Task for junior SWERIK research engineer

Robert Borges, Måns Magnusson, Fredrik Norén

The Riksdagen Records (kammarens protokoll/riksdagsprotokoll) contain minutes of different types of utterances and actions conducted in the parliament chamber. Currently, the SWERIK corpus of records provides relatively little section segmentation of the records. Your task will be to make an attempt at annotating one type of section in the records.

In the Swedish parliament, MPs are able to pose questions to the government. Questions can be posed in written or oral form, but they are answered orally in the chamber, and there is an opportunity within the question format for the MP to respond to the minister's answer.

The task:
- identify these sections containing answers to questions (skriftliga frågor/enkla frågor) + oral responses programmatically
- create a CSV-file with the IDs of the utterances (<u> elements) you identified as a response to a question
- create a line graph showing the number of responses per year in the corpus
- outline the next steps (based on this graph and your experience so far) – How would you take further steps to improve precision and recall of "Svar på fråga" sections?

You should work with the records corpus that can be found here:
https://github.com/swerik-project/riksdagen-records

You should be able to use either the R package here:
https://github.com/swerik-project/rcr

Or the Python library here:
https://github.com/swerik-project/pyriksdagen

Send your code, the CSV file containing the XML IDs, a brief description (max 1 page) of what you did, and an outline of the next steps. The code should be in Python or R and able to run on the corpus to reproduce your results (the CSV file). Also, state whether you are okay with us using your results for the actual corpus.

You are free to use any strategy and packages you like. We are most interested in how you approach the task and evaluate the result. You should spend around 1–2 hours on this task. Results should be submitted to us no later than June 2 at 23.59.