Predicting Authorship for 12 Federalist Papers

Part I: Initial Exploration

- Papers included in the data set include papers written by Hamilton, Jay, Madison, Hamilton & Madison, and Unknown. Narrow down to papers written by Hamilton, Madison, and Unknown.

HAMILTON	MADISON	UNKNOWN
51	14	12

- Order the Papers based on Author, Remove "To the People of the State of New York: " , and establish the corpus and initial DFM Matrix.

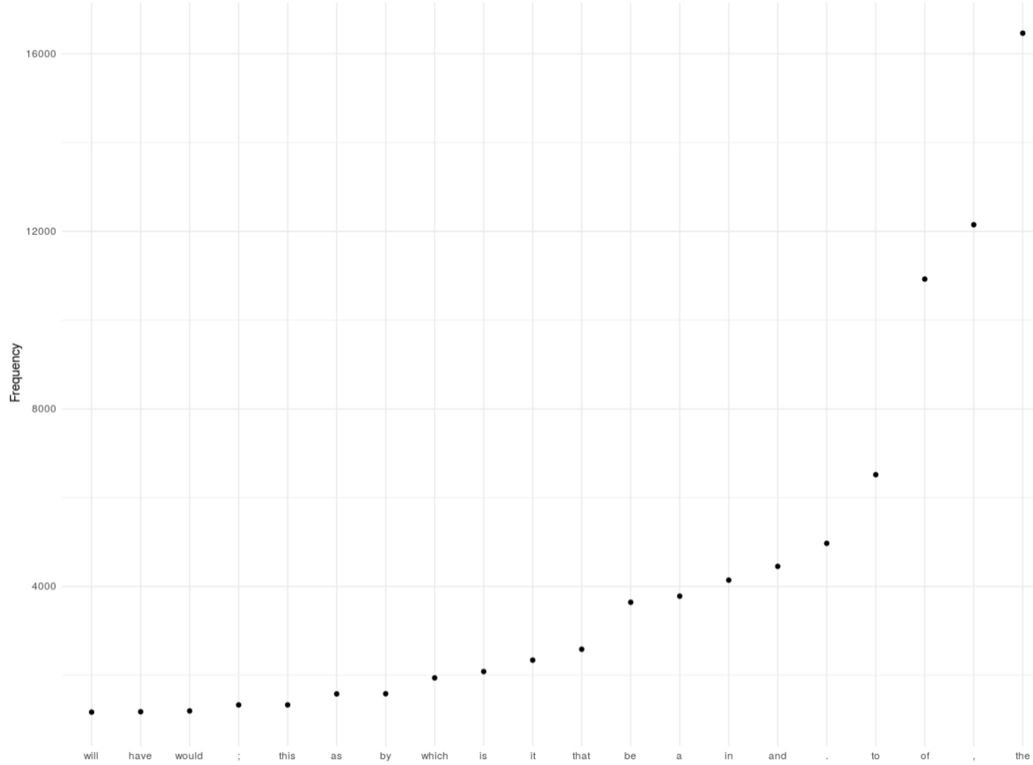
Corpus consisting of 77 documents:

Text	Types	Tokens	Sentences
text1	666	1775	48
text2	844	2321	75
text3	865	2571	84
text4	789	2308	72
text5	785	2222	64
text6	888	2784	87
text7	822	2423	72
text8	401	1052	29
text9	1094	3399	98
text10	751	2218	60
text11	625	1734	42
text12	743	2209	65

- Conduct some simple frequency analysis

	feature	frequency	rank	docfreq	group
1	the	16466	1	77	all
2	,	12151	2	77	all
3	of	10926	3	77	all
4	to	6520	4	77	all
5	.	4973	5	77	all
6	and	4454	6	77	all
7	in	4144	7	77	all
8	a	3783	8	77	all
9	be	3645	9	77	all
10	that	2587	10	77	all
11	it	2342	11	77	all
12	is	2085	12	77	all
13	which	1942	13	77	all
14	by	1585	14	77	all

- Visualize the most frequent terms



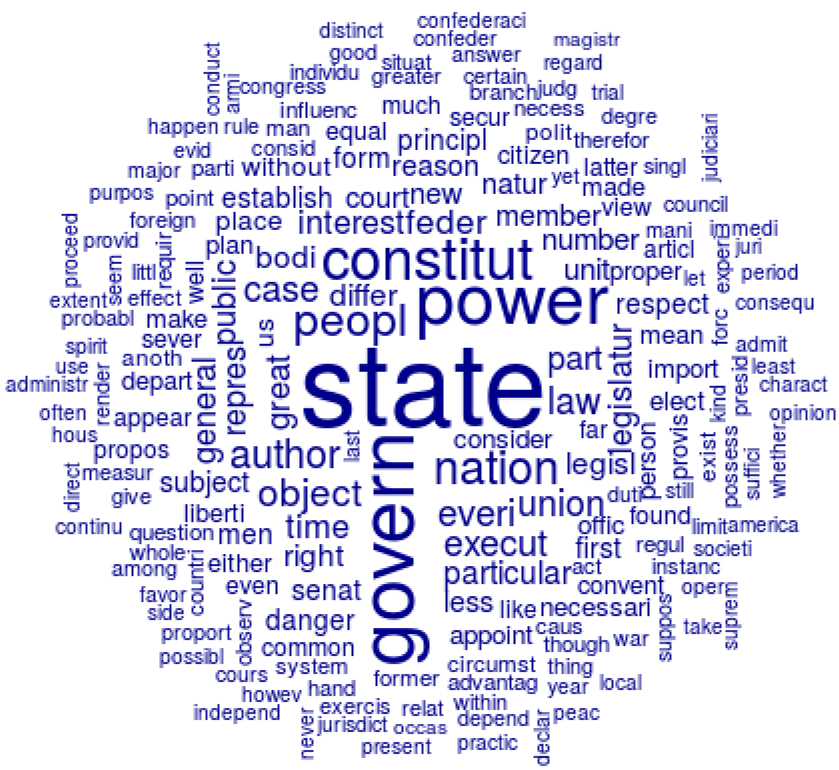
Part II: Exploratory Analyses with Similarity and Clustering

- Remove Stop words and perform stemming

```
topfeatures(myDfm,30)
    state     govern      power        may  constitut     nation        one      peopl
     1524        962        857        771        662        489        487        479
      can       must       upon     author     object      everi       case     execut
      441        428        384        381        362        342        334        333
    union        law      might      feder      great       time       part   interest
      323        322        312        307        292        287        277        276
   public    general     repres legislatur particular     differ
      276        275        269        267        256        253
```

- Add more user-defined stop words based on knowledge of the text and create an updated Word Cloud

```
state     govern      power  constitut     nation      peopl     author     object
 1524        962        857        662        489        479        381        362
everi       case     execut      union        law      feder      great       time
  342        334        333        323        322        307        292        287
 part   interest     public    general     repres legislatur particular     differ
  277        276        276        275        269        267        256        253
 bodi      right     legisl       unit     number        new
  251        250        250        245        239        228
```

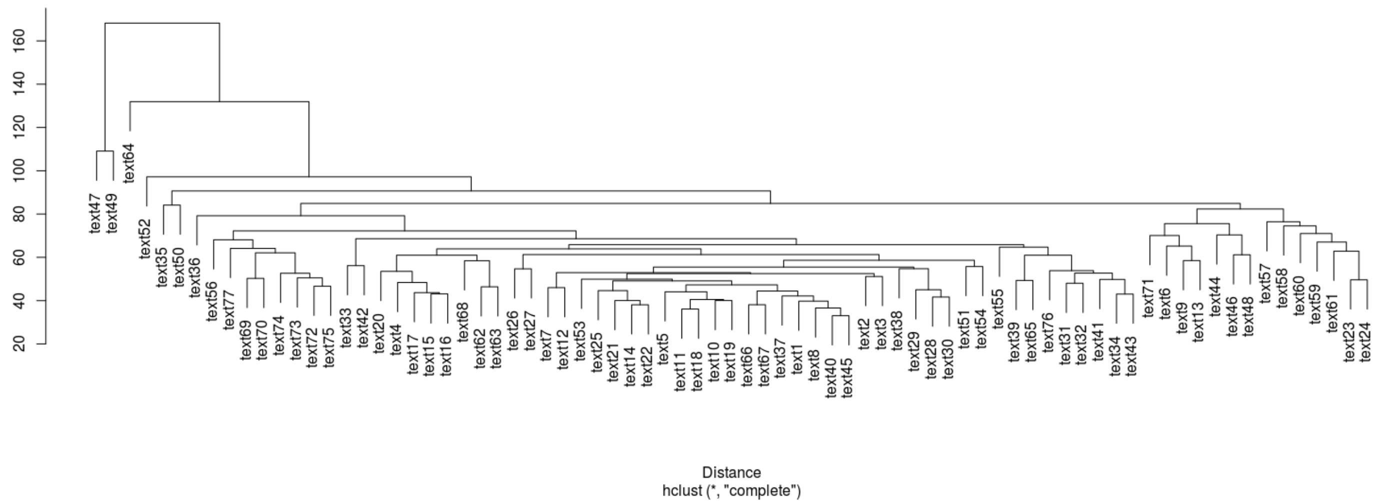


- Remove some very frequent words

author	object	everi	case	execut	union	law	feder
381	362	342	334	333	323	322	307
great	time	part	interest	public	general	repres	legislatur
292	287	277	276	276	275	269	267
particular	differ	bodi	right	legisl	unit	number	new
256	253	251	250	250	245	239	228
member	natur	court	less	reason	subject		
226	223	222	221	221	213		

- Control sparse terms: to further remove some very infrequent words

- Perform document clustering and explore results from clustering analyses



- Explore document similarity for text77 and based on the result, identify who may have written text 77.

text76	text75	text32	text69	text28	text72	text74	text70	text56	text30
0.5495754	0.5165282	0.4834967	0.4827904	0.4802367	0.4717887	0.4582895	0.4304305	0.4164063	0.4141140

Since Hamilton wrote all but one of these texts (56 which is Unknown) we can identify that Hamilton may have been the author of text 77.

- Explore terms most similar to commerc (yes no e at the end)

traffic	trade	intercours	commerci	european	privileg	competit	product
0.8544220	0.8379487	0.7790011	0.7065180	0.6486389	0.5867976	0.5823648	0.5776552

Part III: Topic Modeling

- You can explore with varying k numbers, I chose to show 8, below are the Term-topic probabilities.

```
   topic term          beta
   <int> <chr>        <dbl>
 1     1 unequivoc 3.28e- 11
 2     2 unequivoc 2.15e-166
 3     3 unequivoc 3.43e-  4
 4     4 unequivoc 3.30e-  4
 5     5 unequivoc 1.31e-  7
 6     6 unequivoc 4.54e-  5
 7     7 unequivoc 1.25e-  4
 8     8 unequivoc 3.24e-  9
 9     1 experi    2.23e-  3
10     2 experi    3.49e-  4
```

- Visualize most common terms in each topic

- Document-topic probabilities

```
   document topic     gamma
   <chr>    <int>     <dbl>
 1 text1        1 0.0000663
 2 text2        1 1.000
 3 text3        1 0.925
 4 text4        1 0.0452
 5 text5        1 0.0000529
 6 text6        1 0.636
 7 text7        1 0.856
 8 text8        1 0.755
 9 text9        1 0.102
10 text10       1 0.0000543
```

- View the document Probabilities in a table

	V1	V2	V3	V4	V5	V6	V7	V8
1	6.632242e-05	6.632242e-05	6.632242e-05	1.603136e-01	2.920809e-01	2.592177e-01	6.632242e-05	2.881225e-01
2	9.996393e-01	5.152516e-05	5.152516e-05	5.152516e-05	5.152516e-05	5.152516e-05	5.152516e-05	5.152516e-05
3	9.252825e-01	7.444627e-02	4.520080e-05	4.520080e-05	4.520080e-05	4.520080e-05	4.520080e-05	4.520080e-05
4	4.518565e-02	4.824147e-05	4.824147e-05	4.824147e-05	9.545249e-01	4.824147e-05	4.824147e-05	4.824147e-05
5	5.285772e-05	8.186893e-03	5.285772e-05	5.285772e-05	2.065293e-01	5.285772e-05	5.285772e-05	7.850195e-01
6	6.363545e-01	4.278776e-05	4.278776e-05	3.633888e-01	4.278776e-05	4.278776e-05	4.278776e-05	4.278776e-05
7	8.560977e-01	4.847029e-05	4.847029e-05	1.436115e-01	4.847029e-05	4.847029e-05	4.847029e-05	4.847029e-05

Part IV: Predicting Authorship

- Prepare the corpus by adding the ID and author columns

```
Corpus consisting of 77 documents, showing 10 documents:

   Text Types Tokens Sentences ID   Author
  text1   666   1775        48  1 HAMILTON
  text2   844   2321        75  6 HAMILTON
  text3   865   2571        84  7 HAMILTON
  text4   789   2308        72  8 HAMILTON
  text5   785   2222        64  9 HAMILTON
  text6   888   2784        87 11 HAMILTON
  text7   822   2423        72 12 HAMILTON
  text8   401   1052        29 13 HAMILTON
  text9  1094   3399        98 15 HAMILTON
 text10   751   2218        60 16 HAMILTON
```

- We will first generate SVD columns based on the entire corpus. Pre-process the training corpus, further remove very infrequent words, and weight the predictiv DFM by tf-idf.

- Perform SVD for dimension reduction and choose the number of reduced dimensions as 10

```
        [,1]        [,2]        [,3]       [,4]        [,5]       [,6]        [,7]         [,8]        [,9]         [,10]
text1 0.03722550 -0.05144188 -0.02304636 0.03303964 -0.03405591 0.01784448 -0.02472250  0.008866379  0.01843587  0.0003089497
text2 0.05817127 -0.10299602 -0.06118480 0.03983075  0.04237161 0.12512474  0.01317335  0.011166481  0.06618846 -0.0383376858
text3 0.07062506 -0.09099823 -0.05557816 0.07780988  0.07450674 0.03798801  0.04933225 -0.017613929  0.07198093  0.0379390416
text4 0.05743801 -0.10510110 -0.10824812 0.08162782  0.08453990 0.25890308 -0.18927772 -0.004284296 -0.06599347 -0.1337467266
text5 0.05526979 -0.07966966 -0.01507486 0.04278139 -0.07259775 0.03434550 -0.02614921 -0.079721129  0.04838543  0.0273373394
text6 0.06432966 -0.13140675 -0.12802679 0.14898724  0.32234176 0.13774351  0.49243996  0.056378164 -0.15637921  0.0150327094
```

- Add the author information as the first column (cut off at six to give a better display)

```
        Author         V1          V2          V3         V4          V5         V6
text1 HAMILTON 0.03722550 -0.05144188 -0.02304636 0.03303964 -0.03405591 0.01784448
text2 HAMILTON 0.05817127 -0.10299602 -0.06118480 0.03983075  0.04237161 0.12512474
text3 HAMILTON 0.07062506 -0.09099823 -0.05557816 0.07780988  0.07450674 0.03798801
text4 HAMILTON 0.05743801 -0.10510110 -0.10824812 0.08162782  0.08453990 0.25890308
text5 HAMILTON 0.05526979 -0.07966966 -0.01507486 0.04278139 -0.07259775 0.03434550
text6 HAMILTON 0.06432966 -0.13140675 -0.12802679 0.14898724  0.32234176 0.13774351
```

- Split the data into training & test. Typically we use random data partition, however, given our specific dataset, we manually split the dataset. Training dataset contains papers with known author information and the test dataset contains papers with unknown author information.

- Need to drop the unused unknown level in the training dataset, build a logistic model based on the training dataset, and compare model prediction with known authorships

```
pred.result HAMILTON MADISON
          0       49       2
          1        2      12
```

- Predict authorship for the test dataset and View results.

```
       testData$Author  unknownPred
text66         UNKNOWN 0.0981109586
text67         UNKNOWN 0.0007762791
text68         UNKNOWN 0.9448594225
text69         UNKNOWN 0.0023146463
text70         UNKNOWN 0.1125490132
text71         UNKNOWN 0.0083622307
text72         UNKNOWN 0.3338620602
text73         UNKNOWN 0.0170392029
text74         UNKNOWN 0.0038279540
text75         UNKNOWN 0.0539166864
text76         UNKNOWN 0.9753860293
text77         UNKNOWN 0.9999985888
```