rather than testing a hypothesis?**
  Yes, but make sure to motivate why those methods are important to the study of Computational Social Systems and how understanding them can help us to study

human behavior. For example, your question could be "can we adapt VADER to German?" or "is the Future Orientation Index stable across years?". You can also identify unresolved methodological problems and address them with empirical data in your project.

- **Can our question be about generating a new dataset?**
  Also yes, but explain why that dataset doesn't exist yet and which kind of analyses about human behavior it can enable. Focus your question and analyses on validating the quality of the dataset and providing descriptive statistics and visualizations that can help others to use the dataset. In this case, make sure that you make the dataset accessible in a Github repository or find another way to share the data.

- **Do we have to use data in the project?**
  Projects in this course have to be empirical and use data. Critical discussions are welcome in the last part of the project and theoretical motivations in the first one, but make sure that your project has an empirical and quantitative core.

- **Can we use a statistical method or model not covered by the course?**
  Yes, but make sure that you understand it well. Don't use a method just because someone told you to do so (for example on stack exchange) or because everyone in your field is using it, think critically and make sure that you don't just run some software and get a magical p-value or something like that. You should also briefly explain this method in your report, cite relevant references about it, and explain why you chose to use it.
- **How can we share our code and data with you?**
  You can share code and data by setting up a Github repository and including a link in your report. You will get extra points if the repository is well documented and might help other people in their analyses!

- **Can I use a proprietary dataset I cannot share?**
  That's OK if it answers your question well but you won't qualify to get extra points for open science practices. I often run the code of projects myself and I cannot do that if you can't share the dataset.

- **Can we access historical Twitter data?**
  Twitter API is not availble for now, but historical twitter data might be available from other research projects. Don't depend a project that would require historical Twitter data if you don't have access to it.

- **Can we compare countries on Twitter?**
  You can get some location data from Twitter if you process the "location" field of user profiles with a geocoder package. However, this can take very long and in general it is very hard to compare many countries with Twitter data. Also keep in mind language issues, especially if you plan to analyze text or run sentiment analysis. To compare countries you can use other data sources like Google Trends, Wikipedia page views if languages suit you, or even Facebook data from their official datasets or their ads API.

- **Can I do sentiment analysis in languages other than English?**
  There are resources for multiple languages, for example the translations of the NRC lexicon, the LabMT word lists, SentiStrength versions in many languages, or word lists with valence and arousal in languages like Spanish, French, Italian, or German. Keep in mind the validity of sentiment analysis and provide a careful assessment of the quality of your method for your dataset. If you can't find reliable information about it in previous papers, you can annotate a random sample of a few hundred texts from your dataset to provide an idea of how good the method is.

- **Can I use data from previous papers?**
  Yes, but we also want to see what you learned about data retrieval during the course. If you are going to replicate a previous paper, you will need to produce your own data and the project cannot be based only on data shared by the authors of the original paper. Do not rely on authors sharing any data beyond what is publicly available in a paper.

- **Do I have to use social media data or other Internet-related data?**
  No, but your data has to be relevant to a question about Computational Social Systems. You can also use data from surveys about online behavior, old big data like baby name statistics, or other kinds of quantified human behavior data like Google books or movie subtitles.

Project example

Project title: Political strength in Online Social Impact

Research question(s):

We aim to test the hypothesis of the role of strength in social impact as hypothesized by Social Impact Theory. We will study how the strength of a source on Twitter influences the aggregated impact it has on its followers. To do so, we will focus on US accounts of current and former members of congress and analyze their timelines, measuring strength as a comparison between two groups: current members as strong sources and past members as weaker sources. We will test the following hypotheses:

1. The mean number of retweets per tweet of twitter accounts of US politicians grows as a power of number of followers (replication of exercise).

2. The mean number of retweets per tweet of twitter accounts of current members of congress is higher than former members of congress.

3. The coefficient of the relationship between mean number of retweets per tweet and the number of followers is higher for current the twitter accounts of current members of congress than for former members of congress (multiplicative effect).

Planned data retrieval and analysis to address the questions

We plan the following steps

1. Based on an existing dataset (https://osf.io/mqhgp/), retrieve twitter user profile data and remove from the analysis any accounts that are set to private, have less than 100 followers, wrote less than 100 tweets, or had their last tweet more than a month ago.

2. Retrieve tweets posted in the last year by each of the remaining accounts.

3. Clean tweets by removing retweets, replies, and tweets younger than two days old.

4. Record the impact of each account as the mean number of retweets of their tweets. Produce a data frame per user with a column for the mean number of retweets, a second column for the number of followers, and a third with a code of whether they are current or former congress members.

5. Fit linear regression models to test each of the hypotheses. We will apply permutation tests to estimate the p-value of the hypotheses versus their corresponding null hypothesis and we will use bootstrapping to calculate the confidence intervals around the coefficients we are interested in.