



W-SOURCING

Implementation of a recommendation and classification application for a CV database for HR recruitment needs.

Academic Supervisor

M. Trabelsi Dorra

Professional Supervisor

Mr. Makrem JANNADI

challenge accepted

Guemira Marwa

2019-2020



Table of Contents

Introduction	7
1 Business Understanding and Analytic	9
1. INTRODUCTION	9
1.1 COMPANY PRESENTATION.....	9
2. PROBLEMATIC	9
3. BUSINESS UNDERSTANDING	10
3.1 BUSINESS OBJECTIVES	10
3.2 DATA SCIENCE GOALS	11
4. ANALYTIC APPROACH	11
5. CONCLUSION.....	11
2 Data Collection and Data Preparation	12
1. INTRODUCTION	12
2. DATA PREPARATION	12
2.1 READ DATA FROM MONGODB.....	12
2.2 NATURAL LANGUAGE PROCESSING	13
3. DATA COLLECTION.....	19
3.1 TOOLS.....	20
3.2 DATABASE UPDATING	20
3.3 COMPLEMENTARY DATA.....	22
4. CONCLUSION.....	22
3 Modeling and Evaluation	23
1. INTRODUCTION	23
2. RECOMMENDATION SYSTEM	23
2.1 RECOMMENDATION FILTERING TECHNIQUES	24
3. RECOMMENDATION SYSTEM MODELS	27
3.1 T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING	27
3.1.1 Principle of process.....	27
3.1.2 Application	28
3.2 K-NEAREST NEIGHBORS	29
3.2.1 Principle of process.....	29



3.2.2 Application	29
3.3 COSINE SIMILARITY	30
3.3.1 Principle of process.....	30
3.3.2 Application	31
3.4 K DIMENSIONAL TREE	32
3.4.1 Principle of process.....	32
3.4.2 Application	32
4. EVALUATION.....	33
4.1 CORRELATION TEST	33
4.2 PEARSON CORRELATION	34
4.3 KENDALL CORRELATION.....	34
4.4 SPEARMAN CORRELATION	34
4.5 APPLICATION	35
5. CONCLUSION.....	35
4 Deployment	36
1. INTRODUCTION	36
2. DEVELOPMENT ENVIRONMENT.....	36
3. IMPLEMENTATION OF WEB APPLICATION	37
3.1 LOGIN PAGE	37
3.2 HUMAN RESOURCE AGENT INTERFACE	38
3.3 ADMIN INTERFACE	43
4. CONCLUSION.....	46
Bibliographie	48



List of Figures

2.1	Natural Language Processing	13
2.2	Dictionary of Skills	14
2.3	Result of Professional Skills	14
2.4	Result of Academic Skills	15
2.5	Changes of Experience's duration	15
2.6	Changes of the Feature Language	16
2.7	Changes of the Feature Education degree	16
2.8	Changes of the Feature Education Field	17
2.9	Changes of the Feature University	17
2.10	Changes of the Feature Localisation	17
2.11	Changes of the Feature Volunteering	18
2.12	Job Titles for each profile	18
2.13	Encoding Job Titles	19
2.14	Final result	19
2.15	Selenium logo	20
2.16	BeautifulSoup logo	20
2.17	The extracted Data	21
2.18	Campany's Features	22
3.1	Recommendation filtering techniques	24
3.2	Example of content-based approach	25
3.3	Example of user-item rating matrix	25
3.4	Example of collaborative filtering approach	26
3.5	T-SNE 2 components	28
3.6	T-SNE 3 components	28
3.7	example of K-NN voting	29
3.8	Example of Nearest neighbors similarity result	30
3.9	Calculate Cosine Similarity	31
3.10	Example of Cosine Similarity result	31
3.11	Principle of K Dimensional Tree	32
3.12	Example of K Dimensional Tree result	33
3.13	Result from Cosine Similarity	35
4.1	Django logo	36
4.2	MVC process	37
4.3	Login page	38
4.4	Home page 1	38
4.5	Home page 2	39
4.6	candidate page 1	39
4.7	Candidate page 2	40
4.8	Candidate page 2	40
4.9	example of professional career information	41
4.10	example of educational career information	41
4.11	example of application similarity	42



4.12 linkedin profile	42
4.13 bar chart job title by profile numbers	43
4.14 profile list data table	43
4.15 Profile distribution using tSNE	44
4.16 Visualization profiles distribution using tSNE	44
4.17 Visualization profiles distribution by job title using tSNE 2 dimensions	45
4.18 Visualization profiles distribution by job title using tSNE 3 dimensions	45



List of abbreviations

API	Application Program Interface	CSS
Cascading Style Sheets		
CRISP-DM	Cross Industry Standard Process for Data Mining	
HR	human resources	
HTML	Hypertext Markup Language	IBM
International Business Machines		
IOT	Internet of things	
IT	Information Technology	JD
Job Description		
JSON	JavaScript Object Notation	
KDTree	K Dimensional Tree	
KNN	k-nearest neighbors	
NLP	Natural Language Processing	
REST framework	Representational State Transfer framework	
t-SNE	t-distributed stochastic neighbor embedding	



General Introduction

Human Resources can be described as the talent capital that forms the basis of an organization's play in a domain. It is an investment that Companies make for a better Company, however as expected, there are frequently some problems in the recruitment and selection process. In fact, one of the major struggles that the recruiters these days encounter, is the requirement of resources to be hired in a very short span of time. Also, with multiple projects coming up with huge requirements of resources as the companies are growing enormously in size, it is very challenging for the recruiters to meet the expectations with hiring the required numbers and talent as well.

Based on those circumstances, Wevoo's suggested a project entitled W-SOURCING. Its main objectives are to help a company to make the right decision to choose the best profile. Indeed with aiming to assume some business objectives such as increasing overall revenue and ameliorating the RH development by reducing the time of search and the cost of the recruitment process, helping the RH department to find right items in a large option space, which match their interests using content-based approach that takes into consideration an organization's needs and the skills of candidate and collaborative filtering to find best suitable match.

In order to achieve our business objectives, we will be applying several data science goals. The first important step is to prepare the internal data with Big Data's Distributed Architecture in order to deal with the massive data. Then, we collect our external data and elaborate the new features created by Scrapping. Further more, we implement NLP(Natural Language Processing) for text analytics which means counting, grouping and categorizing words to extract structure and meaning from large volumes of content. After that, we implement machine learning's models to the modeling part where we use some recommended filtering techniques and recommendation system models which will lead us to deal with evaluation part in order to grab the performance of each used model.

As a final step of our project, we will try to build a web application in order to deploy our project in a real environment and to facilitate the use of our models to the endpoint user.

This project as mentioned before is heavily rich with all data science objectives that we assumed to apply from the start to the bottom. There's no doubt that a lot of researches were taken to understand the whole process and a precise commitment to finalize it.

The purpose of this present report is to define and describe the work carried out throughout the project of implementing a recommendation system to find the most profile similarities according to HR recruitment needs. This report will be divided into four main chapters:

- In the first chapter, we are going to define the business objectives and fix our data science goals. We will present also the Wevoo's company.
- In the second chapter, we will describe the steps that we follow to collect the external data and the



steps that we choose in order to prepare our data for the modeling step.

- In the third chapter, we are going to define the recommendation system and its techniques. In addition, we will define and present the different models and methods used to obtain profiles similarities
- In the last chapter, we will present the deployment step as a web application which contains the modeling results.



Chapter 1

Business Understanding and Analytic

1 Introduction

During this first chapter, we will present the company Wevioo and its activity area, we will introduce the business understanding by detailing the problematic, defining the business objectives and fixing the data science goals.

1.1 Company Presentation

The Wevioo group

Since its creation in 1998, the Wevioo group has supported its customers in their digital transformation projects by bringing them its expertise and knowledge on 3 domains:

- Consulting
- Digital
- IOT

Wevioo provides its customers with digital innovation solution perfectly suited to their agility, performance and development challenges.

With a culture of innovation at the heart of its DNA, Wevioo invests in RD to provide its customers and partners with innovative solutions and expertise at the cutting edge of technology.

Wevioo is backed by several hundred demanding projects carried out by its business and technological experts in more than 30 countries in Europe, North America, Africa and the Middle East.

2 Problematic

Human Resources or an organization's people are its most important asset. It is the talent capital that forms the basis of an organization's play in a domain. It is an investment that Companies make for a better Company, but as is the normal course, there are problems in the recruitment and selection process.



It starts with the laying out of the correct Job Description (JD) in co-relation to the demands of the role. More often than not, this step itself is the beginning of recruitment difficulties as the quality of profiles rarely matches the optimal expectation and this is the first stage of facing problems in the understanding the requirement and receiving the right number of quality resumes.

In fact, one of the major struggles that the recruiters these days encounter is the requirement of resources to be hired in a very short span of time. With very little time candidates are made to go through multiple rounds screening and interviews, but the need of good caliber and talent is very important at the same time.

Also, with multiple projects coming up with huge requirements of resources as the companies are growing enormously in size, it is very challenging for the recruiters to meet the expectations with hiring the required numbers and talent as well.

Besides, the most common issues we see in the recruitment process is that the candidates not being able to submit the credentials or academic validations in relation to the roles they became eligible to apply for the jobs at hand. It takes a lot on the part of the recruitment team to actually put across clearly the need for the validations to ensure the authenticity of the candidature. Even considering the credentials have been authenticated with the candidates being offered the Job.

3 Business understanding

Business understanding is the first step of every Data science project using the IBM-master plan or CRISP-DM methodology. In this stage, we need to specify the key variables that are to serve as the model targets and whose related metrics are used to determine the success of the project. We need also to identify the relevant data sources that the business has access to or needs to obtain.

In order to achieve these goals:

- Work with your customer to understand and identify the business problems. Try to formulate questions that define the business goals that the data science techniques can target.
- Identify data sources: Find the relevant data that helps you answer the questions that define the objectives of the project.
- After identifying the main goals of the project, we must verify that they are SMART (Specific, Measurable, Achievable, Relevant, Time-bound) goals.

3.1 Business objectives

The main objectives presented by our project to help the company to make the right decision to choose the best profile are:

- **Increase overall revenue:** By having access to the correct data, we can significantly impact the overall revenue, by:
 - Identifying the most important skills that the company search.
 - Limit the number of employees.
- **Ameliorate the HR development:**
 - Improve the recruitment process by reducing the time of search and the cost of the recruitment process.
 - Helping the HR department to find right items in a large option space, which match their interests.



- Use content-based approach that takes into consideration an organization's needs and the skills of candidate.
- Use of Collaborative Filtering to find best suitable match.

3.2 Data science goals

In order to achieve our business objectives, we identify the following data science goals:

- External Data Collection (Scrapping) and elaborating the new features.
- Profiles Scoring.
- Procedure the Big Data concept in order to deal with the massive data: Distributed Architecture.
- Classify and order the different profiles according to their academic and professional skills.
- Data Cleaning.
- Sentimental Analysis.
- Implementing Machine Learning models.

4 Analytic approach

The Data Scientist can define the analytical approach once a business problem has been clearly identified. To implement this, the problem must be expressed in the context of statistical learning and machine learning techniques. So that the data scientist can identify the appropriate techniques to achieve the desired result such as Regression problem, classification problem, outlier detection problem . . . etc.

In our case, we have to use a Recommendation/Personalization system in order to achieve our goals and our problem will be :

- “ How can we target profiles to a specific job ? ”
- “ How can we classify profiles to a specific Field ? ”
- “ How can we identify the best profiles to a specific job ? ”

5 Conclusion

During this chapter, we have successfully described the business understanding as well as the analytic approach. In addition, we have fixed our data science goals. In the next chapter, we'll move on to getting the required data and preparing them.



Chapter 2

Data Collection and Data Preparation

1 Introduction

During this chapter, we will describe the steps that we used to prepare the internal data. Then, we will present our external data which is created by collecting data.

2 Data preparation

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data.

We can affirm that data preparation is the most important step in a Data Science project.

76% of data scientists say that data preparation is the worst part of their job, but the efficient, accurate business decisions can only be made with clean data. Data preparation helps to :

- Fix errors quickly: Data preparation helps catch errors before processing.
- Produce top-quality data: Cleaning and reformatting datasets ensures that all data used in analysis will be high quality.
- Make better business decisions: Higher quality data that can be processed and analyzed more quickly and efficiently leads to more timely, efficient and high-quality business decisions.

In order to prepare our data, we follow these steps:

2.1 Read Data From MongoDB

Our dataset was originally stored in mongoDB database so we used “ PyMongo “ to get access to Wevioo Data. PyMongo is a Python distribution containing tools for working with MongoDB, and is the recommended way to work with MongoDB from Python.

We used also another way to get Wevioo Data: transform the “Bson” file, which is the main format of MongoDB, into json file. Then, we used “json_normalize” which is provided by Pandas. It is a nice utility



function for flattening semi-structured JSON objects. It has many powerful parameters which make our access to Mongo easier.

After we transform the MongoDB columns into Dataframe, we get this:

We note that we obtain 10000 rows and 7 columns and the columns are in text form. So we need to use Natural Language Processing in order to pick up the most important feature such as professional skills, academic skills and experience.

2.2 Natural Language Processing

Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding.

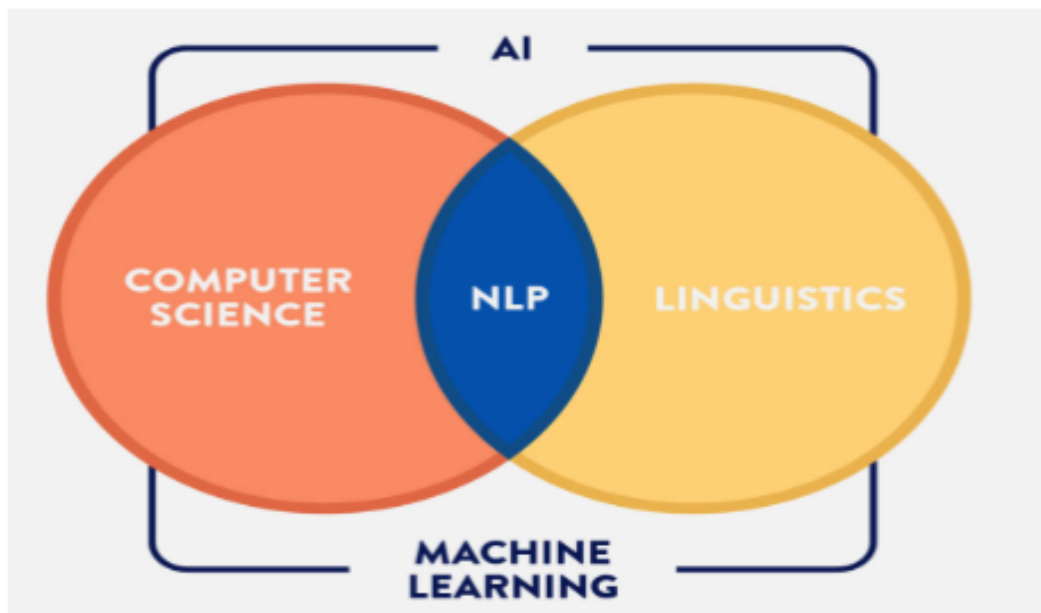


Figure 2.1: Natural Language Processing

NLP capabilities are used often to solve many tasks, such as:

- Content categorization: A linguistic-based document summary, including search and indexing, content alerts and duplication detection.
- Topic discovery and modeling: Accurately capture the meaning and themes in text collections, and apply advanced analytics to text, like optimization and forecasting.
- Contextual extraction: Automatically pull structured information from text-based sources.
- Sentiment analysis: Identifying the mood or subjective opinions within large amounts of text, including average sentiment and opinion mining.
- Machine translation: Automatic translation of text or speech from one language to another.

For our project, we are using NLP for text analytic which means counting, grouping and categorizing words to extract structure and meaning from large volumes of content.



Text analytic is used to explore textual content and derive new variables from raw text that may be visualized, filtered, or used as inputs to predictive models or other statistical methods.

To begin with the part of text analytic, we create skills that are provided and alimented from an Excel file given from the enterprise.

This step makes the dictionary dynamic for the need of enterprise.

```
print(listWeevio)

['javascript', 'sql', 'nosql', 'node.js', 'express.js', 'koa.js', 'hapi.js', 'angularjs', 'react.js', 'jquery', 'bash', 'nginx', 'c', 'c++', 'html5', 'css', 'rest', 'sass', 'postcss', 'webpack', 'gitlab', 'linux', 'embedded c', 'embedded c++', 'java', 'jee', 'microservices', 'intégration continue', 'docker', 'aws', 'nodejs', 'ext.js', 'html', 'mongodb', 'mysql', 'spring', 'soa', 'soap', 'git', 'svn', 'jira', 'confluence', 'spring boot', 'spring security', 'java 8', 'php', 'symfony', 'architecture resful', 'cms', 'drupal', 'scrum', 'analyse fonctionnelle', 'testing', 'rédaction']
```

Figure 2.2: Dictionary of Skills

So after we precise the data requirement especially skills, experiences and creating the dictionary, we used NLP to extract our numerical features:

Dividing skills into two parts professional and academic, is a main part of our project in order to recognize a profile to another. Considering the difference between them, a person who has more professional skills than academic, it's a way more better in term of practiced skills.

(a) Professional Skills:

Based on the created dictionary, we are going to match the jobs description in order to reveal the coherent professional skills.

To be more precise, we used the job description text, which exists in the job section of a LinkedIn profile. Two possible actions, either we extract the skills that each employee practice in that specific job or the skills doesn't appear and Null value will be shown. Then, if any skill in our document appears in job description, we count the number of occurrence

	javascript	sql	nosql	node.js	express.js	koa.js	hapi.js	angularjs	react.js	jquery	bash	nginx	c	c++	html5	css	rest	sass	postcss	webpack
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...
9995	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

10000 rows × 54 columns

Figure 2.3: Result of Professional Skills



(b) Academic skills:

As the same thing with professional skills, we extracted data from the skills and endorsements section. Then, we matched the elements found in every profile with the other included in the dictionary. The value elected for each skill will take one possible chance, one if it has been included in the dictionary else it will take a null value.

```
dft_skill.rename(columns=lambda s: s+'_academic',inplace=True)
dft_skill
```

	javascript_academic	sql_academic	nosql_academic	node.js_academic	express.js_academic	koa.js_academic	hapi.js_academic	angularjs_academic	react
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
2	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
3	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...
9995	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9997	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9999	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

10000 rows x 54 columns

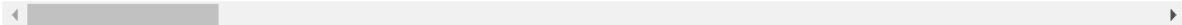


Figure 2.4: Result of Academic Skills

(c) Duration of experience:

To give more meaning to a professional skill, we took its range of time and refined it to be the number of months worked on that experience.

	date_exp	javascript	sql	nosql	node.js	express.js	koa.js	hapi.js	angularjs	react.js	jquery	bash	nginx	c	c++	h
0	232	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	49	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0	3.0	0.0	0.0
2	330	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
4	5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...

Figure 2.5: Changes of Experience's duration



(d) **Language:**

To give way more importance to the language, we considered some essentials "must-have" ones to clarify more Wevioo's needs like French, English and non essentials labeled others.

	arabe	english	français	allemand	espagnol	italien	others
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	1.0	1.0	1.0	0.0	0.0	0.0	0.0
2	1.0	1.0	1.0	0.0	0.0	0.0	0.0
3	1.0	1.0	1.0	0.0	0.0	0.0	1.0
4	1.0	1.0	1.0	0.0	0.0	1.0	0.0
5	1.0	1.0	1.0	0.0	0.0	0.0	0.0
6	1.0	1.0	1.0	0.0	1.0	0.0	0.0

Figure 2.6: Changes of the Feature Language

(e) **Education degree:**

For classifying profiles with their academic background, we worked to identify the education degrees. But, we just took a specific degrees in which they describe more the real grade that Wevioo wants. These take options, one if it is Licence, Master, Engineering.

And to be more deeper in this, we attribute a score with a coefficient for each one of the degrees and to multiplied it with the value elected to the degree.

	Engineer	Master	Licence	Others_Degree	score_diplome
0	0.0	0.0	0.0	1.0	0.5
1	1.0	0.0	0.0	1.0	3.5
2	0.0	0.0	0.0	1.0	0.5
3	1.0	0.0	0.0	1.0	3.5
4	1.0	0.0	0.0	1.0	3.5
5	0.0	0.0	0.0	1.0	0.5
6	0.0	0.0	0.0	1.0	0.5
7	0.0	0.0	0.0	1.0	0.5
8	1.0	0.0	0.0	0.0	3.0
9	1.0	0.0	0.0	1.0	3.5

Figure 2.7: Changes of the Feature Education degree



(f) **Field of studies:**

For classifying profiles with their field of studies, we worked to identify if they applied in IT studies or other fields.

	Informatique	Others_type_Of_Study
0	0.0	1.0
1	1.0	1.0
2	1.0	0.0
3	0.0	1.0
4	0.0	1.0
5	0.0	1.0

Figure 2.8: Changes of the Feature Education Field

(g) **University of Study:**

	ENSI	FST	ENIT	ENISO	INSAT	ESPRIT	OTHERS
0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1	0.0	0.0	0.0	0.0	0.0	0.0	1.0
2	0.0	0.0	0.0	0.0	0.0	0.0	1.0
3	0.0	0.0	0.0	0.0	1.0	0.0	1.0
4	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Figure 2.9: Changes of the Feature University

(h) **Localization:**

As Wevioo's requirements, the localization is divided into two parts: those who lived in Tunisia and foreign countries labeled others.

	Tunisia	Other_Country
0	1.0	0.0
1	1.0	0.0
2	1.0	0.0
3	0.0	1.0
4	1.0	0.0

Figure 2.10: Changes of the Feature Localisation



(i) Volunteering:

To know if that person is active and developing new soft skills by being member of an organisation, university's club or being a volunteer. This one will be honored as an active person.

```
df_vol = pd.DataFrame(list_finale_vol, columns=["Volunteering"])
df_vol.head(100)
```

Volunteering	
0	0.0
1	0.0
2	0.0
3	1.0
4	0.0
...	...
95	0.0
96	0.0
97	0.0
98	0.0
99	0.0

100 rows × 1 columns

Figure 2.11: Changes of the Feature Volunteering

(j) JOB Title:

In order to identify the job title of each profile, we create dictionaries of words based on Wevioo's Jobs. Then, we matched the jobs title with the dictionaries. Meanwhile, if it was true we will append it to a list created and mark it as checked.

job2	
50	Web Back-End
51	Web Back-End
52	0
53	Product Owner
54	Cloud
55	JAVA/JEE

Figure 2.12: Job Titles for each profile



Also, we tried to encode the job title into numerical data.

job		job2
0	0	Technical Lead/ Architecte JEE
1	1	Devops
2	2	Web Back-End
3	3	Data_Science
4	4	Other

Figure 2.13: Encoding Job Titles

Finally we concatenate the job DataFrame with the previous data processed to get the final result.

job	
0	0
1	1
2	2
3	3
4	4
...	...
9995	4
9996	4
9997	2
9998	3
9999	3

10000 rows × 1 columns

Figure 2.14: Final result

3 Data Collection

Data is everywhere nowadays, and it represents gaining new insights that before just weren't possible. Extracting data doesn't just provide valuable information for the business. Also, it can save an enormous amount of time, resources and money by automatically gathering the information you need. Based on that, we involved data scraping in our project from LinkedIn to:

- Update our internal data
- Retrieve relevant data to enhance our database



3.1 Tools

To add a new features we decided to scrap linkedin and we have used two types of tools:

- **Selenium** is a framework which is designed to automate test for web applications

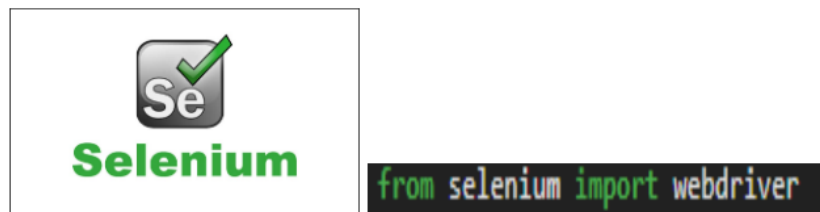


Figure 2.15: Selenium logo

- **BeautifulSoup** is a Python library for pulling data out of HTML and XML files

BeautifulSoup

```
from bs4 import BeautifulSoup
```

Figure 2.16: BeautifulSoup logo

3.2 Database Updating

Over time, information contained in the database will get out of date ,and clutter will accumulate. You should therefore have a process to systematically verify or update each profile's details.

Keep monitoring your database: Just like changes that take place in the industry, it is possible that prospective profiles might make changes to their systems, information, experiences, etc.

It is important to periodically monitor the database and keep it up to date.

In this part we used both of Selenium and BeautifulSoup for the updating . We basically Focused on the most relevant features:

- Experience
- Skills
- Location
- Certifications
- Recommendations



First we used ChromeDriver as an open source tool for automated testing of web apps across many browsers. It provides capabilities for navigating to web pages.

The main purpose of the ChromeDriver is to launch Google Chrome. Without that, it is not possible to execute Selenium test scripts in Google Chrome as well as automate any web application. This is the main reason why we needed ChromeDriver to run test cases on Google Chrome browser.

Second we used CSS selectors to extract our features from the HTML page.

CSS Selectors are used to identifying a user desired HTML web element. This fits into an element locator strategy of automated test development where the primary aim is to interact with page elements through different types of locators. While there are several other methods to identify element locator such as id, name, class name, link text, partial link text, XPath, tag name, etc. More than CSS selectors in Selenium, we prefer the CSS way due to below benefits:

- Faster Identification and reduced test execution time – Compared to XPath CSS selectors would tend to identify the elements better as most used browsers such as Chrome and Firefox are tuned for better performance with CSS selectors.
- Enhanced readability.

```
{'url': 'https://www.linkedin.com/in/habib-morchedi-75562178', 'recommendations': [], 'certifs': [('Lean Six Sigma Yellow Belt', 'L2M Tunis Nord'), ('IELTS certificate', 'IELTS BRITISHCOUNCIL'), ('PTC Creo Parametric', 'B2P Engineering')]}
{'url': 'https://www.linkedin.com/in/omar-ghazouani-4a6890151', 'recommendations': [], 'certifs': []}
{'url': 'https://www.linkedin.com/in/hamza-kefi-95209465', 'recommendations': [], 'certifs': [('MSA / MSP', 'Issued Sep 2010 Expiration Date'), ('Advanced Product Quality Planning (APQP)', 'TÜV Rheinland Maghreb'), ('Formation MRPG', 'TÜV Rheinland Maghreb'), ('Customized training on Creo Parametric 3.0', 'B2P Engineering'), ('Calypso Curve, freeform et CAO', 'Carl Zeiss S.A.S Division Métrologie')]}
{'url': 'https://www.linkedin.com/in/maroua-ellouze-1ba8a528', 'recommendations': [], 'certifs': [('CSM', 'Scrum Alliance')]}
{'url': 'https://www.linkedin.com/in/ines-mahjoub-353ba130', 'recommendations': [], 'certifs': []}
{'url': 'https://www.linkedin.com/in/talmoudi-khouloud-b76b4437', 'recommendations': [], 'certifs': []}
{'url': 'https://www.linkedin.com/in/zied-brahmi-67787b48', 'recommendations': [], 'certifs': []}
{'url': 'https://www.linkedin.com/in/kammoun-walid-34b73171', 'recommendations': [], 'certifs': []}
{'url': 'https://www.linkedin.com/in/nourhelmi', 'recommendations': [], 'certifs': []}
{'url': 'https://www.linkedin.com/in/bmhamza', 'recommendations': [], 'certifs': [('Applied Machine Learning in Python', 'Coursera'), ('Getting and Cleaning Data', 'Coursera'), ('Junior Level Linux Certification (LPIC-1)', 'Linux Professional Institute'), ('SUSE Certified Linux Administrator', 'Novell')]}
{'url': 'https://www.linkedin.com/in/ali-daboussi-b5a5979b', 'recommendations': [], 'certifs': []}
{'url': 'https://www.linkedin.com/in/hanen-ben-ali-82617591', 'recommendations': [], 'certifs': []}
{'url': 'https://www.linkedin.com/in/fdebbabib', 'recommendations': [], 'certifs': [('AWS Certified Solutions Architect - Associate', 'Amazon Web Services'), ('CCNA1', 'Cisco'), ('CCNA2', 'Cisco'), ('CCNA3', 'Cisco'), ('Fondamentaux du Marketing digital - Digital Active', 'Google'), ('Introduction to Virtualization', 'Pluralsight')]}
{'url': 'https://www.linkedin.com/in/nadabensaid', 'recommendations': [], 'certifs': []}
{'url': 'https://www.linkedin.com/in/manel-bayoudh-a068a839', 'recommendations': [], 'certifs': []}
{'url': 'https://www.linkedin.com/in/cassandra-fabre-227400153', 'recommendations': [], 'certifs': []}
{'url': 'https://www.linkedin.com/in/moncefbetaieb', 'recommendations': [], 'certifs': [('Machine Learning', 'Coursera'), ('M101J: MongoDB for Java Developers', 'MongoDB, Inc.'), ('Exploratory Data Analysis', 'Coursera'), ('Getting and Cleaning Data', 'Coursera'), ('R Programming', 'Coursera'), ('The Data Scientist's Toolbox', 'Coursera'), ('Programming in HTML5 with JavaScript and CSS3 Specialist', 'Microsoft'), ('Kaggle R Tutorial on Machine Learning', 'DataCamp')]}
```

Figure 2.17: The extracted Data

Finally, we updated our Database using PyMongo as we mentioned before.



3.3 Complementary Data

In this part, we tried to extract a complementary data to enhance our data and to be more specific. This extra data provide for us the necessary information for the companies which mentioned in the experience feature. Also to evaluate the experience of a profiles based on that company. We used too Selenium as our extracting tool.

Basically we extracted all a company's informations:

- Location
- Speciality
- Type
- Number of employees
- Founding Date

```
dft2.head(20)
```

	description	type	site	secteur	nbemploye	fondation	specialisation
0	1	0	0	0	0	0	0
1	0	1	0	0	0	0	0
2	Infor builds business software for specific in...	New York, NY	http://www.infor.com	Logiciels informatiques	10 001 employés et plus	Société civile/Société commerciale/Autres type...	2002
3	Créée en 2007, Talan Tunisie est le centre de ...	Société civile/Société commerciale/Autres type...	http://www.talan.tn	Technologies et services de l'information	201-500 employés	2007	ingénierie informatique
4	0	0	0	0	1	0	0
5	0	Djerba, Medenine	http://www.ido-developers.com/	Technologies et services de l'information	2-10 employés	2018	0
6	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0
11	About us\nEmira Travel was set up more than tw...	Société civile/Société commerciale/Autres type...	http://www.emira-travel.com	Hôtellerie et hébergement	11-50 employés	1994	DMC
12	OCTO Technology, the very proud recipient of t...	Paris, -	http://www.octo.com	Technologies et services de l'information	201-500 employés	Société cotée en bourse	1998
13	0	0	0	0	0	0	0
14	Sofrecom est une entreprise de conseil et d'in...	Société civile/Société commerciale/Autres type...	http://www.sofrecom.com/	Technologies et services de l'information	501-1 000 employés	2011	0
15	Mersen est un expert mondial des spécialités é...	Société cotée en bourse	https://www.mersen.com	Ingénierie mécanique ou industrielle	5 001-10 000 employés	1891	high temperature applications, anticorrosion...

Figure 2.18: Company's Features

4 Conclusion

During this chapter, we have detailed the steps that we used in data preparation and data collection phase and we have mentioned the different tools and techniques that we used to obtain a DataFrame ready for building a Recommending System.



Chapter 3

Modeling and Evaluation

1 Introduction

After finishing the step of data collection and data processing, this chapter is going to focus on modeling and evaluation step.

First, we are going to mention the categories of recommendation systems, their differences and their fields of application. Then, we will list the different machine learning algorithms and methods that we used in order to achieve our objective. Finally, we will present the different obtained results.

2 Recommendation system

Recommender system is defined as a decision making strategy for users under complex information environments. Furthermore, a recommender system was defined from the perspective of E-commerce as a tool that helps users search through records of knowledge which is related to users's interest and preference. [1]

It is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item and to predict users similarity based on users' historical behaviors. They are primarily used in commercial applications.

Recommender systems are used in a variety of areas and are most commonly recognized as playlist generators for video and music services like Netflix, YouTube and Spotify, product recommenders for services such as Amazon, or content recommenders for social media platforms such as Facebook and Twitter. [1]



2.1 Recommendation filtering techniques

In order to use efficient and accurate recommendation techniques, it is very important to understand the final result and the features that you possess in your data. In the figure below, we mention the different recommender system techniques:

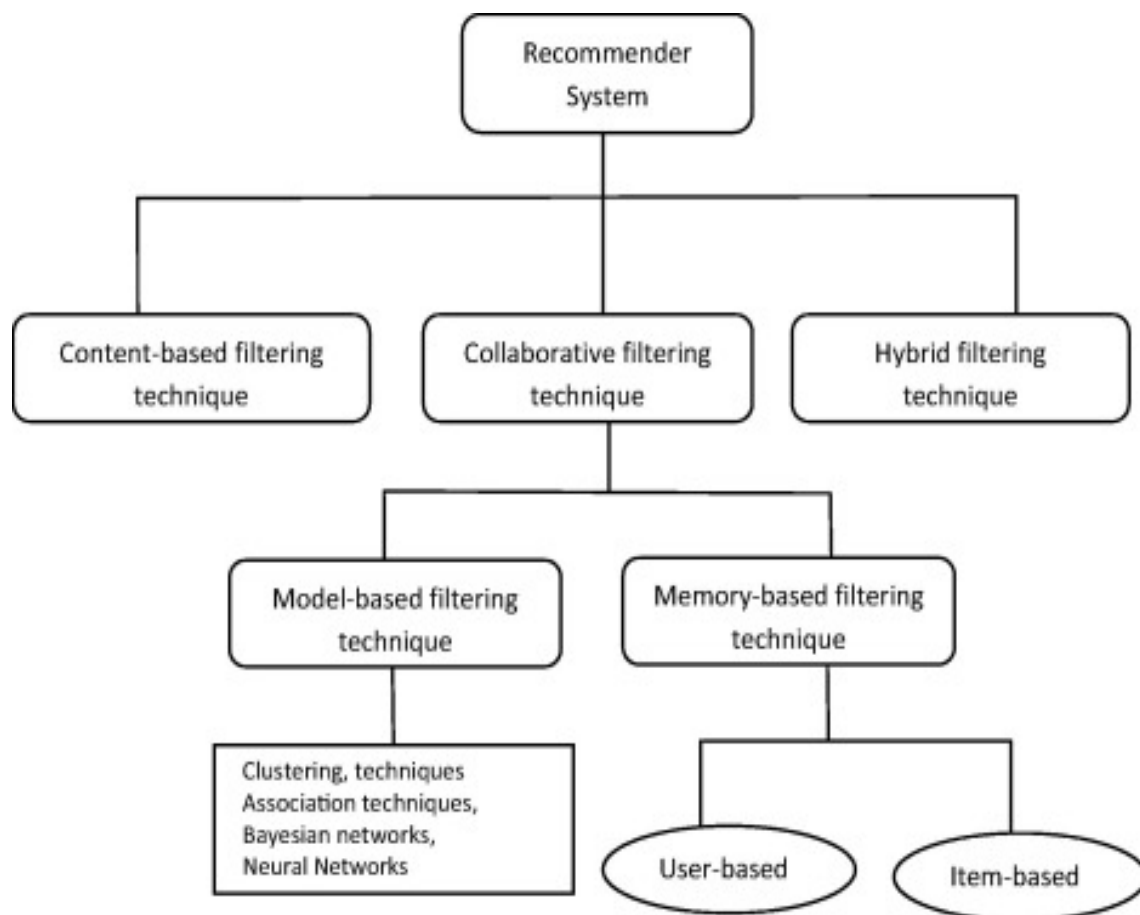


Figure 3.1: Recommendation filtering techniques

To build a recommender system, the most two popular approaches are Content-based and Collaborative Filtering :

- **Content-based** approach is a domain-dependent algorithm and it emphasizes more on the analysis of the attributes of items in order to generate predictions. It requires a good amount of information of items' own features, rather than using users' interactions and feed backs.

For example, When documents such as web pages, publications and news are to be recommended, content-based filtering technique is the most successful.



This figure below shows an example of content-based filtering [1] :

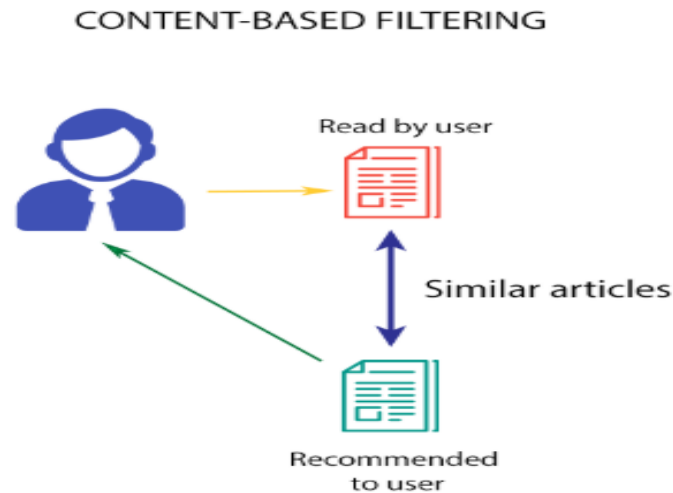


Figure 3.2: Example of content-based approach

• **Collaborative filtering** is a domain-independent prediction technique for content that cannot easily and adequately be described by metadata such as movies and music. Collaborative filtering technique works by building a database (user-item matrix) of preferences for items by users. It then matches users with relevant interest and preferences by calculating similarities between their profiles to make recommendations.

The figure below shows an example of user-item matrix of collaborative filtering [2] :

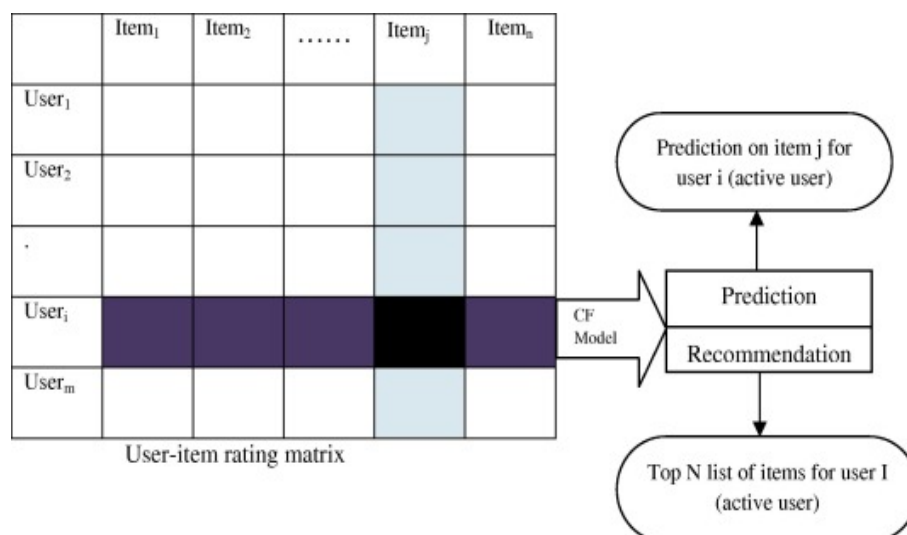


Figure 3.3: Example of user-item rating matrix



As we see in this figure, the result of collaborative filtering can be used either in prediction such as predict rating or to make recommendation by similarities which is the main goal of our project. The technique of collaborative filtering can be divided into two categories: memory-based and model-based. The figure below shows the operating principle of collaborative filtering.

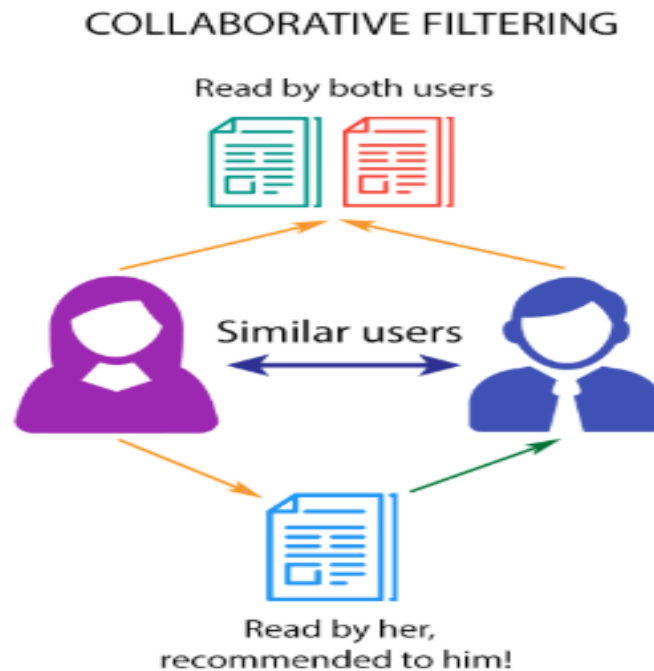


Figure 3.4: Example of collaborative filtering approach

In our project, we tried different methods such as nearest neighbours, cosine similarities that we will explain their main concept in the next section.



3 Recommendation system models

3.1 t-distributed stochastic neighbor embedding

3.1.1 Principle of process

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets [3].

It is extensively applied in image processing, NLP and speech processing. Here's the steps of t-SNE algorithm :

1. The algorithms starts by calculating the probability of similarity of points in high-dimensional space and calculating the probability of similarity of points in the corresponding low-dimensional space. The similarity of points is calculated as the conditional probability that a point A would choose point B as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian (normal distribution) centered at A.
2. It then tries to minimize the difference between these conditional probabilities (conditional probabilities represents similarities) in higher-dimensional and lower-dimensional space for a perfect representation of data points in lower-dimensional space.
3. To measure the minimization of the sum of difference of conditional probability ,t-SNE minimizes the sum of "Kullback-Leibler divergence" of overall data points using a gradient descent method.

In simpler terms, t-Distributed stochastic neighbor embedding (t-SNE) minimizes the divergence between two distributions: a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding [3].

In this way, t-SNE maps the multi-dimensional data to a lower dimensional space and attempts to find patterns in the data by identifying observed clusters based on similarity of data points with multiple features.

However, after this process, the input features are no longer identifiable, and you cannot make any inference based only on the output of t-SNE. Hence, it is mainly a data exploration and visualization technique.



3.1.2 Application

We implement t-SNE to visualize the distribution of the candidates profile, to view the similarity between the profiles and even the similarity of jobs. The figure below shows the visualization of t-SNE with two components: Depending on the visualization, we can affirm that the profiles of job "Techni-

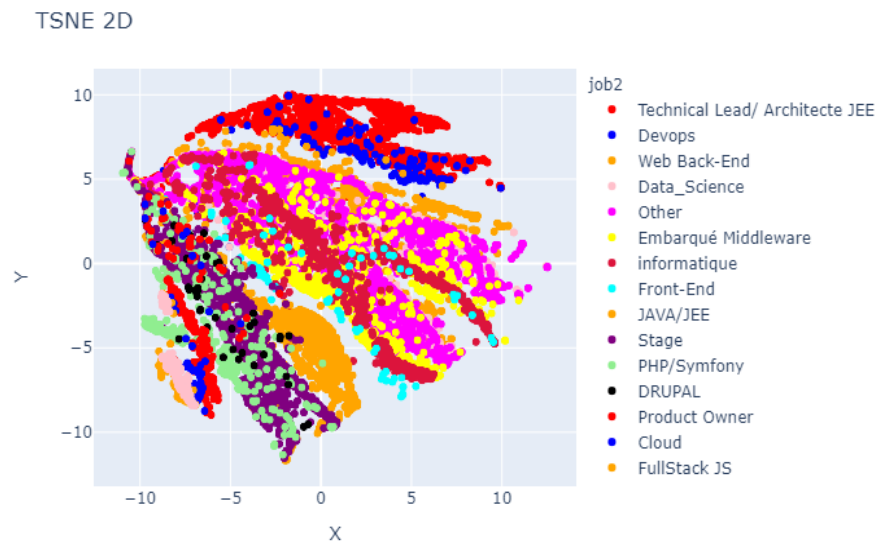


Figure 3.5: T-SNE 2 components

cal lead/ Architecte JEE" and "Devops" are too close. Also, it seems that the profiles "DRUPAL" and "PHP/Symfonyare" are similar. To be more confident about this visualization, we attempt with another t-SNE model using this time 3 components in order to display the distribution in 3 dimensions. We obtained this result as mentioned in this figure:



Figure 3.6: T-SNE 3 components



3.2 K-Nearest neighbors

3.2.1 Principle of process

KNN is used for both classification and regression problems. In classification problems to predict the label of a instance we first find k closest instances to the given one based on the distance metric and based on the majority voting scheme or weighted majority voting(neighbors which are closer are weighted higher) we predict the labels [4].

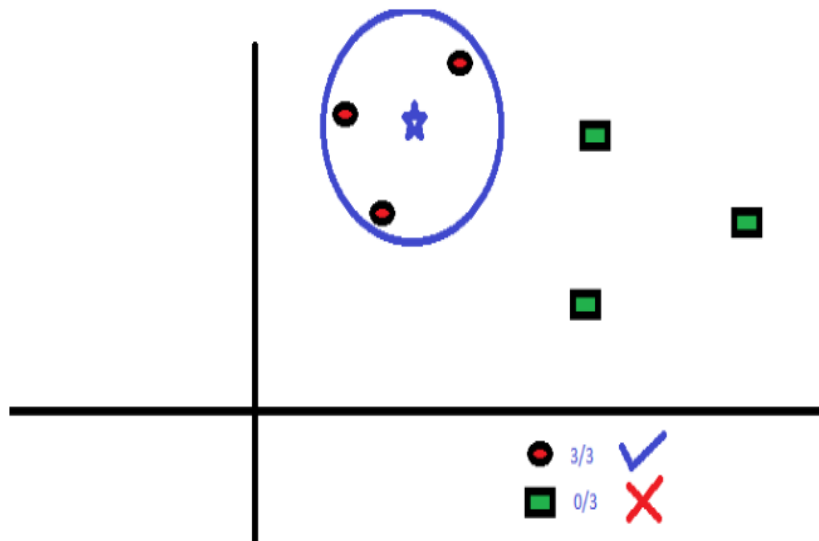


Figure 3.7: example of K-NN voting

This algorithm is also very useful to find similarities based on the distances metrics between the points. There are several distance calculation functions, in particular, the Euclidean distance, the distance from Manhattan, the distance from Minkowski, that from Jaccard, the distance from Hamming. . . etc. We choose the distance function according to the types of data we are handling.

3.2.2 Application

In order to identify the most 10 similar profile, we follow these steps:

1. Extract the numpy array containing the features of all profiles from the dataframe already processed in the previous step
2. Fit the KNN model from scikit learn to the numpy array data and calculate the nearest neighbors for each distances. In this case, we've used the unsupervised Nearest Neighbors method for implementing neighbor searches. Note that we used $k=11$ as a parameter because the first neighbor that the KNN returns is always itself since the distance of an instance to itself is 0 and we can't use that. Also, we used the algorithm "balltree" as a metric to calculate distances.
3. Use a function that aim to get the most 10 similar profiles to a specific one profile either by given the name or the id of the correspond profile that we want to look to his similar profiles.



The final result is stored into dataframe which contains the name, job of each profile , the distance between the two profiles and the LinkedIn profile link to have further information. This figure shows an example of the obtained result:

	name	job2	distance	url
682	Fethi Krout	Front-End	0	https://www.linkedin.com/in/fethi-krout-b6557948
4888	hamdi ghaoui	Front-End	7.72651	https://www.linkedin.com/in/hamdi-ghaoui-68210250
11	walid H.	Front-End	8.08572	https://www.linkedin.com/in/walid-h-188a673a
932	Mahmoud Nbet	Front-End	8.23483	https://www.linkedin.com/in/mahmoud-nbet
138	Fares Doghri	Front-End	9.30303	https://www.linkedin.com/in/faresdoghri
5739	Bilel Bekkouche	Front-End	9.3829	https://www.linkedin.com/in/bilel-bekkouche-05a00483
1577	Mourad Akremi	Front-End	9.46448	https://www.linkedin.com/in/mourad-akremi-93168562
3401	Anouar El Heni	Front-End	9.47715	https://www.linkedin.com/in/anouar-el-heni-632261117
3099	Houdhaifa Hamza	Front-End	9.5	https://www.linkedin.com/in/houdhaifa-hamza
8134	Abdelhak El Mahdaouy	Front-End	9.5	https://www.linkedin.com/in/abdelhak-el-mahdaouy-187bb283
3710	Maher Soua	Front-End	9.67917	https://www.linkedin.com/in/mahersoua

Figure 3.8: Example of Nearest neighbors similarity result

3.3 Cosine Similarity

3.3.1 Principle of process

Cosine similarity is a metric used to determine how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space.

In this context, the two vectors I am talking about are arrays containing the word counts of two documents.

Values range between -1 and 1, where -1 is perfectly dissimilar and 1 is perfectly similar.

Cosine similarity is computed using the following formula:

$$\text{Similarity}(A, B) = \frac{A \cdot B}{\|A\| * \|B\|}$$



We implement this model to find the similarity between profiles

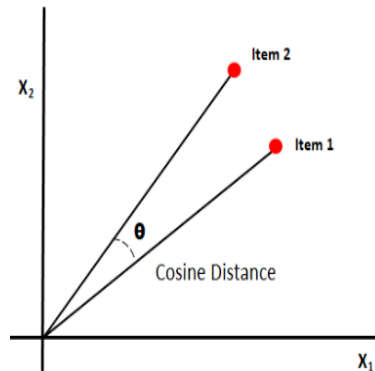


Figure 3.9: Calculate Cosine Similarity

3.3.2 Application

The main goal in applying cosine similarity is to identify the most 10 similar profiles compared to a given one, we followed these steps:

1. calculate the cosine similarity using its theoretical formula.
2. Store all the calculated similarity in a dataframe.
3. Compute a function that allows us to put the correspondent name, job, similarity and url for the 10 closest similarity profile.

The final result is stored into a DataFrame which contains the name, job, the calculated distance and the LinkedIn profile link of each person. This figure 3.10 illustrates an example of the obtained result:

	name	job2	similarity	url
8985	khaled ben jannet	Product Owner	1.000000	https://www.linkedin.com/in/khaled-ben-jannet-64987920
6111	Maimoun BEN TAHER, PMP®	Product Owner	0.944460	https://www.linkedin.com/in/maimoun-ben-taher-pmp®-32244150
4867	Ghazi TEKAYA	Product Owner	0.936830	https://www.linkedin.com/in/ghazitekaya
7867	Nabil LAADHARI	Product Owner	0.936620	https://www.linkedin.com/in/nabillaadhari
9844	Manel Hammouda CSPO	Product Owner	0.936400	https://www.linkedin.com/in/manel-hammouda-cspo-5b52051b
8585	hassen feki	Product Owner	0.932270	https://www.linkedin.com/in/hassen-feki-75726b60
4951	Imen Laabidi	Product Owner	0.930980	https://www.linkedin.com/in/imen-laabidi-73033087
6512	Imededdine HOSNI	Product Owner	0.929300	https://www.linkedin.com/in/imededdinehosni
6940	Mohamed Salama ZIADI	Product Owner	0.926920	https://www.linkedin.com/in/mohamed-salama-ziadi-1b6719a
1134	Hamza Zayani	Product Owner	0.924480	https://www.linkedin.com/in/zayanihamza
3075	Feki Firas	Product Owner	0.924030	https://www.linkedin.com/in/feki-firas
7774	Meriem Chakroun	Product Owner	0.922270	https://www.linkedin.com/in/meriem-chakroun-a315b845

Figure 3.10: Example of Cosine Similarity result



3.4 K Dimensional Tree

3.4.1 Principle of process

The kDTree is a binary tree in which each node is a k-dimensional numerical point, and each node on the tree represents a hyperplane which is perpendicular to the coordinate axis of the current division dimension and divides the space into two parts in the dimension.

The given figure 3.11 clarifies more the principle: One part is in its left subtree and the other part is in its right subtree. That is, if the division dimension of the current node is d , the coordinate values of all points on the left subtree in the d dimension are smaller than the current value, and the coordinate values of all points on the right subtree in the d dimension are greater than or equal to the current value, and the definition is Any child nodes are established.

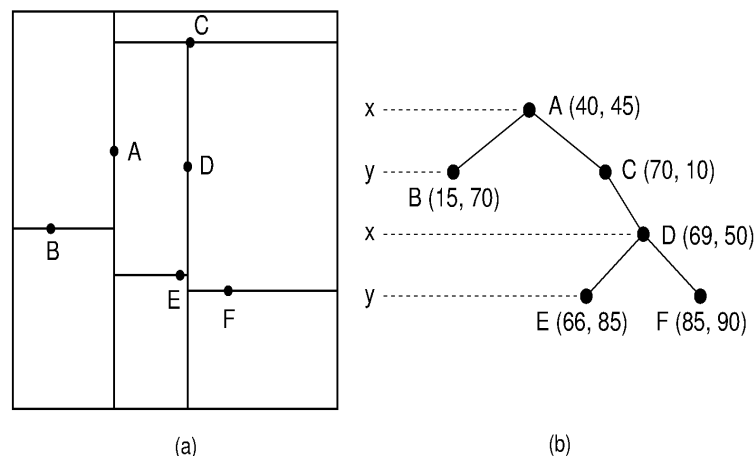


Figure 3.11: Principle of K Dimensional Tree

3.4.2 Application

In order to identify the most 10 similar profiles compared to a given one, we tried to apply K Dimensional Tree and we followed these steps:

1. Use the K Dimensional Tree machine learning algorithm to identify the distances between the profiles.
2. Identify the closest 10 indices distance for each employee.
3. Identify the closest 10 profile distances for each employee.
4. Prepare the functions that allows to identify the index from name from the K Dimensional Tree algorithm.
5. Define a function that allows us to have a dataframe that contains the name, job, distance and URL for the 10 closest profile using K Dimensional Tree algorithm.



The final result is stored into a DataFrame which contains the name, job, the distance and the LinkedIn profile link of each person. This figure 3.12 illustrates an example of the obtained result:

	name	job2	distance	url
3604	Malek Massoudi	DRUPAL	0	https://www.linkedin.com/in/malek-massoudi-8822aaa2
4361	Sofiene Chaari	DRUPAL	5.99657	https://www.linkedin.com/in/sofiene-chaari
5155	haythem hammami	Stage	7.39587	https://www.linkedin.com/in/haythem-hammami-06129466
5626	Dhafer Ben Slama	PHP/Symfony	7.50417	https://www.linkedin.com/in/benslama
1522	Walid Guesmi	DRUPAL	7.55241	https://www.linkedin.com/in/wguesmi09
5487	Tlili Achref	PHP/Symfony	7.78373	https://www.linkedin.com/in/tlili-achref-951857a2
3812	Rami KESSENTINI	PHP/Symfony	7.80618	https://www.linkedin.com/in/rami-kessentini-bb525967
9779	Ahmed BAKLOUTI	PHP/Symfony	7.86209	https://www.linkedin.com/in/ahmedbaklouti
5239	Ben wanes Mohamed Ali	PHP/Symfony	7.92543	https://www.linkedin.com/in/benwanesmohamedali
451	chaibi issam	PHP/Symfony	7.95692	https://www.linkedin.com/in/chaibi-issam-53275724

Figure 3.12: Example of K Dimensional Tree result

4 Evaluation

Based on the database provided by Wevioo, it's not required with a target value. So, we had to implement another metric to deal with evaluation between two values in order to get the performance of the model chosen.

Meanwhile, correlation tells us about the relationship and also the similarity occurring between two profiles.

4.1 Correlation test

Correlation is a bi-variate analysis that measures the strength of association between two variables and the direction of the relationship.

In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1.

A value of ± 1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker.

To verify the result of Cosine Similarity and KDTree Model we have implemented a different type of correlation test.



4.2 Pearson Correlation

Pearson r correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables.

The Pearson's correlation coefficient is calculated as the covariance of the two variables divided by the product of the standard deviation of each data sample

$$\text{Pearson's correlation coefficient} = \text{covariance}(X, Y) / (\text{stdv}(X) * \text{stdv}(Y))$$

The result of the calculation, the correlation coefficient can be interpreted to understand the relationship. A value of 0 means no correlation.

The value must be interpreted, where often a value below -0.5 or above 0.5 indicates a notable correlation, and values below those values suggests a less notable correlation.

4.3 Kendall Correlation

Kendall correlation is a non-parametric test that measures the strength of dependence between two variables

$$\text{Kendall's correlation coefficient} = N_c - N_d / \sqrt{(N_c + N_d + T_x)(N_c + N_d + T_y)}$$

where:

N_c and N_d denoting the number of concordant pairs and the number of discordant pairs

Concordant/Discordant describe if the ranks of two samples are ordered in the same way

T_x denoting the number of pairs tied for the first response variable only

T_y denoting the number of pairs tied for the second variable only

4.4 Spearman correlation

Spearman correlation is a non-parametric test that is used to measure the degree of association between two variables.

Instead of calculating the coefficient using covariance and standard deviations on the samples themselves, these statistics are calculated from the relative rank of values on each sample.

$$\text{Spearman's correlation coefficient} = \text{covariance}(\text{rank}(X), \text{rank}(Y)) / (\text{stdv}(\text{rank}(X)) * \text{stdv}(\text{rank}(Y)))$$

As with the Pearson correlation coefficient, the scores are between -1 and 1 for perfectly negatively correlated variables and perfectly positively correlated respectively.

Taken together, in regards to tolerance of outliers and discrepancies in data, Kendall's correlation is the most robust measure, followed by Spearman's correlation while Pearson's correlation is the most sensitive one.



4.5 Application

Evaluating model performance with the data, we had to check the correlation between two profiles. Taking for example two similar profiles classified as "Technical Lead/ Architecte JEE " and we will test the correlation between them.

	name	job2	similarity	url
0	<u>Imen Hammi</u>	Technical Lead/ Architecte JEE	1.000000	https://www.linkedin.com/in/imen-hammi-3102919a
8320	Kais Mbarki	Other	0.999800	https://www.linkedin.com/in/kais-mbarki-26b20932
5956	Mohamed Lrb RAOUAFI	Stage	0.999660	https://www.linkedin.com/in/mohamed-lrb-raouafi-30959730
333	Imed loukil	Other	0.999580	https://www.linkedin.com/in/imed-loukil-3b559335
2420	ali belhaj	Technical Lead/ Architecte JEE	0.999400	https://www.linkedin.com/in/ali-belhaj-0bb6411a
2934	<u>taoufik jaghmoun</u>	Technical Lead/ Architecte JEE	0.999050	https://www.linkedin.com/in/taoufik-jaghmoun-819347159
4699	Thouraya Hammami Bekri	Data_Science	0.998950	https://www.linkedin.com/in/thouraya-hammami-bekri-4703aba
5209	Touhami Saiidia	Technical Lead/ Architecte JEE	0.998940	https://www.linkedin.com/in/touhami-saiidia-7873b1104
273	Fourat Mamoghli	Other	0.998930	https://www.linkedin.com/in/fourat-mamoghli-aa24a137
1770	Hatem Bacha	Other	0.998930	https://www.linkedin.com/in/hatem-bacha-92007910

Figure 3.13: Result from Cosine Similarity

As a result from the tests, we got with the kendall and Spearman Correlation a value worth 1 and with Pearson test we had 0.999. Likewise, we can say that there is a strong strength correlation between this two profiles

5 Conclusion

During this chapter, we used many different methods like visualization and models. Provided that to build our recommendation system in which we will represent it as a web application in the next chapter.



Chapter 4

Deployment

1 Introduction

In this section we reach to the final step of our work which is the Deployment phase. In order to use the project in a real environment by all users we build a web application. In this chapter we will explain the various functions of this application and describe the architecture and the tools used to create our deployment.

2 Development environment

In order to deploy our recommendation system models, we choose Django as a web development tool. Django is a high-level Python Web Development framework that encourages rapid development and clean, pragmatic design. It has been built by experienced developers, and takes care of much of the hassle of Web development. It is also free and open source [7].



Figure 4.1: Django logo

Django REST Framework is a powerful and flexible toolkit for building Web APIs which can be used to Machine Learning model deployment and since Django is written in Python it makes it a great choice of web framework for deploying machine learning models [6]. With the help of Django REST framework, complex machine learning models can be easily used just by calling an API endpoint. A few recognizable websites that use Django include Instagram, Pinterest, YouTube, and Spotify, and many others.



Django is based on the Model-View-Controller software design pattern as it's effective way of structuring a dynamic website .

The MVC concepts are:

- Model : the model handles the dynamic data structure
- View : the view is what the user can see and interact with
- Controller : the controller is the middle man that accepts inputs and converts it to commands for the model or view

This figure below illustrates how Django work using the MVC design pattern :

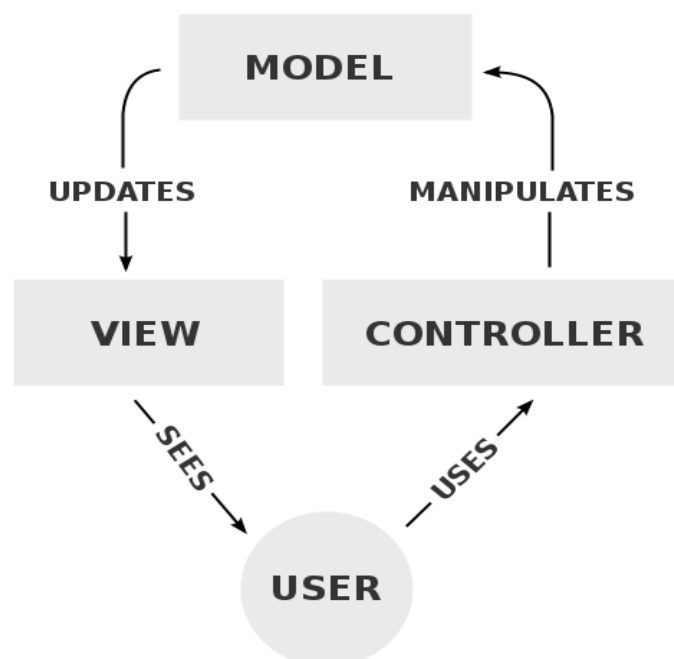


Figure 4.2: MVC process

3 Implementation of web application

We decided to create a web applications for two types of users depending on the type of endpoint user. The result is mainly for the human resource agent that he/she can easily select the desired profile and view the most similar profile with detailed information. But, we think that we should create another interface for a data scientist employee or any employee in the IT field in order to evaluate the results and understand better the errors in the modeling phase.

3.1 Login page

As any web application which wants to restrict the access between the user of the application and secure data, we start our application by the login page which let the access only for the correct name user and password.

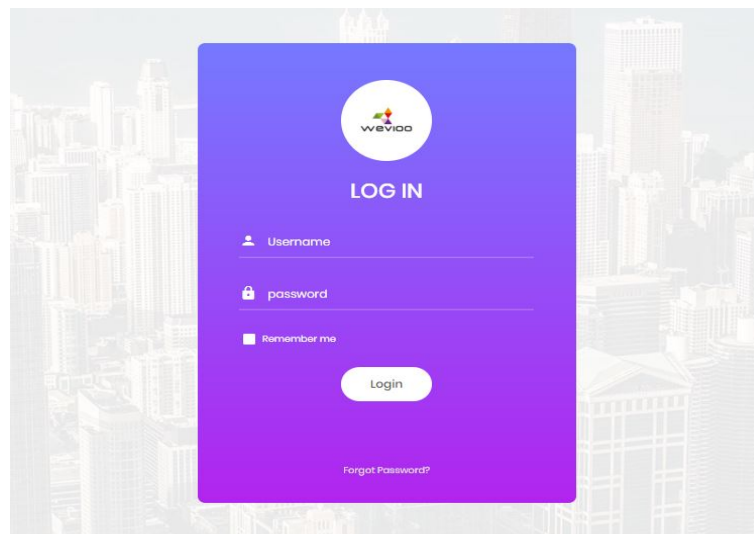


Figure 4.3: Login page

3.2 Human resource agent interface

After login as long as human resource agent user, you will find this " Home page ".

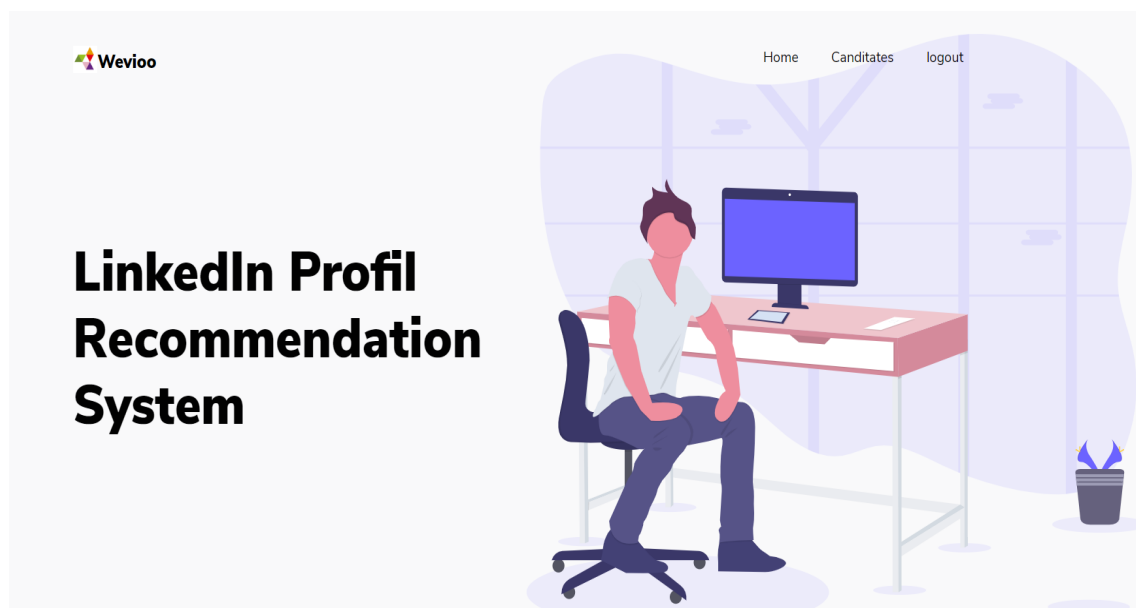


Figure 4.4: Home page 1

While scrolling in the " Home page ", we can find a section which contains the jobs which have the most number of employees. In addition, we can find the number of profile of each job from the original database which contains 10000 profiles. As we see in the figure below, the most number of profile is for the job data science then we find java EE job ...

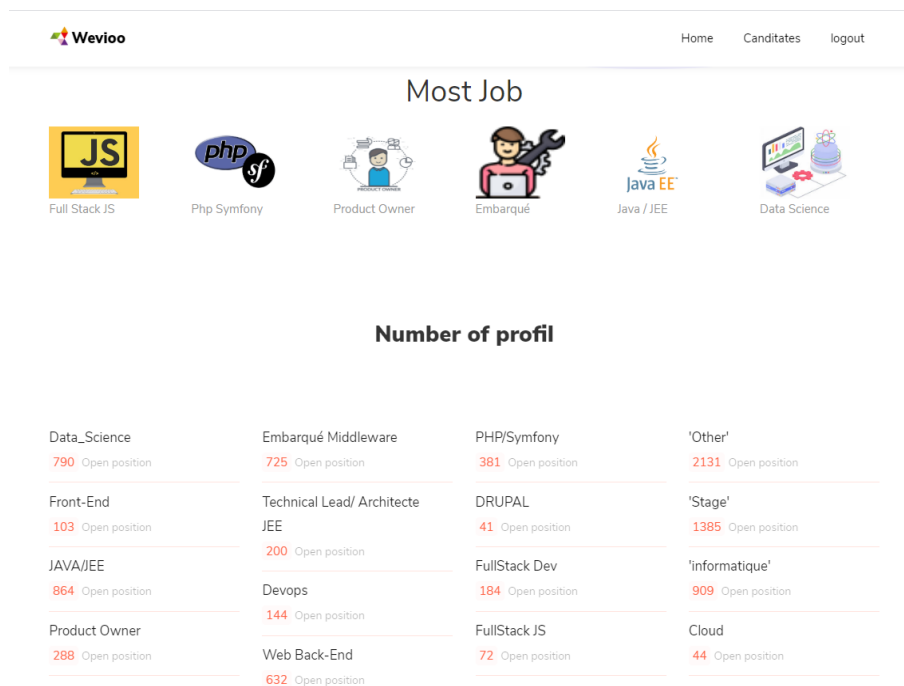


Figure 4.5: Home page 2

Moving to the "Candidate page", we find all the 10000 profiles that exists in the original database with the possibility of filtering the profiles by the job as mentioned in this figure :

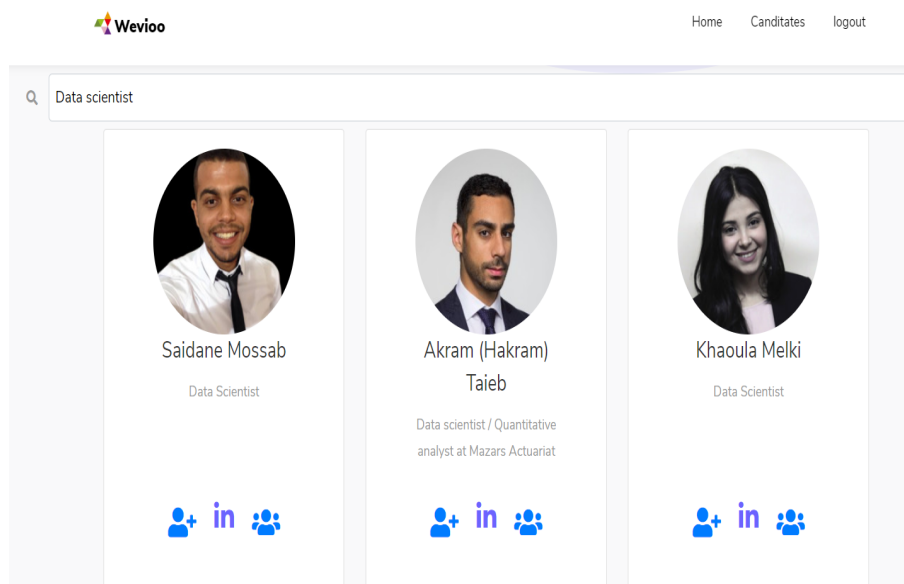


Figure 4.6: candidate page 1



Also, we can filter by the name of the profile as mentioned in this figure below:

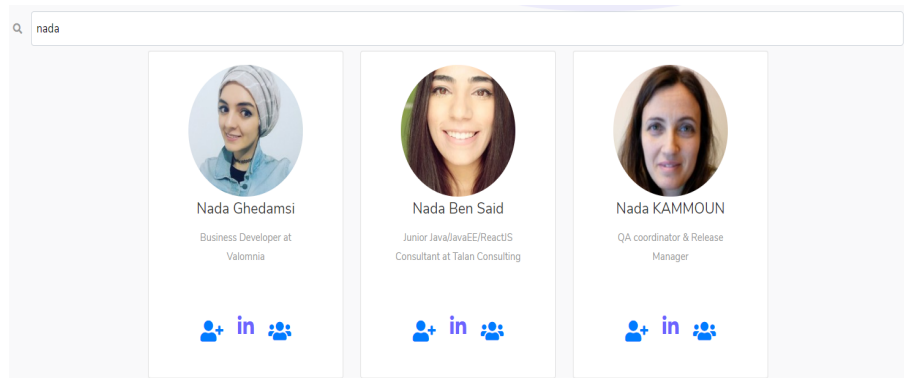


Figure 4.7: Candidate page 2

Each candidate card has a button to have more information. In this page you can find all details that you will find in the LinkedIn profile such as the experience duration, professional skills, the actual job...etc.

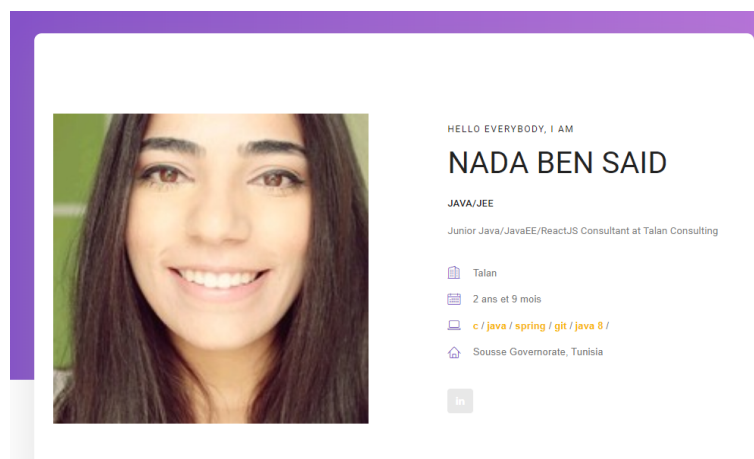


Figure 4.8: Candidate page 2

Furthermore, you can visualize the professional career and even the educational career for each profile. These two figures below show an example of a profile professional and educational career.

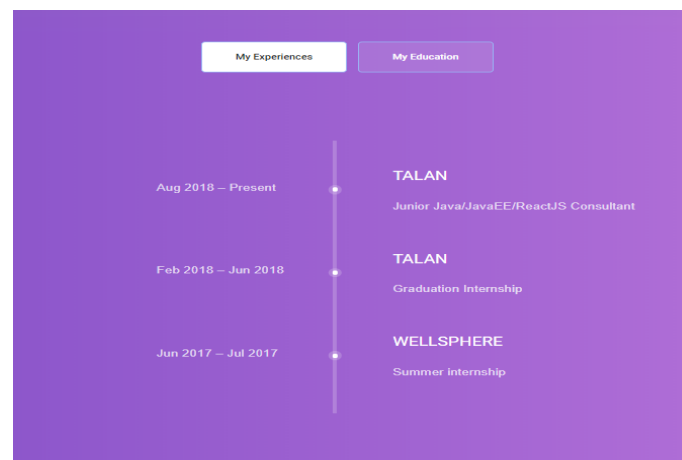


Figure 4.9: example of professional career information

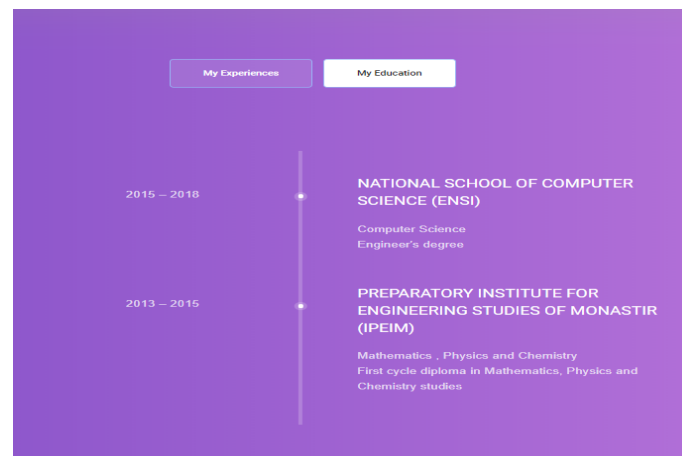


Figure 4.10: example of educational career information

Now, we move to the main objective of the web application: to find similarities of different profiles. we choose the model of Nearest neighbors to display the result of the recommendation system which give us a good result in the previous step " Modeling and evaluating ". we use pickle to save the model and to use it in the django application.



By clicking on the right button on any profile card you will find as a result the most 10 similar profile according to a specific profile. We should notice also that the first profile is the profile itself because it's distance equivalent at 0. The figure below shows an example of similarity of a FullStack JS profile: If

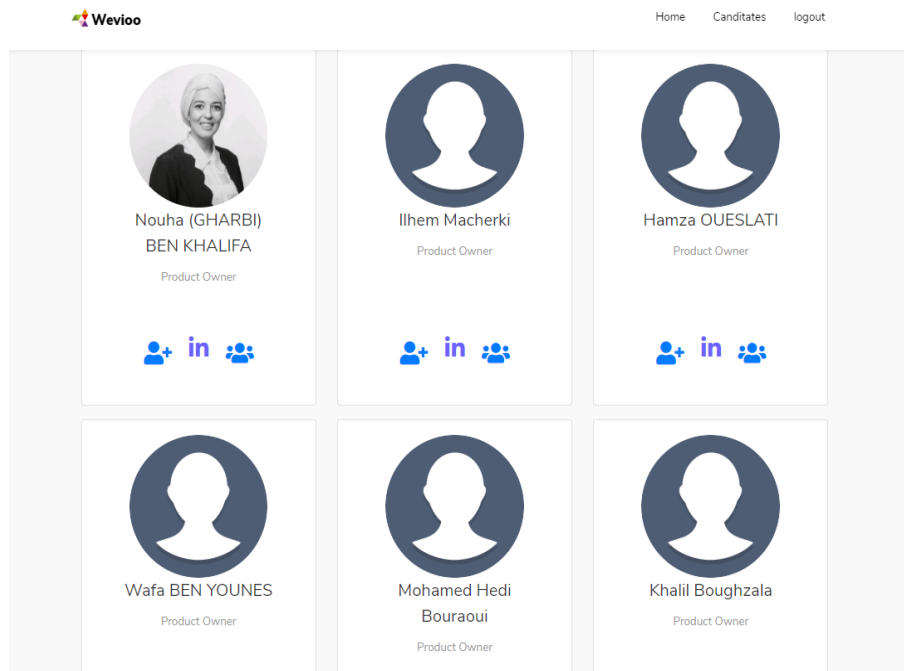


Figure 4.11: example of application similarity

you want to verify the information or to have more specific information, you can go to any profile linkedin using the middle button "In" as mentioned in this figure:

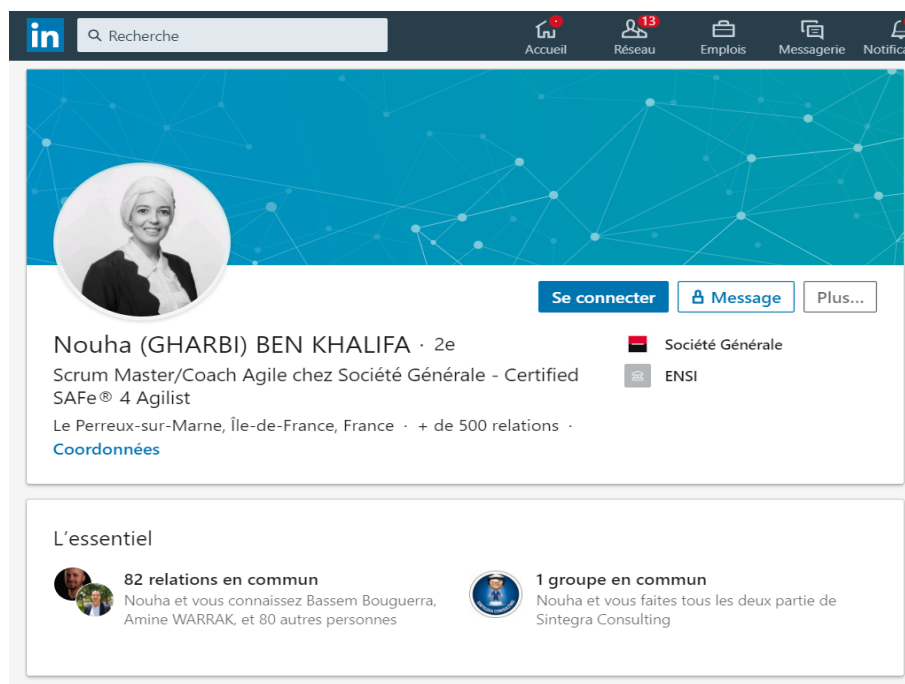


Figure 4.12: linkedin profile



3.3 Admin interface

Now, we move to the admin interface where we can find more information about the profiles similarities with many possible visualisation.

At the home page we find a bar chart which represents the job title by the profile number.

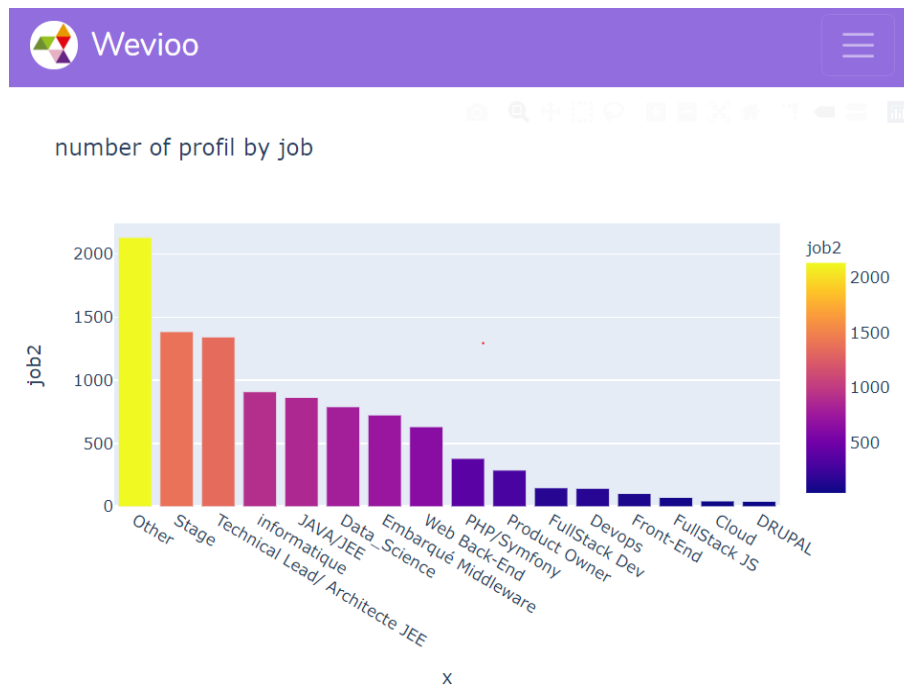


Figure 4.13: bar chart job title by profile numbers

After that, we find a data table which contains all the linkedin profile form the original database. Each row contains job title, name of the employees and the linkedin url for the correspondent profile. The figure below shows the profile list data table.

Profil list & Job

List

Check Similarity

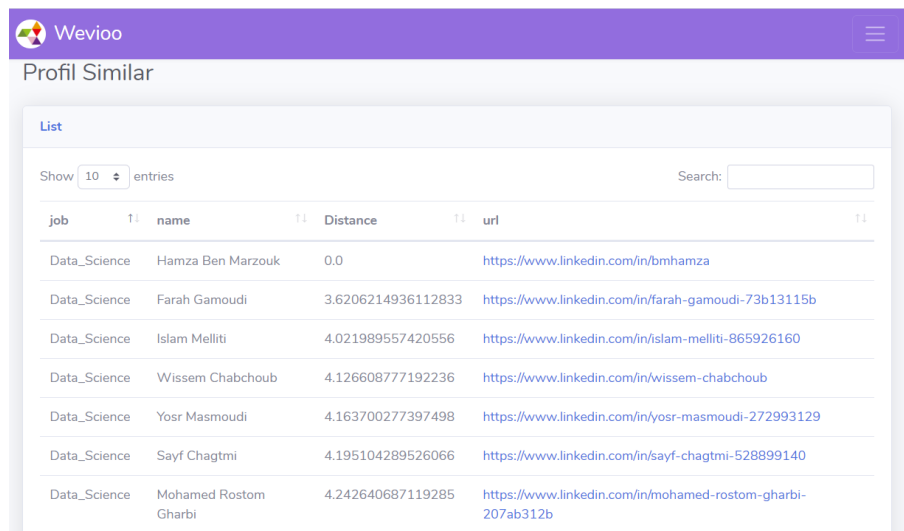
Show 10 entries Search:

id	job	name	url
120	PHP/Symfony	ons essaddi	https://www.linkedin.com/in/ons-essaddi-bb6639130
121	Other	Gayth Mliki	https://www.linkedin.com/in/gayth-mliki-36a665152
122	Embarqué Middleware	Mohamed Anis BOURIGA	https://www.linkedin.com/in/mohamed-anis-bouriga-9292b84b
123	PHP/Symfony	hechmi shimi	https://www.linkedin.com/in/hechmi-shimi-2a6374ba
124	Embarqué Middleware	Mourad Bellili	https://www.linkedin.com/in/mourad-bellili-91a0bb111

Figure 4.14: profile list data table



When we click on a row , we get the its id and we can then check the profile similarities. The result is like we show in the figure below an example similarity of one profile : we have the most 10 similarity profiles and each row contains as features job title, profile name , the distance between the two profiles and the linkedin url profile.



The screenshot shows the 'Profil Similar' section of the Wevioo application. It includes a search bar and a table with 10 entries. The table columns are job, name, Distance, and url. The data is as follows:

job	name	Distance	url
Data_Science	Hamza Ben Marzouk	0.0	https://www.linkedin.com/in/bmhamza
Data_Science	Farah Gamoudi	3.6206214936112833	https://www.linkedin.com/in/farah-gamoudi-73b13115b
Data_Science	Islam Melliti	4.021989557420556	https://www.linkedin.com/in/islam-melliti-B65926160
Data_Science	Wisse Chabchoub	4.126608777192236	https://www.linkedin.com/in/wisse-chabchoub
Data_Science	Yosr Masmoudi	4.163700277397498	https://www.linkedin.com/in/yosr-masmoudi-272993129
Data_Science	Sayf Chaghtmi	4.195104289526066	https://www.linkedin.com/in/sayf-chaghtmi-528899140
Data_Science	Mohamed Rostom Gharbi	4.242640687119285	https://www.linkedin.com/in/mohamed-rostom-gharbi-207ab312b

Figure 4.15: Profile distribution using tSNE

Moreover, the admin can view many types of the profile distribution depending on their features using the previous result of tSNE and plotly as a visualization tool.

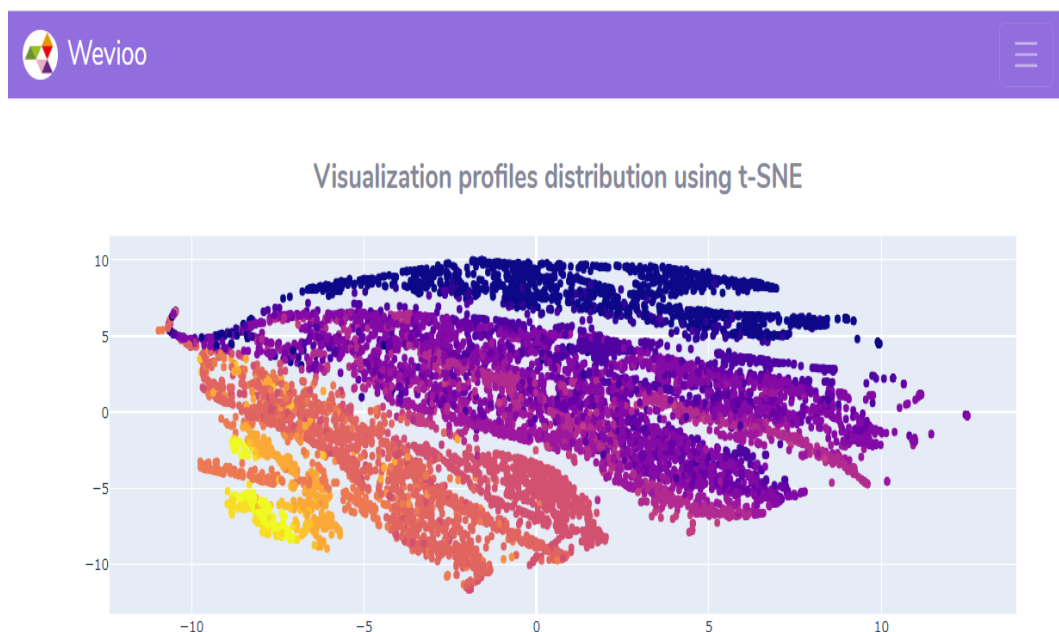


Figure 4.16: Visualization profiles distribution using tSNE



we can also view the profile distribution by their job title using 2 components of the tSNE dimension reduction as we mention in this figure :

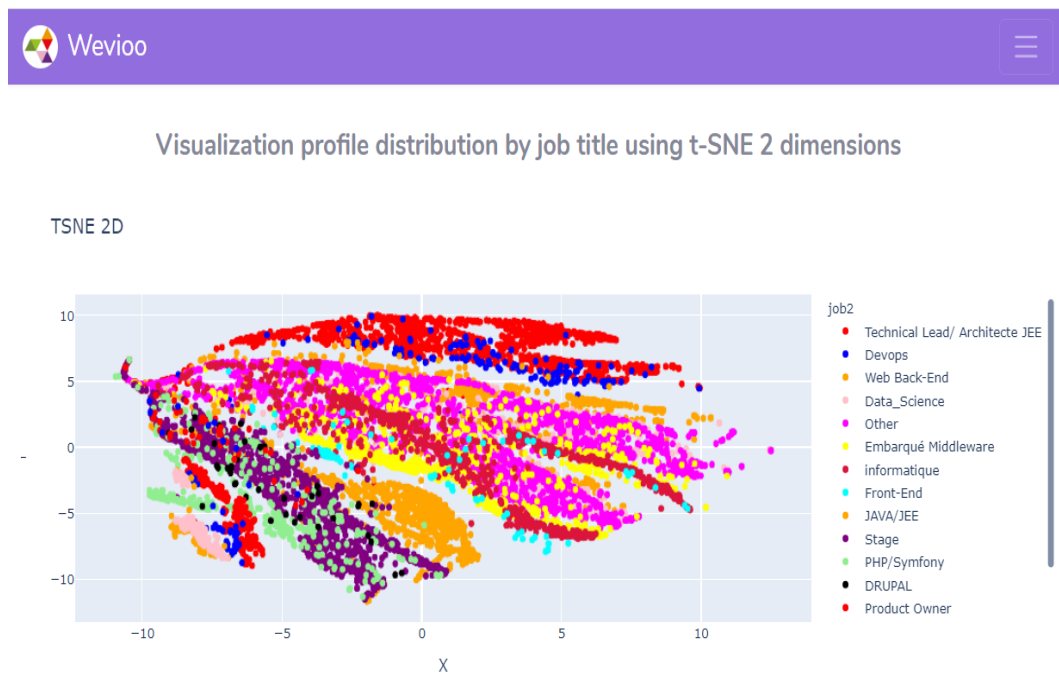


Figure 4.17: Visualization profiles distribution by job title using tSNE 2 dimensions

For a better visualization, there is another visualisation using 3 components of the tSNE algorithm in order to display the profiles distribution in 3 dimensions.

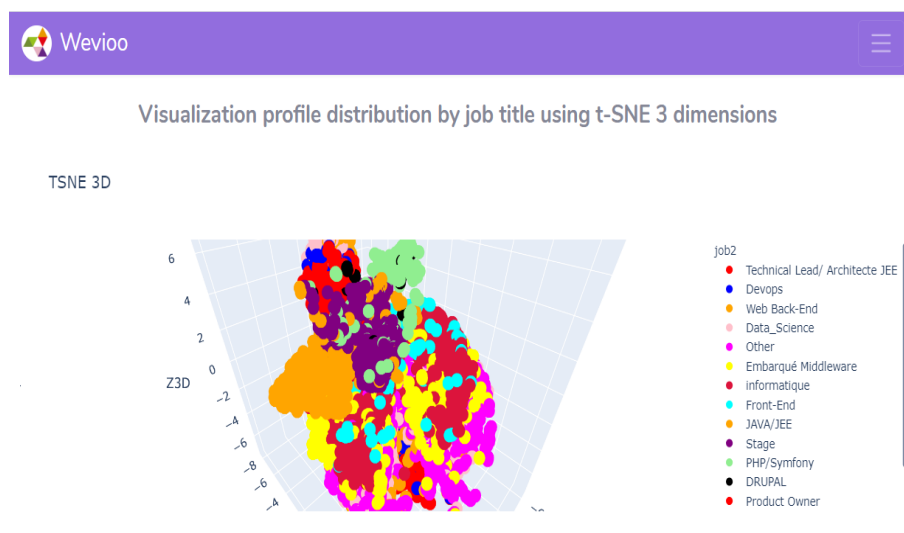


Figure 4.18: Visualization profiles distribution by job title using tSNE 3 dimensions



4 Conclusion

In this final chapter, we have successfully deployed our modeling results into a web application which makes easy for any type of user to find the most 10 similar profiles such as a Human resource agent and the possibility to view more information for any profile. In addition, we create a second interface for admin which contains add to the first interface more visualization and details about the profiles and job distributions using the tSNE model.



General Conclusion

Information overloaded is a serious problem in information retrieval systems and that's the same problem for a recruitment process. Recommendation system resolve the problem of information overloaded and open new opportunities of retrieving individualized information on the internet and access the profiles of the candidates.

In this project, we have described and realized a collaborative filtering recommendation system and enhancing the execution of the system. We also discussed various algorithms to generate recommendation and measuring the quality, performance of the recommendation system.

In the whole project, we respected all the steps of a data science life cycle project. At the beginning, we had an understanding for the problem [Business Understanding] because getting clarity allows us to determine which data will be used to answer the core question. Then we chose our analytic approach to limit the algorithms that will be used.

Once that is done, we identified [Data Collection Data Preparation] the necessary data content, formats, the most relevant features mainly using NLP and also sources for initial data collection, which is a representative of the problem we want to solve [Updating Data Base, Complementary Data].

Next, The modeling step includes two types that depends on the business understanding that we dealt with at first: Descriptive and classification. These models are based on the analytic approach that was taken and we tried to evaluate the accuracy of them.

Once valid, the model was deployed and a feedback phase will be launched in order re-evaluate it from a customer point of view.

This project could be implemented in many companies that need a recommendation system for its HR department and it could help them to take the right decision about choosing the suitable profile for their needs.

Lastly, to enriching our work we are planning to add as a perspective work many details. One of the ideas, we are figuring to make like a confidence interval for the recommended profiles and just presenting the ones included in that interval. Accordingly, we could adjust the number of the recommended profiles according to that interval.

Furthermore, we were thinking why we had been restricted just to LinkedIn. So, we could ameliorate the data base with more profiles from many other professional social networks and add supplemental profiles to work with in order to increase the performance of our Recommendation System. Eventually, for the list given by wevioo, we want to extend it and add extra elements to be more specific in jobs.



References

- [1] *Introduction to Recommender System*
<https://towardsdatascience.com/intro-to-recommender-system-collaborative-filtering-64a238194a26>
- [2] Recommendation systems: Principles, methods and evaluation :
<https://www.sciencedirect.com/science/article/pii/S1110866515000341>
- [3] *Introduction to t-SNE*, <https://www.datacamp.com/community/tutorials/introduction-t-sne>
- [4] *Introduction to k-Nearest Neighbors*,
<https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/>
- [5] *Prototyping a Recommender System Step by Step*,
<https://towardsdatascience.com/prototyping-a-recommender-system-step-by-step-part-1-knn-item-base>
- [6] *Correlation Test*
<https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between->
- [7] *Introduction to Deploying Machine Learning Models with Django*,
<https://www.mlq.ai/django-machine-learning/>