

# Text Mining Techniques for Knowledge Extraction from Technical Documents

*Filippo Chiarello*

*2018-08-31*



# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Goal . . . . .	5
1.2	Problem . . . . .	5
1.3	Solutions . . . . .	5
1.4	Challenges: Understanding and Programming . . . . .	5
1.5	Research Questions . . . . .	5
1.6	Stakeholders . . . . .	5
<b>2</b>	<b>State of the Art</b>	<b>7</b>
2.1	Phases, Task, and Techniques . . . . .	8
2.2	Documents . . . . .	8
<b>3</b>	<b>Methods</b>	<b>11</b>
3.1	Patents . . . . .	11
3.2	Papers . . . . .	11
3.3	Projects . . . . .	11
3.4	Wikipedia . . . . .	11
3.5	Twitter . . . . .	11
3.6	Job Profiles . . . . .	11
<b>4</b>	<b>Applications and Results</b>	<b>13</b>
4.1	Patents . . . . .	13
4.2	Papers . . . . .	13
4.3	Projects . . . . .	13
4.4	Wikipedia . . . . .	13
4.5	Twitter . . . . .	13
4.6	Job Profiles . . . . .	13
<b>5</b>	<b>Future Developments</b>	<b>15</b>
5.1	Marketing . . . . .	15
5.2	Research and Development . . . . .	15
5.3	Design . . . . .	15
5.4	Human Resources . . . . .	15
<b>6</b>	<b>Conclusions</b>	<b>17</b>



# Chapter 1

## Introduction

### 1.1 Goal

### 1.2 Problem

### 1.3 Solutions

### 1.4 Challenges: Understanding and Programming

#### 1.4.1 Understanding

#### 1.4.2 Programming

### 1.5 Research Questions

### 1.6 Stakeholders

Marketing

Research and Development

Design

Human Resources

Other Stakeholders



## Chapter 2

# State of the Art

The analysis of technical documents require the design of processes that rely both on programming and Natural Language Processing techniques and on the understanding and knowledge of field experts. While the first techniques are codified and explicit, the second are sometimes implicit and always harder to systematize. In this section i treat these two groups of techniques in the same way to give to the reader a sistematic litterature review on these topics. For this reason the sections of this chapter has the sequent structure:

- At a first level we have two sections 2.1 and 2.2, reviewing respectively the process of *programming and Natural Language Processing* and of *undestanding and knowldege of field experts application*;
- Section 2.1 has a subsection for each of the *phases* showed in figure tot. These subsections goes from 2.1.1 to 2.1.7;
- Each subsection from 2.1.1 to @refref(sotatoolscommunicate) contains the relative Natural Language Processing *task* that are relevant for the analysis of technical documents, for example Document Retrieval 2.1.2.1, Part-Of-Speech-Tagging; #ref(sotatoolstransformpos) or Named Entity Recognition 2.1.5.4.
- Each task subsection describes the relevant *techniques* to perform that task;
- Since the second section 2.2 describes less systematics phases, task and techniques this section opens with a first subsection 2.2.1 that focuses on the studies of the problems of using expert knowledge in an analytical process and which are the techniques to convert this knowledge in a format that is usable in a Natural Language Processing workflow.
- Finally, always section 2.2 has a subsection for each of the technical *documents* I analyzed (aggiungi gancio con introduzione). These subsections goes from 2.2.2 to 2.2.7;

## 2.1 Phases, Task, and Techniques

### 2.1.1 Program

### 2.1.2 Import

#### 2.1.2.1 Document Retrieval

### 2.1.3 Tidy

### 2.1.4 Transform

#### 2.1.4.1 Stemming

#### 2.1.4.2 Lemmatisation

#### 2.1.4.3 N-Grams

#### 2.1.4.4 Part-of-Speech Tagging

#### 2.1.4.5 Regular Expressions

### 2.1.5 Model

#### 2.1.5.1 Document Classification

#### 2.1.5.2 Network Analysis

#### 2.1.5.3 Sentiment Analysis

#### 2.1.5.4 Named Entity Recognition

#### 2.1.5.5 Vector Semantics

#### 2.1.5.6 Topic Modelling

### 2.1.6 Visualize

### 2.1.7 Communicate

## 2.2 Documents

### 2.2.1 Understand

Remember to modify the DS workflow



2.2.1.1 The problem of byases

2.2.1.2 The Importance of Lexicons for Technical Documents Analysis

2.2.2 Patents

2.2.3 Papers

2.2.4 Projects

2.2.5 Wikipedia

2.2.6 Twitter

2.2.7 Job Profiles



## Chapter 3

# Methods

In this chapter I describe the methods applied for the analysis of different types of documents containing technical information. The methods are ensemble of Natural Language Processing (NLP) and Text Mining techniques described in @ref(sota\_tools), re-designed depending on the analyzed document and the analysis goal.

Table tot summarise the relations between the documents under analysis (introduced in section @ref(sota\_documents)) and the NLP techniques.

Table documents vs tools

Table algorithms vs tools

### 3.1 Patents

### 3.2 Papers

### 3.3 Projects

### 3.4 Wikipedia

### 3.5 Twitter

### 3.6 Job Profiles



## Chapter 4

# Applications and Results

Some *significant* applications are demonstrated in this chapter.

### 4.1 Patents

### 4.2 Papers

### 4.3 Projects

### 4.4 Wikipedia

### 4.5 Twitter

### 4.6 Job Profiles



## Chapter 5

# Future Developments

5.1 Marketing

5.2 Research and Development

5.3 Design

5.4 Human Resources





## Chapter 6

# Conclusions

We have finished a nice thesis