Assignment 1

(updates will be posted on Piazza).

Objective: "Do something." Experiment with generating research ideas and designing simple measurements. Experience the challenges of working with conversational data. Get used to working within an open-source framework that encourages collaboration and exchange of resources. Kickstart productive research-oriented discussions.

Due dates (all deadlines refer to 5:00pm unless otherwise specified)

- By Wednesday, August 24, Enroll on the course Piazza page: https://piazza.com/cornell/fall2022/cs6742 the piazza password will be provided on the first day of class (and later listed in the CMS class description, request it by email if needed).
- Friday, August 26, 2:30pm: **Deliverable for A**
- Optional: Friday, August 26, Midnight: **B.1**
- ~~Thursday~~ [corrected Aug 25] Tuesday, August 30: **B.3** form groups on CMS. CMS group formation requires invitations and acceptance of invitations via the system, i.e., action by two people per person added; please check the official CMS documentation or this more graphically-oriented guide for instructions. I need the group information from CMS to schedule the group presentations.
- Thursday, September 1: **B.4**
- Wednesday, September 7: **C**
- Tuesday, September 13: **D**
- Tuesday, September 20: **E**
- Thursday, September 22: in-class presentations of results (more details will be announced, depending on the number of groups)

General notes:

- Don't panic!
- I realize this is abrupt and that many of you don't have any NLP experience. Grades will be based on creativity and good-faith effort to engage in this exercise, as well as on the quality of the discussions that emerge. It does not matter if the hypothesis you are proposing is finally disproved.
- It is neither required nor expected that your proposal for this assignment will relate to your final course project.
- Strive to post your first piazza post (Deliverable for A) in advance of the actual due date to (a) give time to your classmates to read your proposal and post feedback; (b) since you are encouraged to work in groups, early posting will facilitate linking up with classmates having similar interests.
- Continue to monitor and participate on Piazza. Example things to post: feedback on other people's proposals; some oddity of the datasets you've found that is worth alerting others to; unexpected early results that are interesting or that you need help interpreting.
- Basically, I would like us all to act as a team; *we're all in this together!*

Tasks:

A. Dataset exploration and thinking about questions (individual part of the assignment, although it's OK to start as a group if you already know who you want to work with)

A.1 Explore the list of conversational datasets provided on Piazza and skim the respective papers (you can focus on the dataset description and skip the technical details in the respective papers for now). Note that you don't need to download the dataset, just understand its characteristics.

A.2 Formulate a *guess* of how politeness might interact with the available metadata in *one* of the datasets (e.g., I expect short people to be more polite than long people in the Chats-Between-Tall-And-Short-Folks dataset). Explain why you made this guess, either based on your experience/intuition or based on literature you happen to know about (we do not expect you to go looking for literature at this point, but if you know of some then do include it).

A.3 Think about another hypothesis that might tie the metadata available in the dataset you selected with a signal (other than politeness) that might be present in the text (focus on a signal that you think would be easy-to-extract) . Explain why you have this intuition. This hypothesis can be inspired by the discussions we have in class, by your own experience, or by literature (including

the papers associated with the respective datasets). Make sure to give proper attribution to the source of this hypothesis.

Deliverable: Post a note (not a question) on Piazza under the folder "A1A" with the title:

"A1A. -- [name of dataset for A.2] -- [<5 words description of hypothesis in A.3]"

In the content of the note, include a short description of the dataset, a discussion of your A.2 explanation, a discussion of your A3 intuition, and initial thoughts on how you would extract the respective signal from the text. Make connections to the dataset specifics and (if applicable) to the discussions we have in class.

The length expectation is 3+ paragraphs; these paragraphs don't have to be long.

B. Group formation

B.1 Read other A1A piazza notes. Optionally, although strongly encouraged, post a reply to at least one note. Think about why you agree/disagree with the guess expressed within (e.g., does the post remind you of some of your own experience or about some literature?), do you think it's feasible to extract the respective signals? Propose joining forces where appropriate (there is an early-bird advantage to both posting your initial note early, and commenting early).

B.2 Participate in the in-class discussion summarizing these initial proposals.

B.3 Create groups of on CMS (size 2-3 is recommended). Each group should focus on a single dataset and one hypothesis. Take into account the feasibility of automatically extracting the signals from the dataset, given your coding skills and the timeline of this assignment.

B.4 Post a piazza note in the folder A1B as a group, clarifying which dataset you are working on, what is the hypothesis you want to investigate (and why) and propose a *simple* measure for the signal that is involved in your hypothesis. Try to make sure your proposal is different enough from other proposals (again an early-bird advantage).

Note: I really mean the measure should be simple, and this might require refining your hypothesis. It's OK to use measures that are mentioned in your respective dataset paper, but you will need to reimplement them from scratch. (Reproducibility is an important aspect of this field, so checking whether you can reproduce some existing results about the dataset is valuable.) You can not use a measure that is already implemented in ConvoKit.

C. Re-format dataset

C.1 Familiarize yourself with ConvoKit toolkit by reading the Core Concepts and High-Level Tutorial part of the documentation (https://convokit.cornell.edu/documentation/).

C.2 Transform your corpus into the ConvoKit format (example script is provided in ConvoKit) and test its integrity by comparing it to the original. On CMS submit a zip file of the corpus in ConvoKit format and a zip file with the (documented) code you used for the re-formatting.

C.3. Thoroughly document your corpus following the example documentation format used ConvoKit documentation file in .rst format. On CMS submit this .rst file.

Note: Since your re-formatted dataset will be shared with other students in the class, it is important to thoroughly check it and clearly document it.

D. Explore your corpus

D.1 Extract simple statistics (e.g., number of conversations involving at least k tall people). Post a follow-up on your A1B Piazza post discussing the statistics that are most relevant to your hypotheses (as well as any inconsistencies you might discover).

D.2 Apply the ConvoKit politeness transformer to your data and explore the relationship between the extracted strategies and the metadata. Post a follow-up on your A1B Piazza post discussing these results and how they match/do not match your intuition.

E. Implement measure(s) and apply it (them) to two datasets

E.1 Implement and apply your measure(s) to your dataset (We recommend implementing your measure as a ConvoKit transformer, this will encourage re-usability in the class and beyond.)

E.2 Apply your measure to your dataset and post a followup on your A1B Piazza post.

E.3 Pick another ConvoKit-formatted dataset prepared by another group that you think will play well with your measure (I will distribute these following the deadline for C). If everything was well documented and implemented correctly, applying your new measure to another dataset should be seamless. Briefly discuss your observations as a follow-up on your A1B piazza post.

Note: For this assignment is fine (and even expected) to update your measure and hypotheses as you progress in your understanding of the data. Make sure to document these changes (keeping the previous versions) by updating your Piazza post. You'll need to discuss the reasons for these changes in your presentation as well.

Academic Integrity Academic and scientific integrity compels one to properly attribute to others any work, ideas, or phrasing that one did not create oneself. To do otherwise is fraud.

Certain points deserve emphasis here. In this class, talking to and helping others is strongly encouraged. You may also, *with attribution*, use the code from other sources. The easiest rule of thumb is, *acknowledge the work and contributions and ideas and words and wordings of others*. Do not copy or slightly reword portions of papers, Wikipedia articles, textbooks, other students' work, Stack Overflow answers, something you heard from a talk or a conversation or saw on the Internet, or anything else, really, without acknowledging your sources. See "Acknowledging the Work of Others" in The Essential Guide to Academic Integrity at Cornell and http://www.theuniversityfaculty.cornell.edu/AcadInteg/ for more information and useful examples.

This is not to say that you can receive course credit for work that is not your own — e.g., taking someone else's report and putting your name at the top, next to the other person(s)' names. However, violations of academic integrity (e.g., fraud) undergo the academic-integrity hearing process on top of any grade penalties imposed, whereas not following the rules of the assignment "only" risks grade penalties.