Description

Hey Maxime. Happy to hear that you are interested in this workshop. I compiled some info on the dataset/corpus and what could be interesting to delve into during the workshop.

**Info on “Aktstykker”**

Aktstykker are how ministries get funding for specific projects that are not already specified on the “Finanslov”.

[Aktstykke / Folketinget (ft.dk)](https://www.ft.dk/da/leksikon/Aktstykke)

They contain metadata such as a grant date and the ‘sector’ also represented in the dataset.

[Aktstk. 122 - 2023-24 (oversigt): Tidligere fortroligt aktstykke D af 11. november 2021 om anskaffelse af nyt it-system til SU / Folketinget (ft.dk)](https://www.ft.dk/samling/20231/aktstykke/aktstk.122/index.htm) (metadata from page)

In the actual aktstykke you’ll see both some context for the policy/project at hand, expected project costs and some tables with valuable information (although I am not sure what the tables look like in the dataset my colleagues have gathered).

[Aktstykke nr. 122, Folketinget 2023-24, Tidligere fortroligt aktstykke D af 11. november 2021 om anskaffelse af nyt it-system til SU (ft.dk)](https://www.ft.dk/RIPdf/samling/20231/aktstykke/aktstk122/20231_aktstk_anmeldt122.pdf) (actual aktstykke, the text here is the corpus)

**What could be relevant for the workshop?**

There are many applications for quantitative text analysis in our corpus, but we’ll have to narrow it down for the workshop. These tasks could be relevant to us, among others.

* Extracting specific information such as references to other ‘aktstykker’, dates, tables, account numbers, and so on.
  + Some of this is substantial information we are interested in
  + Others (such as account numbers or references) is useful in order to merge these datasets with data from other soures.
* Classifying – is this going to fund a building project, an IT infrastructure project, and so on?
* Which actors (institutions, committees, etc) are present in the corpus?

No matter which information we extract/analyze, some preprocessing or “bread & butter” work would be useful. What is good practice in terms of stemming, lowercasing, and so on? which R-packages are good to use here?

I think we are very much open to which tasks we focus on and which methods we use. What’s most important is that it is applicable and accessible to the participants. There is a distribution on how R- savvy the participants are. While some are very proficient and have been doing research and big projects in R, others are a rusty and have only used it a little bit. It’s probably good to make the coding simple when possible, but otherwise to strive for a ‘medium’ difficulty. We can discuss all this ☺