UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'ENERGIA DEI SISTEMI
DEL TERRITORIO E DELLE COSTRUZIONI

RELAZIONE PER IL CONSEGUIMENTO DELLA
LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE

The evolution of jobs: automatic job titles extraction based on a text mining approach

SINTESI

RELATORI

Prof. Ing. Gualtiero Fantoni
Dipartimento di Ingegneria Civile e Industriale
Università di Pisa

Dott. Pietro Manfredi

IL CANDIDATO

Francesca Mattolini

francescamattolini@ymail.com

Sessione di Laurea Magistrale del 18/06/2020

Anno Accademico 2019/2020

Consultazione NON consentita

Abstract

Nowadays, the job market is continuously evolving, and elementary professional figures are progressively diversifying into more specialised profiles in numerous application fields. To categorise every occupation and its related skills and technologies, worldwide job databases have been developed. These frameworks share common difficulties concerning data completeness and upgrades, including manual and sporadic updates. By exploiting text mining techniques, this thesis focused on the development of an automatic method for the identification and extraction of job titles from text sources. Firstly, a set of clues was defined and collected to uniquely allow the identification of job titles within any text files. The collection of these markers was performed by analysing scientific papers, occupational databases, online dictionaries, and multimedia vocabularies. In a second step, the identified set of clues was utilised to automatically extract compound job titles. This extraction was achieved by generating a list of rules able to recognise any kind of complement that specifies the application area of each base job title. To demonstrate its broad effectiveness, the implemented algorithm was tested in the meaningful topic of the green economy. Lastly, a demo-web application was created and published online to allow the user to assess the algorithm functionalities. Overall, we believe this work could constitute a useful approach for worldwide database authorities, international recruitment companies, and institutions involved in professional formation and training courses.

Sommario

Oggi il mercato del lavoro è in continua evoluzione e le figure professionali elementari si stanno progressivamente specializzando in vari campi di applicazione. In tutto il mondo sono stati creati database lavorativi per classificare le professioni e le loro relative competenze e tecnologie. Le difficoltà caratterizzanti questi framework riguardano la completezza e l'aggiornamento manuale e sporadico dei dati. Sfruttando le tecniche di text mining, questa tesi si è concentrata sullo sviluppo di un metodo automatico per l'identificazione e l'estrazione di titoli di lavoro da fonti testuali. In primo luogo, attraverso l’analisi di articoli scientifici, banche dati professionali, dizionari online e vocabolari multimediali, è stata definita e raccolta una lista di marcatori capaci di individuare titoli lavorativi nei file di testo. Nella seconda fase, l'insieme di identificatori collezionati è stato utilizzato per estrarre automaticamente i titoli lavorativi composti, mediante la generazione di regole per il riconoscimento dei complementi che specificano l'area di applicazione dei titoli lavorativi elementari. Per dimostrarne la sua efficacia, l'algoritmo è stato testato attraverso lo sviluppo di un caso studio su un argomento attualmente di spicco: la green economy. Infine, è stata creata e pubblicata online un'applicazione per consentire all'utente di valutare le funzionalità dell'algoritmo. Si ritiene che il presente lavoro possa costituire un approccio utile per le autorità dei database mondiali, le società di reclutamento internazionali e le istituzioni coinvolte nella formazione e nel training professionale.

1. Introduction

According to historians, only thirty-six different job types were present during the Chinese Tang Dynasty (618-907)[1]. Curiously, this age marks the origin of a famous Chinese saying that *"every trade has its master"*. Today, the job market is evolving so fast that it is tough to give an exact number of occupations affecting our lives. As professional profiles are continuously changing, disappearing, and emerging, the collection and structuring of data related to job titles has become complicated. For instance, in recent years, the figure of the *manager* has diversified into more specialised profiles like *PI managers*, *IT managers*, *project managers*, and *intergenerational engagement managers*[2].

Today, the need for a complete collection and continuous update of the various occupations into job databases has arisen to match supply and demand in job markets. To exhaustively categorise job profiles, a description of roles, competencies, experience requirements and education levels would be desirable[3]. The creation of worldwide databases with the cited characteristics would allow not only a better understanding of the existing jobs but also the development of more sophisticated systems able to perform increasingly complex services with employment data. Among different results, this approach could allow overcoming current difficulties in career planning, job search, trend identification, and policy design[4].

Currently, the worldwide job databases (*O*NET*[5], *ESCO*[6], *ChinaJob*[7]) share three common issues concerning data upgrade: databanks are updated manually, every given period, and using unclear methods. The enormous amount of online-collected information has not yet been organised according to a common strategy by the various worldwide databases. Notably, this point concerns the persistence of ambiguous categories in the existing databanks, the diversity of detail in the available data (i.e., distinct grainy nature), and incompleteness of the provided information. The second problem which links big data collection and the labour market is characterised by a difficulty in defining a standard nomenclature for all the occupations. Even if the same professional figure can be sought by recruiting companies and offered by employment agencies, it is essential to stress that this figure may be defined in a slightly different way from an involved player to another one[8]. For this reason, the occupations collected in the job databases are not exhaustive, as they do not include a large variety of synonyms in the definition of each single job title. That being said, this work focused on solving there issues by developing an automatic and structured method to continuously extract and collect job titles

[1] Nienhauser W. H., "*Tang Dynasty Tales: A Guided Reader*", World Scientific Publishing Co Pte Ltd, 2010.

[2] LI W., Dong and SHI Kan, "*A brief introduction to the development of the U.S. national standard occupational classification system and its implications to China*", 2006.

[3] Borgman C.L., Case D.O., "*Database guides - are they doing a good job?*", 2019.

[4] Niederman F.A., Sumner M., "*Resolving the IS Skills Paradox: A Content Analysis of a Jobs Database*", 2019.

[5] https://www.onetonline.org/

[6] https://ec.europa.eu/esco/portal/home

[7] http://www.chinajob.com/

[8] Paliotta A.P., Lovergine S., "*Web data mining e costruzione di profili professionali: Il Business analyst nelle inserzioni di lavoro online*", SINAPPSI, 2017.

from publications and other text sources. Being an automated approach, it is possible to process large amounts of data in just seconds, also avoiding errors which would be usually encountered in manual processes[9]. Furthermore, the process would leave the possibility for governments and authorised institutions to define recognised structures at a later stage, by merely listing job titles and without any categorisation[10]. Finally, in contrast to today's updates, the present methodology can be performed regularly and continuously.

The developed process start from the analysis of the world databases currently in use, and proceeded by defining a set of structured rules for the identification of job titles within text files, exploiting advanced *Natural Language Processing*[11] and *Text Mining*[12] techniques. Through the defined rules, it was then possible to develop a robust methodology for the automatic extraction of job profiles from text files. Mainly, the work focused on collecting as many job profiles as possible, in all their variants and terminological definitions, without any previous classification of professions.

2. Methodology

2.1 Definition of job title

As shown in *Figure 1*, the first step consisted in gaining an understanding on what a job title really is, thus defining it. This phase is crucial for the whole process, since a good explanation leads to good rules. For this reason the *Cambridge Dictionary* definition of job title as "the name of a particular job in an organisation" was adopted[13]. As a synonym of job profile, this interpretation includes a description of the exact task involved in a specific job, and of the skills, experience, and personality a person would need to do the job.

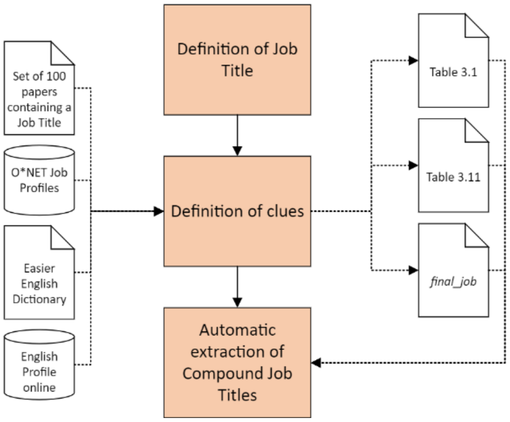


Figure 1. Activities map of the "Methodology" Section.

2.2 Definition of clues

After defining job titles, the second step consisted in analysing how documents mention them. The analysis of how job titles are mentioned within texts helps us identifying what are the recurring rules to give as input to the machine for the process automation. To identify as many job titles as possible, it was necessary to determine how they were mentioned and described within the different available sources (e.g., scientific publications, global employment websites, online databases). This process is shown in *Figure 2* and better described in the following sub section.

[9] Bonaccorsi A., Apreda R., Fantoni G., "*Cognitive and Motivational Biases in Technology Foresight, Technological Forecasting and Social Change*", 2017.

[10] Bruni M., Paliotta A.P., Tagliaferro C., "*Classificare le professioni. Una proposta metodologica, Professionalità*", 31, n. 66, pp. 29-38, 2001.

[11] Nawab K., Ramsey G., Schreiber R., "*Natural Language Processing to Extract Meaningful Information from Patient Experience Feedback*", Applied Clinical Informatics, Volume 11, Issue 2, pp. 242-252, 2020.

[12] Waegel D., "*The Development of Text-Mining Tools and Algorithms.*", Ursinus College, 2006.

[13] European Commission, "*A new skills agenda for Europe, Communication from the commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions*", COM, pp. 381-final, 2016.

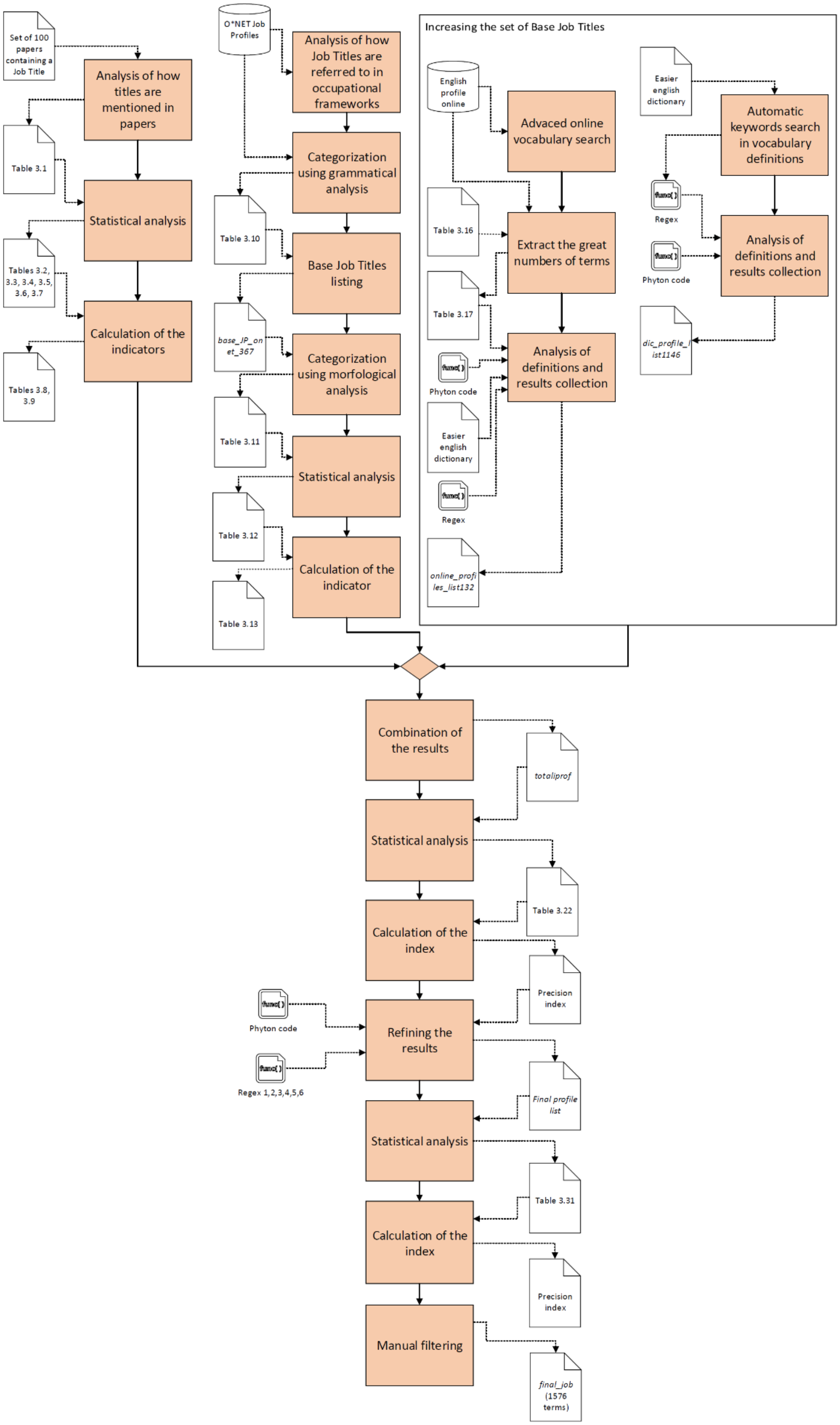
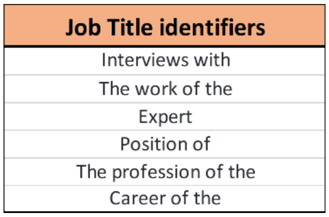

Set of 100 papers containing a Job Title
Analysis of how titles are mentioned in papers
Table 3.1
Statistical analysis
Tables 3.2, 3.3, 3.4, 3.5, 3.6, 3.7
Calculation of the indicators
Tables 3.8, 3.9
O*NET Job Profiles
Analysis of how Job Titles are referred to in occupational frameworks
Categorization using grammatical analysis
Table 3.10
Base Job Titles listing
base_JP_onet_367
Categorization using morfological analysis
Table 3.11
Statistical analysis
Table 3.12
Calculation of the indicator
Table 3.13
Increasing the set of Base Job Titles
English profile online
Advaced online vocabulary search
Easier english dictionary
Automatic keywords search in vocabulary definitions
Table 3.16
Extract the great numbers of terms
Regex
Phyton code
Analysis of definitions and results collection
Table 3.17
dic_profile_list1146
Phyton code
Analysis of definitions and results collection
Easier english dictionary
Regex
online_profiles_list132
Combination of the results
totaliprof
Statistical analysis
Table 3.22
Calculation of the index
Precision index
Phyton code
Refining the results
Regex 1,2,3,4,5,6
Final profile list
Statistical analysis
Table 3.31
Calculation of the index
Precision index
Manual filtering
final_job (1576 terms)

Figure 2. Activities map of the "Definition of clues" Section.

2.2.1 Analysis of how titles are mentioned in papers

This first analysis focused on how job titles are mentioned in sentences of papers. About 1000 phrases of 200 different papers containing a job profile have been analysed, and it was noticed that job titles were often preceded by recurring identifiers (i.e., adverbs, nouns, or short sentences), which could determine the presence of a job title (*Table 1*).



Job Title identifiers
Interviews with
The work of the
Expert
Position of
The profession of the
Career of the

Table 1. Job title identifiers in papers.

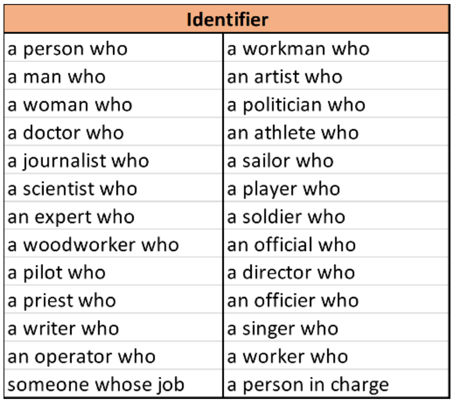
2.2.2 Analysis of how job titles are referred to in occupational frameworks

The complete list of job titles currently registered on *O*NET* framework was downloaded, and a logical analysis was conducted. It was noticed that the data could be divided into two macro-categories: *base job titles* (i.e., job titles made of a single noun, for example, *carpenters*), and *compound job titles* (i.e., occupations composed of a base job title preceded and/or followed by a limited number of adjectives and nouns specifying the area of application of the work, for example, *agricultural engineers*). This analysis was built on the idea that a new job title could derive from the combination of a *base job title* and a new term that made it more specific. Also, each equivalent *base job title* was extracted from the *compound job title*. Once all these base job titles were collected, a morphological analysis was carried out, and we noticed that lots of job titles ended with the suffixes *er*, *or*, *ist*, and *ian*. An analysis of word suffixes was then performed.

We then decided to automate the process by searching of compound titles constituted by one of the base job titles belonging to the created set, and by prefixes and/or objective complements within documents. The hypothesis behind this decision is that any job title can be defined as a combination of already known words (i.e., base job titles) and new adjectives/nouns referring to the base job titles, which specify and define new areas of application.

2.2.3 Increasing the set of base job titles using vocabularies

To extract as many compound job titles as possible, it was necessary to increase as much as possible the number of base job titles which were then searched in the texts through a code written in Python. Thus, two methods were identified, corresponding to a "top-down" and a "bottom-up" methodology. Both methods were based on the use of the English dictionary. For this reason, an analysis of how job titles definitions were created within the vocabulary was developed. A set of identifiers (*Table 2*), was found through a set of 100 job titles definitions analysed in the vocabulary. Also, it was observed that a limited number of words was often present between the indefinite articles (i.e., "a" or "an") and the names (e.g., "a doctor"). These words focused on the specific field in which the title would go to operate. To group all these identifiers by



Identifier	
a person who	a workman who
a man who	an artist who
a woman who	a politician who
a doctor who	an athlete who
a journalist who	a sailor who
a scientist who	a player who
an expert who	a soldier who
a woodworker who	an official who
a pilot who	a director who
a priest who	an officier who
a writer who	a singer who
an operator who	a worker who
someone whose job	a person in charge

Table 2. Job title identifiers in in the dictionary definition.

considering the presence of up to 3 words between the indefinite articles and the names, a regex[14], which could be given as input to the Python code, was developed. The regex is: *((a|an)\s\w{3,}\s(who|in charge)|someone whose job)*.

Advanced online vocabulary search

The first method (i.e., the top-down one) aimed to initially extract the greatest number of terms with defined characteristics and subsequently analyse their meaning. This search was developed using a tool that allows the selection of the following parameters in the online vocabulary *EnglishProfile*[15]: Topic: *Work*; Part of speech: *Noun*; Category: *Words*; Grammar: *Countable nouns*; Suffixes: *er*, *or*, *ist*, *ant*. In the end, the 132 results from the searches were collected.

Automatic keywords search in vocabulary definitions

The second method was based on the idea that each job title should have a similar description in the monolingual dictionary. Differently from the first one, the bottom-up approach was defined by starting from the definitions and going back to the terms themselves. By using the regex previously shown, the algorithm searched the identifiers in the *Definitions* column of the xlsx file of *Easier English* dictionary[16]. As soon as one of the clues in the term definition was found, the corresponding word in the *Words* column of *Easier English* dictionary was extracted. All 1146 terms, whose meanings contained the regex pattern, were collected.

2.2.4 Refining the results

Finally, the above rules were used to identify all the base job titles which were then collected in a single set, and the duplicates were eliminated. As it can be expected, not all the collected terms refer to job titles. A screening of the whole collected data was done, using data cleansing techniques also eliminating those terms which are not job titles (e.g., *villager: a person who lives in a village*; *lover: a person who likes or enjoys a particular thing*).

A semantic analysis was carried out on the first hundred words in alphabetical order of the titles set. A precision index ζ, which indicates the percentage of extracted terms referring to job titles, was calculated:

$$\zeta = \frac{n^\circ\ of\ extracted\ terms\ referring\ to\ job\ title}{n^\circ\ of\ extracted\ terms} \times 100 = 79\%$$

Then, to define rules that could enhance this measure, a common pattern detection in the definitions of those terms not labelled as a job title was studied. The terms classified as *non-job titles* were divided into categories, to identify rules to be given as input to the code, that would eliminate the incorrect terms.

[14] https://regexr.com/

[15] https://englishprofile.org/wordlists/evp

[16] https://www.academia.edu/36399186/Easier_English_Student_Dictionary

Rule 1) Part of speech: The terms were divided into *nouns*, *pronouns*, *adjectives* and *verbs*. This classification was carried out because indeed a job title is described as a noun but not as a verb, pronoun or adjective; therefore, the last three categories of terms could be eliminated immediately.

Rule 2) Verb type: The terms were divided according to the type of verb that followed the definition identifier. This classification was made to eliminate those words whose definitions contained *thought* and *feeling* stative verbs, indicating ideas, emotions, and characteristics of the person, instead of working activities. Definitions with *transition* verbs (e.g., sit, stay, live), which do not identify activities performed by a job title, were deleted too.

Rule 3) Verbal time: The terms were classified according to the verbal time in the definition. This classification was made because it was noted that job titles definitions contain *present* tenses, whereas some titles describing past features include *past* tenses.

Rule 4) Sentence type: The terms were classified into *positive* and *negative* sentences. This step was done because we noticed that job titles are described with positive sentences, and thus we could eliminate the negative ones.

Rule 5) Verbal form: The terms were classified according to their *active* or *passive* verbal form because job titles are described by active verbal form. Thus, we could eliminate the terms defined by passive verbal form.

Rule 6) Pronoun type: The terms whose definition contained an identifier (e.g., a person) followed by "*whom*" or "*whose*" were deleted, except where the pronoun was followed by the word "job".

Then, we defined regexes that generalise the above rules given as input to the code. Using the regexes, we obtained a new profiles list. The precision index was calculated again by analysing the first hundred words in alphabetical order of the new profile list. Thanks to these six cleaning methods, the ζ index increased to 91%. Later, a manual filtering was performed to eliminate those terms that certainly did not refer to job titles, but which were still in the list as they were not eliminated through the six previously defined rules. Before the words were eliminated, their literary meaning was verified, using the *Easier English* digital vocabulary. In the end, a manual method that generated all the possible suffixes and prefixes of the collected terms was developed (e.g., from *guard* we obtain *coastguard*, *safeguard*, *bodyguard*, *lifeguard*, *guardian*, *guardsman*).

Finally, all the original and derived job titles were collected. Furthermore, all the plural forms of all the terms previously identified were generated. All singular and plural job titles (1576 terms) were collected in the *final_job* file. As the final set was considered satisfying enough, the *final_job* list represented the final set on which the automatic process of job titles extraction from scientific publications was built in the next parts of the work. Then, it will be possible to automate the entire search process by searching for the single term within the text, preceded and succeeded by a "space" character.

2.3 Automatic extraction of compound job titles

The developed method was divided into several activities (*Figure 3*). A Python code was initially written to create the OR query of all the base titles collected in *final_job.* The resulted query was used to search and collect papers in a txt file from the *Web of Science* database. This file was given as input to a second Python code that split the sentences, extracted and collected only those sub-phrases containing a base job title. The code generated an xlsx file and a txt file. The obtained xlsx file was used to collect a set of rules able to identify the presence of a compound job title within a sentence and to develop a statistical analysis of the goodness of the extracted sentences. The rules were generated using *Part.of.speech.Info*[17], *DisplaCy Dependency Visualizer*[18], and *Rule-based Matcher Explorer*[19]. The obtained txt file containing these rules was then used to write a final Python code capable of identifying compound job titles within any given text file. Once the rules were defined, through the analysis of the extracted sentences, the rules were classified and expanded, trying to explain all the possible complements, also by cross-referencing compound job title on other papers. The final code able to identify, within the text file given as input, the compound job titles, defined through the set of rules given as input, was developed. It creates as output the text file, which contains the various job titles highlighted with a different colour, according to the corresponding base job title. It should be noticed that the entire developed process can be reproduced for any text file, of any type and size.

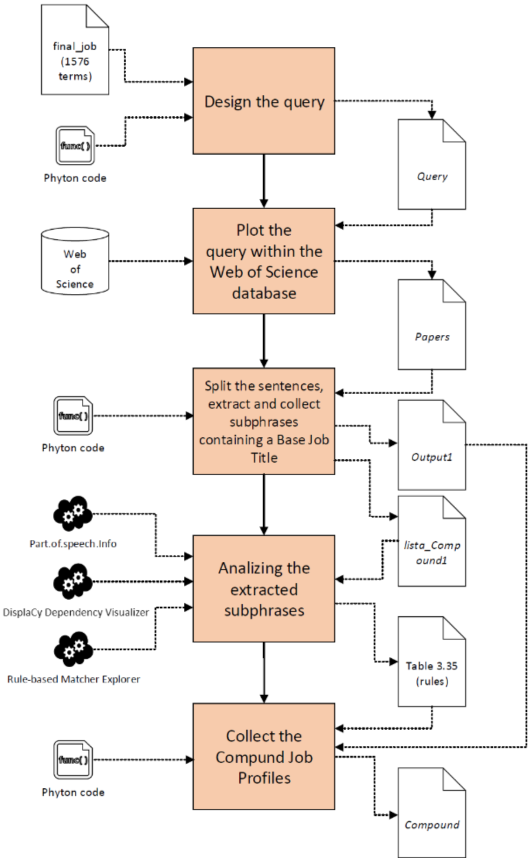

final_job (1576 terms)
Phyton code
Design the query
Query
Web of Science
Plot the query within the Web of Science database
Papers
Phyton code
Split the sentences, extract and collect subphrases containing a Base Job Title
Output1
lista_Compound1
Part.of.speech.Info
DisplaCy Dependency Visualizer
Rule-based Matcher Explorer
Analizing the extracted subphrases
Table 3.35 (rules)
Phyton code
Collect the Compund Job Profiles
Compound

Figure 3: Activities map of the "Automatic extraction of compound job titles" Section.

Finally, to widely clarify the work, the developed methodology and code were tested in a case study presented in the following *Chapter*.

[17] https://parts-of-speech.info/
[18] https://explosion.ai/demos/displacy
[19] https://explosion.ai/demos/matcher

3. A case study on the green economy

This *Chapter* presents a case study of the methodology developed in *Chapter 2*. In detail, the process was applied for a current topic of relevant interest: the green economy (*Figure 4*).

UN experts recommend focusing on the creation of a new green economy, with an increasing role for the state and inter-state bodies in economic governance, promotion of business growth based on new green technologies and greening of industrial branches of the economy[20].

Well-educated workers, with sufficient knowledge and skills to be able to adapt well to technology-intensive industries, make a real difference among the many factors influencing industrial retraining, and the working figures will change[21].

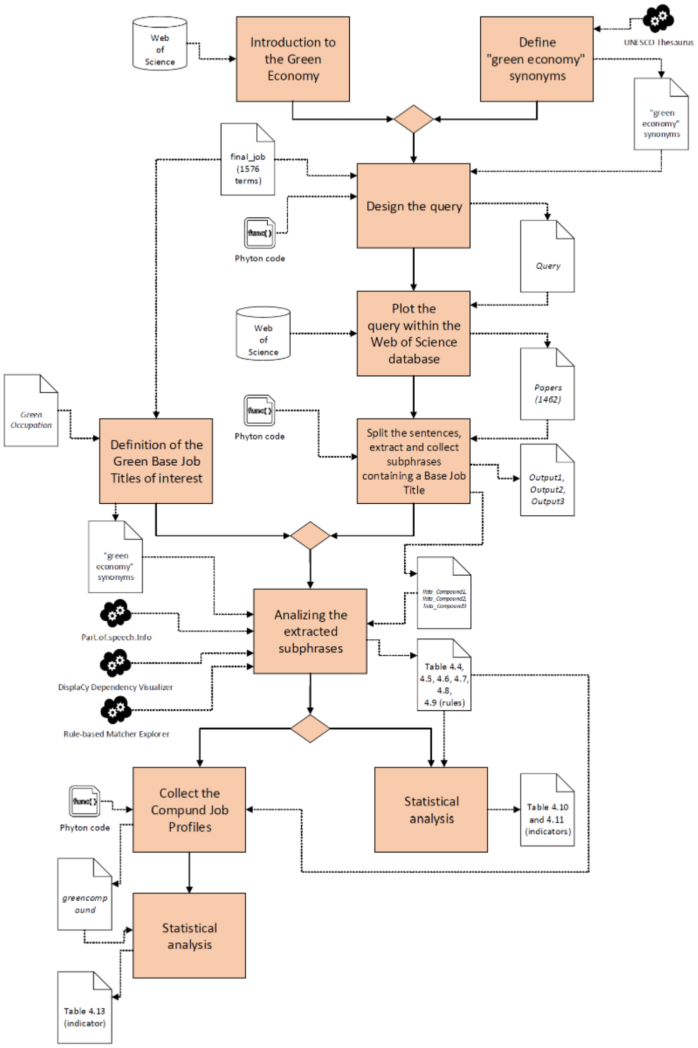

Web of Science
Introduction to the Green Economy
Define "green economy" synonyms
UNESCO Thesaurus
"green economy" synonyms
final_job (1576 terms)
Design the query
Phyton code
Query
Web of Science
Plot the query within the Web of Science database
Papers (1462)
Green Occupation
Definition of the Green Base Job Titles of interest
Phyton code
Split the sentences, extract and collect subphrases containing a Base Job Title
Output1, Output2, Output3
"green economy" synonyms
lista_Compound1, lista_Compound2, lista_Compound3
Part.of.speech.Info
Analizing the extracted subphrases
DisplaCy Dependency Visualizer
Table 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 (rules)
Rule-based Matcher Explorer
Phyton code
Collect the Compund Job Profiles
Statistical analysis
Table 4.10 and 4.11 (indicators)
greencomp ound
Statistical analysis
Table 4.13 (indicator)

Figure 4: Activities map of "A case study on the green economy" Section.

The present study started with identifying those papers related to the green economy and that contained a base profile from the list defined in *Chapter 2*. This was achieved by creating a query made of synonyms generally used to address the topic and the complete list of base profiles linked through the OR logical condition. The search, carried out on *Web of Science*, generated 1462 papers.

Since many base job titles resulted from the above-developed analysis, it was difficult to focus on the extraction of compound profiles on the entire set of base job titles. The six terms (10% of the extracted base job titles) with the highest percentage of quotations in the *Green Occupation* file by *O*NET* were selected.

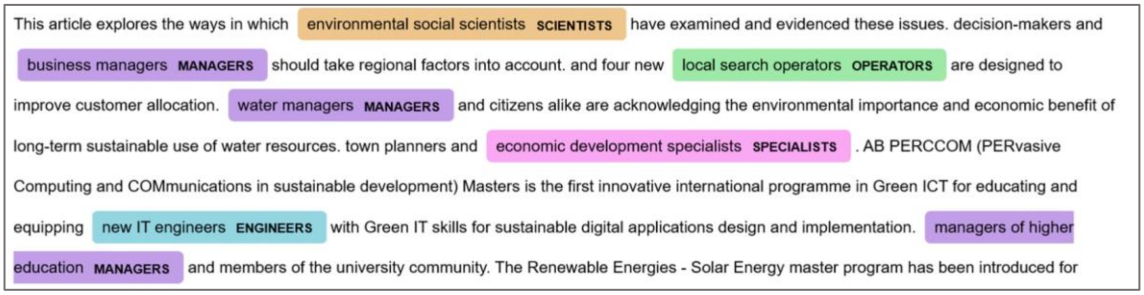
Of course, everything that was done for the six green base job titles under consideration can be reproduced on the entire sample of base job titles previously collected.

Then, all the sub-phrases containing one of the selected green base job titles were extracted and analysed finding the related rules for the extraction of compound job titles. Once the rules were defined, they were

[20] Mamedov G., Babych N., "*GREEN ECONOMY: GLOBAL DEVELOPMENT PRIORITIES*", Volume 6, Issue 1, pp. 87-91, 2020.

[21] Wenjuan G., Xiaohao D., Ran C., Weifang M., "*An Empirical Study of the Role of Higher Education in Building a Green Economy*", MDPI, 2019.

classified and expanded, trying to explain all the possible complements, also by cross-referencing compound job titles on other papers. Using *Jupyter Notebook*, the *Python* code creates as output the text file, which contains the various job titles highlighted with a different colour, according to the corresponding green base job title (*Figure 5*).



This article explores the ways in which environmental social scientists SCIENTISTS have examined and evidenced these issues. decision-makers and business managers MANAGERS should take regional factors into account. and four new local search operators OPERATORS are designed to improve customer allocation. water managers MANAGERS and citizens alike are acknowledging the environmental importance and economic benefit of long-term sustainable use of water resources. town planners and economic development specialists SPECIALISTS . AB PERCCOM (PERvasive Computing and COMmunications in sustainable development) Masters is the first innovative international programme in Green ICT for educating and equipping new IT engineers ENGINEERS with Green IT skills for sustainable digital applications design and implementation. managers of higher education MANAGERS and members of the university community. The Renewable Energies - Solar Energy master program has been introduced for

Figure 5. Example of the results obtained by plotting the Python code on Jupyter Notebook.

To evaluate the novelty level of the collected job titles, the resulted 102 terms were compared with *O*NET* currently registered green job titles. Collected job titles were manually classified between titles already registered in the green occupations of the *O*NET* database, and new green job titles. The ω index, which describes the percentage of collected entities that were new green compound job titles, was calculated.

$$\omega = \frac{n° \; of \; new \; green \; job \; titles}{n° \; of \; green \; compound \; job \; titles \; entities} \times 100 = 64\%$$

From the result, we could notice that a high percentage of the identified titles is not currently recorded on *O*NET*. This demonstrates the high value of the developed methodology, as the algorithm allows the identification and extraction of job titles not yet registered in the worldwide databases.

Finally, to demonstrate the continuous evolution of job profiles, an evolutionary model of the professional figure of the *manager* has been created. To develop the model, all the declinations of manager registered in *O*NET* and all the declinations of manager identified through the study developed in the following *Section* were examined. The professional figures were searched within *Web of Science* and, for each of them, the publication year of the first paper citing the specific job title was collected.

We noticed that in the first years of the 20th century, only the elementary figure of the *manager* existed; only since the 70s, some declensions of the manager were born. As we might have expected, specific green economy titles (e.g., *wind farm manager*, *solar energy installation manager*) have developed only over the last decade, while less detailed titles (e.g., *business manager*, *account marketing manager*, and *sales manager*) have emerged firstly.

4. Streamlit application

To present the results obtained through the methodology developed in this thesis, a *Streamlit* application has been created (*Figure 6*). First of all, the application was designed by defining the main features: the possibility to insert a text file of your choice, the possibility to choose one or more base job titles of interest, the possibility to display within the inserted txt file the compound job titles corresponding to the selected base job titles of interest, and the possibility to download an xlsx file containing the compound job titles corresponding to the selected base job titles of interest. By using the *Streamlit* functions, the *Python* code and the previously defined rules, the *Job Visualiser* application code has been generated and run in *Sublime Text*. Finally, the application was published online. It is, therefore, possible to view and test it at the link: https://frozen-brook-75436.herokuapp.com/

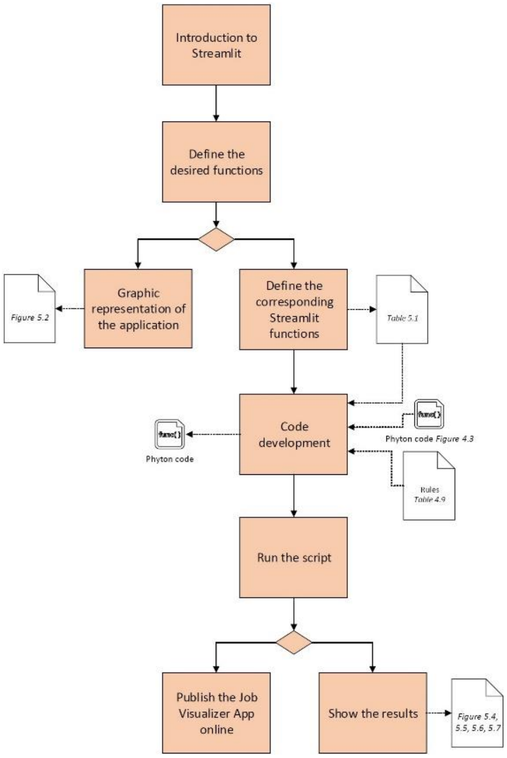

Introduction to Streamlit
Define the desired functions
Graphic representation of the application
Figure 5.2
Define the corresponding Streamlit functions
Table 5.1
Code development
func()
Phyton code
func()
Phyton code Figure 4.3
Rules Table 4.9
Run the script
Publish the Job Visualizer App online
Show the results
Figure 5.4, 5.5, 5.6, 5.7

Figure 6: Activities map of the "Streamlit application" Section.

5. Conclusions and future works

In literature, to our knowledge, no automatic methods exist for extracting job titles from text files. In this thesis, firstly, a semi-automatic algorithm able to generate a set of 1576 clues that allows the univocal identification of job titles within the text were developed. Once the rules were obtained, a second automatic method was carried out for the extraction of compound job titles from text files. By applying the methods to the central topic of the green economy, 53 new green figures were extracted. Finally, to allow the testing of the developed algorithm, a web interface was created. The application highlights and extracts compound job titles from text files, by uploading any txt file and selecting one or more base job titles of interest.

In the future, to improve the process level, the generated base job titles list will be refined by utilizing a more recent and detailed English vocabulary. Also, further semantics terms research will be carried out to build new base job titles using the identifiers collected in *Section 2.2.1*. The automatic extraction process will be improved by defining all the logical analysis rules for the identification of the specification complements, as implemented for the green base job titles. In addition, by combining this thesis with other literature works, it will be possible to develop an automatic method to update information about the nomenclature of profiles and the study programmes, skills, and technologies related to them. Overall, we believe the approach introduced in this work constitutes a solid foundation to create an update service for the worldwide database authorities, international recruitment companies, and institutions involved in professional formation and training courses.