Stylometry

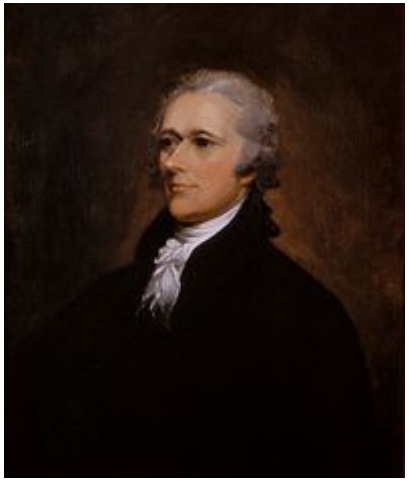
NLP Project

Stylometry

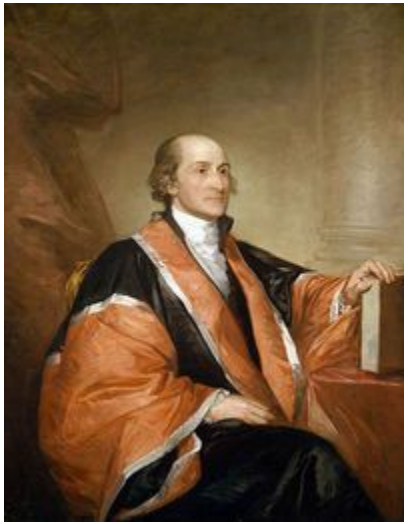
- Quantitative study of literary style through computational distant reading methods
- Consistent + unique way of writing
- Style
  - Vocabulary
  - Short / long sentences
  - Punctuation
  - Importance of function words

➢ Men / women differences
➢ Plagiarism
➢ Different style over time
➢ Authorship attribution

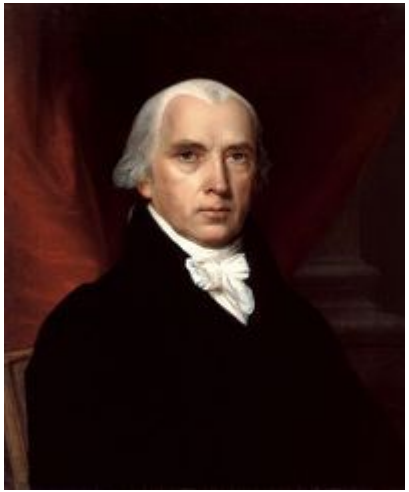
The Federalist Papers



Alexander Hamilton



John Jay



James Madison

→ “Publius”

1877-1878

➢ common test case for machine learning algorithms

Dataset

	label	text
0	Madison	10\n\nThe Same Subject Continued (The Union a...
1	Madison	14\n\nObjections to the Proposed Constitution...
2	Madison	37\n\nConcerning the Difficulties of the Conv...
3	Madison	38\n\nThe Same Subject Continued, and the Inc...
4	Madison	39\n\nThe Conformity of the Plan to Republica...
...	...	...
80	Disputed	57\n\nThe Alleged Tendency of the New Plan to...
81	Disputed	58\n\nObjection That The Number of Members Wi...
82	Disputed	62\n\nThe Senate\n\nFor the Independent Journ...
83	Disputed	63\n\nThe Senate Continued\n\nFor the Indepen...
84	TestCase	64\n\nThe Powers of the Senate\n\nFrom The In...

85 rows × 2 columns

Labels (y):

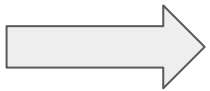
- Madison
- Hamilton
- Jay
- Disputed & Shared

Text (X):

- Papers

Data processing

Preprocessing:
- Separating text into words
- Chunking text into smaller chunks
- Vectorizing text



Test-Evaluation-Train Split:
- **Test**: Extract Test Cases
- **Train**: 80% of remaining data
- **Evaluation**: 20% of remaining data

Chunking & Vectorization

	label	text
0	Madison	10\n\nThe Same Subject Continued (The Union a...
1	Madison	14\n\nObjections to the Proposed Constitution...
2	Madison	37\n\nConcerning the Difficulties of the Conv...
3	Madison	38\n\nThe Same Subject Continued, and the Inc...
4	Madison	39\n\nThe Conformity of the Plan to Republica...
...	...	...
80	Disputed	57\n\nThe Alleged Tendency of the New Plan to...
81	Disputed	58\n\nObjection That The Number of Members Wi...
82	Disputed	62\n\nThe Senate\n\nFor the Independent Journ...
83	Disputed	63\n\nThe Senate Continued\n\nFor the Indepen...
84	TestCase	64\n\nThe Powers of the Senate\n\nFrom The In...

85 rows × 2 columns

Chunking

	label	text
0	Madison	10\n\nThe Same Subject Continued (The Union a...
1	Madison	14\n\nObjections to the Proposed Constitution...
2	Madison	37\n\nConcerning the Difficulties of the Conv...
3	Madison	38\n\nThe Same Subject Continued, and the Inc...
4	Madison	39\n\nThe Conformity of the Plan to Republica...
...	...	...
80	Disputed	57\n\nThe Alleged Tendency of the New Plan to...
81	Disputed	58\n\nObjection That The Number of Members Wi...
82	Disputed	62\n\nThe Senate\n\nFor the Independent Journ...
83	Disputed	63\n\nThe Senate Continued\n\nFor the Indepen...
84	TestCase	64\n\nThe Powers of the Senate\n\nFrom The In...

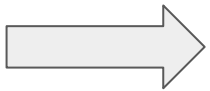
85 rows × 2 columns

	label	text
0	Madison	[, 10\n\nThe, Same, Subject, Continued, (The, ...
1	Madison	[of, public, and, private, faith,, and, of, pu...
2	Madison	[actuated, by, some\ncommon, impulse, of, pass...
3	Madison	[a, reciprocal, influence, on, each, other;, a...
4	Madison	[fall, into, mutual\nanimosities,, that, where...
...	...	...
912	TestCase	[the, advice, and\nconsent, of, the, Senate,, ...
913	TestCase	[the, judicial., It, surely, does\nnot, follow...
914	TestCase	[as, the, consent, of, both, was, essential, t...
915	TestCase	[equal, degree, of, influence, in, that\nbody,...
916	TestCase	[if, it, should, ever, happen,, the, treaty, s...

917 rows × 2 columns

Data processing

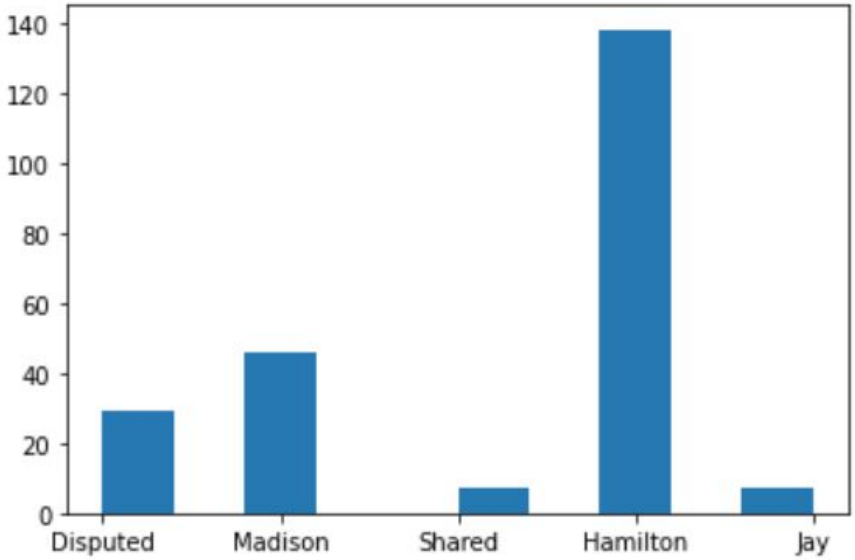
Preprocessing:
- Separating text into words
- Chunking text into smaller chunks
- Vectorizing text



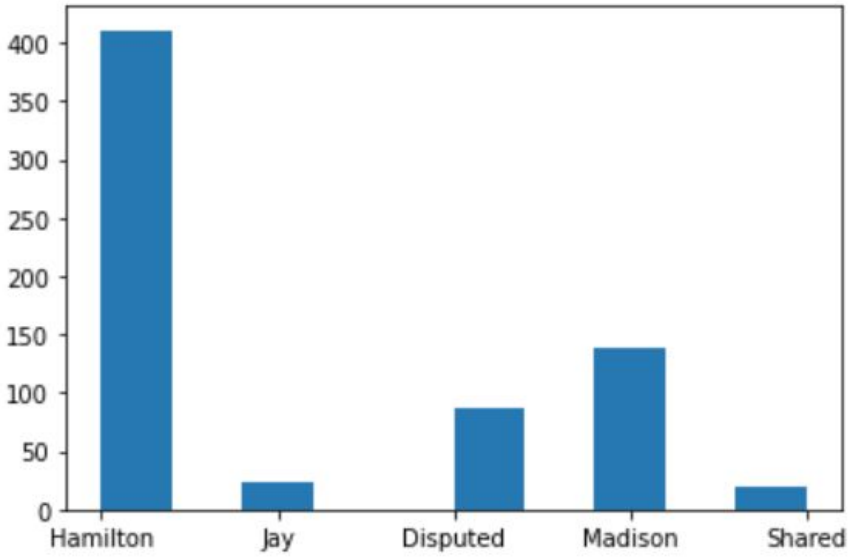
Test-Evaluation-Train Split:
- **Test**: Extract Test Cases
- **Train**: 80% of remaining data
- **Evaluation**: 20% of remaining data

Test-Train Split with stratification

Test



Train



Method: Burrows Delta

One the most prominent stylometric measures for authorship attribution

$$Z_i = \frac{C_i - \mu_i}{\sigma_i}$$

Figure 7: Equation for the z-score statistic.

Relative frequency
- word frequencies
- vocabulary richness

Result

In [202]:
```
clf = BurrowsDelta()
clf.fit(X_train, y_train)
y_pred = clf.predict(X_eval)

print(classification_report(np.array(y_eval), y_pred))
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-202-6c9e52ce96ca> in <module>
      1 clf = BurrowsDelta()
----> 2 clf.fit(X_train, y_train)
      3 y_pred = clf.predict(X_eval)
      4
      5 print(classification_report(np.array(y_eval), y_pred))

<ipython-input-200-20070a36b00c> in fit(self, X, y)
      9
     10         self.chosen_words = np.ravel(X.sum(axis=0)).argsort()[::-1][:self.num_words]
---> 11         sX = X.T[self.chosen_words].toarray()
     12
     13         ### YOUR CODE BELOW

AttributeError: 'numpy.ndarray' object has no attribute 'toarray'
```

Limitations

- Disputed & Shared authorship

- Delta’s Burrow has no clear theoretical foundation