Applied Data Analysis – Le Temps Dataset Word Frequency Prediction

Axel De La Harpe, Thomas Vetterli, Lukman Olagoke – January 2017

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Extract Transfom Pipeline

The Dataset

200 years of daily articles from:

Gazette de Lausanne
ET JOURNAL SUISSE

Publication dates: 1798 – 1998

JOURNAL DE GENÈVE

Publication dates: 1826 – 1998

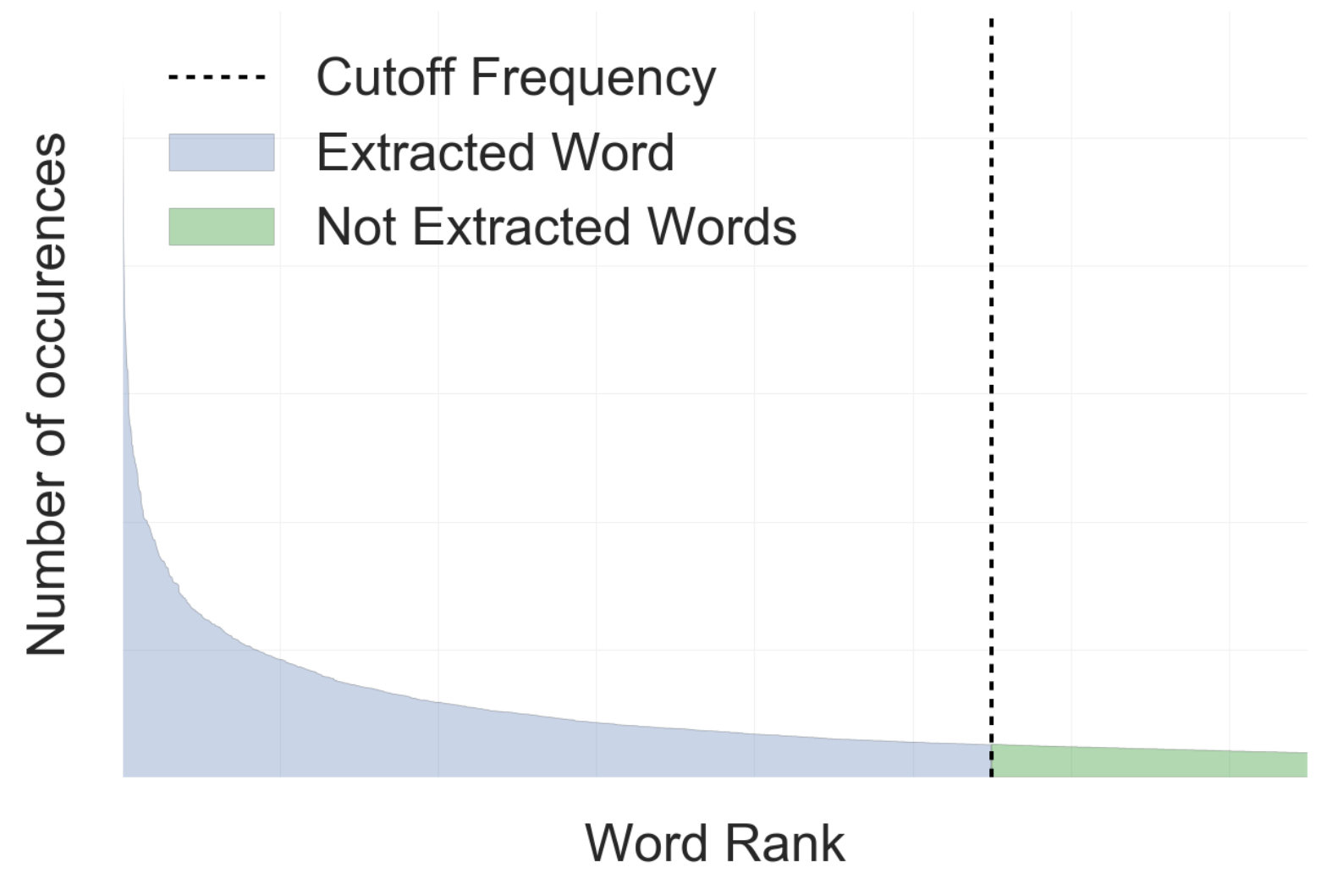
Extraction: Counting the 3000+ most frequent words per month

Data Extraction

1. Removal of punctuation / numbers
2. Removal of French stop words
3. Custom NLTK processing:
   - Singular / Plural
   - Masculin / Féminin
   - Verbs and their conjugations
   - Adverbs + Noun
4. Cutoff Frequency: We only save the 3000+ most frequent word per months

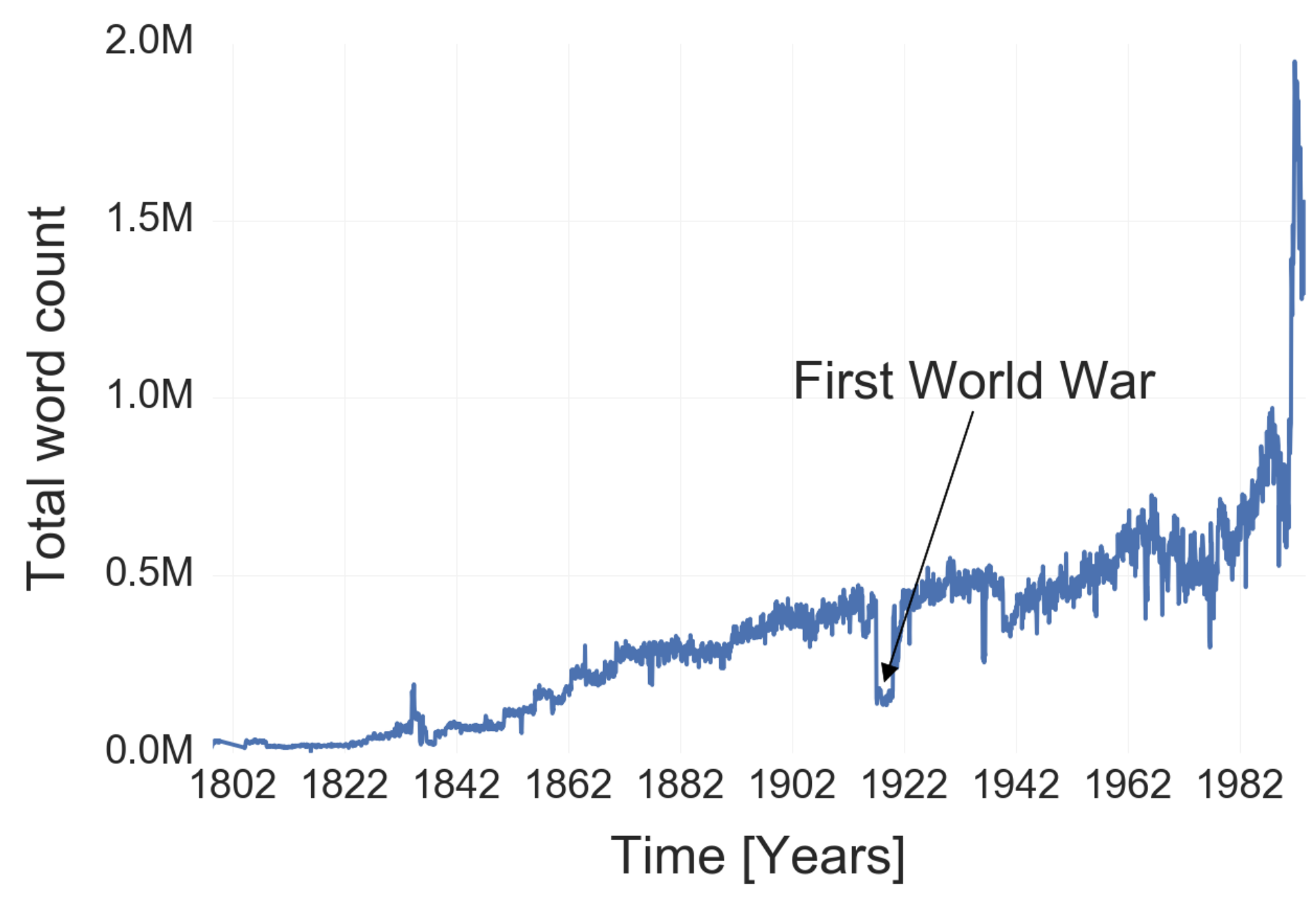
Result: Time serie of the frequency of each word
Resolution: 1 time point per month

Long tail distribution of word frequency: Due to our cutoff frequency, we miss a part of the data:



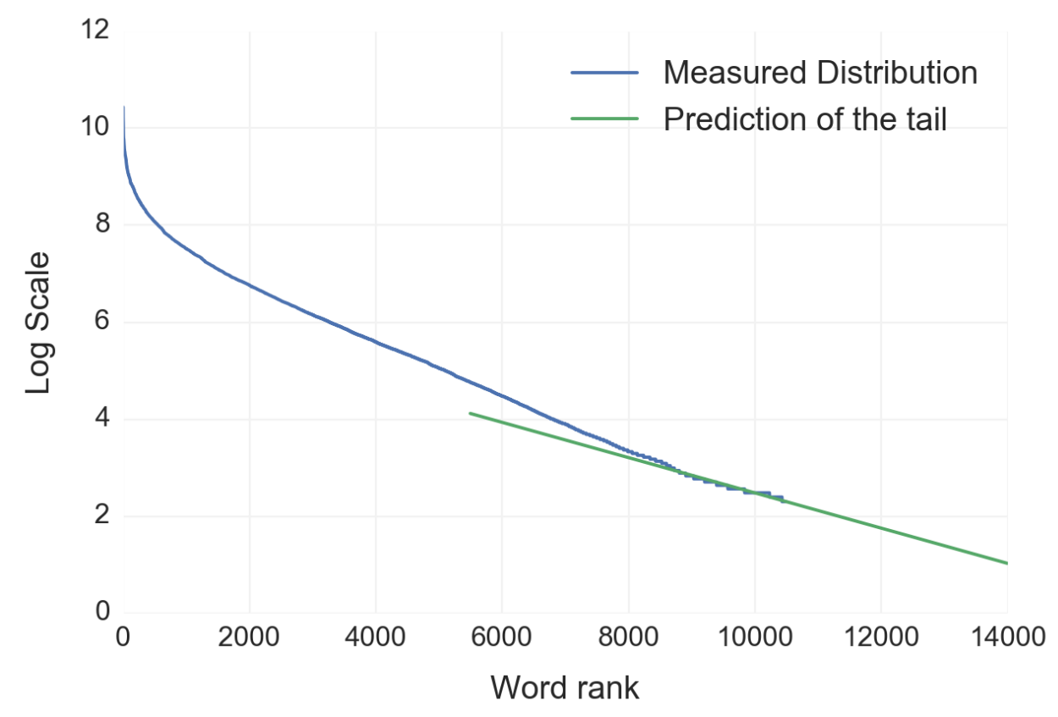
Data Visualization

Total word count over the years



How much of the data was not extracted?

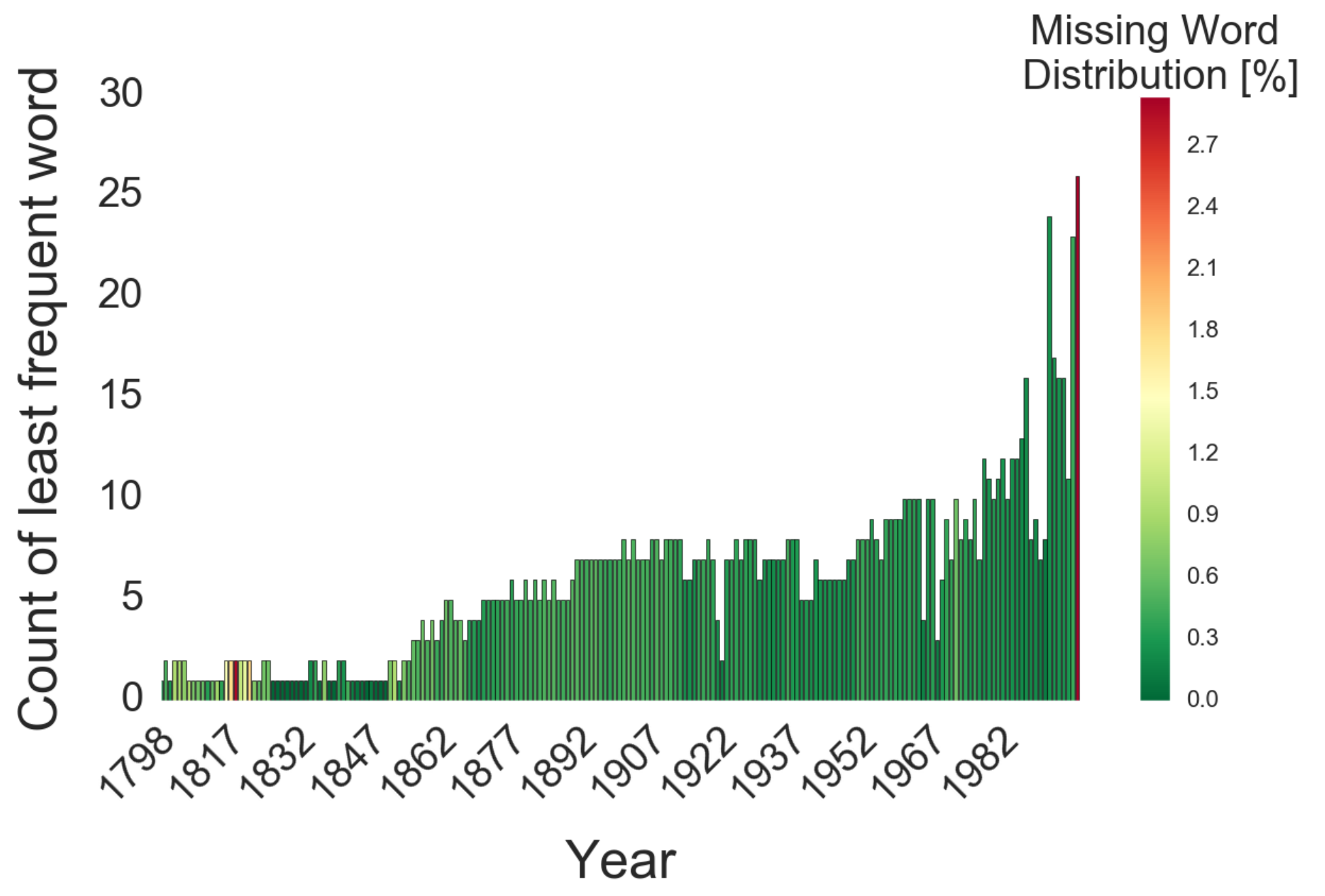
- Long tail distribution is linear in log-lin scale
- Linear regression: We predict the tail of the distribution.



We can predict the percentage of the distribution that we did not extract.

We can see that in theory we did not miss an large part of the word distribution (graph on the right)

Number of occurrences of the least frequent word with percentage of the data that was missed



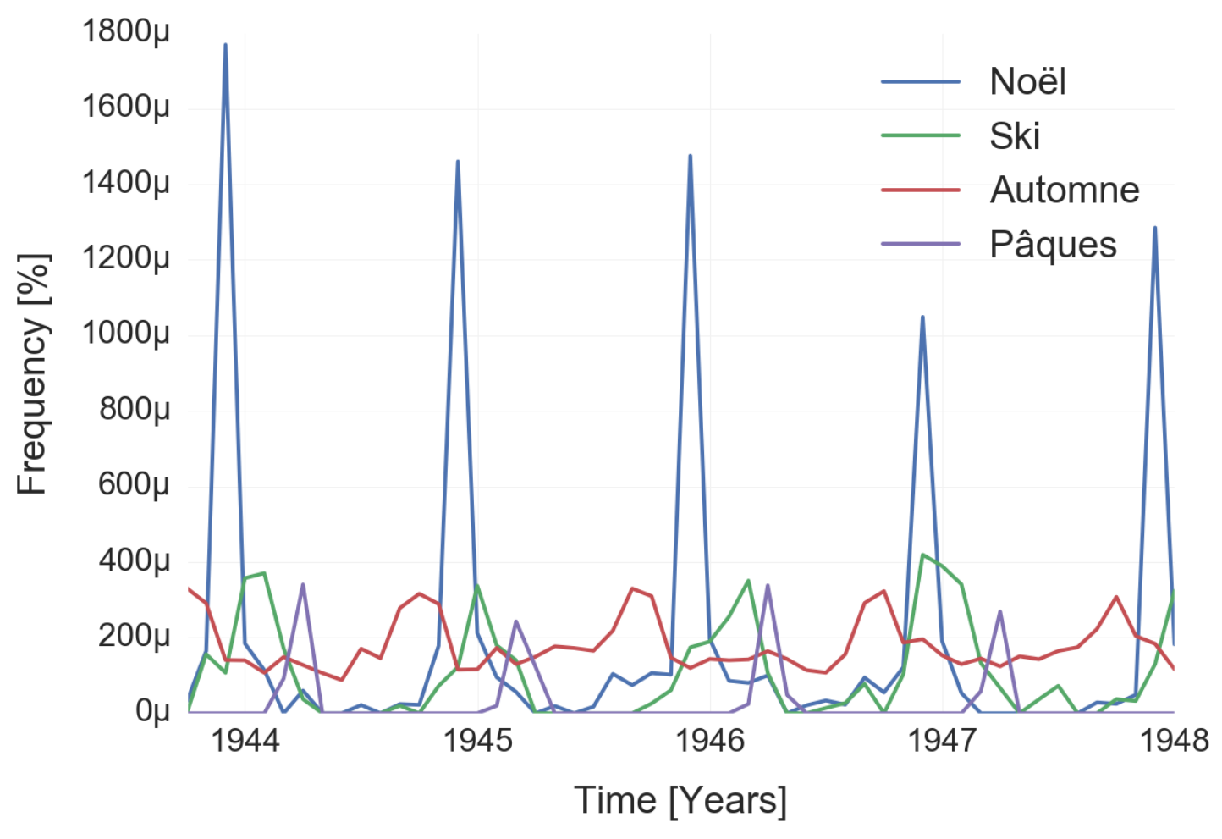
Words with interesting time series

To find relevant time series, several methods were used:

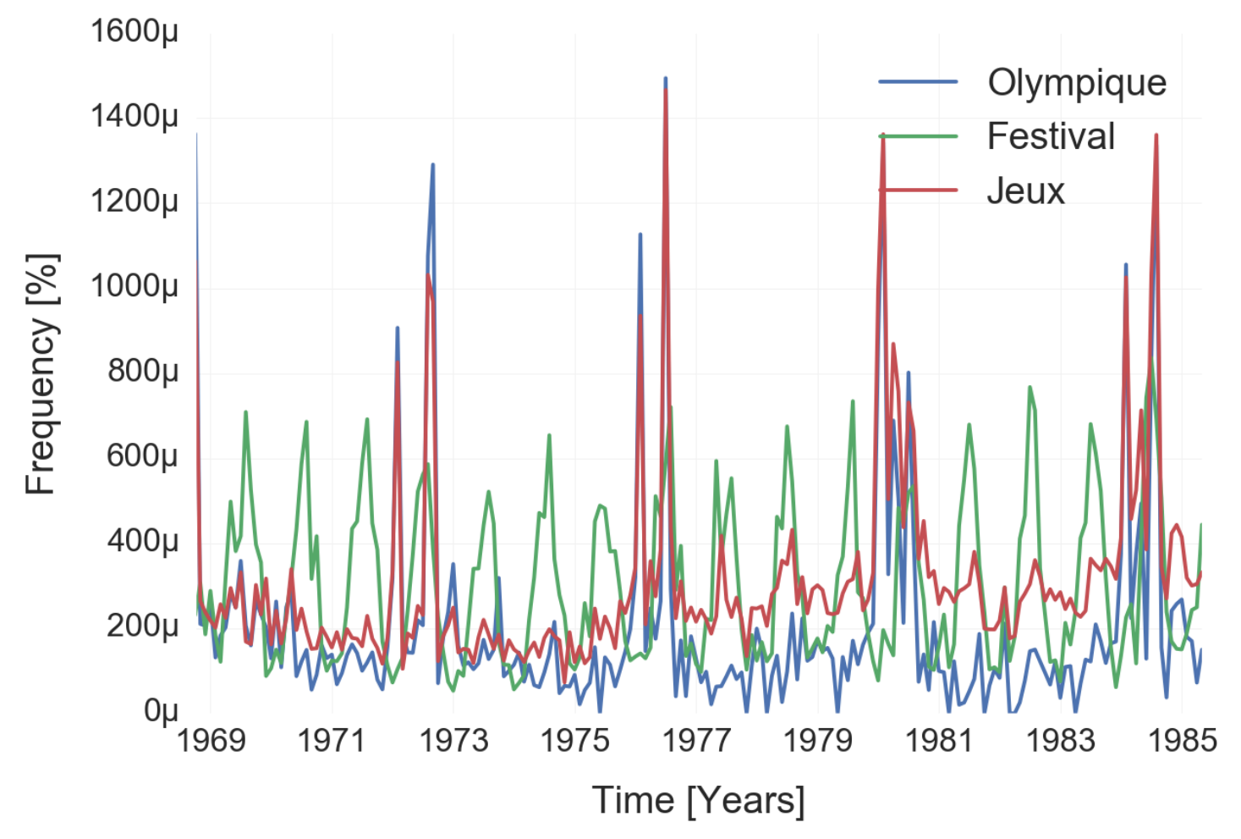
- Pearson correlation : Computing similarity between time series
- Fourier transform : Finding words with periodicity
- Gradient : Finding decreasing and increasing time series
- Dendogram clustering
- Frequency ranking
- Manual Search
- Working with smoothed time series : Rolling mean, Interpolation

We show below a few interesting results:

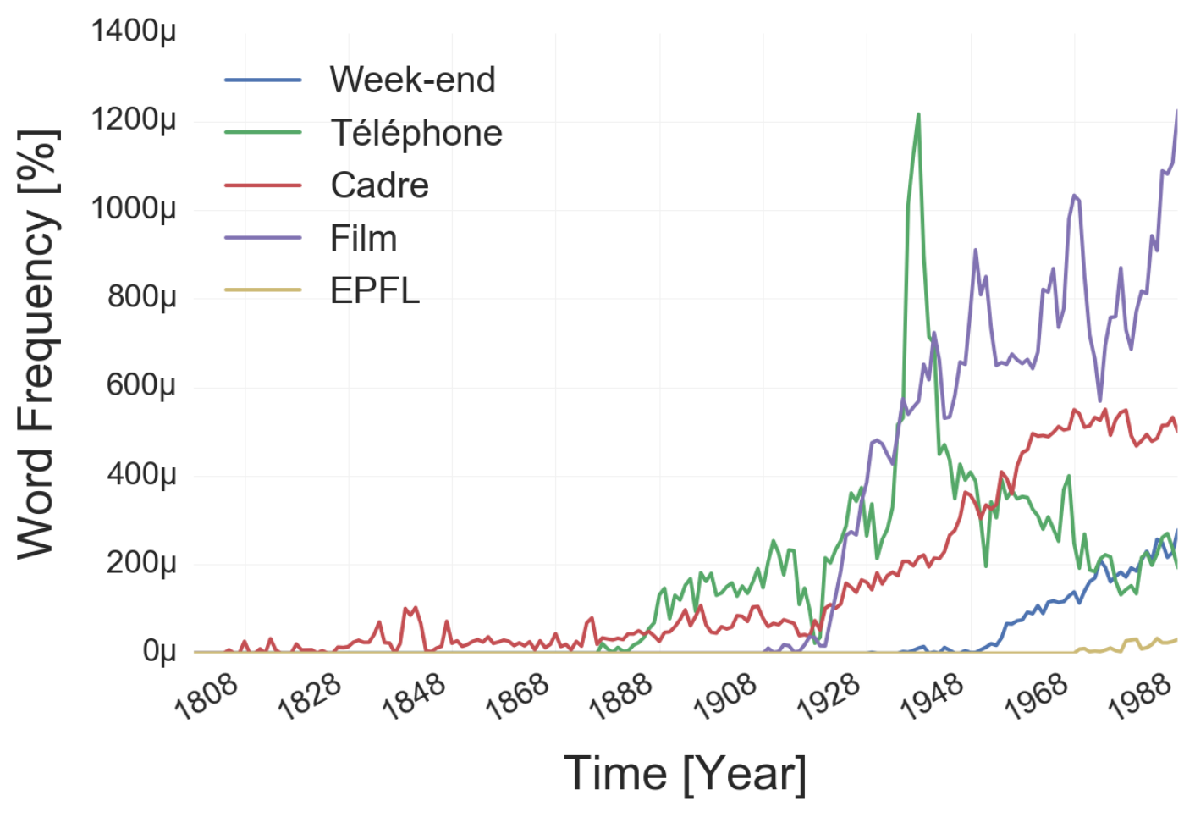
Words with monthly periodicity



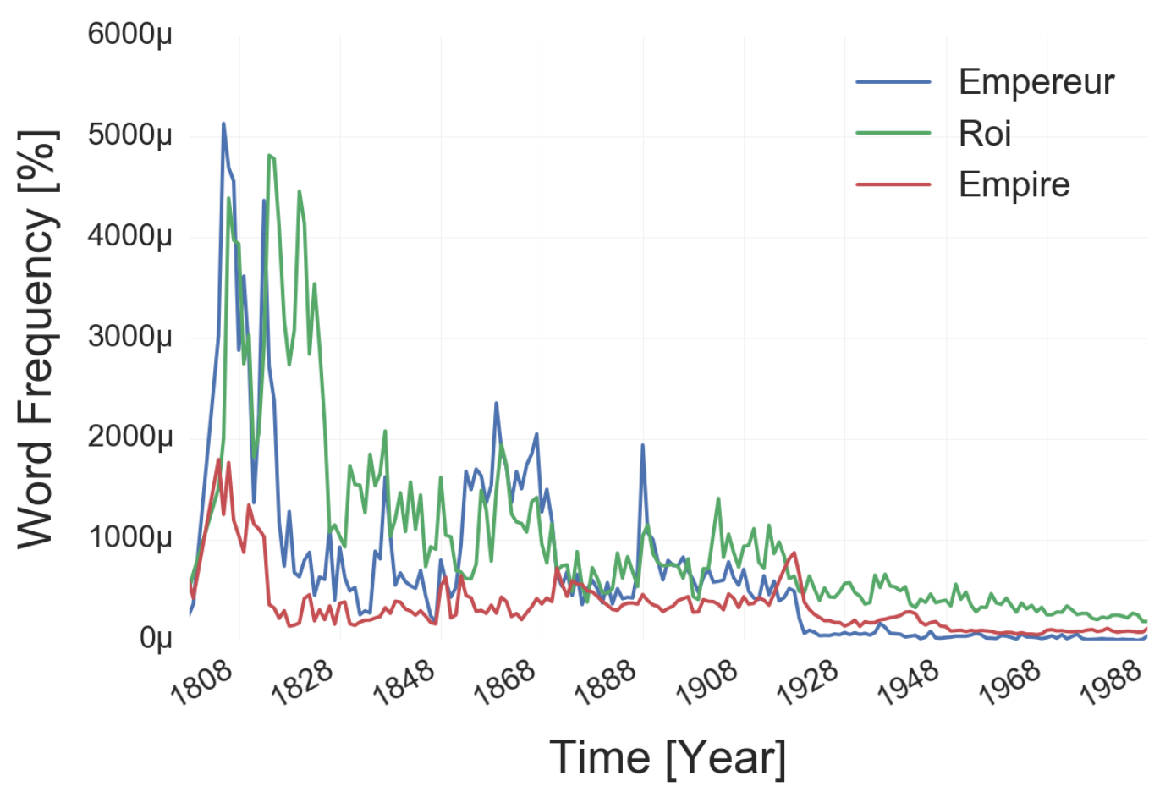
Words with multi year periodicity



Appearing Words

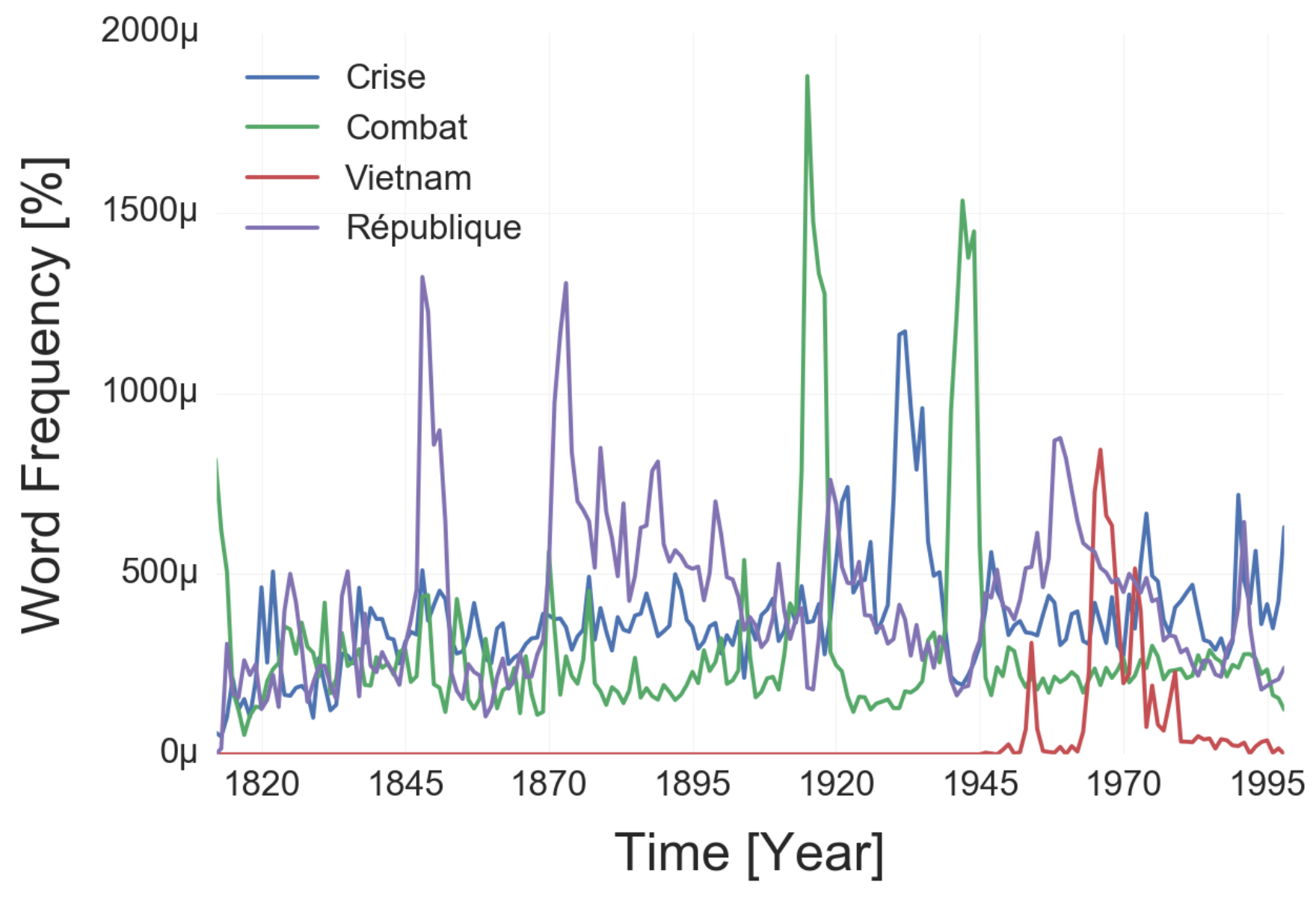


Disappearing Words



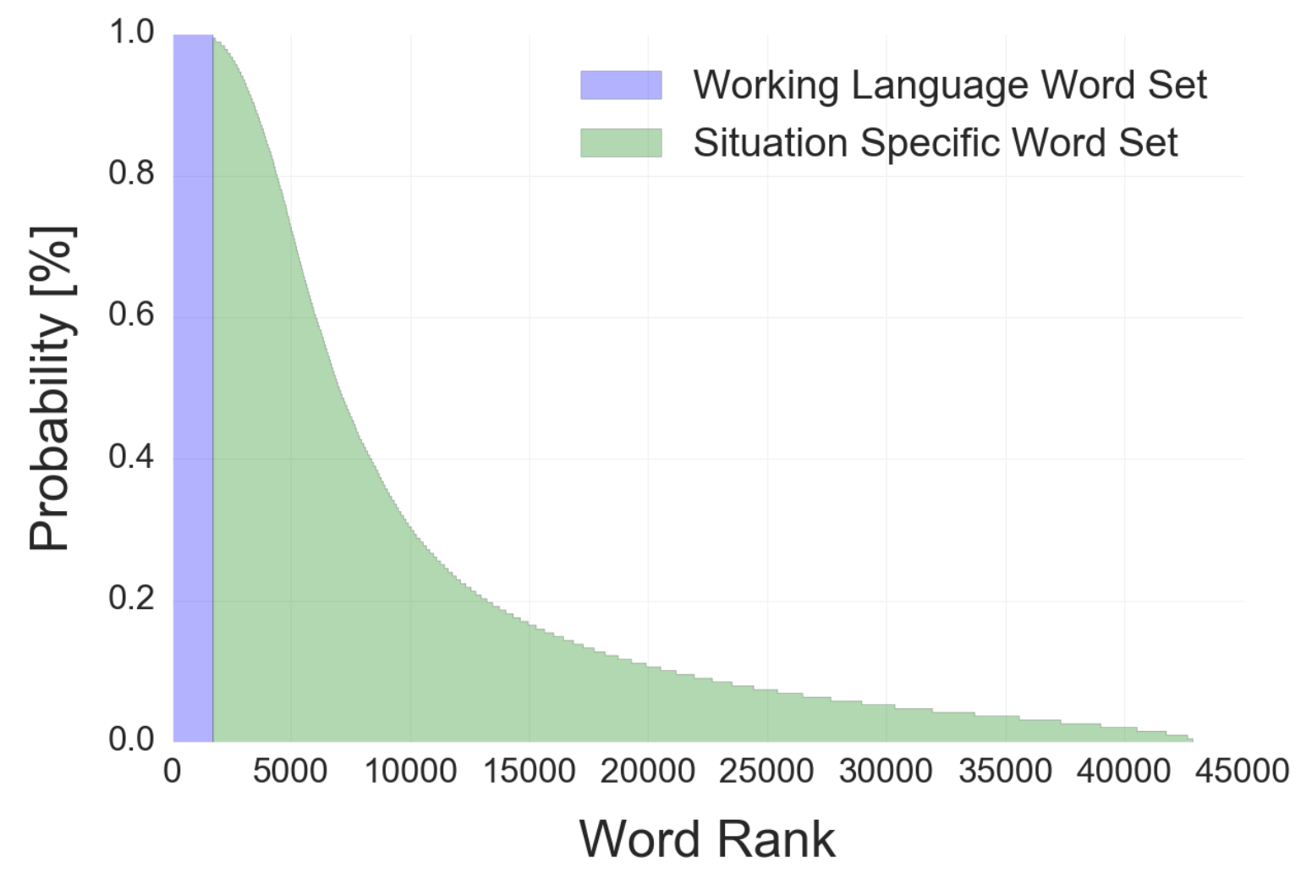
Historic Words

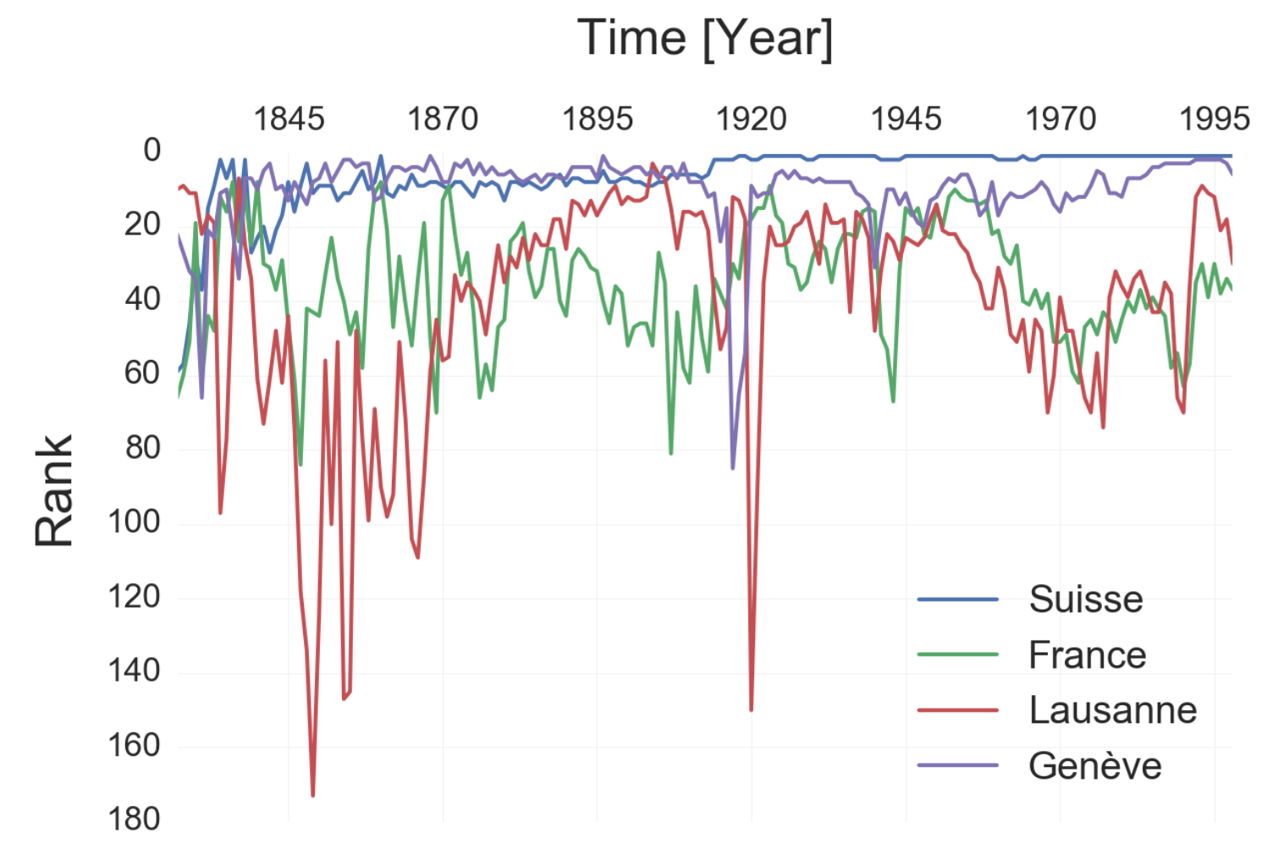
The time series of certain words follow historic events.
1848: Second republic of France
1870: Third republic of France
1918: First World War
1929 - 31: Financial Crisis
1939-1945: Second World War
1955 – 1975: Vietnam War



Ranking of the words:

We have analyzed how the set of 15'000 most frequent words evolved over time. We plot on the left the probability of finding the nth most frequent word in a given year. We see that the 1700 first words are always present (working language, P = 1) and then the other words are situation specific, as the curve takes some time to reach 0, we see that the word set in these two journals is very large. On the right we see the evolution of the rank of certain words over time (1 = most frequent).



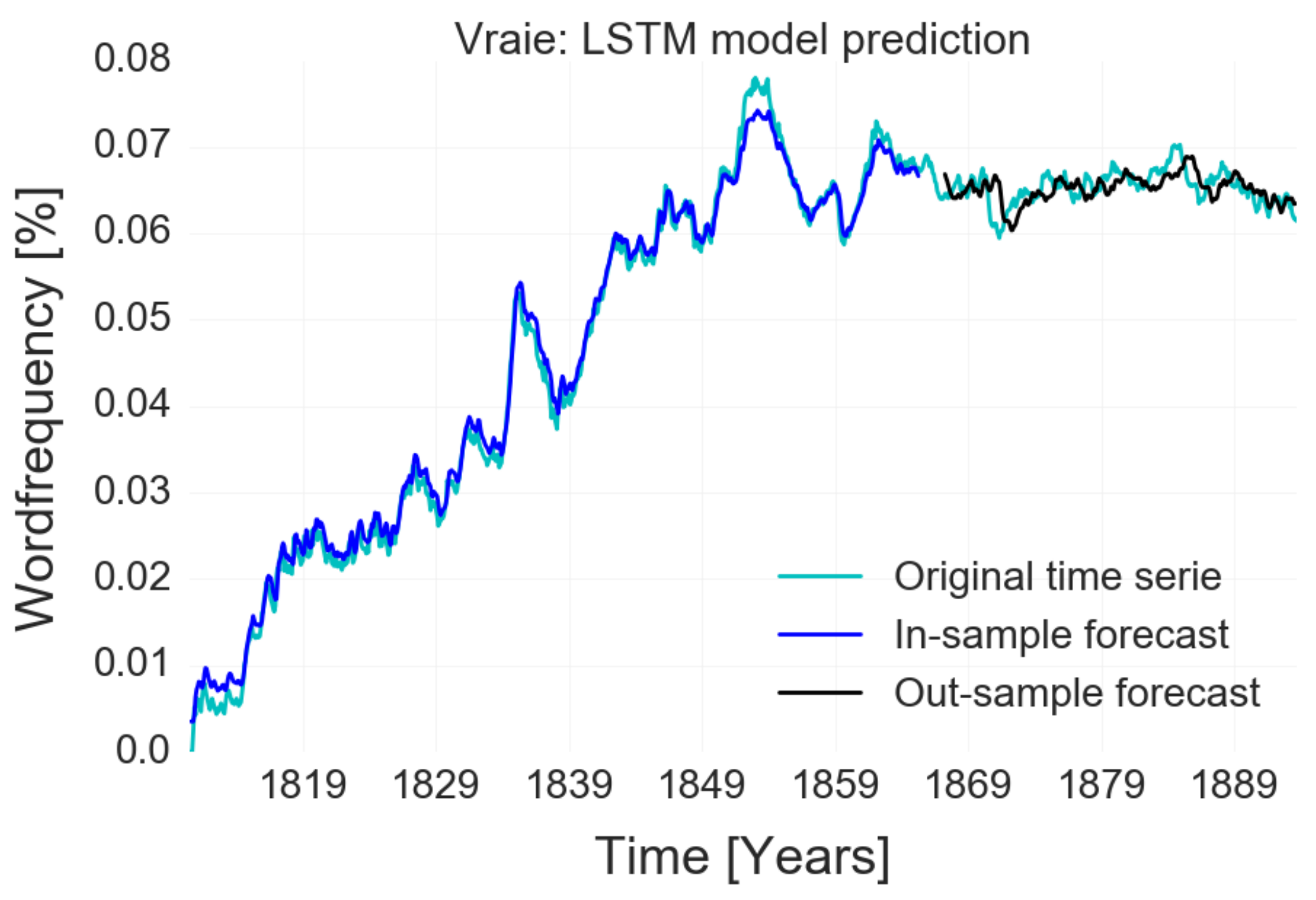


Word Frequency Prediction

LSTM model prediction:

Model: A simple LSTM (long short term memory). The problem is formed as a regression task with a RNN (Recurent neura network).

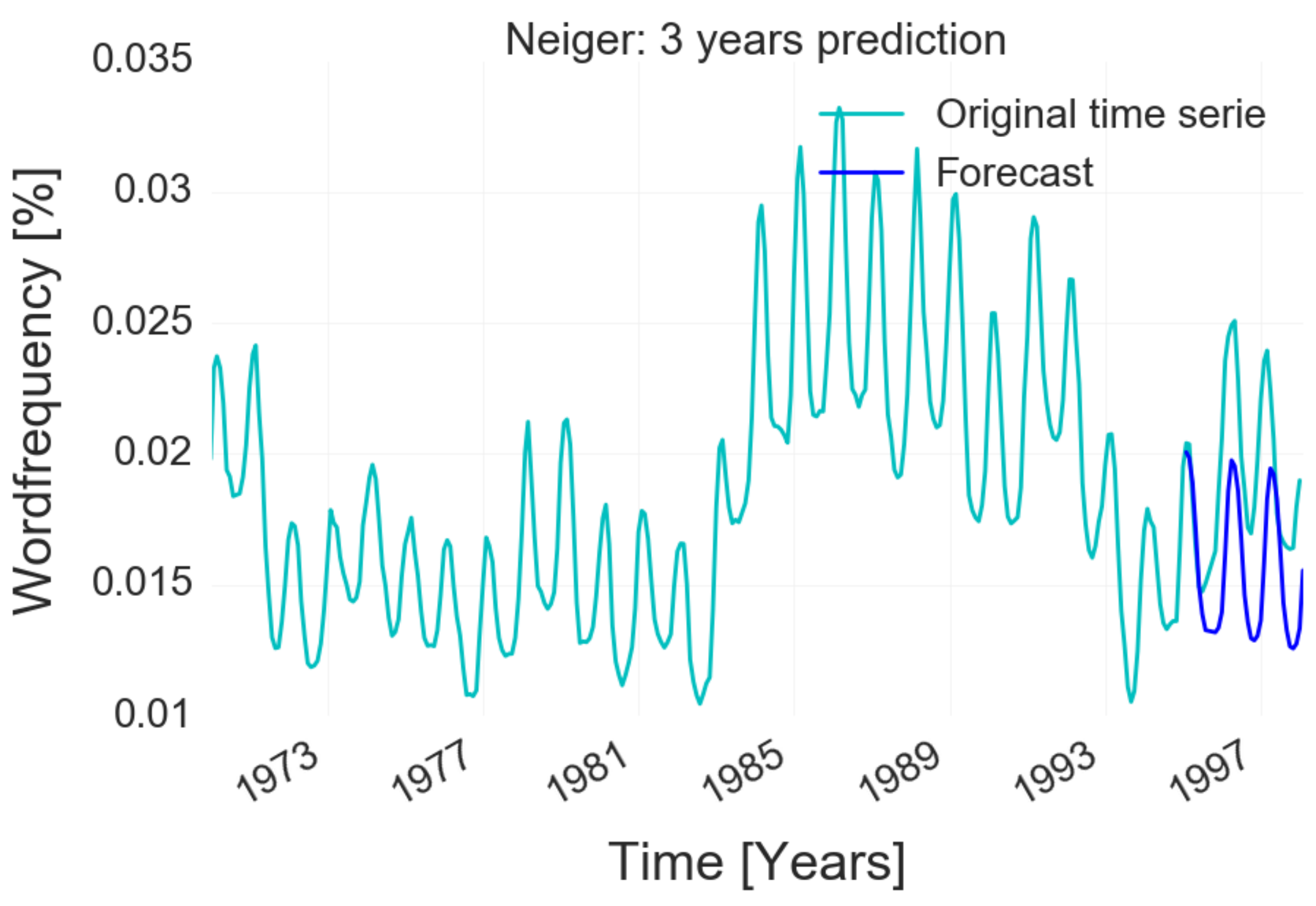
Result: The model is not making a true forecast. It has simply learnt to output the previous time value with some minor changes. In other words, it simply mimicks the time serie. It make sense as the model is trying to reduce the error and the previous time value are not too far away from the future time value.



SARIMA model prediction for seasonal words:

Model: A SARIMA (Seasonal autoregressive integraded moving average) model is a combination of autoregression with moving average component plus seasonal component in order to predict the future time values.

Result: The model is able to predict the correct seasonality of the word and outputs a coherent local trend. It is not able to integrate changing trend which can be regarded as random movement. The output is a repetitive sequence in the same direction. The reliability of the prediction decreases as we increase the length of the prediction.


Neiger: 3 years prediction
Original time serie
Forecast
Wordfrequency [%]
Time [Years]