

### Assignment 0

**Submission Deadline: preferably a day before we meet on this topic. Worst case: 31st October 2020**

**Description:** This assignment consists of two parts. First part involves reading a paper and answering a question about it. Second part is a open question.

**Submission Instructions:** Since we are all \*trustworthy\* adults, to make it simple, you can all upload your assignments as \*\*a single pdf file\* with the filename in "Your last name"\_"Assignment0.pdf" in the folder Assignments/Assignment0 in our shared, private google drive (whose link is communicated separately via email). Everyone can see everyone's submissions. Your submissions should be no more than 2 pages long, and in a legible/big enough font.

1. Jena Organization Corpus (JOCO) is a freely available corpus of annual reports (ARs) and corporate social responsibility (CSR) reports of US American, British and German business organizations. It is intended to be used as a community effort to improve NLP techniques for economics domain. More details on the corpus are in the paper described below:

*Sebastian G.M. Hndschke, Sven Buechel, Jan Goldenstein, Philipp Poschmann, Tinghui Duan, Peter Walgenbach, and Udo Hahn. 2018. A Corpus of Corporate Annual and Social Responsibility Reports: 280 Million Tokens of Balanced Organizational Writing. In ECONLP 2018 Proceedings of the First Workshop on Economics and Natural Language Processing @ ACL 2018. Melbourne, Australia, July 20, 2018. Pages 2031. <https://www.aclweb.org/anthology/W18-3103/>*

Your task in this assignment is to read the paper, and explore the corpus if you want (it is in our shared space in a folder called Corpora), and write how you think this corpus can be useful for research at the intersection of economics and NLP.

Note: JOCO corpus is shared with you with a permission for use in this course. Do not share it with others or use it later. If you want to continue using it, sign the (free) license agreement with Uni Jena's team and get it ([https://www.orga.uni-jena.de/en/jena+organization+corpus+\(joco\)](https://www.orga.uni-jena.de/en/jena+organization+corpus+(joco)))

2. What are some sources of textual data for economics researchers? If you don't have a dataset at hand, how will you go about compiling your corpus for a given research problem needing NLP?