1. Data PreProcessing:

(1) I download the lobbying data that is in XML format from United States Senate web site:
http://www.senate.gov/legislative/Public_Disclosure/database_download.htm.

(2) I convert each individual XML file into JSON file. I employ XML2JSON in python to finish this job, and the github link for this tool is:
https://github.com/hay/xml2json.

(3) One problem for this lobbying data resource is there are tens of small XML files for each quarterly data, so I write a shell script "tojson.sh" that can convert them to JSON files at first. After that, I write another shell script file "cat.sh" that can concatenate small JSON files as a big JSON ARRAY and save the array in a single JSON file – "lobbying.json".

2. Analysis with IBM Bluemix Apache Spark -- Lobbying.ipynb

The analysis processes and results are shown in ipython notebook file -- Lobbying.ipynb.