Assignment: Senators

Wednesday January 18, 2017

Technologies: bokeh, Jupyter, k-means – whatever else you can think of

In today's assignment, we will look at historical voting records of US senators. The final goal of the assignment is to use bokeh and k-means clustering (or other ML tools) to create a (well-documented) Jupyter notebook that visualizes the historical development of the political divide of the US.

You have already seen how to use bokeh in its simplest form in the tutorial. Further examples of how bokeh can be used are shown here. We also covered how to use sklearn to cluster – other clustering methods in sklearn are described here, but we're unlikely to need any of them, except for maybe spectral clustering. Let us know look at a (much) bigger data-set. Unfortunately, we will have to work a little harder to get it.

Parsing

We will use data from Voteview.com. For each senate, the data-set contains a text-file that may look confusing at first – the specifications of the files are given on each site, e.g. here. You should use the Python-module urllib2 to download each data-set.

Recording

While our data-set is much larger than the ones we have used before, it is still small enough for sqlite3 to handle. You should create the following tables:

1. a table of all members of the senate
2. a table of all roll call votes

The first should have an ID as well as information about each session each senator was in and what party the senator was affiliated with in each session (be aware that senators sometimes change party affiliation – for the purpose of this assignment, it should be enough to find an affiliation within each session). The latter should have an ID for each roll call and the information in which session it happened.

Finally, create a table with the votes for each roll call: for each senator who was in session for a particular roll call, this table should contain the information whether the senator voted yea, nay, or abstained.

Visualizing

This part of the exercise is open-ended. I want you to use the data you have to investigate the question *if the US (Senate) is the most divided it has ever been*, as was reported by Forbes in 2013. While there may be no right or wrong answer to this question, there certainly are *wrong ways* of answering the question. If you are unsure about what trends you want to show, talk to me before starting to code. Further, try to adhere to the following principles in your visualization:

1. Do not use more *dimensions* (color, shape, size, etc.) than you need. Having circles appear both larger and darker due to more divisiveness is redundant.

2. Convey *only* relevant information – think of the message your graphic is meant to present; for each piece of information, ask yourself if the graphic would work equally well without!

3. Make use of bokeh's functionalities!

This assignment is meant to take you a while – correspondingly, I expect high quality solutions. Aim for a notebook that could be added to the bokeh gallery!