Text Analysis of Senate Bills

# Background

The federal government of the USA is divided into three branches by the US Constitution in order to establish a system of “checks and balances.” The three branches of government are the Congress, Executive and Judicial, each of which executing its own roles in the creation and passage of legislation. Congress, which is divided into the House of Representatives and the Senate, is the branch responsible for scripting bills before they are assessed and voted on by each individual branch[[1]](#footnote-1).

While the public may be interested in keeping up with legislation being introduced by Congress, bill text is often written in a manner that is inaccessible to general comprehension. Further, due to corruption in Congress, authors of bills will on occasion include riders[[2]](#footnote-2), additional provisions that have little or nothing to do with the subject matter of the bill, in their bills in order to pass additional legislation or prevent the passage of the bill[[3]](#footnote-3). Finally, a large percentage of introduced bills do not become laws. Therefore, for this project, I wanted to explore the use of statistics and text analytics in predicting the success of a bill and therefore creating a model to predict bills to which citizens should pay more attention. This project is divided into three components: exploratory statistics, predicative analysis using text analytics and summarization.

## Exploratory statistics

For the exploratory statistics component of this project, I wanted to explore whether or not certain features were correlated with bill passage, as well as if there are trends in bill introduction by state and political party. Some of the questions explored include:

* Are senators from one party more prolific in writing bills than the others?
* Which states produce the most bills?
* Are certain features predicative of a bill being passed?

## Predicative statistics using text analytics

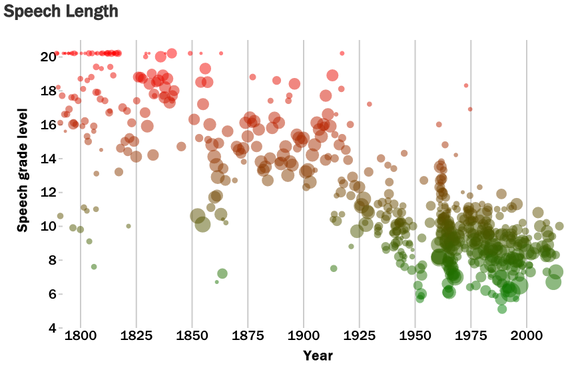
While some features of a bill are already known to predict the probability of passage of a bill[[4]](#footnote-4), I wanted to explore the use of text analytics and machine learning in bill passage prediction. Additionally, can text analytics be used to find differences between bills written by each party?

## Summarization and Readability Score

Time permitting, there are additional analyses of that may be of interest and use. One such application would be to create a summarization tool, using algorithm similar to that of autotldr[[5]](#footnote-5), for the bills. This would provide interested parties with a quick and readable summary of each bill, while removing potentially confusing jargon and formatting.

Another potential extension of interest would be using algorithms developed by Flesch and Kincaid to conduct a readability test of on the bills and determine how they have changed over time. As noted in an essay published in The Atlantic, while presidential election speeches use to be given at a college level, the length and complexity of presidential speeches have decreased to roughly a 8th grade reading level, as measured using the Flesch-Kincaid readability test[[6]](#footnote-6) (Figure 1). For this analysis, a new dataset would have to be collected, containing a smaller, random sample of bills from each year available, rather than all available bills for a smaller subset of years, which will be used for the aforementioned analyses and discussed further in the Dataset section below.

#### Figure 1. Reading level of presidential speeches throughout history, as reported by *The Atlantic*



### Problem Statement

Can text analytics be used to predict the success or failure of a bill passed by the Senate since 2000? What other features are predicative of the success of a Senate bill? Can summarization be used to synthesize Senate bills?

# Dataset

While multiple APIs are available to obtain government data, I used a dataset obtained from the [Congressional Bills Project](http://www.congressionalbills.org/download.html). This dataset contains data on Congressional bills written from Senate number 93 to 113 (current senate number is 114). The variables contained within this dataset are outlined in Table X. The Congressional Bills Project dataset codebook can be found in Appendix 1.

##### Table 1. Dataset variables used

|  |  |
| --- | --- |
| ID | Description |
| BillID | In the form "Cong-BillType-BillNum" |
| BillNum | Bill number |
| BillType | Bill type ("HR" or "S"). |
| ByReq | Introduction by request of an agency? |
| Chamber | 0 for House, 1 for Senate |
| Commem | Commemorative bill? |
| Cong | Congress of introduction |
| Cosponsr | Number of cosponsors |
| IntrDate | Date of introduction |
| Major | Major Policy Agendas Project topic code |
| Minor | Minor Policy Agendas Project topic code |
| Month | Month of introduction |
| Mult | Multiple referrals? |
| MultNo | Number of referrals |
| PassH | Passed House? |
| PassS | Passed Senate? |
| PLaw | Public law (passed both and signed)? |
| PLawDate | Date bill became public law |
| PLawNum | Public law number of bill |
| Private | Private issue bill? |
| ReferArr | Referred to numbered committee. Committee codes come from the Policy Agendas Project committee codebook |
| ReportH | Reported by a House committee? |
| ReportS | Reported by a Senate committee? |
| Title | Full bill title |
| Veto | Vetoed by the President? |
| Year | Year of introduction |
| Age | Age of member in year |
| Class | Class of introducing Senator |
| ComCArr | Chair of numbered committee? Committee codes come from the Policy Agendas Project committee codebook. Note: This variable may contain errors. The ComC variable is constructed from another more reliable source. |
| ComC | Chair of any committee? |
| ComMArr | Member of numbered committee? Committee codes come from the Policy Agendas Project committee codebook. Note: This variable may contain errors. The Mref variable is constructed from another more reliable source. |
| ComRArr | Ranking member of numbered committee? Committee codes come from the Policy Agendas Project committee codebook. Note: This variable may contain errors. The ComR variable is constructed from another more reliable source. |
| ComR | Ranking member of any committee? |
| CumHServ | Cumulative years of House service at intro. |
| CumSServ | Cumulative years of Senate service at intro. |
| Delegate | Delegate (from Guam, D.C., etc.)? |
| District | District of House member |
| DW1 | First dimension DW-NOMINATE score |
| DW2 | Second dimension DW-NOMINATE score |
| FrstConH | First Congress of House service |
| FrstConS | First Congress of Senate service |
| Gender | 0 for male, 1 for female |
| LeadCham | Leader of a chamber (leader or whip)? Double check this variable before using. |
| LeadSubC | Leader of any subcommittee? Double check this variable by crosstabilting with the SubChref variable. |
| Majority | Member of majority party? |
| MemberID | In the form "PooleID-Cong-Party" |
| MRef | Member of any committee to which the bill was referred? |
| NameFirst | First name |
| NameFull | Full name |
| NameLast | Last name |
| Party | Party code |
| PooleID | Poole's revised ICPSR number |
| Postal | Postal code of home state (e.g. WA) |
| State | Numeric code of home state |
| URL | link to Congress.gov information about the bill |
| ChRef | Chair of any committee to which the bill was referred? 101st Congress to present (from CQ.com) |
| RankRef | Ranking Member of any committee to which the bill was referred? 101st Congress to present (from CQ.com) |
| SubChRef | Subcommittee chair of any committee to which the bill was referred? 101st Congress to present (from CQ.com) |
| SubRankRef | Ranking member of any committee to which the bill was referred? 101st Congress to present (from CQ.com) |

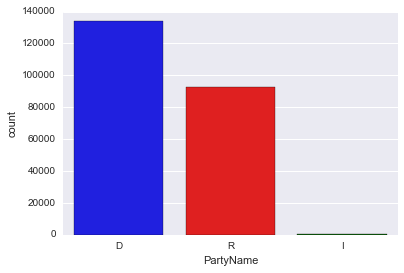
While this dataset contained pertinent information on the bills’ author, passage, title, etc., it does not contain data on the bill text. In order to collect the text from each individual bill, a web scraper was written in order to obtain the text for each individual bill from <https://www.congress.gov/>. Due to the large volume of data (each Senate number produced between 3000 and 4500 bills) collected using the web scraper, I decided to limit the web scraping to Senate number 106 to 109 for my analysis. These bills were then stripped of newline and other formatting characters and appended to the original dataset.

# Findings

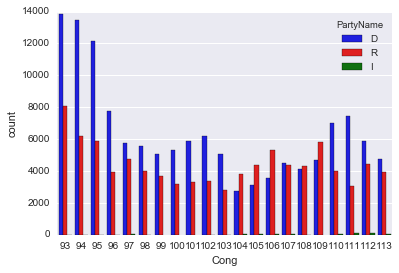
## Exploratory Statistics

* Analyses of bill features by time and if bill passed or not.
* Are senators from one party more prolific in writing bills than the others?

**Figure X. Bill Authorship by Political Party, All Data 1993-2014**



**Figure X. Bill Authorship by Political Party, Congress Number**



Map of USA with number of bills produced

* Are certain features predicative of a bill being passed?

Findings of Logistical regression analyses.

## Predicative Statistics

* Details of my Natural language processing analysis

# Challenges and Successes

## Data Collection and Web Scraping

By far, the most significant challenge of this project thus far has been the acquisition of the actual bill text. I explored the potential of using an API in order to acquire the bill data (excluding bill text), as two APIs were found: the Sunlight API and the US Congress API. I opted not to use the APIs and instead found the data that I was looking for compiled at the aforementioned [Congressional Bills Project](http://www.congressionalbills.org/download.html).

Building the web scraper took multiple iterations until I was able to arrive at one that produced the exact data that I needed- and of course, the final iteration was less than 10 lines of code.

# Possible Extensions

* Possible extensions or business applications of your project

# Conclusions

## Appendix 1. Dataset Codebook

From <http://www.congressionalbills.org/codebooks.html>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Id |  | Unsigned integer | | Description |
| BillID |  | Text |  | In the form "Cong-BillType-BillNum" |
| BillNum |  | Unsigned integer | | Bill number |
| BillType |  | Text |  | Bill type ("HR" or "S"). |
| ByReq |  | Boolean |  | Introduction by request of an agency? |
| Chamber |  | Boolean |  | 0 for House, 1 for Senate |
| Commem |  | Boolean |  | Commemorative bill? |
| Cong |  | Unsigned integer | | Congress of introduction |
| Cosponsr |  | Unsigned integer | | Number of cosponsors |
| IntrDate |  | Date |  | Date of introduction |
| Major |  | Unsigned integer | | Major Policy Agendas Project topic code |
| Minor |  | Unsigned integer | | Minor Policy Agendas Project topic code |
| Month |  | Unsigned integer | | Month of introduction |
| Mult |  | Boolean |  | Multiple referrals? |
| MultNo |  | Unsigned integer | | Number of referrals |
| PassH |  | Boolean |  | Passed House? |
| PassS |  | Boolean |  | Passed Senate? |
| PLaw |  | Boolean |  | Public law (passed both and signed)? |
| PLawDate |  | Date |  | Date bill became public law |
| PLawNum |  | Text |  | Public law number of bill |
| Private |  | Boolean |  | Private issue bill? |
| ReferArr |  | Array of booleans | | Referred to numbered committee. Committee codes come from the Policy Agendas Project committee codebook |
| ReportH |  | Boolean |  | Reported by a House committee? |
| ReportS |  | Boolean |  | Reported by a Senate committee? |
| Title |  | Text |  | Full bill title |
| Veto |  | Boolean |  | Vetoed by the President? |
| Year |  | Unsigned integer | | Year of introduction |
| Age |  | Unsigned integer | | Age of member in year |
| Class |  | Unsigned integer | | Class of introducing Senator |
| ComCArr |  | Array of booleans | | Chair of numbered committee? Committee codes come from the Policy Agendas Project committee codebook. Note: This variable may contain errors. The ComC variable is constructed from another more reliable source. |
| ComC |  | Boolean |  | Chair of any committee? |
| ComMArr |  | Array of booleans | | Member of numbered committee? Committee codes come from the Policy Agendas Project committee codebook. Note: This variable may contain errors. The Mref variable is constructed from another more reliable source. |
| ComRArr |  | Array of booleans | | Ranking member of numbered committee? Committee codes come from the Policy Agendas Project committee codebook. Note: This variable may contain errors. The ComR variable is constructed from another more reliable source. |
| ComR |  | Boolean |  | Ranking member of any committee? |
| CumHServ |  | Floating point | | Cumulative years of House service at intro. |
| CumSServ |  | Floating point | | Cumulative years of Senate service at intro. |
| Delegate |  | Boolean |  | Delegate (from Guam, D.C., etc.)? |
| District |  | Unsigned integer | | District of House member |
| DW1 |  | Floating point | | First dimension DW-NOMINATE score |
| DW2 |  | Floating point | | Second dimension DW-NOMINATE score |
| FrstConH |  | Unsigned integer | | First Congress of House service |
| FrstConS |  | Unsigned integer | | First Congress of Senate service |
| Gender |  | Boolean |  | 0 for male, 1 for female |
| LeadCham |  | Boolean |  | Leader of a chamber (leader or whip)? Double check this variable before using. |
| LeadSubC |  | Boolean |  | Leader of any subcommittee? Double check this variable by crosstabilting with the SubChref variable. |
| Majority |  | Boolean |  | Member of majority party? |
| MemberID |  | Text |  | In the form "PooleID-Cong-Party" |
| MRef |  | Boolean |  | Member of any committee to which the bill was referred? |
| NameFirst |  | Text |  | First name |
| NameFull |  | Text |  | Full name |
| NameLast |  | Text |  | Last name |
| Party |  | Unsigned integer | | Party code |
| PooleID |  | Unsigned integer | | Poole's revised ICPSR number |
| Postal |  | Text |  | Postal code of home state (e.g. WA) |
| State |  | Unsigned integer | | Numeric code of home state |
| URL |  | Hyperlink |  | link to Congress.gov information about the bill |
| ChRef |  | Boolean |  | Chair of any committee to which the bill was referred? 101st Congress to present (from CQ.com) |
| RankRef |  | Boolean |  | Ranking Member of any committee to which the bill was referred? 101st Congress to present (from CQ.com) |
| SubChRef |  | Boolean |  | Subcommittee chair of any committee to which the bill was referred? 101st Congress to present (from CQ.com) |
| SubRankRef |  | Boolean |  | Ranking member of any committee to which the bill was referred? 101st Congress to present (from CQ.com) |

1. https://www.usa.gov/branches-of-government [↑](#footnote-ref-1)
2. https://en.wikipedia.org/wiki/Rider\_(legislation) [↑](#footnote-ref-2)
3. https://www.quora.com/Why-are-some-seemingly-irrelevant-issues-piggybacked-onto-bills-going-through-the-voting-process-in-Congress [↑](#footnote-ref-3)
4. https://www.govtrack.us/blog/2012/04/08/even-better-bill-prognosis-now-with-real-probabilities/ [↑](#footnote-ref-4)
5. https://np.reddit.com/r/autotldr/comments/31bfht/theory\_autotldr\_concept [↑](#footnote-ref-5)
6. http://www.theatlantic.com/politics/archive/2014/10/have-presidential-speeches-gotten-less-sophisticated-over-time/381410/ [↑](#footnote-ref-6)