Applied Data Analysis – Le Temps Dataset Word Frequency Prediction

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Extract Transfom Pipeline

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

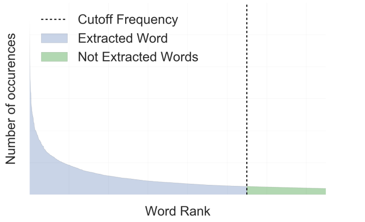
Extraction: Counting the 3000+ most frequent words per month

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. Cutoff Frequency: Removal of words that were not present enough

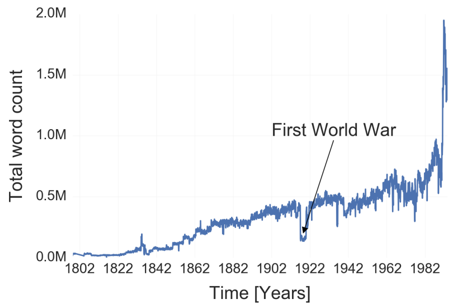
Result: Time serie of the frequency of each word

Long Tail Distribution of words:
Due to our cutoff, we miss a part of the data:



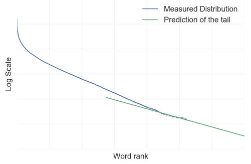
Data Visualization

Total word count over the years



How much of the data was not extracted?

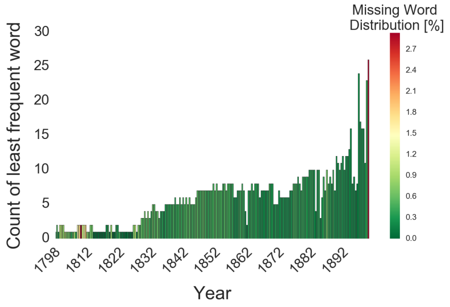
- Power Law distribution is linear in log-lin scale
- Linear Regression of the distribution to predict the non extracted part



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

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

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

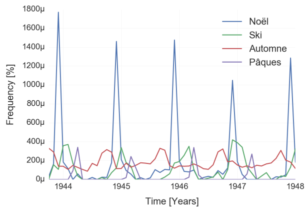


Words with interesting time series

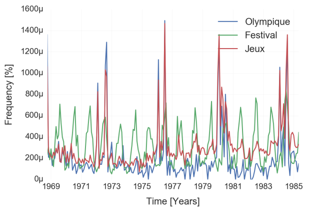
To find relevant time series, several methods were used:

- Pearson Correlation : computing similarity between word
- Fourier Transform : Finding words with periodicity
- Gradient : Finding decreasing and increasing time series
- Dendogram clustering
- Frequency ranking
- Manual Search
- Search of the smoothed out series (rolling mean)

Words with monthly periodicity



Words with multi year periodicity



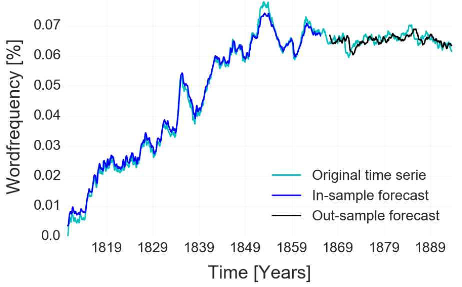
Word Frequency Prediction

LSTM model prediction:

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

Result: The model is not making true forecast. it has simply learn to output the previous time value with some minor change. In oder words, it simply mimick the time serie. It make sense as the model is trying to reduce the error and the previous time value are not to far away from the future time value.

Vraie: LSTM model prediction

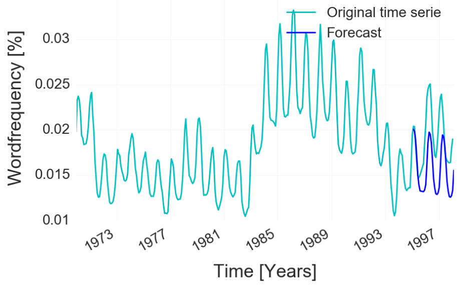


SARIMA model prediction for seasonal words:

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

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

Neiger: 3 years prediction


Wordfrequency [%]
Original time serie
Forecast
Time [Years]