Prof Spirling
Text-as-Data: DS-GA 1015
March 12, 2020

Final Paper: A Guide

Final Paper (50%): There will be a final written paper of not longer than 10 double spaced pages of text, which explores an original research project or idea. This may be substantive or technical in nature. You are encouraged to work in teams of up to two people on this paper. The deadline for the paper will be May 8, 2020 with no extensions or exceptions.

Some possibilities for the final paper include, but are not limited to:

- Writing a paper that applies some of the methods discussed in class to a substantive social science question with real data.

- Building an original social science dataset from textual data and conducting (at least) some preliminary analysis.

- Developing a new model or procedure for a social science problem or subject area or dataset. This can include obtaining substantial improvement over a previous approach to a problem, or uncovering new results in a given area.

We will <u>not</u> distribute data sets, but you can ask us to help you find relevant data. Alternatively, you can use your own, or one that already exists and that you have access to at NYU. For the paper 'social science' can be thought of very broadly: we mean the study of human beings interacting with each other. Thus, anything from linguistics, to politics and political science, to economics, to marketing, to anthropology, to sociology is acceptable subject matter. But, whatever you choose, you must connect (i.e. cite correctly) to a literature in social science and data science.

In general, papers will ask questions like "how does X affect Y?", "what is the best way to measure X when studying X?", "what can we learn from X, Y and Z about the behavior of these types of people?", "how can we improve our understanding of the relationship between X and Y".

There is no requirement that you find a 'positive' (or statistically significant) result: the idea of the paper is to make you think carefully about text-as-data, data science and social science—not to obtain a bullet-proof finding.

It will be helpful to think about your paper as having some variation of the sections below:

- *Introduction*: what is the topic of interest and why is it of interest? Broadly, what is the paper about, and what will it suggest?

- *Literature*: what do we know about this topic so far? Who has worked on this, and how does their work relate to yours? how do we usually approach this problem methodologically? what types of data have been used?

- *Theory and Hypotheses*: if this is a regression-style inference problem, how do you think the dependent and independent variables relate? What precisely—i.e. what type of relationship—will you be testing? If you are attempting to simply explore or summarize the data, what do you expect to find in terms of latent structure? If it's a methods paper, how or why to you expect your (new) idea for modelling to work (better)?

- *Data and Methods*: describe your data and what methods/techniques/approaches you are using to test your theory or explore the data. What's new about the data and methods? Have they been applied to this type of problem before? If not, why not? If yes, why were they unsatisfactory?

- *Results*: what did you find? How does it relate to your hypotheses/priors?

- *Discussion*: summarize the findings and explain why they matter. What further work needs to be done? how have you contributed to the debate or literature in this part of data science and/or social science?

The paper should be double-spaced, 12pt font, 1-inch (or so) all round margins, **no more than 2700 words**. You must print a word count on the first page of your paper.