# Scientific texts analysis and classification by using machine learning methods

Nicolas LAFITTE

September 6, 2019

**Abstract**

The project is part of the Artificial Intelligence (AI) and Machine Learning (ML) training course held at Sorbonne University (Paris, France) during 2018-2019 academic year. As final project, a work on the analysis and classification of scientific texts have been chosen.
The initial motivations were to discover and learn mathematical algorithms and methods used in AI and ML in order to evaluate any application for the Fluigent European project HoliFAB. It aims at adapting an existing pilot line for the production of microfluidic instruments, and develops hardware and software strategies for the optimization of the production and system. Some mathematical developments at the beginning of the project led to the implementation of regression functions that auto-place and auto-wire a system layout.

However the motivation for the current project raises from the frustration as a researcher to not be able to read and study all papers relevant of a field. Making a bibliography on a 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, the task often lacks an intensive study of the domain.

Text mining, which is the task of extracting meaningful information from text, is developed here on a database of scientific papers. More precisely Latent Dirichlet Algorithms (LDA) are studied in order statistically categorize text and extract topics. The tool helps to find and name topics and applied on different database of different years check for evolution in the scientific research and next innovation that will probably lead the market.

**Keywords:** mining, classification, clustering, information extraction, topic extraction, information retrieval, Latent Dirichlet Algorithm (LDA)

# Contents

# Chapter 1

# Introduction

The amount of text that is generated every day is increasing dramatically. This tremendous volume of mostly unstructured text cannot be simply processed and perceived by computers. 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.

# Chapter 2

# Implementation

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 processing, which requires large hand-labelled datasets for training.

## 2.1 Pipeline

## 2.2 Data extraction

### 2.2.1 Sources

### 2.2.2 Converting pdf to dataframe

## 2.3 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.

### 2.3.1 Tokenization

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

### 2.3.2 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.

### 2.3.3 Stemming

## 2.4 Algorithm

### 2.4.1 Latent Dirichlet Algorithm

### 2.4.2 Others

# Chapter 3

# Results

# Chapter 4

# Conclusion and perspectives

# Chapter 5

# Annexes

# Bibliography

[1] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. 2017.