Uncovering Hidden Structure in the *De Re Publica*

A Computational Analysis of the Text

Marc E. Canby

Lati 318 – Cicero: *De Re Publica*
Rice University

Outline

1. Getting and Cleaning the Data
2. Exploratory Text Analysis
3. Underlying Word Structure: Word2Vec

Outline

1. Getting and Cleaning the Data
2. Exploratory Text Analysis
3. Underlying Word Structure: Word2Vec

Getting and Cleaning the Data

- Text obtained from *The Latin Library* at the sentence level:
  - `['nempe', 'ab', 'iis', 'qui', 'haec', 'disciplinis', 'informata', 'alia', 'moribus', 'confirmarunt', ',', 'sanxerunt', 'autem', 'alia', 'legibus', '.']`
- Cleaned up messy elements of data:
  - Line numbers: `[1,2,...,71]`
  - Angle brackets: `['&', 'lt', ';', 'im&gt', ';', 'petu', 'liberavissent', ',', 'nec',...]`
    - `&lt;` should be <   `&gt;` should be >
  - Hyphens encoded as `&#`
  - English words: `['Cicero', 'The', 'Latin', 'Library', 'The', 'Classics', 'Page']`

Outline

1. Getting and Cleaning the Data
2. Exploratory Text Analysis
3. Underlying Word Structure: Word2Vec

Exploratory Text Analysis: Tokenization and POS Tagging

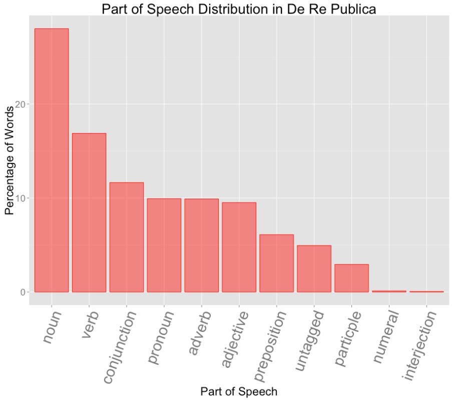
- Map each word to its base form (*lemma* or *token*) and its POS
- Often ignore *stop words* ('et', 'sum', etc.) – highlighted in red

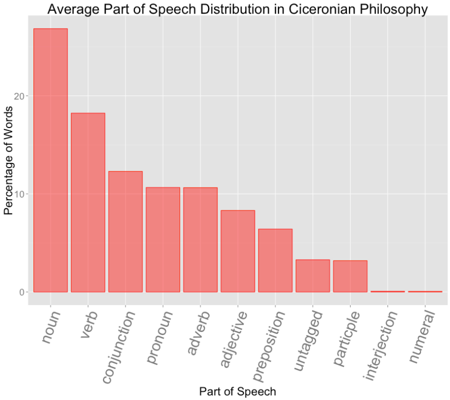
```
['nempe', 'ab', 'iis', 'qui', 'haec', 'disciplinis',
'informata', 'alia', 'moribus', 'confirmarunt', ',',
'sanxerunt', 'autem', 'alia', 'legibus', '.']
```

```
[('nempe', 'adverb'), ('ab', 'preposition'), ('is',
'pronoun'), ('qui', 'pronoun'), ('hic', 'pronoun'),
('disciplina', 'noun'), ('informo', 'noun'), ('alius2',
'adjective'), ('mos', 'noun'), ('confirmo', 'verb'),
('sancio', 'verb'), ('autem', 'conjunction'), ('alius2'
```

Exploratory Text Analysis

- Number of characters: 109,777 (*average*: 136,893)
- Number of words: 20,067 (*average*: 24,924)
- Number of sentences: 820 (*average*: 1,059)


Part of Speech Distribution in De Re Publica


Average Part of Speech Distribution in Ciceronian Philosophy

Outline

1. Getting and Cleaning the Data
2. Exploratory Text Analysis
3. Underlying Word Structure: Word2Vec