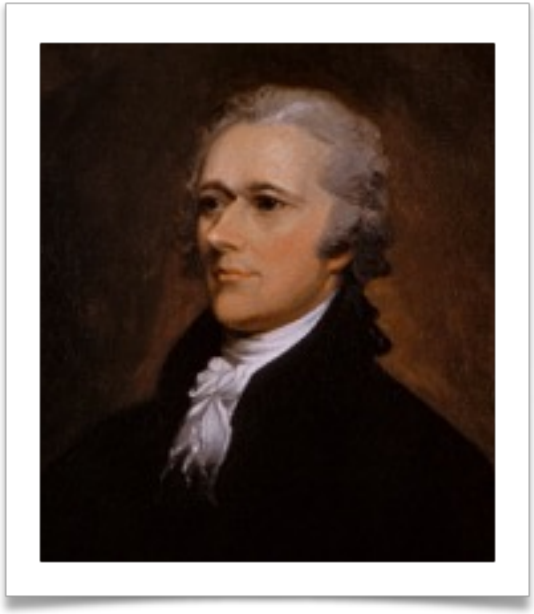
Université de Montréal

Analyse numérique de la stylométrie des *Federalist Papers*

François Dominic Laramée
Doctorant, Département d'histoire de l'Université de Montréal
19 mars 2016

Les *Federalist Papers*

- Document fondateur de la science politique américaine
- 85 articles publiés entre octobre 1787 et mai 1788
- Trois auteurs, un pseudonyme: Publius
- Une « paternité » contestée
  - La liste de Hamilton
  - La réplique de Madison
  - Des styles étonnamment similaires
  - Le curieux cas du *Federalist 64*
- Une cause célèbre dans l'historiographie



Portrait d'Alexander Hamilton par John Trumbull (1806)

La stylométrie et les *Federalist Papers*

- La signature secrète de l'humble grammaire
  - Conjonctions, articles, ponctuation

- *Federalist Papers:* des preuves pro-Madison majoritaires…
  - Mesures de fréquences de *may, also, his, an, any…* (Mosteller et Wallace 1964, Mosteller 1987)
  - Comparaison *on-upon* (Merriam 1987)

- … Mais pas unanimes
  - Collaborations cachées? (Collins et al. 2004)

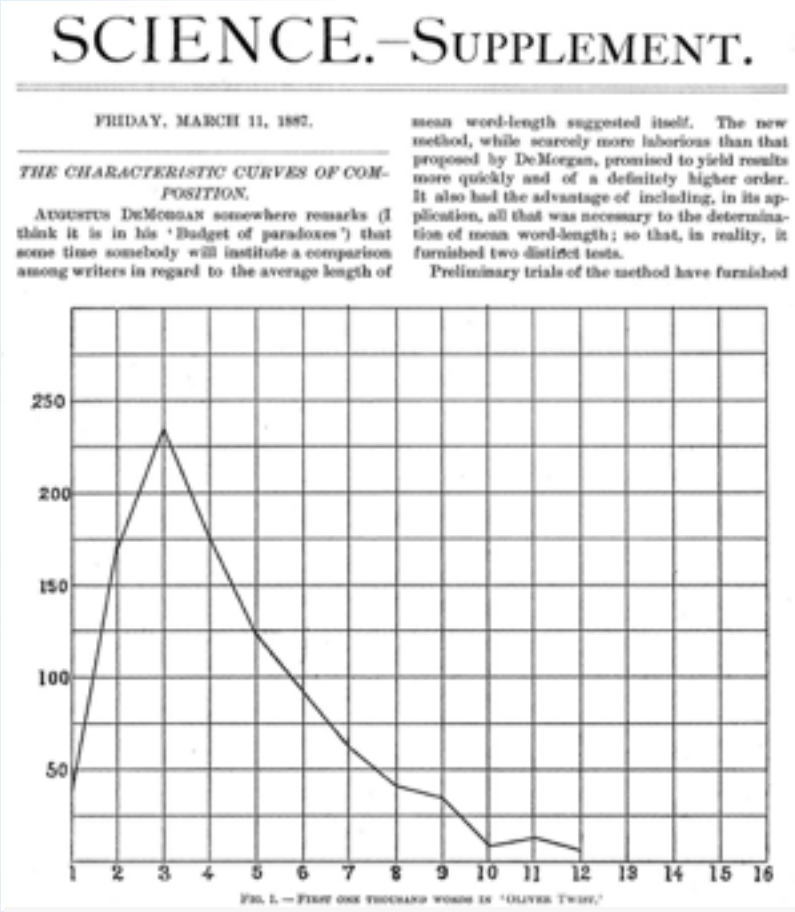
Portrait de James Madison par John Vanderlyn (1815)

Le projet

- Corpus: Projet Gutenberg
- Madison vs Hamilton: trois tests
  - Courbe caractéristique de Mendenhall
  - Examen du vocabulaire et des n-grams
  - Chi-carré de Kilgariff
- Le cas du *Federalist 64*:
  - Le delta de Burrows

Portrait de John Jay par Gilbert Stuart (1793)

Hamilton vs Madison #1: Courbes caractéristiques



SCIENCE.–SUPPLEMENT.

FRIDAY, MARCH 11, 1887.

THE CHARACTERISTIC CURVES OF COMPOSITION.

AUGUSTUS DEMORGAN somewhere remarks (I think it is in his 'Budget of paradoxes') that some time somebody will institute a comparison among writers in regard to the average length of mean word-length suggested itself. The new method, while scarcely more laborious than that proposed by DeMorgan, promised to yield results more quickly and of a definitely higher order. It also had the advantage of including, in its application, all that was necessary to the determination of mean word-length; so that, in reality, it furnished two distinct tests.

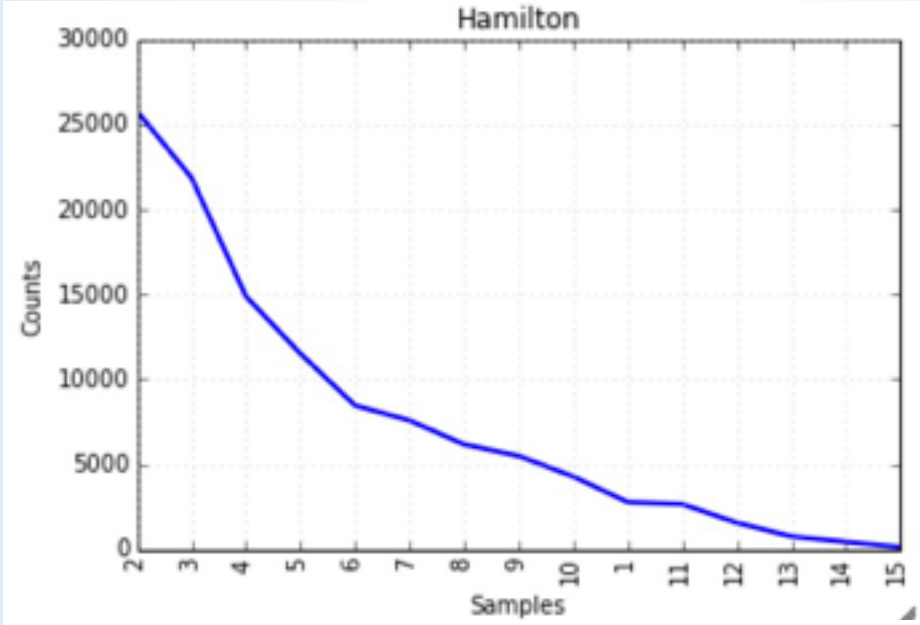
Preliminary trials of the method have furnished

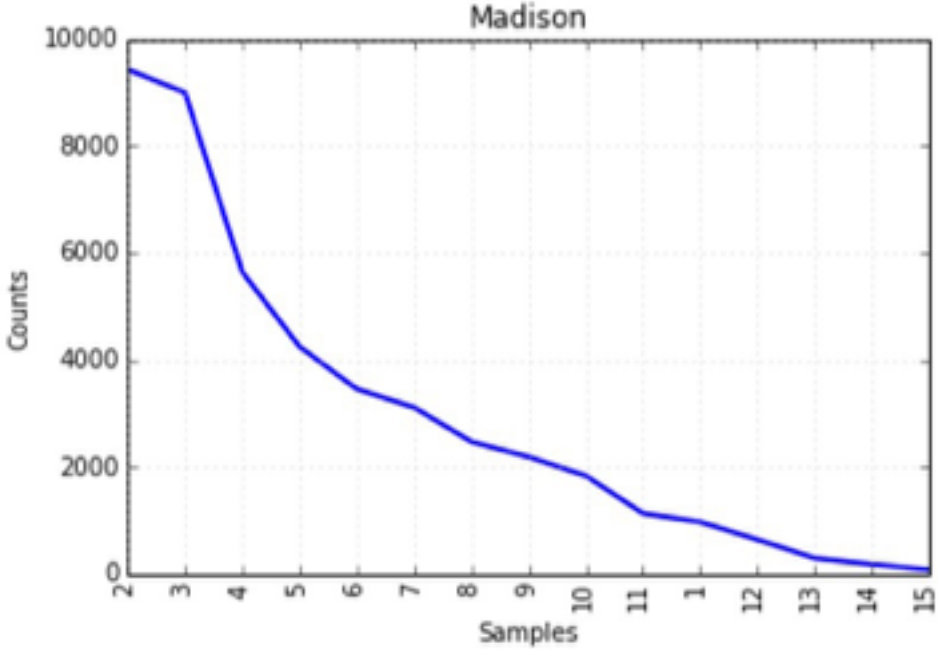
FIG. 1.—FIRST ONE THOUSAND WORDS IN 'OLIVER TWIST.'

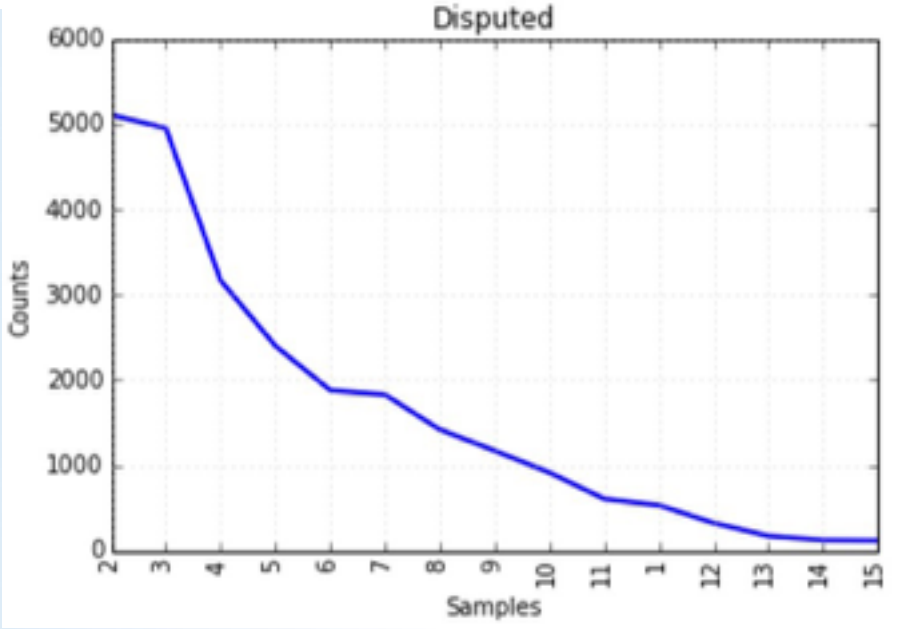
- « Signature » d'auteur détectable dans la fréquence d'utilisation de mots d'une certaine longueur
- En Python: 3 lignes de code
- Technique publiée… en 1887.

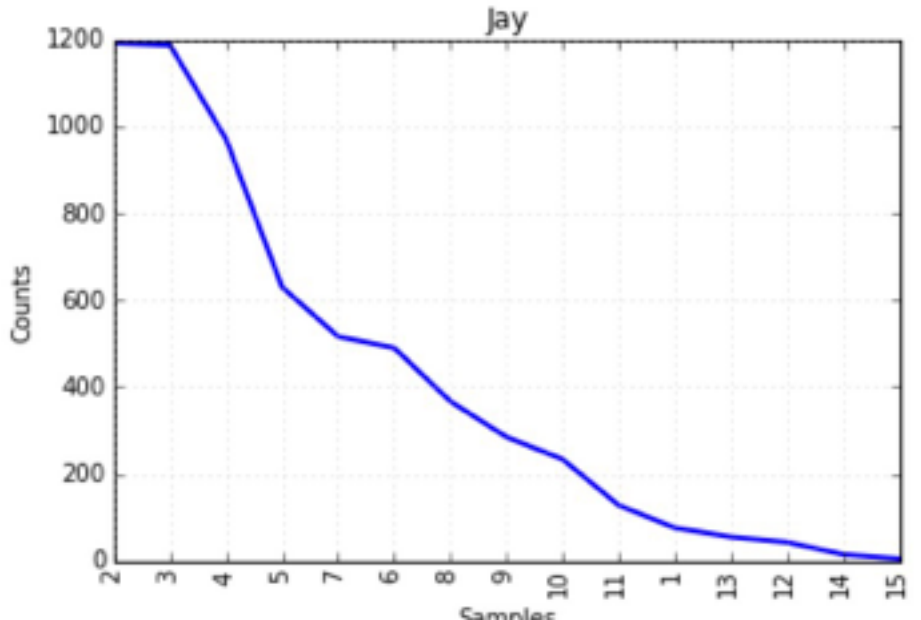
```
# Get a distribution of token lengths
tokenLengths = [ len( token ) for token in federalistByAuthorTokens[ subcorpus ] ]
federalistByAuthorLengthDistributions[ subcorpus ] = nltk.FreqDist( tokenLengths )
federalistByAuthorLengthDistributions[ subcorpus ].plot( 15, title = subcorpus )
```

Hamilton vs Madison #1: Courbes caractéristiques









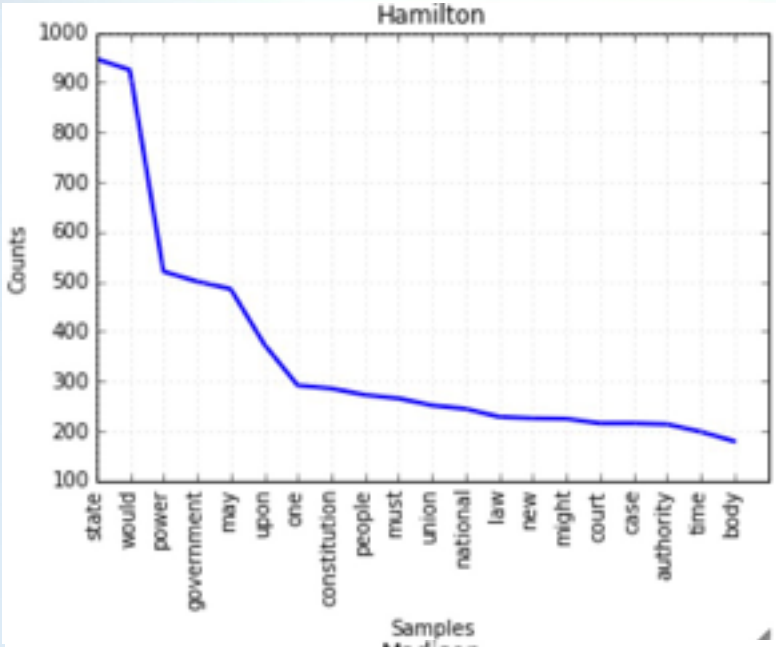
Hamilton vs Madison #2: Vocabulaire et N-grams: mots les plus fréquents

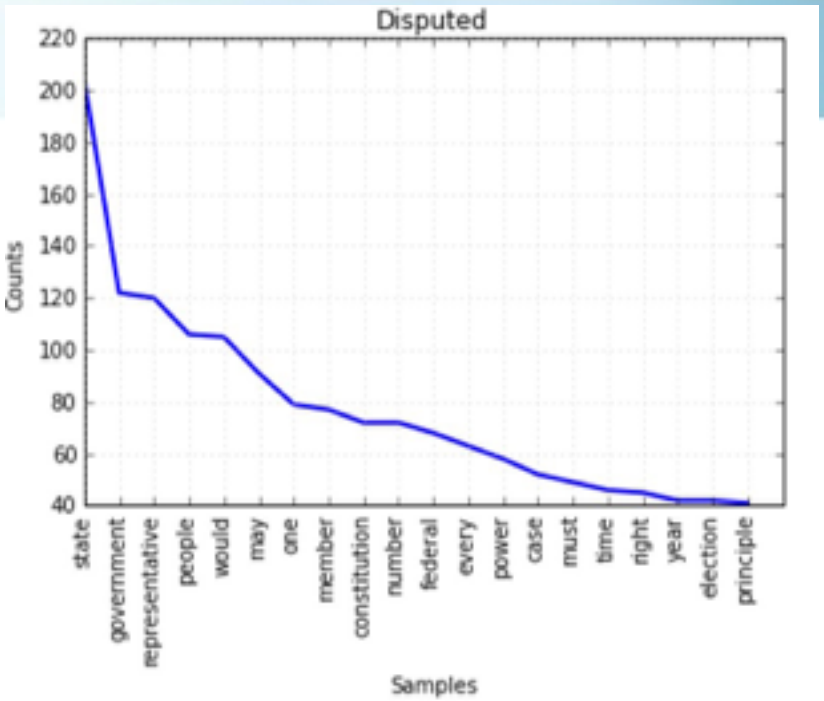
Madison	Hamilton	Disputés
the (4435)	the (10 598)	the (2454)
of (2668)	of (7370)	of (1488)
to (1435)	to (4614)	to (758)
and (1306)	in (2833)	and (671)
in (926)	and (2730)	in (538)
a (904)	a (2507)	be (491)
be (876)	be (2300)	a (488)
that (627)	that (1717)	that (299)
it (568)	it (1549)	it (295)
is (554)	is (1330)	which (275)

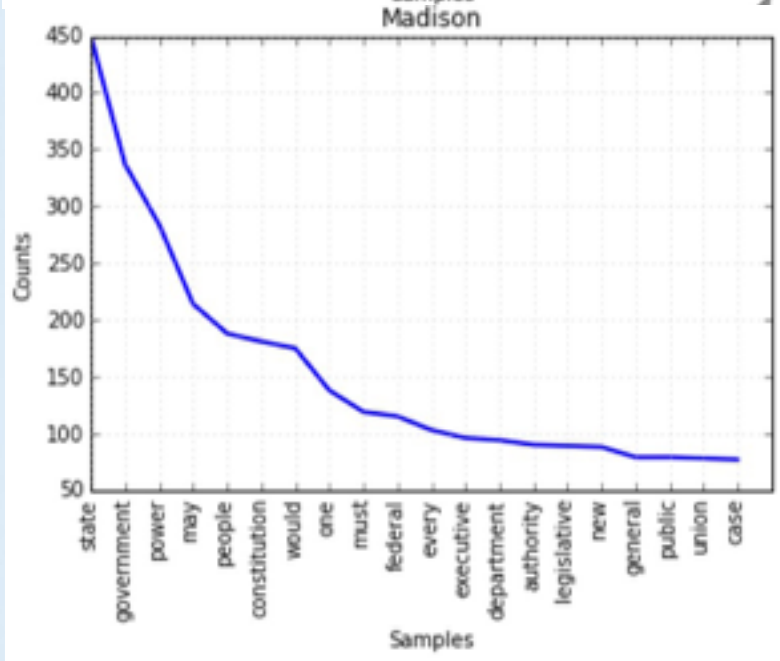
Hamilton vs Madison #2: Vocabulaire et N-grams - Trigrammes

Madison	Hamilton	Disputés
of the people (71)	of the state (135)	of the people (37)
of the state (54)	of the union (132)	the house of (31)
the united states (49)	the united states (125)	of the state (29)
the people of (42)	of the people (104)	house of representatives (28)
of the states (39)	the power of (103)	to the people (25)
of the union (38)	of new york (83)	the people of (25)
the federal government (36)	the people of (77)	the number of (23)
members of the (34)	of the united (75)	of the government (23)
of the federal (33)	of the national (71)	ought to be (20)
the state governments (32)	to the people (67)	people of the (17)

Hamilton vs Madison #2: Vocabulaire et N-grams - Mots de contenu







- Intersection Madison/Disputés:
  - 33 mots sur 50
  - Distance: 1169
- Intersection Hamilton/Disputés:
  - 31 mots sur 50
  - Distance: 1233

Hamilton vs Madison #3: Chi-carré de Kilgariff

- Question: Deux corpus proviennent-ils de la même « population » statistique?

- **Méthode:**
  - Soient deux corpus A et B
  - Trouver les N mots les plus courants dans (A + B)
  - Calculer les fréquences F_A et F_B de ces N mots dans A et dans B
  - Calculer les fréquences attendues F'_A et F'_B si A et B provenaient de la même population
  - Pour chacun des N mots, calculer un chi-carré: $(F_A - F'_A) * (F_A - F'_A) / F'_A$ OU $(F_B - F'_B) * (F_B - F'_B) / F'_B$

- ***Plus la somme des chi-carrés est petite, plus les deux corpus sont statistiquement semblables.***

Hamilton vs Madison #3: Chi-carré de Kilgariff

```
chisquared = 0
for word, jointCount in mostCommonInJointCorpus:

    # How often do we really see it?
    candidateCount = federalistByAuthorTokens[ candidate ].count( word )
    disputedCount = federalistByAuthorTokens[ "Disputed" ].count( word )

    # How often should we see it?
    expCandidateCount = jointCount * candidateShareInJointCorpus
    expDisputedCount = jointCount * ( 1 - candidateShareInJointCorpus )

    # Add the word's contribution to the chi-squared statistic
    chisquared += ( candidateCount - expCandidateCount ) * \
                  ( candidateCount - expCandidateCount ) / expCandidateCount

    chisquared += ( disputedCount - expDisputedCount ) * \
                  ( disputedCount - expDisputedCount ) / expDisputedCount

print( "The Chi-squared statistic for candidate", candidate, "is", chisquared )
```

```
The Chi-squared statistic for candidate Hamilton is 2997.8
The Chi-squared statistic for candidate Madison is 1533.56
```

Federalist 64: Delta de Burrows

- Pour identifier l'auteur d'un texte T parmi plusieurs candidats:
  - Trouvez les mots les plus courants dans l'ensemble des textes
  - Calculez comment les textes de chaque candidat se distinguent de la « moyenne »
  - Calculez comment le texte T se distingue de la moyenne
  - Comparez T à chacun des candidats
- Conçu pour un grand nombre de candidats (~25)
  - Ici, seulement 4
- ***Plus la statistique est petite, plus les « signatures » correspondent***

Federalist 64: Delta de Burrows

```
Test case z-score for feature ('the', 'DT') is -0.5715580470247853
Test case z-score for feature ('of', 'IN') is -1.530570023015507 7
Test case z-score for feature ('to', 'TO') is 0.8614793246103539
Test case z-score for feature ('and', 'CC') is 0.9923420641528881
Test case z-score for feature ('in', 'IN') is 0.4020915007213734
Test case z-score for feature ('a', 'DT') is -0.9117144930479929
Test case z-score for feature ('be', 'VB') is 3.211624164136822
Test case z-score for feature ('it', 'PRP') is -0.4225688536389689
Test case z-score for feature ('is', 'VBZ') is -1.0961185455958937
Test case z-score for feature ('which', 'WDT') is -1.861735205161324 7
Test case z-score for feature ('that', 'IN') is 3.4609374852197483
Test case z-score for feature ('by', 'IN') is 1.495841660671324 7
Test case z-score for feature ('as', 'IN') is 9.90787889195006
```

```
for candidate in candidateList:
    delta = 0
    for feature in featuresList:
        delta += math.fabs( testCaseZScores[ feature ] - featureZScores[ candidate ][ feature ] )
    delta /= len( featuresList )
    print( "Delta score for candidate", candidate, "is", delta )
```

```
Delta score for candidate Hamilton is 2.2728145489103126
Delta score for candidate Madison is 2.101912545169431
Delta score for candidate Jay is 1.9528815383895344
Delta score for candidate Disputed is 1.9810301574357099
```

Conclusions

- Les 12 articles contestés par Hamilton et Madison?
  - Courbes caractéristiques de Mendenhall: avantage Madison
  - Vocabulaire et N-grams: (petit) avantage Madison
  - Chi-carré de Kilgariff: avantage Madison
- Le *Federalist 64*?
  - Delta de Burrows: avantage Jay
  - « Candidature » de Hamilton: la moins probable
- **Résultats en accord avec la majorité des chercheurs**

Merci!

François Dominic Laramée
Carnet iPython: www.francoisdominiclaramee.com
Courriel: fdl@francoisdominiclaramee.com
@fdlaramee sur Twitter