Information Theoretic Approach to Identifying Vietnamese Stopwords
Natural Language Processing Project and Reading Notes
July 7, 2022

Contents

1 Introduction

Natural language processing (NLP) is an important application of machine learning (ML) in data science. Prior to processing data with with NLP algorithms, words of insignificance, or *stop words*, are usually removed. For example, words like *the* and *an* do not add context of the sentence with respect to some NLP algorithms, and thus are commonly removed in the data preparation period. Now, *how do we decide the significance of a word?* This question was proven to be difficult, as there is no agreed upon list of stop words among NLP experts. Stop words lists vary depending on the NLP algorithm used and the language at hand. Furthermore, these lists are usually curated manually, and as a result, there is an inherent bias towards more popular languages like English. **Gerlach, Shi, and Amaral (2019)** proposed an universal information theoretic approach to the identification of stop words, which was used to identify stop words in English, Portuguese, Chinese, and German. In this project, I will apply this method to generate a stop word list from a curated Vietnamese News corpus found here https://github.com/binhvq/news-corpus.

2 Method

In this section, a summary of the information theoretic method highlighted in [1] is presented.

2.1 Shannon Entropy

2.1.1 Parameters

Consider a corpus $\mathcal{C}$ with $\mathcal{D}$ documents in total.

- $n(w, d)$: the number of occurrences (tokens) of a word w in document d

- $n(d) = \sum_w n(w,d)$: the number of tokens in document d
- $n(w) = \sum_d n(w,d)$: the frequency of word w
- $N = \sum_{w,d} n(w,d)$: the total number of tokens in the entire corpus
- $p(w) = \frac{n(w)}{N}$: the relative frequency of a word

For every word w, its distribution over all documents is

$$p(d|w) = \frac{p(w,d)}{p(w)} = \frac{n(w,d)}{n(w)}$$

2.1.2 Formal definitions

In information theory, **Shannon entropy** H measures the uncertainty of a discrete random variable X. Let $P(X)$ be the probability mass function and $I(X) = -\log(P(X))$ be information content of X, then its Shannon entropy is defined as

$$H(X) = E[I(X)] = E[-\log(P(X))] = -\sum_{i=1}^{n} P(x_i)\log_b(P(x_i))$$

Moreover, the **conditional entropy** of random variables X and Y is

$$H(X|Y) = -\sum_{i,j} p(x_i, y_j)\log\left(\frac{p(x_i, y_j)}{p(y_j)}\right)$$

where $p(x_i, y_j)$ is the probability that $X = x_i$ and $Y = y_j$.

2.1.3 Conditional Entropy Model

Again, let w be a word and d be a document in corpus $\mathcal{C}$, then the conditional entropy is

$$H(w|\mathcal{C}) = -\sum_d p(d|w)\log p(d|w) = -\sum_d \frac{n(w,d)}{n(w)}\log\frac{n(w,d)}{n(w)}$$

If we randomly draw a token w, this conditional entropy quantifies (in bits) the amount of *uncertainty* that w provides about the document d it occurs in.

2.2 Random Null Model

Zipf's law for word frequencies: most words occur with a very low frequency, so $p(d|w)$ is expected to be undersampled.
Null model : ??

- $\tilde{H}(w|\mathcal{C})$: null model to estimate the expected entropy of randomly distributed words
- $\tilde{n}(w,d)$: random distribution of words across all documents after shuffling tokens across documents. Note that $n(w)$ and $n(d)$ are preserved.
- $\tilde{H}(w|\mathcal{C}) \propto \log(1 - e^{-n(w)/D})$

2.3 Information Content

- $\langle \tilde{H}(w|\mathcal{C}) \rangle$: the average over different realizations of the null model
- Information content of a word: $I(w|\mathcal{C}) \coloneqq \langle \tilde{H}(w|\mathcal{C}) \rangle - H(w|\mathcal{C})$. Low values of $I(w|\mathcal{C})$ can be used to identify stop words.

LaTeX [**?**] is a set of macros built atop TeX [1].

References

[1] Martin Gerlach, Hanyu Shi, and Luís A. Nunes Amaral. A universal information theoretic approach to the identification of stopwords. *Nature Machine Intelligence*, 1(12):606–612, 2019.