GPU application

Requested specifications

How much data will be stored?

- Input file: ~ 1GB
- Output files: < 100 MB

For how long should the data be stored?

- 1 month

How many cpus?

- 8 kernels

How much ram?

- ~ 16 GB

Collaborators

Gustav Huitfeldt Kragelund Helms
BSc student in Cognitive science
201907348@post.au.dk
AU639185

Morten Ravnsgaard Gade
BSc student in Cognitive science
201706692@post.au.dk
AU592075

Project title

Measuring political polarization in parliamentary debates from the Danish Folketing

Abstract

In this paper we firstly investigate political polarization in the Danish Folketing and secondly the intergroup behaviour of the parties. Political polarization is extrapolated by using a SGD classifier and intergroup behavior with the wordshoal method.

Description of the specific research activities

We plan to do two separate types of calculations with the GPU. Firstly, we would like to conduct a SGD classifier and secondly we would like to use the GPU to calculate wordshoal estimates. We have scraped transcripts of debates in the folketinget from 1990 - 2021. These 31 years of transcripts approximately contain ~800.000 speeches with a file size of ~1 GB.

We will first use these speeches to calculate political polarization within the Danish Folketing. Most recent literature within the field of measuring political polarization from transcripts argues that the best method for obtaining political standpoints is to use a machine learning (ML) classifier to classify party labels from speeches . The accuracy of the classifier will then be the estimated level of polarization (Goet, 2019; Peterson & Spirling, 2018). The method has been used in the British parliament (Goet, 2019), the Irish Dail and the US senate (Peterson & Spirling, 2018) with promising results. The method has not yet been conducted on the Danish Folketing which is what we intend to do in this project. We have so far conducted the analysis on a subset of the data but plan to scale it up to all the data. The analysis is conducted in python with Jupyter Notebook. We use the scikit-learn library to implement a ML model with a stochastic gradient descent (SGD) classifier. We use a stratified k-fold to split the data 3 times into train and test sets for each year. The model then uses a grid search to optimize hyperparameters. There are in total five different settings that are cross validated 10 times each. The model is optimizing the hyperparameter on Cohen's-kappa score. This means that the model is fitted and validated 10 times * 5 settings = 50 times per fold, which in total gives 150 fit per year. This gives in total 150 * 31 years = 4650 model fits. The summarized and exported results will have a file size of a couple of MB.

The second thing we intend to do is to conduct a Wordshoal analysis on the scraped speeches. Wordshoal is a method that distributes all speakers on a one-dimensional spectrum based on overlap in word frequencies. The estimates provided by the wordshoal method can be interpreted as how different the parties speak compared to the other parties but also how alike they are within the parties (Peterson & Spirling, 2018). The estimates then make it possible to investigate how these relationships change over time and whether they are interrelated. We plan to use this to test predictions made from the optimal distinctiveness theory (ODT) on intergroup behaviour (Leonardelli et al., 2010). The wordshoal analysis will be conducted in R using the package *wordshoal,* which was developed by the authors of the paper that first presented the method (Peterson & Spirling, 2018). The *wordshoal* package is built as an extension model to the *quanteda* framework from which it is also operationalized. The output file of the wordshoal estimates will only have a file size of less than 50 MB.

List of software needed:

- Python: *pandas, numpy, scikit-learn*

- R: *tidyverse, corpus, lubridate, quanteda, wordshoal, quanteda.textmodels, quanteda.corpora.*

Literature

Goet, N. D. (2019). Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811–2015. *Political Analysis*, *27*(4), 518–539. https://doi.org/10.1017/pan.2019.2

Leonardelli, G. J., Pickett, C. L., & Brewer, M. B. (2010). Optimal distinctiveness theory: A framework for social identity, social cognition, and intergroup relations. In *Advances in experimental social psychology* (Vol. 43, pp. 63–113). Elsevier.

Peterson, A., & Spirling, A. (2018). Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems. *Political Analysis*, *26*(1), 120–128. https://doi.org/10.1017/pan.2017.39