Question 1

For the beginning to make a language model of a collection we needed to define separators. To make that we used the most common separators, like punctuation marks and numbers. Then we added more as we continued.

After that we removed all the words with length 2 or less, because there were a lot of single letter words that didn’t make sense. We used several functions to create Language models and to get results that we can analyze. Laplace smoothing is necessary for the model like this to be able to process queries. To analyze the 20 most “popular”(highest probability according to the model) terms were taken from each document.

First results weren’t good. Almost all the documents had “the”, “and” as the top 2 most popular terms. Majority of the documents models consisted mostly of the stop words.

A Model of Political Bias in Social Science Research1.txt:['the:0.014713', 'and:0.0059516', 'news:0.0035842', 'bias:0.00354', 'that:0.0030754', 'for:0.0029647', 'media:0.002301', 'coverage:0.0019027', 'from:0.0016815', 'The:0.0015709', 'Groeling:0.0012169', 'stories:0.0011947', 'this:0.0011947', 'are:0.0011726', 'partisan:0.001062', 'which:0.0010177', 'Polit:0.0009956', 'more:0.0009956', 'org:0.0009956', 'www:0.0009956']

A Model of Political Bias in Social Science Research2.txt:['the:0.0073354', 'and:0.0070104', 'that:0.0035052', 'social:0.0018571', 'are:0.0016946', 'for:0.0016713', 'doi:0.0016713', 'bias:0.001416', 'Jussim:0.0013232', 'The:0.0011839', 'political:0.0011374', 'biases:0.0011374', 'this:0.001091', 'not:0.0010446', 'from:0.0009517', 'Psychology:0.0008821', 'conservatives:0.0008589', 'can:0.0008357', 'may:0.0008125', 'findings:0.0007892']

AMERICANS’ VIEWS OF POLITICAL BIAS IN THE ACADEMY AND ACADEMIC FREEDOM.txt:['the:0.0068452', 'and:0.004297', 'percent:0.0037224', 'that:0.0021485', 'are:0.0018237', 'professors:0.0016988', 'with:0.0016239', 'those:0.001374', 'college:0.0012491', 'for:0.0012241', 'say:0.0011492', 'respondents:0.0009493', 'this:0.0008744', 'The:0.0008494', 'have:0.0008494', 'about:0.0008244', 'higher:0.0007994', 'Americans:0.0007745', 'education:0.0007745', 'political:0.0007745']

An Empirical Study of Political Bias in Legal Scholarship.txt:['the:0.0106614', 'and:0.006317', 'that:0.0050758', 'articles:0.0026155', 'law:0.0024603', 'our:0.0022387', 'professors:0.0021057', 'are:0.0021057', 'for:0.0017954', 'The:0.0017954', 'liberal:0.0015737', 'not:0.0014407', 'ideological:0.0013964', 'Law:0.0013964', 'with:0.0013077', 'political:0.0013077', 'scholarship:0.0012856', 'legal:0.0012634', 'net:0.0011969', 'bias:0.0011083']

Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity.txt:['the:0.0058141', 'and:0.0031449', 'bias:0.0027749', 'for:0.0010571', 'media:0.0010307', 'articles:0.0009778', 'that:0.0008457', 'political:0.0007664', 'this:0.0007135', 'with:0.0006607', 'are:0.0006607', 'The:0.0006607', 'biased:0.0006343', 'news:0.0006343', 'level:0.0006078', 'text:0.0005286', 'article:0.0005021', 'three:0.0005021', 'unfairness:0.0004493', 'corpus:0.0004493']Then we decided to get rid of them. As we saw, results changed drastically. Now we saw less stop words(but still had them) and we saw a lot of words like “political”, “bias/biased”.

A Model of Political Bias in Social Science Research1.txt:['news:0.003791', 'bias:0.0037442', 'media:0.0024337', 'coverage:0.0020125', 'The:0.0016615', 'Groeling:0.0012871', 'stories:0.0012637', 'partisan:0.0011233', 'org:0.0010531', 'Polit:0.0010531', 'www:0.0010531', 'one:0.000936', 'content:0.000936', 'annualreviews:0.0009126', 'use:0.0008658', 'For:0.0008658', 'party:0.0007956', 'Sci:0.0007956', 'political:0.0007722', 'July:0.0007488']

A Model of Political Bias in Social Science Research2.txt:['social:0.0019421', 'doi:0.0017479', 'bias:0.0014808', 'Jussim:0.0013837', 'The:0.0012381', 'biases:0.0011895', 'political:0.0011895', 'Psychology:0.0009225', 'conservatives:0.0008982', 'may:0.0008497', 'findings:0.0008254', 'research:0.0008011', 'liberals:0.0007768', 'one:0.0007526', 'scientific:0.0007283', 'science:0.0007283', 'Social:0.000704', 'https:0.0006555', 'Psychological:0.0006312', 'psychology:0.0006312']

AMERICANS’ VIEWS OF POLITICAL BIAS IN THE ACADEMY AND ACADEMIC FREEDOM.txt:['percent:0.0038709', 'professors:0.0017666', 'college:0.001299', 'say:0.0011951', 'respondents:0.0009872', 'The:0.0008833', 'higher:0.0008313', 'education:0.0008054', 'Americans:0.0008054', 'political:0.0008054', 'high:0.0007794', 'bias:0.0007274', 'confidence:0.0006495', 'views:0.0006495', 'universities:0.0005975', 'school:0.0005715', 'tenure:0.0005456', 'colleges:0.0005456', 'university:0.0005456', 'classroom:0.0005196']

An Empirical Study of Political Bias in Legal Scholarship.txt:['articles:0.0027653', 'law:0.0026013', 'professors:0.0022263', 'The:0.0018982', 'liberal:0.0016639', 'Law:0.0014764', 'ideological:0.0014764', 'political:0.0013827', 'scholarship:0.0013592', 'legal:0.0013358', 'net:0.0012655', 'bias:0.0011718', 'conservative:0.0011249', 'results:0.0010077', 'Republican:0.000914', 'LEGAL:0.0008905', 'research:0.0008905', 'See:0.0008671', 'sample:0.0008671', 'BIAS:0.0008202']

Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity.txt:['bias:0.0028362', 'media:0.0010535', 'articles:0.0009994', 'political:0.0007833', 'The:0.0006753', 'biased:0.0006483', 'news:0.0006483', 'level:0.0006213', 'text:0.0005402', 'article:0.0005132', 'three:0.0005132', 'corpus:0.0004592', 'unfairness:0.0004592', 'com:0.0004322', 'Bias:0.0004052', 'word:0.0003782', 'levels:0.0003782', 'granularity:0.0003782', 'analysis:0.0003782', 'sentence:0.0003512']

To delete the stop words totally from our lists, and to group same words, like “Bias” and “bias” in the last example, we used case folding to lowercase all the words.

A Model of Political Bias in Social Science Research1.txt:['news:0.0051633', 'bias:0.0051098', 'media:0.0034244', 'coverage:0.0023543', 'groeling:0.0014982', 'partisan:0.0014982', 'stories:0.0014714', 'www:0.0012039', 'org:0.0012039', 'polit:0.0012039', 'content:0.0011236', 'political:0.0010701', 'one:0.0010701', 'annualreviews:0.0010434', 'use:0.0010166', 'party:0.0009899', 'sci:0.0009096', 'july:0.0008561', 'population:0.0008561', 'story:0.0008561']

A Model of Political Bias in Social Science Research2.txt:['social:0.0030068', 'doi:0.0020324', 'bias:0.0019489', 'political:0.001921', 'psychology:0.001754', 'jussim:0.0015869', 'science:0.0014756', 'biases:0.0013642', 'research:0.0011972', 'conservatives:0.0010858', 'one:0.0010301', 'findings:0.0010023', 'psychological:0.0010023', 'may:0.0009744', 'scientific:0.0009188', 'liberals:0.0008909', 'review:0.0008352', 'studies:0.0007517', 'https:0.0007517', 'crawford:0.000696']

AMERICANS’ VIEWS OF POLITICAL BIAS IN THE ACADEMY AND ACADEMIC FREEDOM.txt:['percent:0.0045732', 'professors:0.0022264', 'college:0.0015946', 'respondents:0.0015043', 'education:0.001384', 'say:0.001384', 'higher:0.0012937', 'political:0.0011132', 'high:0.0009628', 'bias:0.0009628', 'americans:0.0009628', 'confidence:0.0009026', 'views:0.0008725', 'university:0.0008424', 'universities:0.0007823', 'colleges:0.0007522', 'liberal:0.0007221', 'differences:0.0006619', 'tenure:0.0006619', 'school:0.0006619']

An Empirical Study of Political Bias in Legal Scholarship.txt:['law:0.0047451', 'articles:0.0038872', 'legal:0.0028417', 'political:0.0028417', 'professors:0.0027881', 'scholarship:0.002654', 'bias:0.0024395', 'net:0.0022787', 'liberal:0.0021178', 'ideological:0.0017962', 'conservative:0.0016353', 'results:0.0012064', 'research:0.0011796', 'see:0.0011796', 'sample:0.0010455', 'republican:0.0010455', 'donations:0.0010187', 'donors:0.0009919', 'donor:0.0009919', 'professor:0.0009919']

Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity.txt:['bias:0.0037366', 'media:0.00157', 'articles:0.0011932', 'political:0.001099', 'news:0.0009106', 'biased:0.0008164', 'level:0.0007222', 'unfairness:0.0006594', 'text:0.000628', 'analysis:0.0005966', 'corpus:0.0005966', 'article:0.0005966', 'three:0.0005966', 'com:0.0005024', 'levels:0.000471', 'granularity:0.000471', 'word:0.000471', 'sentence:0.0004396', 'set:0.0004082', 'classifiers:0.0004082']

Now we see less changes. Finally, no more stop words. We still get some similar words and to deal with it we used Porter Stemming technique. And our final result is:

A Model of Political Bias in Social Science Research1.txt:['news:0.0062579', 'bia:0.0061931', 'media:0.0041503', 'polit:0.0031452', 'coverag:0.0028533', 'stori:0.0027885', 'use:0.0022049', 'partisan:0.001913', 'newspap:0.0018158', 'groel:0.0018158', 'parti:0.0016212', 'organ:0.0015564', 'studi:0.0015239', 'org:0.0014591', 'www:0.0014591', 'research:0.0014591', 'content:0.0014267', 'select:0.0013942', 'event:0.0013294', 'one:0.0013294']

A Model of Political Bias in Social Science Research2.txt:['social:0.0038802', 'psycholog:0.0033356', 'polit:0.0030293', 'scienc:0.0025868', 'doi:0.0024847', 'bia:0.0023826', 'research:0.0020422', 'jussim:0.0019401', 'bias:0.0019401', 'conserv:0.001838', 'liber:0.0018039', 'find:0.0018039', 'review:0.0015657', 'one:0.0012594', 'measur:0.0011913', 'studi:0.0011913', 'may:0.0011913', 'scientif:0.0011572', 'http:0.0010211', 'stereotyp:0.000953']

AMERICANS’ VIEWS OF POLITICAL BIAS IN THE ACADEMY AND ACADEMIC FREEDOM.txt:['percent:0.0056931', 'professor:0.0031836', 'colleg:0.0029215', 'educ:0.0021724', 'respond:0.00206', 'univers:0.0019851', 'say:0.0018353', 'polit:0.0017978', 'american:0.0017604', 'higher:0.0016105', 'differ:0.0014233', 'view:0.0014233', 'liber:0.0013484', 'high:0.001236', 'bia:0.0011985', 'confid:0.0011611', 'conserv:0.0011236', 'cluster:0.0009738', 'compar:0.0008989', 'academ:0.0008989']

An Empirical Study of Political Bias in Legal Scholarship.txt:['law:0.0058185', 'articl:0.0052659', 'professor:0.0045508', 'polit:0.0039007', 'ideolog:0.0035431', 'legal:0.0034456', 'scholarship:0.003218', 'bia:0.002958', 'net:0.002763', 'liber:0.0027305', 'code:0.0025354', 'conserv:0.0023729', 'donor:0.0023729', 'republican:0.0020478', 'use:0.0019178', 'school:0.0018853', 'write:0.0017228', 'donat:0.0016903', 'democrat:0.0016578', 'result:0.0015928']

Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity.txt:['bia:0.0047019', 'articl:0.0022127', 'media:0.0019756', 'polit:0.0018966', 'level:0.0014619', 'news:0.0011458', 'unfair:0.0011063', 'bias:0.0010273', 'sentenc:0.0009088', 'text:0.0008693', 'word:0.0008297', 'three:0.0007507', 'analysi:0.0007507', 'corpu:0.0007507', 'label:0.0006717', 'portal:0.0006322', 'use:0.0006322', 'com:0.0006322', 'detect:0.0005927', 'granular:0.0005927']

We didn’t manage to reach the perfect result. Still some words are used in the same list. But if we compare all the results, we did a great job. In our lists we see a lot of internet links words like “www”, “net”, “com” etc. But if we don’t include them, we see that these articles grouped by “Political bias” for a reason. Most of them contain these words and words that are connected to education, law, geography and news.

Now let’s see the numbers of tokens and vocabulary size in each model:

|  |  |  |
| --- | --- | --- |
| Model | Number of Tokens | Vocabulary size |
| Collection model | 361972 | 34722 |
| No stop words | 276910 | 34612 |
| Case folding | 269614 | 29490 |
| Stemming | 269614 | 22952 |

As expected, after removing stop words number of tokens decreased by ~25% and 110 stop words were removed from vocabulary. After case folding we removed additional stop words like “The” and number of tokens decreased again, but only by ~2.5%. Although we see significant decrease in vocabulary size(~15%), because a lot of words were grouped up. In the final step we didn’t remove any words, so number of tokens remained the same, but we see even bigger difference in vocabulary size(~22%), because we grouped up a lot of similar words.

If we look at probabilities of the model they make sense. Each time we improved our models, we saw an increase in probability of the words. For example, when we look at the first article and ‘news’ probability we get:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Collection | No stop words | Case folding | Stemming |
| news | 0.0035842 | 0.003791 | 0.0051633 | 0.0062579 |

Because we used Laplace smoothing, decrease in vocabulary size has a lot of effect on our probabilities. If we used mixture model, our results wouldn’t be the same.