Data Preparation

Collect all Wikipedia biographies and create a comparable base of female and male politicians

1. Clean and prepare data
2. Extract life and career subsection
3. Match data on relevant variables to make it more comparable
   - Year of birth
   - Party
   - Duration in Office
   - Aggregated page views
   - Important offices
   - Exact matching on Session

Descriptive Analysis

Conduct an analysis of descriptive indicators

1. Log(2) transform data due to outliers that skew the means
2. Compare means across gender for indicators of interest:
   - Text length (overall/career section/life section)
   - Number of links in the biography
   - Aggregated number of Edits of bibliography

PMI Analysis

Conduct a PMI analysis, outputting top 100 words associated with each gender

1. Tokenize; remove stop words and politician's names; replace gender-specific job titles; apply stemming
2. Create a Document-feature-matrix to obtain the vocabulary and only keep words that appear in both genders:

	Word 1	Word 2	Word 3
Female	1	4	~~0~~
Male	3	6	~~3~~

3. Calculate PMI values, normalize them and output top 100 words per gender, adding a threshold, as PMI overemphasize rare words.

$$\mathrm{PMI}(c, w) = \log\left(\frac{p(c,w)}{p(c)p(w)}\right)$$

4. Annotate words manually with one of the following categories:
   - Family
   - Gender
   - Relationship
   - Other