Project Proposal
AlexElin = Alex Hermansson and Elin Samuelsson

Problem description

We want to use clustering to find correlations in how the members of the Swedish parliament have been voting using clustering.

For instance, we expect to see some correlation between a member's votes and which party he/she belongs to. We also expect some parties to have more similar voting patterns than others. In this project we aim to validate these hypotheses and possibly draw other conclusions.

Tools

We plan to use Spark SQL to create features of interest, which in this case are the votes (Yes = 1, Refrain = 0, No = -1) of each member over several polls. Thus, we have the ability to use both the procedural and relational API for flexibility and efficiency. After that, we will use MLlib to do the actual clustering.

Data

Our plan is to use data from the swedish parliament. The dataset consists of votes from members of the parliament along with metadata such as political party, name etc. All data is spread over multiple files formatted in JSON, one file for each voting session. It can be found at: https://data.riksdagen.se/data/voteringar/

Methodology and Algorithm

We will use a clustering algorithm from the Spark MLlib. We will have to investigate which specific algorithm is more appropriate, either Gaussian Mixture or K-Means.

When we have obtained results, we will either plot or display as tables using standard python libraries such as Seaborn or Matplotlib.