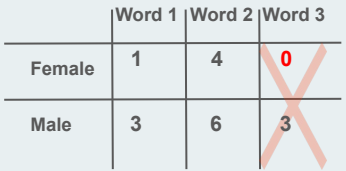
	(A) Data Preparation	(B) Descriptive Analysis	(C) PMI Analysis
Objective	**Collect Wikipedia texts and create a comparable base of female and male politicians**	**Conduct analysis of three descriptive indicators**	**Conduct PMI analysis, to retrieve top 100 words associated with each gender**
Conducted Steps	1. Get Wikipedia texts of German Politicians 2. Clean and prepare data 3. Extract life and career subsections 4. 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	1. Log(2) transform data to address 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 the biography	1. Prepare data for PMI analysis: 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:  3. Calculate PMI values, normalize PMI values, rank and output top 100 words per gender $\mathrm{PMI}(c,w) = \log\left(\frac{P(g,w)}{P(g)P(w)}\right)$ 4. Annotate words with the following categories: - Family - Gender - Relationship - Other