Scientific texts analysis and classification by using machine learning methods

Nicolas LAFITTE

August 10, 2019

Abstract

The project is part of the "formation" Artificial Intelligence (AI) and Machine Learning (ML) held at Sorbonne University (Paris, France) during 2018-2019 academic year. As final homework, a work on the analysis and classification of scientific texts have been chosen.
Initial motivations of the student were to discover and learn mathematical algorithms and methods used in AI and ML in order to evaluate any application for the Fluigent project HoliFAB, which in part aim at optimizing microfluidic system layout. First works on this point led to the implementation of linear and non-linear regression functions.

However the motivation of the student for the current project raises from the frustration as a researcher to not be able to read and study all relevant papers of a field. Making a bibliography on specific topic is common and easy to perform thanks to different database and search engine available on the internet. However, when it comes to study a broad scientific field for a global understanding or new research investigation or market analysis, it is

bla bla

Contents

Chapter 1

Introduction

Chapter 2

Methods

The main source of machine-interpretable data for the materials research community has come from structured property databases. Beyond property values, publications contain valuable knowledge regarding the connections and relationships between data items as interpreted by the authors. To improve the identification and use of this knowledge, several studies have focused on the retrieval of information from scientific literature using supervised natural language processing3–10, which requires large hand-labelled datasets for training.

2.1 Data extraction

2.1.1 Sources

2.1.2 Converting pdf to dataframe

2.2 Text transformation

The texts need some transformation to be relevant for algorithms or models applied afterwards. Typically models sensitive to term appearance frequency to determine importance of the words will be biased by recurrent term like in English: *I, and, or ...* that do not carry much meaning.
A transformer is an abstraction that includes feature transformers and learned models: ie.A transformer implements a method, which converts one dataframe into another, generally by appending a new column.

2.2.1 Tokenization

Tokenization is the process of taking the text (such a sentence) and breaking it into individual terms (usually words).

2.2.2 Stop words

Stop words process takes as input a sequence of strings (e.g.the output of the tokenization) and drops all the stop words.
Stop words are words which should be excluded because the words appears frequently and carry as much meaning.

2.2.3 Stemming

2.3 Algorithm

2.3.1 Latent Dirichlet Algorithm

2.3.2 Others

Chapter 3

Results

Chapter 4

Conclusion and perspectives

Chapter 5

Annexes