Applied Data Analysis – Le Temps Dataset Word Frequency Prediction

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

The Dataset:

200 years of daily articles from:

Gazette de Lausanne ET JOURNAL SUISSE FONDÉE EN 1798 — Publication dates: 1798 – 1998

JOURNAL DE GENÈVE — Publication dates: 1826 – 1998

Extraction: Counting the 3000+ most frequent words per month

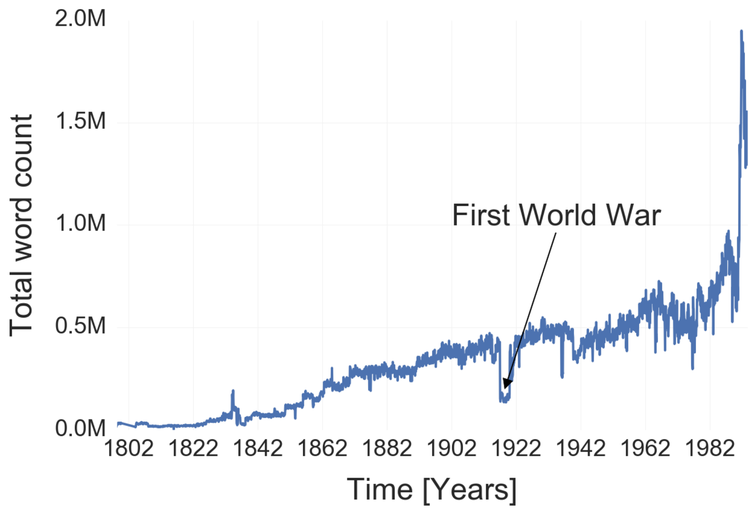
Data Extraction:

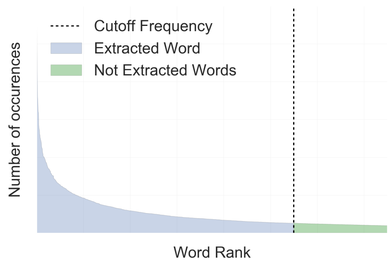
1. Removal of punctuation
2. Removal of French stop words
3. Custom NLTK processing:
   - Singular / Plural
   - Masculin / Féminin
   - Verbs and their conjugations
   - Adverbs + Noun
4. Removal of words that were not present enough

Result: Time serie of the frequency of each word

Data Visualization and Processing

Total word count over the years





How much of the data was not extracted?