Automatic Information Extraction and Relevance Evaluation of Epidemiological Texts Using Natural Language Processing

Auss Abbood, University Osnabrück and Robert Koch Institut

Oktober 2018 - April 2019

Abstract

Pizza (Pizza et al. (2000)) is an understudied yet widely utilized implement for delivering in-vivo *Solanum lycopersicum* based liquid mediums in a variety of next-generation mastications studies. Here we describe a de novo approach for large scale *T. aestivum* assemblies based on protein folding that drastically reduces the generation time of the mutation rate.

Acknowledgment

Introduction

Motivation

The course goal at the beginning was to utilize natural language processing (NLP) in a dashboard for epidemiologists. The Signale group at the Robert Koch Institute where I wrote my thesis, focused to create custom solutions for different fields in epidemiology. Two groups at the RKI, the Epilag and the INIG needed to parse many texts for their work and thus we decided to look into their work to figure out how NLP could aid them in their everyday work. Both group's expertise is epidemiological surveillance. While the Epilag focuses on events within Germany, INIG is mainly interested in international outbreak news. The Epilag group exists for a longer time now and has clear working routines that incorporate several health departments of the Kreise and Länder of Germany. Even if I would have found a helpful use of NLP in their work-flow I would have been difficult to consider so many participants of this group that however are spread across Germany and have different access to resources of the RKI. INIG however is a young project that much more depends on a variety of text. They have much less work-power that can be spared to do tedious work opposed to Epilag where every health department needs to report to the RKI where then *only* the bigger picture is put together. Currently one person of the INIG team is responsible to read a fixed set of articles and filter out outbreak news that are important. These are then put into a database. This costs around 30 minutes every day, that should be spent writing assessments on the outbreaks for the German Ministry of Health. Thus, the idea developed to automate this process that least required expertise. Therefore, putting key points of a outbreak article into a database was the first goal of the thesis. Second, being able to describe the text based on this condensed form lead to the second goal: Use these keywords to determine the relevance of an article and then use this knowledge to write and recommendation system.

Signale

The Robert Koch Institute is the public health institute of Germanny- It is split into several Abteilungen. Most of them are interested in epidemiology. These Abteilungen are then split into several Fachbereiche. That is necessary so that every Fachbereich can focus on a specific topic and publish papers in this area. FG31 where I have been working is the IT FG and the only one that is working on Software solutions for the RKI. We stick out, since we are the interface between statistics and application. We create dashboards and are relatively independent from the other research and work done.

INIG

Also within the same Abteilung there are many groups that combine their expertise of certain diseases classes to form a super groups that does international surveillance. Their advantage is that they have expertise knowledge in different parts of infectious disease medicine- When they are reading news articles and they are unsure whether and article is interesting or not, they have the opportunity to consult each other which then leads to a very educated decision whether an outbreak article is interesting or not.

Background

Working At The RKI

Natural Language Processing

Stop words

Assuming we want to analyze text on the level of words, it can be very different which words are actually important for further analyzes. If a syntactical analysis would be done, then we would care for (almost) all words. Every word in this context has some information about the grammar and thus should be taking into consideration. But, if we should be focusing on semantics then we might end searching for this semantic information within only a handful of words. It could be that we are only interest in numbers, URLs, or company names. In this case it would only distract our learner, classification algorithm, or other machine learning tool, if the vast majority of words would not transmit the information of interest. Therefore, it is common practice to remove certain words that occur frequently in a language.

Tokenization

A token is just an abstraction of a piece of information. In NLP this can be a single character, word, punctuation, or sentences. The goal if tokenization is to split text into meaningful chunks that then can be analyzed on their own. Sentence tokenization already becomes important if a text is slightly longer and includes logical parts that need to be handled differently. Word-tokenization is important when you want to apply POS-tagging, followed by NER. These two methods depend on the being given words.

Bag Of Words

In a machine learning task we use the amount of data as a leverage to avoid doing explicit feature engineering, but to let the algorithm pick the feature from a set of potential

(Disease) Name Entity Recognition

Name entity recognition (NER) is based in the middle of a NLP pipeline. After sentences and words have been tokenized, and position-of-speak tagging (POS-tagging) was applied, it might be important for certain learning algorithms to infer the entity of the word at hand. Common examples are the distinction of ambiguous words as apple. In the beginning of a text it could be the fruit or a billion dollar company.

In the medical field such ambiguities rise not because the proper names are so indistinguishable form other common words, but because there are many form how to write a disease name and equally many abbreviations. Thus, disease-NER is a very important processing step.

Machine Learning

Naive Bayes Classifier

The naive Bayes classifier (NBC) is a probabilistic classifier. It describes a set of algorithms capable to learn to infer a label given a set of features. These algorithms are trained in a supervised fashion.

Methods

NLTK

> This is a block quote. This paragraph has two lines.
>
> 1. This is a list inside a block quote.
> 2. Second item. # Results ## Diagram

Algorithm

$$f(x) = pizza^2$$

```
lol = [rofl + 1 for rofl in roflcopter]
```

Discussion

References

Pizza, Mariagrazia, Vincenzo Scarlato, Vega Masignani, Marzia Monica Giuliani, Beatrice Arico, Maurizio Comanducci, Gary T Jennings, et al. 2000. "Identification of Vaccine Candidates Against Serogroup B Meningococcus by Whole-Genome Sequencing." *Science* 287 (5459). American Association for the Advancement of Science: 1816–20.

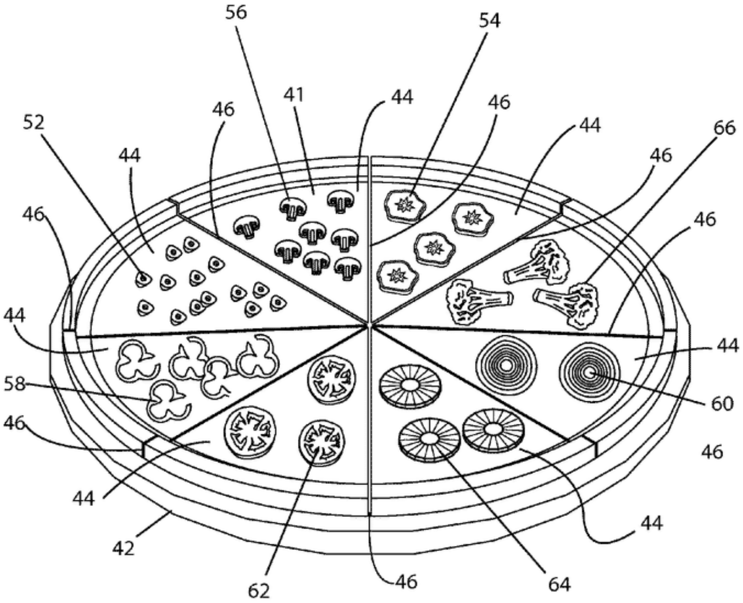

Figure 1: It's Pizza