An AI based talent acquisition system for matching resume and job posts.

MISHRA Rudresh, PORTILO Valentin and RODRIGUEZ Ricardo M.

Supervisors: Yannis HARALAMBOUS and Nicolas JULLIEN

Department of Information Technology, IMT Atlantique France

Contents

1 Excecutive Summary **4**

2 Code **5**

2.1 Github . . . 5

2.2 DataBase . . . 5

3 Abstract **5**

4 Introduction **5**

5 Motivation **6**

6 Problem Identification and Objectives **7**

6.1 Problem . . . 7

6.2 Current market solution . . . 8

6.3 General and specific aims . . . 9

7 Business Understanding **9**

7.1 What is important in a CV ? . . . 9

7.2 Regular structure on a CV - Metadata . . . 10

7.3 Recruiter points of views . . . 10

7.4 Related work on Parsing and Matching . . . 11

7.5 Recruiters Criteria to match the CV to a job post . . . 12

7.5.1 HAY criteria . . . 12

7.6 Open Source Knowledge Bases . . . 12

7.6.1 Rome . . . 12

7.6.2 O-NET . . . 12

7.6.3 ESCO . . . 13

7.7 Strategy/Plan . . . 13

7.7.1 Summary generated from CV's . . . 13

7.7.2 Feedback from CV's to the candidate . . . 13

7.7.3 Recommendation . . . 13

7.7.4 Stages, algorithm flow . . . 14

7.7.5 Metrics . . . 15

7.8 Ontology . . . 21

7.9 Model Proposition (Input/Output) . . . 21

7.9.1 No data model proposal . . . 21

7.9.2 Data model proposal . . . 24

8 Implementation **25**
8.1 Create ontology's . . . 25
8.1.1 Technical Skill ontology . . . 25
8.1.2 CSO generation . . . 26
8.1.3 Domain Skill ontology . . . 27
8.1.4 Domain skill ontology generation . . . 27
8.1.5 Cultural values ontology . . . 27
8.2 Matching . . . 30
8.2.1 Creating Skill Graph from Ontologies . . . 30
8.2.2 Cultural Match . . . 32
8.2.3 Education Match . . . 33
8.2.4 Required Skill Match . . . 33
8.3 Multiple CVs to Job Post matching . . . 33
8.3.1 Filtering . . . 33
8.3.2 Sorting . . . 34

9 Evaluation **35**
9.1 Use case 1 - OneToMany Matching . . . 35
9.1.1 Result from test 1: 5 CVs and Business Oriented Job Post . . . 35
9.1.2 Result from test 2: 5 CVs and Technical Oriented Job Post . . . 36
9.2 Use case 2 - OneToOne Matching . . . 37
9.2.1 Result from test 3: CV1 and Business Oriented Job Post . . . 37
9.2.2 Result from test 4: CV3 and Technical Oriented Job Post . . . 37

10 Conclusions **38**

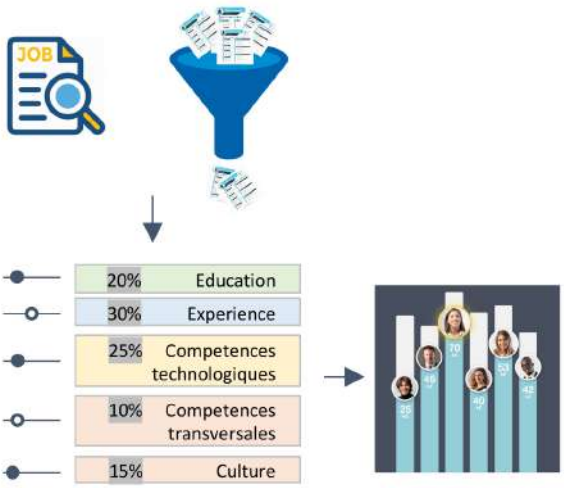
11 Future Prospective **38**

1 EXCECUTIVE SUMMARY

Our CV - Job Posting matcher proposal is basically the most innovative and effective product to accomplish a successful hiring process. Such exploit is backed up by artificial intelligence, operational research algorithms that empowers the user to tune its hiring process by making it accurate, simple, fair and effective. Why just try to simply hire the best candidate if you can hire the right one.

As you know, finding the right candidates for your organization is a serious task. You might be needing the one that is the best on the domain but also the one that will be the best fit for a non cohesive team or maybe, the one that would let you do some high impact in the team productivity. Furthermore, the hiring team is constrained to do very repetitive tasks for a long period of time, instead of directing this team to higher impact tasks for your enterprise. Besides, let's face it, a person can't handle a lot of information at the same time (in fact, maximum, seven dimensions), so "less accurate" should be added to the process flaws. Actually, since these tasks are of high impact to your business, having good human capital totally justifies the sacrifice of all the flaws, time and resources invested to accomplish the best result.

This could have been the way to go before. But technology, and more specifically, artificial intelligence and language processing open new horizons to boost the process to which we were restricted before. So, our CV - Job Posting matcher proposal breaks through this barrier and gives the opportunity to automate these tasks for any domain such as construction or business and any level such as entry or senior roles. By using our tool, recruiters would be able to tune parameters that will adjust to the company context and goals of the recruiting process for any role. Organizational culture, domain and transversal skills, education level, role coherency and several other dimensions that your recruiter might be interested in, could be tuned. All these dimensions reduced to 5 dimensions which will handle the rest, the 5 dimensions that will let you hire someone with "more that kind of profile" rather than "this other".



As Data Scientists, we understand the art of artificial intelligence, natural language processing, and the nuances of computer science that enables us to develop and end-to-end product. Besides, as one of the basis of the project, several recruiters have been interviewed from several domains and experience causing the product to have strong basis to develop the theory of the needs, and the process. Furthermore, experts and researchers in the domains of language processing and management and data science guided us through the process.

We would love to see our product empowering enterprises and creating a more healthy and right environment in the organizations. Increasingly having more right candidates in the companies, and RH team members assigned to more and more different high impact and rewarding tasks.

2 CODE

2.1 GITHUB

** See instructions, installing and usage inside*

`https://github.com/RudreshMishra/S5_CvParserMatcher`

2.2 DATABASE

** To visualize the code, you can do it direcly from **mongod** command line, or using **MongoDB Compass** (recommended, easy to use)*

`mongodb+srv://user_imt:2020@s5resumesdb-ppukj.azure.mongodb.net/test`

3 ABSTRACT

In a recruitment industry, selecting a best cv from particular job post within pile of thousand cv's is quite challenging. Finding a perfect candidate for an organization who can be fit to work within organizational culture is a difficult task. In order to help the recruiters to fill these gaps we leverage the help of AI. we propose two models to solve these problems , model with the data and one without the data. In this report our approach is to perform the business understanding in order to justify why such problems arise and how we intend to solve these problems. we limit our project only to solve the problem in the domain of computer science industry.

4 INTRODUCTION

Unlike traditional recruitment methods, such as employee referrals, CV screening, and face-to-face interviews, AI is able to find patterns unseen by the human eye. it could be used to find the right person for the right role faster and more efficiently than ever before. In order to rapidly improve talent management and take full advantage of the power and potential AI offers, we need to shift our focus from developing more ethical HR systems to developing more ethical AI. McKinsey's Global Institute model predicts that approximately 70 percent of companies will adopt some form of AI by 2030 .When it comes to identifying talent or potential, most organizations still play it by ear. Recruiters spend just a few seconds looking at a resume before deciding who to "weed out" [**erecruit_cv_form**]. Often when hiring is made its very important to know the current strength of the organization and based on it if hiring is made candidate is referred to be a good fit for an organization [5]. There's an increasing evidence that AI could overcome this trade-off by deploying more dynamic and personalized scoring algorithms that are sensitive as much accuracy as to fairness to an organization.

AI has power to provide deep hiring efficiencies, increase talent mobility and will ensure that the scores that come out of the hiring process are both maximally predictive of outcomes that matter to employers and also free from all types of bias and provides best fitting candidate as per organizational work environment .

In this report, we present an analysis of the problem of recruitment, followed by a proposition of the model. In Section 2, we will cover a general understanding of the recruitment field, and we will cover the problem statement of this project and the proposed formulations.

5 MOTIVATION

According to several surveys, the recruiting field is one of the main concerns of many CEO's [4, 25]. In fact, according to the Society for Human Resource Management [4], employers spend an enormous amount on hiring: an average of $4,129 per job in the United States.

We observe that the recruiting process is not an easy task. It contains several stages, as is represented in Fig. 1. The average time to fill an open position is approximately 42 days [1] and even with this long process, most of the timer recruiters are not sure that they choose the right candidate [4].

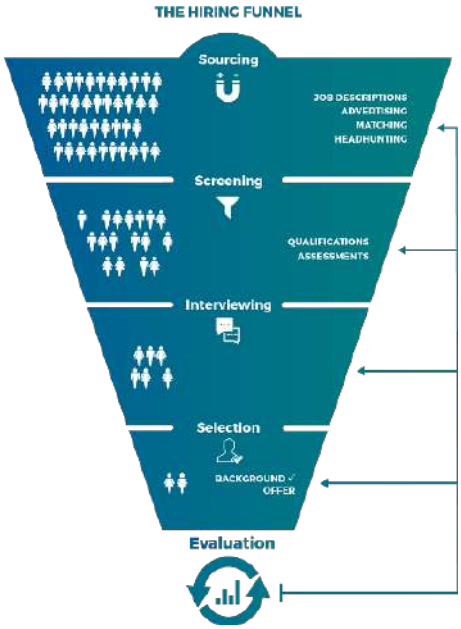


Figure 1: The hiring funnel. Image taken from [5]

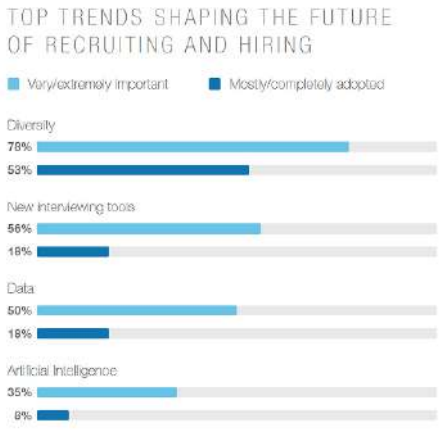
To overcome this issue, many innovations in the recruiting field have arisen, such as video interviews analysis, accurate CV parsers, AI personality tests, AI candidate recommendation, among others. According to an analysis made by the Linkedin talent blog in 2018 [16], there exists four trends that will shape the future of recruiting:

- **Diversity**: Refers to the fact that changing demographics are diversifying communities, shrinking talent pools for companies that don't adapt. This trend is relevant since diverse teams are more productive, more innovative, and more engaged also makes it hard to ignore.
- **New interviewing tools**: This tools try to improve ineffective traditional way of interviewing . New tools are concentrated on online soft skills assessments, job auditions, casual interviews, among others.
- **Data** : Refers to data informing talent decisions, such as prediction of hiring outcomes, or smarter recruiting decisions based on data analysis.

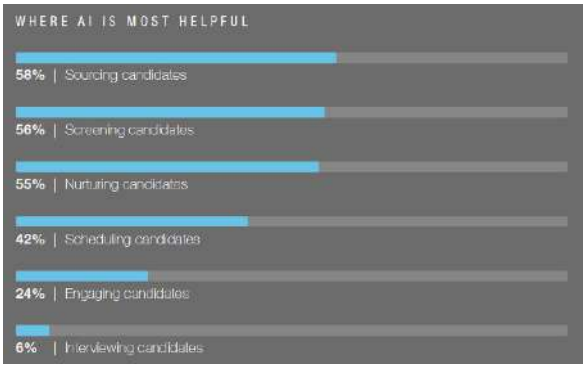
– **Artificial Intelligence**: It is focused on automated candidate searches and quickly find prospects that match specific criteria. There are also technologies that help to screen candidates before even speaking to them. The development of chat bots can respond to candidate questions so recruiters don't have to.

A graph showing statistics of this trend is shown in Fig. 2a. We observe that all this trends have almost similar importance, however Artificial Intelligence is the field less adopted. This is seen as an opportunity to develop many projects.

In Fig. 2b we observe the details of the trends in AI. According to recruiters, the top helpful applications could be sourcing candidates, screening and nurturing.



(a)



(b)

Figure 2: (a) Four trends were identified based on numerous expert interviews and a survey of 9,000 talent leaders and hiring managers across the globe [15] **(b)** Details on hiring trends.

While Sourcing candidates is the process to contact as much as candidates as possible, screening candidates refers to the problem on select a candidate based on its cv. It makes sense that the screening is one of the fields where to innovate since Profile/CV matching is a multi dimensional task, where the human eye is not enough to compare precisely many CV's in a multi dimensional way.

6 PROBLEM IDENTIFICATION AND OBJECTIVES

6.1 PROBLEM

Recruitment can be a very demanding and tough process for a company and their recruiters. Many a times, recruiters end up hiring a not so competent candidate which eventually renders all the efforts put through a recruitment process as waste [19]. Having a perfect fit for a job position is as tough as finding that perfect fit and that entire process of finding one can be very demanding at times. it is very important for a recruiter to pick a candidate whose competent matches with current organization strength. In addition to this, there are a lot of difficulties which candidates faces while searching for their dream job. Starting from finding a trusted platform to look for job roles, to tracking their application and feedback, the entire process has a lot of roadblocks which renders the entire recruitment process as very time

consuming and as a frustrating one [2]. The root problem is because profile matching field is multi dimensional and it is very difficult as an individual to cover all dimensions and to select the best candidate and to justify the reason for the selection and rejection of the candidate.

As a way out, the candidates or recruiter often make use of third parties to reach out for their desired job roles and positions. they have a team dedicated to a more manually approach to do the matching. This eventually results in a major chunk of their salary being lost to the third party facilitators and targeted problem cannot be solved.

Every organization works as per their unique values and strength and its very hard to generalize the common matching solution to a candidate which can fit best for all the organization. There are many solutions existing in the market to automate the matching such as creating the recommendation systems which are based on keyword matching which often results in poor recommendations. Also their are many AI related solution exist which provides solution to the problem, however if the candidate is recruited without considering into account the organization values and strength , it becomes hard for a candidate to survive and give their best to an organization.

6.2 CURRENT MARKET SOLUTION

There exist many platforms in the market who are already providing their service to the client in order to improve their work force [26] .

Existing online market tool which are providing the service to businesses in various ways.

We define online talent platforms based on data usage and functionality

	Digital tools that enable users to…	Example platforms, 2015
Matching individuals with traditional jobs	• Post full-time or part-time jobs • Create online resumes of individuals • Search for talent or work opportunities based on extended matching attributes • Provide transparency into company or worker reputations, skills, and other traits	Careerbuilder Glassdoor Indeed LinkedIn Monster Vault Viadeo Xing
Online marketplaces for contingent work	• Connect individuals with contingent or freelance projects or tasks • Facilitate transactions by providing transparency on reputation and ratings	Amazon Home Services Angie's List TaskRabbit Uber Upwork
Talent management	• Assess candidates' attributes, skills, or fit • Personalize onboarding, training, and talent management • Optimize team formation and internal matching • Determine the best options for training and skill development	Good.co PayScale Pymetrics beta ReviewSnap

Note: The landscape of providers and solutions is evolving rapidly. These examples reflect a snapshot as of May 2015.

SOURCE: McKinsey Global Institute analysis

Figure 3: Existing online talent platforms

The below AI driven talent platform has been assisting the enterprise with the features.

Metric	Description
Text kernel	ML (DL) for document understanding, Web Mining external sources, Synonyms,Software understands & searches unstructured data, Fuzzy text matching through OCR, Ontology Mining, Machine-learned ranking(MLR).
CVScan	Free service ,Scan CV and job description and compare keywords &frequencies & match rate , Includes top skills per industry (weighted).
Untapt	Talent-matching based on Natural Language (not keywords), Identify future leaders based on custom data analysis, White label solution or branded, AI-driven hiring decisions.
Google - Cloud Talent Solution	Talent Solution uses ML technology to better understand job content and jobseeker intent, Talent Solution can interpret the vagueness of any job description, job search query, or profile search query., includes military occupational specialty code translation (MOS, AFSC, NEC).
Zoho Recruit	A candidate's match score is calculated using their Skills and Qualifications, Contact the matched person through the platform, Semantic Search, Radius Search (location), Integrates with Linkedin, Parse CV, Large CV database
DaXtra	Offered as a component deployment or hosted service, Rich structured data output, Skills taxonomy extraction, Geographical and multilingual coverage, Social media awareness, Highly accurate, continually updated

6.3 GENERAL AND SPECIFIC AIMS

With the intervention of AI, recruitment process may be completely disrupted to a new future revolution. Our proposed solution is to create a personalized recruitment product for an organization based on its workforce strength with which we can provide the scoring of the candidate along with the feedback (acceptance/rejection) purely driven by AI. The product which could have the ability to think multi-dimensional having ability to take care of all the aspects between candidate and job post which can help complete the entire recruitment process with more efficiency, effectiveness, and ultimately fit between potential candidates and recruiters.

7 BUSINESS UNDERSTANDING

7.1 WHAT IS IMPORTANT IN A CV ?

As a first glance, the recruiter will be already evaluating the quality of a CV and its organization. Fact that could help or harm the overall result. Although, we won't take into account this aspect of the CV

and recruiter first encounter. Instead we're going to extract the content and study it.

So, for the content, the CV is a structured document that can be separated into several sections. A CV could have or not each of these parts depending on the experience, the type of applicant (researcher, private/public) and simply if whether or not they follow a standard structure. But whatever the CV received is, objectively, the recruiter will search for some specific parts in order to understand who is the applicant, and if he/she passes this first filter, a reading, understanding and making sense of the CV.

In order to understand the profile, an understanding of each section should be done. The different sections are: contact information, personal details, skills, professional experience, academic experience, projects, recognitions and awards, publications, certifications and references. And an overall review on spelling and grammar is important too.

7.2 REGULAR STRUCTURE ON A CV - METADATA

As mentioned before, the list of expected input is limited, each one has its own objective on helping defining the candidate's profile.

- Contact information: To contact the candidate.
- Personal details like birthday, nationality, social networks, blogs or github: In order to go further the CV if desired. (a further work can be web crawling to discover some traits of the applicant).
- Skills: Have an overview of the candidate's values and personal characteristics.
- Professional experience: Understand the relevance of the professional path regarding the offer and the enterprise.
- Academic experience: Extract the basis of the human capital and see if it's pertinent and use it as an indicator.
- Projects: Depending of the type, professional or academical, they tell about the experience or motivation
- Recognitions and awards: In order to differentiate from the others.
- Publications: In a research context, their role is to describe the potential of a candidate.
- Certifications: If necessary for the job, otherwise, a recognition
- References: Get further feedback of the candidate, speaks about candidate's values.

7.3 RECRUITER POINTS OF VIEWS

We met a recruiter of our school from the human resources department, who has been involved in the recruitment field in the engineering industry for more than 15 years, in order to have a better understanding on the recruitment process.

As an initial observation, she told us that this approach was her strategy and which is used most of the time in the recruiting field. They divide the work in two parts, extracting features from the required

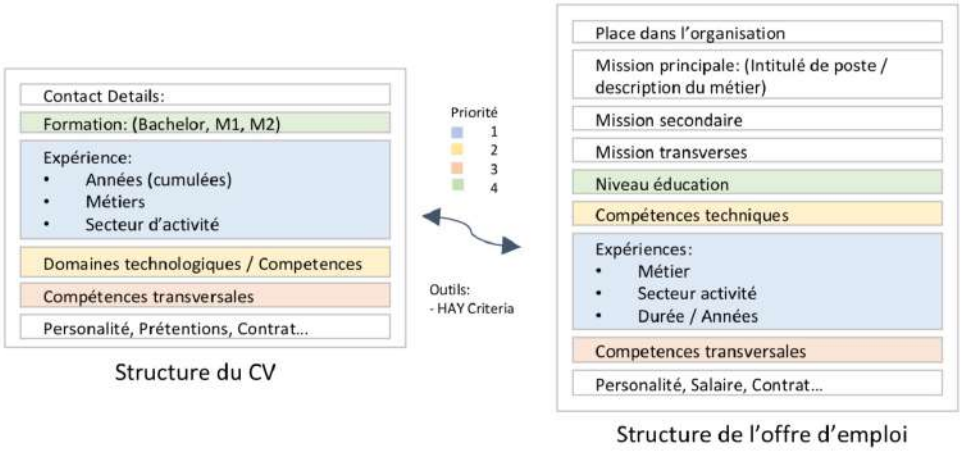


Figure 4: Basic match made by recruiter

job posts and extracting features from the CV. From this two As interesting points we find that recruiters point of view can vary, depending on the country and depending on the company. Some workers that are selected for a company, may not fit for another similar post in another company. Then, we observe a sign that the culture of a company is important to set parameters into the recruiting process.

7.4 RELATED WORK ON PARSING AND MATCHING

The domain of job matching has been researched since decades. AI has become talk of an hour by many researchers and business enterprises. The researchers are creating new algorithm in the field of talent acquisition which can help the business to find the best candidates without introducing any algorithm bias[23].

In the literature, the research is sparse and not a lot of specific domain studies have been done. For instance, finding "how to parse a CV and match/recommend/class it to a job posting" is not at all available. This type of study might have been interesting because a CV is a structured document where information of different categories could be extracted and analysed in parallel to extract information and get more accurate results for each sub-structure. A similar approach to our desired project is a recommendation model by implementing a genetic algorithm that uses recruitment records to establish the users demand model [23]. From the researcher perspective the matching problem has been tackled in different ways for eg Recommender systems are broadly accepted in various areas to suggest products, services, and information items to latent customers. Yi et al. used structured relevance models (SRM) to match résumés and jobs[23].Drigas et al. presented an expert system to match jobs and job seekers, and to recommend unemployed to the positions. The expert system used Neuro-Fuzzy rules to evaluate the matching between user profiles and job openings. they also proposed a fuzzy logic based expert system (FES) tool for online personnel recruitment. The system uses a fuzzy distance metric to rank candidates' profiles in the order of their eligibility for the job[23].

7.5 RECRUITERS CRITERIA TO MATCH THE CV TO A JOB POST

The primary steps involved for a recruiters before matching a CV to job post is understanding the job post. it is very essential for the recruiter to understand what they expect from the particular job post. In the process of which each job post is evaluated based on certain defined criteria and the candidates are accessed if they meet those criteria. the most popular defined criteria used by most of the recruiters

7.5.1 HAY CRITERIA

The HAY system is based on measuring the job against three elements which are deemed to be common in all jobs [9, 10]. These elements are:

- KNOW HOW - This measures the range of technical, planning, organising, controlling and communicating/influencing skills required in order to be able to perform the job competently.
- PROBLEM SOLVING - This measures the degree of complexity involved in carrying out the job.
- ACCOUNTABILITY - This measures the influence that the job has and the decisions made in achieving the end result.

Each job is measured against these three elements. A numeric score for each is calculated, using charts provided by HAY Management Consultants. The total of the three scores (job units) identifies the grade into which the job falls.

7.6 OPEN SOURCE KNOWLEDGE BASES

Many countries have made their data open source for the purpose of study on job openings, hires, and separations, providing an assessment of the availability of unfilled jobs, and information to help assess the presence or extent of labor shortages.

7.6.1 ROME

In France, ROME (Répertoire Opérationnel des Métiers et des Emplois) is a tool for professional mobility and the matching of offers and candidates. The ROME was built by the Pôle emploi teams with the contribution of a large network of partners (companies, branches and professional unions, AFPA...), based on a practical approach: inventory of the most common job titles/jobs, analysis of activities and skills, job grouping according to a principle of equivalence or proximity.

7.6.2 O-NET

The O*NET Program is the nation's primary source of occupational information. Valid data are essential to understanding the rapidly changing nature of work and how it impacts the workforce and U.S. economy. From this information, applications are developed to facilitate the development and maintenance of a skilled workforce [18].

7.6.3 ESCO

European Skills, Competences, Qualifications and Occupations (ESCO) is a multilingual classification of European Skills, Competences, Qualifications and Occupations. ESCO is part of the Europe 2020 strategy. The ESCO classification identifies and categorises skills, competences, qualifications and occupations relevant for the EU labour market and education and training. It systematically shows the relationships between the different concepts. ESCO has been developed in an open IT format, is available for use free of charge by everyone and can be accessed via the ESCO portal.

The table 1 is an example of referenced LaTeXelements.

Section	Detail	Comments
Personal Information	Name (Format Names can vary)	
	Picture	
	Address (Country, State)	
	Picture	
	Picture	

Table 1: Table to test captions and labels

7.7 STRATEGY/PLAN

7.7.1 SUMMARY GENERATED FROM CV'S

Each recruiter has to search in each part some highlights in order to get an overview of facts that would help him understand if the candidate's is adequate for the role, and then, some that would show characteristics that most save time to the recruiter.

7.7.2 FEEDBACK FROM CV'S TO THE CANDIDATE

We propose as a further step, to give feedback to the applicant about it's CV, like if the role he's searching for could be not very suitable for him, propose him some roles. Also, tell him where he is according to the job needs, if he should train on something else and how adequate he is in respect to similar job offers.

7.7.3 RECOMMENDATION

In order to do the scoring and matching we need to understand how we're going to do it. From some research and recruiters feedback, we have come with some metrics to extract from the CV. On thing to take into account is that for several metrics, it's existence is not certain so this fact must be taken into account. A "must have" note will be then proposed in order to mitigate the possible missing values that are not mandatory, but still, give them some importance to the fact that they exist if ever present.

7.7.4 STAGES, ALGORITHM FLOW

In order to create the final product, we want to follow a recruiter based evaluation logic in order to optimize the processes. This would permit the flow to ignore and class CVs. So, these two are the main stages to do the recommendation.

In the case of doing the whole process of recommendation we would already know what the recruiter is searching for, so we would be able to apply the HAY job evaluation criteria in order to offer two things: drop and score. Since the HAY criteria offers us a way to see the immediate relationship between two roles, and to understand how much the candidates experience is adequate for the job role the company is proposing.

As a first step, remove. We could search for the minimum requirements the recruiter is searching for, the "must have" ones in order to do a direct match with the job posting and drop candidates who don't have these minimal skills. Then, we would apply the HAY criteria in order to know how much related the job position is to the experience and roles the candidate has had. If we are not able to extract this information from the candidate we would return as feedback for him to add it to the CV and reapply. If the information is "blurry", we would simply not delete the candidate but assign a high score to the dropping criteria. Also, as a further step, we would require the recruiter feedback in order to improve this analysis when "blurry".

As a second step, classify. Once the non at all related profiles has been ignored, we could use the HAY evaluation done as a first input to the classifying algorithm. Then, we would use the metrics in order to do a classification among the candidates. For this, we would also apply the recruiters point point of view in order to give higher or smaller scores to the metrics results.

7.7.5 METRICS

Below are the different metrics which recruiters look for while matching the CV to a job post.

1. Professional Experience

- **Periods**

Metric	Description	Type
total_years_of_experience	from the first job to last job	number
experience_occupation _percentage	percentage of the time of experience that actually has been invested in working	number [0, 1]
experience_shifts_behavior	note describing the pause behavior between works: is it random? each time is one year? it has reduced over time?	number [-1, 1] (don't take into account if 0 or less)
experience_total_occupation _time_jobs_ratio	ratio of time per job	number [0, 1]
experience_gap_limit_ repetitions	count how many times the pauses between jobs were bigger than 9 months (tolerance + 5 days)	number

- **Company**

Metric	Description	Type
activity_sector	activity sector (civil engineering, computer science ...)	name
country	idem.	name

– **Activities**

Metric	Description	Type
experience_action_words _list	Keep list for job matching relevance evaluation (Created,received,deployed …)	action_list
experience_important_words _list	Keep important words list for further usage in job matching (Managed Optimised Reduced Developed Increased Supported Negotiated Presented Resolved Improved...)	important_list
experience_activities_skills_list	Deduce type of activities from overall skills: management, abstraction, scientific framework....	skill_list (all types)

– **Role**

Metric	Description	Type
experience_role_type	title of the role (would work to see the distance to the actual job position)	name
career_continuity	where the successive roles related?	mapping of career sectors

2. Academic Experience

Metric	Description	Type
academic_institution_title	Name	name
academic_institution_country	country	name
academic_experience_period	period	date
academic_experience_total	cumulated years	number
academic_degree	degree	number
academic_major	major	name
academic_grades	grades	number
academic_institution_score	score	number

3. Academic Projects

- **Type**

Metric	Description	Type
aca_project_types	in acadamical purpose?, professional? entrepreneurship?]	name

- **Subject**

Metric	Description	Type
aca_project_subjects	Distance measuring between role activities, and type can be done.	name

- **Period**

Metric	Description	Type
aca_project_duration_list	projects duration	number
aca_project_start _date_lists	projects start date	date

- **Name**

Metric	Description	Type
aca_project_name_list	[in acadamical purpose?, professional? entrepreneurship?]	name
aca_project_count	number of projects	number
aca_project_count_if_relevant	relevant projects count	number

– **Words and skills**

Metric	Description	Type
aca_skills_list	skills list obtained from the project if any	skill_list (all types)
aca_action_words	action words list	action_list

4. Personal Information

Metric	Description	Type
cand_name	name, last name	name
cand_picture	picture	blob
cand_linkedin	linkedin	name
cand_github	github	name
cand_facebook	facebook	name
cand_twitter	twitter	name
cand_blog	blog page / web page	name
cand_nationality	nationality	name

5. Contact Information

Metric	Description	Type
cand_mail	mail	name
cand_phone	phone	name

6. Grammar, spelling and congruence

Metric	Description	Type
grammar_mark	quantity of errors in phrases regarding total phrases words	number
spelling_mark	quantity of errors in phrases regarding total phrases words	number

7. Skills

Metric	Description	Type
soft_skills	soft skills	list
transversal_skills	transversal	list
language_skills	languages	list

8. Recognitions/awards

- **General**

Metric	Description	Type
award_type	type	name
award_year	year	number
award_name	name	name

- **Score (normalized)**

Metric	Description	Type
award_total	count all of them	number
award_freq	form start to beginning, divided by the total	number
award_behavior	has it been constant? is it increasing? is it decreasing?	[-1,1]

9. Publications

Metric	Description	Type
pub_type	type/subject	name
pub_year	year	number
pub_name	name	name
pub_magazine	magazine	name
pub_impact	impact [local, national, international]	name
pub_coworkers	coworkers	name

10. Certifications

Metric	Description	Type
cert_type	type (match if any certification needed	name
cert_name	name	name
cert_year	year	number
cert_date_validity	until year	number

11. References

Metric	Description	Type
ref_job	correspondent job	name/number
ref_tone	tone (negative, positive, can't say)	-1,0,1
ref_match_job	correspondence with activities	1,0
ref_skills	skills	list
ref_name	name	name
ref_contact_info	contact info	name

12. Candidate's summary

Metric	Description	Type
content	summary, very variate, not structured, for the moment, just identify it, and pass it as it is to the recruiter	name

7.8 ONTOLOGY

The ontology insist to build HR ontology which is composed of thirteen modular ontologies : Competence, Education, Job Offer, Job Seeker, Language, Occupation, Skill and Time. The main sub ontologies are the Job Offer and Job Seeker, which are intended to represent the structure of a job posting and CV respectively. While these two sub ontologies were built taking as a starting point some HR-XML [11] recommendations, the other sub ontologies were derived from the available international standards (like NACE, ISCO-88 (COM), FOET, etc.) and ES classifications and international codes (like ISO 3166, ISO 6392, etc.) that best fit the European requirements.

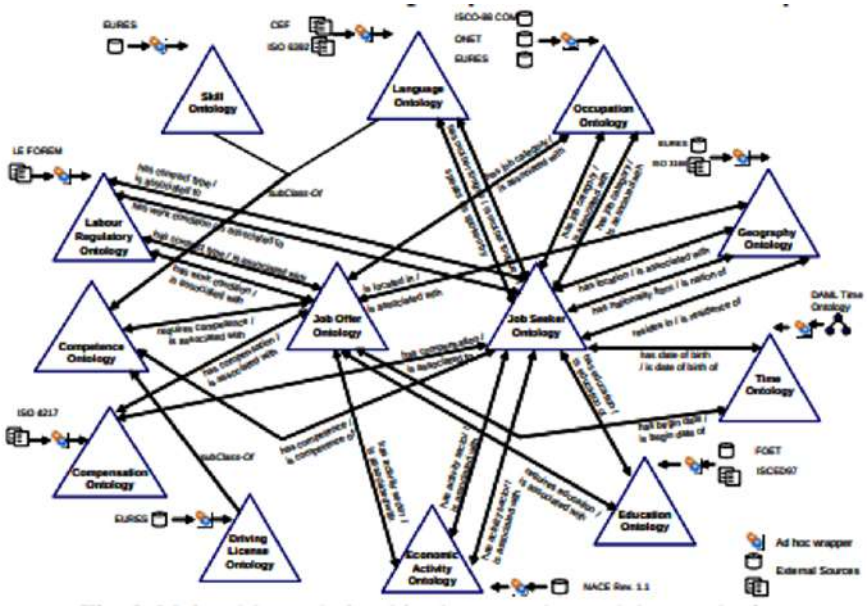


Figure 5: Main ad-hoc relationships between the modular ontologies.

details of the ontology is explained well in “Reusing Human Resources Management Standards for Employment Services” [8].

In the scope of our project we build a basic job skill ontology based on online available technology, to build an ontology we intent to build these ontologies using the reference provided in [13]. The flow chart shown below depicts how to build basics taxonomies which can be further converted into ontologies.

7.9 MODEL PROPOSITION (INPUT/OUTPUT)

7.9.1 NO DATA MODEL PROPOSAL

Approach to take: divide and conquer plus expertise dependence. Since the recruitment process is singular for each sector and even for each enterprise or recruiting framework, try to do a general approach

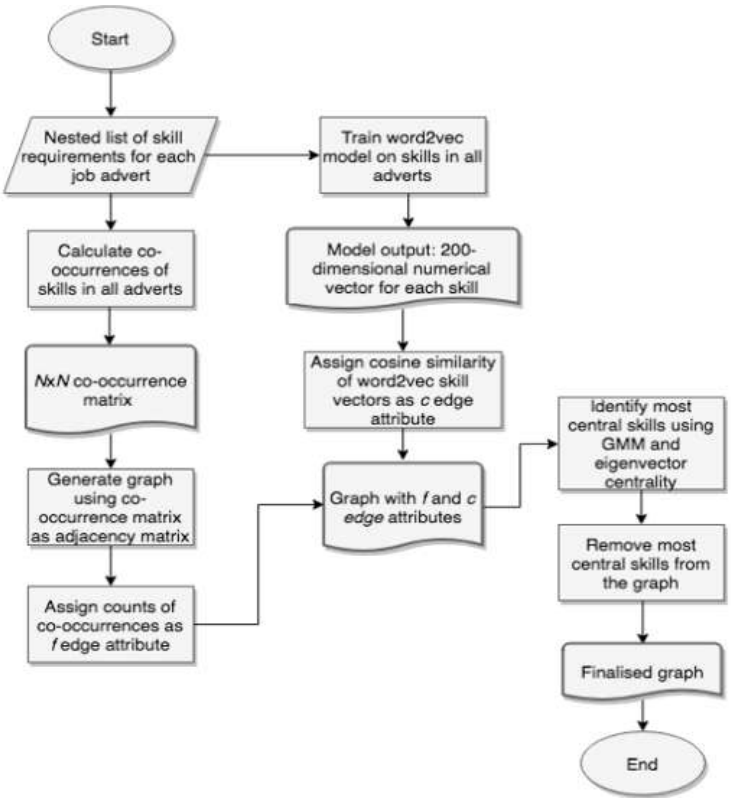


Figure 6: Flow chart for building and preparing the skills graph

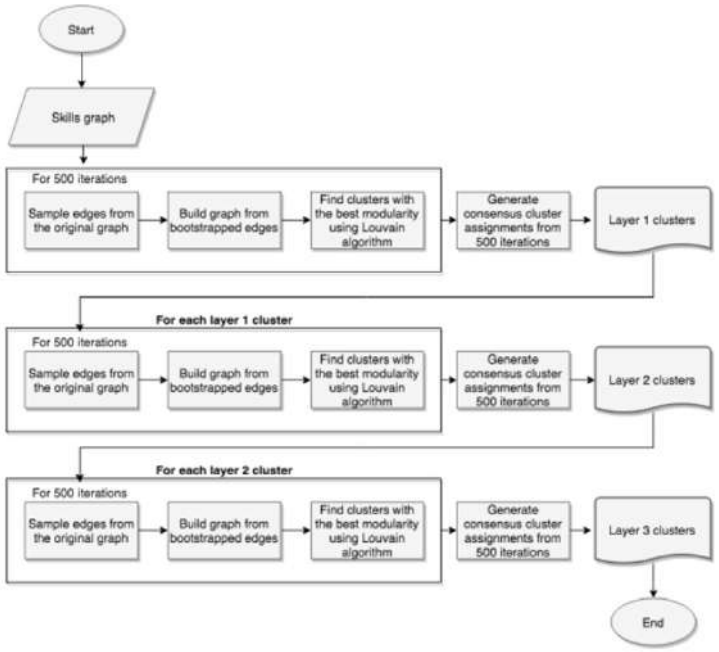


Figure 7: Flow chart for detecting hierarchical communities in the skills graph

remains exhaustive and even impossible to tune. An alternative to attack this issue is to let the recruiting area of the enterprise (our final customer) the freedom to tune some metrics that would be introduced to our final algorithm. This would permit a broader impact in the market since, behind the scenes, the algorithm would remain the same without need for us to adapt it trying to handle the biggest amount of study cases. Of course, this could lead to a "difficult product" so the tuning parameters should remain reduced. This would remain the more delicate "issue" for the customer's point of view since with a data based model, he would not need to do any tuning, so the objective would be to enlighten its possibilities and advantages.

In order to solve each small challenge, each of the metrics will be taken into account. A "very important", "important" and "not important" categories would be proposed for the main categories and maybe for each category if the recruiter needed it. This, he could adapt in order to comply with the needs of the role to be fulfilled, including all the cultural and organizational characteristics that the company must be seeking.

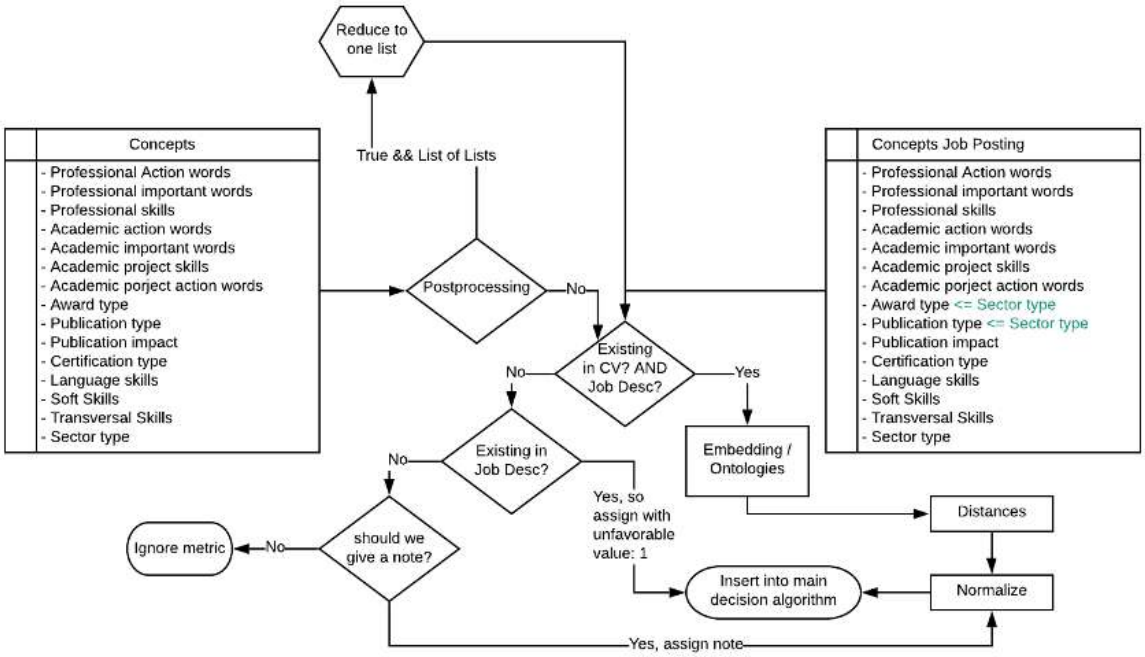


Figure 8: Example, flowchart for concepts treatment

As explained in the figure above, each type of metrics numbers, names, concepts would be treated with an algorithm to extract value that could be inserted into the "main algorithm" that would help to score.

First, a filter would be done so there would be a list that when found, extracted and matched less than required, the treatment of CV would stop by justifying the reason so that it still appear in the list of candidates' CVs but as "not qualified, reason: X".

Next, the extraction of these metrics would continue and contribute to feed the "main algorithm". The "main algorithm"model is not decided yet. The options regarding this proposal are: simply a decision rule according to the existence and the importance the recruiter gives to the different features and the mark given by the evaluation, so a simple formula. Another, would be to create a cost function in order to

satisfy the most the recruiter and a genetic algorithm to optimize. And finally, a multi criteria approach. The formula calculation could be the simplest one and maybe enough but we have to study more and to speak with experts in the field to understand what solution would fit.

7.9.2 DATA MODEL PROPOSAL

In the second approach we aim to build a proof of concepts. We use the help of the ontology's in order to do the match between CV's and job post. The key concept are identified from the text using text processing and is matched with the nodes of the ontology's in order to get detailed information related identified concept. In addition we can take any additional requirements from the organization which can enhance their work culture. For eg if the organization emphasize much on soft skill, creativity and other culture fit criteria needed with in.

The parser parse's the cv and get all the particular details from the CV's and job post. At the first layer the filter tries to reject the CV's based on basic requirements from the job post. The basic requirements can include years of experience, degree, contract type. At the second layer the model will exploit to match every section extensively like experience ,skills, soft skills, academic projects, motivation. for eg for the skill section ,it will generate a skill graph for both the job post and CV and measure the similarity between them. Further by using a tunable algorithm such as MR sort final score is obtained.

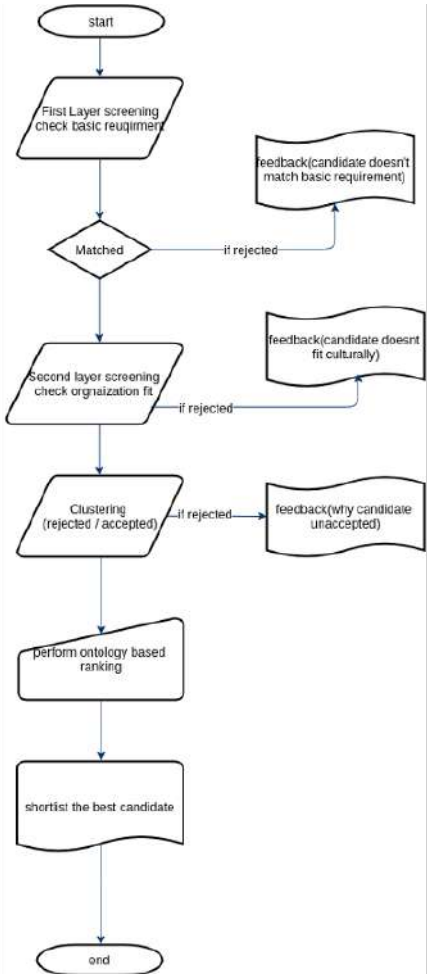


Figure 9: Example, flowchart for concepts treatment based on data

8 IMPLEMENTATION

For the implementation, some goals had to be decided leaving some ideas an processes out of this first iteration, that we propose as the minimal viable product. Functionalities like the parser, and more dimension evaluations had to be ignored in order to complete a first functional product. So, from the ***Model Propositions*** *done, we actually finished by doing a mix. Even if we didn't have any data, we managed to get existent ontologies and embedding, so we were able to work out functionalities related with the proposed model for data, and in other cases, a "no data" approach had to be taken by creating dependencies manually. Moreover, data was created by using an existing software that organized already all CV structured data so that we could exploit it directly. So, next chapters correspond to the development of what we call the "first iteration".*

8.1 CREATE ONTOLOGY'S

The structure of ontology's borrows a lot from graph theory, and For instance, when considering competencies, each competency is a 'node' and each relationship between competencies is an 'edge'. Ontology's are represented as undirected graphs .

In order to create an ontology for the skill development, Based on the research which we made on the online available ontology's. we chooses to work with CSO Ontology's. We also created manually the domain specific ontology from the crawled job posts.

8.1.1 TECHNICAL SKILL ONTOLOGY

The Computer Science Ontology (CSO) is a large-scale ontology of research areas that was automatically generated using the Klink-2 algorithm on the Rexplore dataset, which consists of about 16 million publications, mainly in the field of Computer Science [https://cso.kmi.open.ac.uk/home].The Klink-2 algorithm combines semantic technologies, machine learning, and knowledge from external sources to automatically generate a fully populated ontology of research areas.It also includes Linguistics, Geometry. The current version of CSO includes 26K topics and 226K semantic relationships.

It includes five semantic relations:

- relatedEquivalent, which indicates that two topics can be treated as equivalent for the purpose of exploring research data (e.g., Ontology Matching and Ontology Mapping). For the sake of avoiding technical jargon, in the CSO Portal this predicate is referred to as alternative label of
- skos:broaderGeneric, which indicates that a topic is a super-area of another one (e.g., Semantic Web is a super-area of Linked Data). This predicate is referred to as parent of in the portal. The inverse relation (child of) is instead implicit
- contributesTo, which indicates that the research output of one topic contributes to another. For instance, research in Ontology Engineering contributes to Semantic Web, but arguably Ontology Engineering is not a sub-area of Semantic Web – that is, there is plenty of research in Ontology Engineering outside the Semantic Web area.
- rdf:type, this relation is used to state that a resource is an instance of a class.For example, a resource in our ontology is an instance of topic.

– rdfs:label, this relation is used to provide a human-readable version of a resource's name.

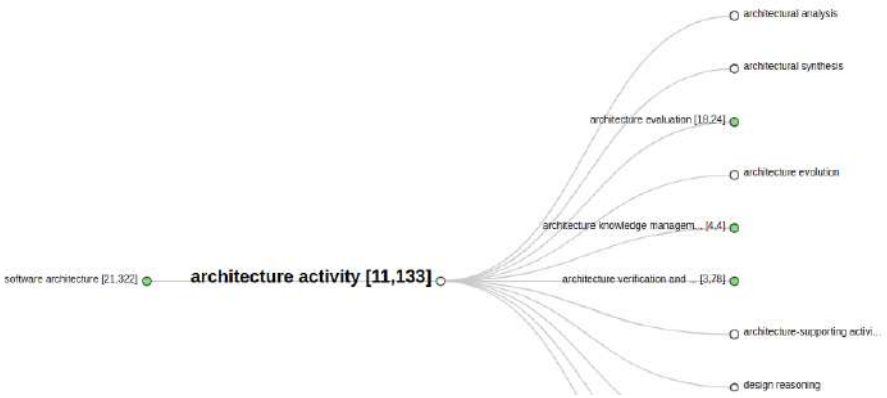


Figure 10: CSO ontology overview

8.1.2 CSO GENERATION

The working of Klink2 algorithm [7] takes as input a set of keywords and investigate their relationship with the set of their most co occurring keywords. The algorithm tries to find the semantic relationship between keyword x and y by the means of three metrics which are hierarchical relationship, temporal relationship and similarity . The first two used to to detect skos:broaderGeneric and contributesTo relationships, while the latter is used to infer relatedEquivalent relationships. The pseudo code of the klink algorithm used to generate the CSO is below.

```
Input  : List of keywords keywords, User feedbacks feedbacks
Output: Ontology CSO
relationships={};                          // Initialise an empty set
while some keywords yet to process do
    foreach k1 in keywords do
        candidates = GetCandidates(k1, feedbacks);
        foreach k2 in candidates do
            relationship = InferRelationship(k1, k2, feedback,
             relationships);
        end foreach
    end foreach
    relationships = RemoveLoops(relationships);
    new.keywords = MergeAndSplitKeywords(keywords, relationships);
    keywords = AddNewKeywords(new.keywords);
end while
keywords = FilterTopics(keywords, feedbacks, relationships);
CSO = GenerateSemanticRelationships(relationships);
return(CSO);
```

Figure 11: Data science doomain skill ontology

8.1.3 DOMAIN SKILL ONTOLOGY

In order to create a domain skill ontology, we collected the job post of the particular domain for e.g we focused to create a ontology in the domain of data science. As it helps to find the key term which exist related to the domain in such job post. For example the cso ontology lag the term such as algorithms or tools which are explicitly related to particular domain .Here we build a hierarchical based ontology where the nodes of the same type have some special semantics for defining parent/child relationships as this is a very common relationship necessary to express existing child and parent relationship frameworks. A node defined as a parent generally is a broader version of all of its children, having many shared attributes s. For instance, ESCO defines an 'advanced nurse practitioner' and 'specialist nurse' as both being children of 'nursing professionals'. These occupations understandably share many competencies, and it is easy to imagine experience in any type of nursing professional occupation as being broadly applicable to other nursing professional occupations. This parent/child hierarchy is necessary, but not itself sufficient for defining rich ontology's capable of expressing the relationships between other nodes or child

8.1.4 DOMAIN SKILL ONTOLOGY GENERATION

We created a large text corpus where we collected all the job post of data science skill domain. The job post was crawled from the Dice platform (https://www.dice.com/) . We collected 10000 job post. To generate the ontology's we employ machine learning methods, such as word embedding and clustering algorithms.

All the stop words existing in the job post corpus is removed and then we created a concept based on the number of occurrence of words in whole corpus based on n-gram technique. We used the library word cloud [1] which choose the top 200 originated concept. After creating the vectors using the nltk vectorize model, the cluster is being formed using the K-Mean Algorithm. After exploring the obtained cluster thoroughly we create a basic ontology using the protege software [2]. The ontology can be accessed using `http://owlgred.lumii.lv/online_visualization/lli4`. Below figure shows the brief overview of ontology.

8.1.5 CULTURAL VALUES ONTOLOGY

In order to understand the culture that a text could explain by itself from an enterprise point of view we approached theory that studies this as concepts that permit a classification of the enterprises on a cultural level. So to start with, to locate this problem, we're going to make some remarks: we're treating CVs and job postings, so general and domain specific terms would be found, and the context happens in an enterprise-like environment, an organization, whether public or private. These kind of documents, are not very explainable texts since they don't have a lot of complete phrases that would need a deeper sense analysis to get better results. This being said, we're going to explain how we managed to approach the extraction of what we call "organizational culture" from CVs and job postings.

[1] https:/github.com/amueller/word$_cloud$

[2] https://protege.stanford.edu

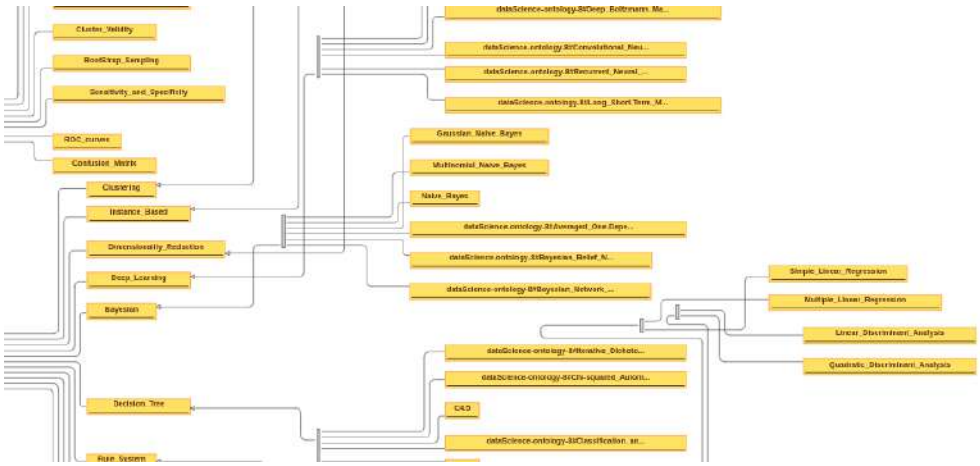


Figure 12: Data science domain skill ontology

To start with, we based our understanding in proposed organizational culture understanding theory [14] where several researchers propose a way of characterizing an organization. We use a global view of the characterization. As a first step, four levels can characterize culture in an organization {*Symbols*, *Heroes*, *Rituals*, *Values* } from which we're just interested in the values since they are only ones that can be extracted from text. The other levels imply abstraction or more inside-company behaviors and traditions that can't be forcefully extracted from CV or a job posting. As a second step, values have been distinguished in six different concepts { *Power Distance*, *Individualism*, *Uncertainty Avoidance*, *Masculinity & Femininity*, *Long Term Orientation*, *Indulgence Vs Restraint* } and each of these concept is subdivided in two "antonym" set of values describing its parent *(showed in the next table)*.

Organizational Culture Dimension	Concepts Comparison
Power Distance	Small <-> Large
Individualism	Individualism <-> Collectivism
Uncertainty Avoidance	Weak <-> Strong
Masculinity & Femininity	Masculinity <-> Femininity
Long Term Orientation	Short <-> Long
Indulgence Vs Restraint	Indulgence <-> Restraint

Table 2: Organizational Culture Dimensions

Each of this "antonym" concepts has other concepts that describe it (for more information see [14]). For example, for Power Distance, we have the next descriptors: *decentralization/centralization*, *management by experience/management by rules*, *autonomy of employee/order directed employee*, *pragmatic superior relationships/emotional superior relationships*, *no privileges/privileges*.

Thus, in order to make this information useful from a practical point of view, and create CVs and job posting profiles at the cultural level, we proposed a graph. The approach was to develop a directed graph that would handle a multiple level division of concepts until the very end where terms would describe concepts following the next logic: Culture -> Values -> Organizational Culture Dimensions -> Concepts Comparison ("antonyms") -> Concepts Definition Concepts (descriptors) -> terms (descriptors' terms).

So, a tree-like graph where each of these descriptors has terms that describe them, let's call them "descriptors' terms". These descriptors' terms were proposed by us, so improvements can be made. The way to assign the terms was to do a limited and definition directed list of terms referring to the parent node searching for definitions and extracting the most coherent and related terms.

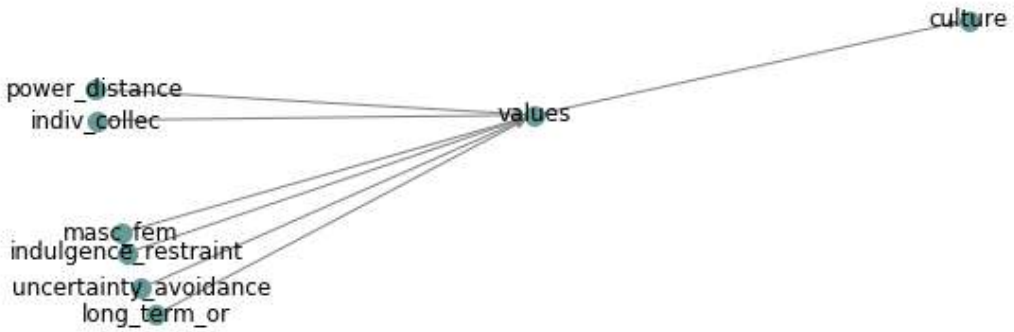


Figure 13: 3 First levels of culture graph

Furthermore, the idea to use a directed graph with no inter-related terms or concept resides in a practical decision where simplicity and the nature of the task intervenes. Since the task, "extract culture profile", means that the vocabulary to use is general and not domain specific, already trained models could be implemented. To this end, an embedded model has been chosen: the GloVe model [12]. This model was chosen against the word2vec model because of four reasons. The first one is because it was stated that somehow it maintains a better analogy. The second one, by doing some tests, the results given by simulating the application of terms (as it would happen with CVs and job postings) threw congruent and meaningful terms. And the third one, word2vec model was trained in a set of news texts (google News), so could be news context specific, and glove in wikipedia corpus, so a broader set of contexts. And fourth, because when searching for similarities, of terms against a set of words we want to be antonyms by less similar and for the word2vec model, the antonyms used to be more similar than in GloVe model. Still, this model could be changes, simply, it has to respect the *gensim [21] word2vec* standards.

GloVe	Word2Vec
'centralised', 'decentralized', 'hierarchical', 'decentralised', 'bureaucracy'	'decentralized', 'centralizing', 'centralize', 'Centralized', 'centrally_managed'

Table 3: Example of results similar to "centralized"

Finally, now that the structure of the graph and the why's of the structure have been answered, we can conclude by saying that the idea for this ontology is simply a term list (descriptors' terms) that represent concepts (descriptors). These concepts are sets of concepts (descriptors) that are part of a main

concept definition and the ensemble of this main concept definitions are somehow antonyms that belongs to wider concepts (organizational culture dimensions) which describe the values of the culture. At the end the part of the graph that will participate on the matching will be the leaves (descriptors' terms).

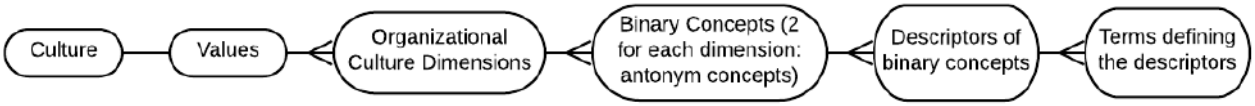


Figure 14: Structure of Cultural Graph

8.2 MATCHING

In order to do fair matching between the job post and CV'S . We create a graph with the help of the above created ontology's for both job post and CV for each section(general skill, domain skill and culture).

Similarity matrix is calculated between the each section graphs obtained from both the job post and CV. The obtained matrix is normalized to a matching score. The library used to measure the similarity between the two graphs are known as GMatch4py. GMatch4py follows the algorithm of graph edit distance by combining Hausdorff matching and greedy assignment [3]. After we receive the score from the different section such as General skill match, domain skill match and cultural match. we aggregate it to the common score using MR sort algorithm (explained in section 7.3.2)

8.2.1 CREATING SKILL GRAPH FROM ONTOLOGIES

In order to create a graph the algorithm takes the job description or the candidate work experience text as the input and outputs a list of relevant concepts from the job and CV's. For the Skill graph generation we followed the similar approach followed by CSO classifier [22].

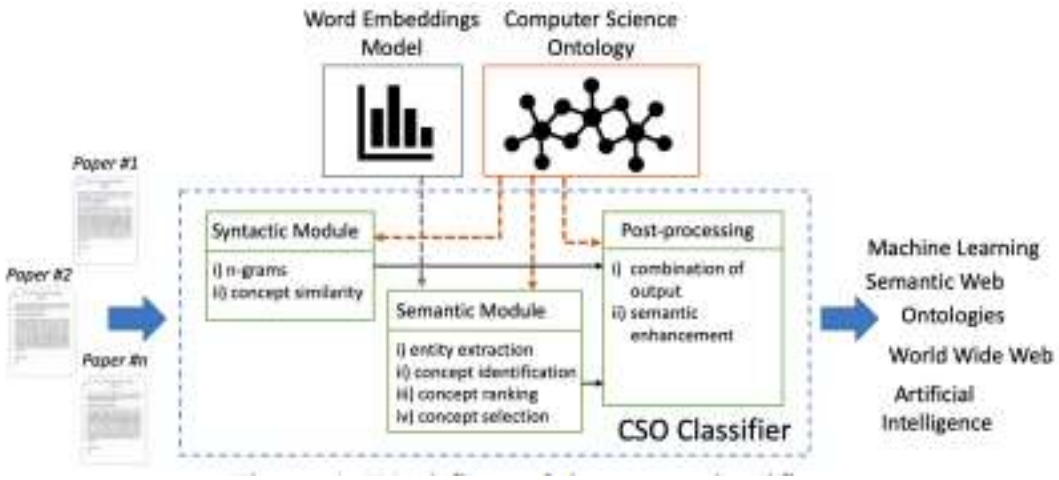


Figure 15: Workflow of CSO Classifier

It consist of two main components: (i) the syntactic module and (ii) the semantic module.

The syntactic module parses the input documents and identifies concept that are explicitly referred in the document. The semantic module uses part-of-speech tagging to identify promising terms and then exploits word embeddings to infer semantically related topics. Finally, the graph combines the results of these two modules and enhances them by including relevant super-areas.

Syntactic Module The syntactic module maps n-gram chunks in the text to concepts. The algorithm removes the stop words and collect the unigrams, bigrams and trigrams chunks. Then for each n-gram, it computes the levenshtein similarity with the labels of the topic in ontology's. the minimum similarity level can be set manually and it has been set to 0.94. This value allows us to recognize many variations between concept and ontology's.

Semantic Module The semantic module was designed to find topics that are semantically related to the text. These topics are explicitly not mentioned in the text. Here it require the word embeddings produced by word2vec to compute the semantic similarity between the terms in the text and the ontology's.

it follows four step .

- Entity extraction.
- Ontology concept identification.
- Concept ranking.
- Concept selection.
- Combined generated graph

The word embedding model was created by CSO using word2vec model. The model is trained on text collected from the technical research paper in the domain of computer science.

Entity extraction The concepts can be represented either by nouns or adjectives followed by nouns. The classifier tags each word according to its part of speech (e.g., nouns, verbs, adjectives, adverbs) and then applies a grammar-based chunk parser to identify chunks of words, expressed by the grammar.

Concept identification The extracted concept in the entity extraction stage are decomposes further into n-grams. Then similarity is measure between the n-grams with the ontology. The score with the top 10 similar word are identified as the concepts.

Concept ranking Since its possible that above step may develop lot of topics from the ontology with the help of n grams similarity to the nodes in which there may be the concepts which might not be related to the topic we are dealing with. It means many of the identified topic could be unrelated. So in order to choose the concept which are really important relevance score is calculated which is the product between the number of times it was identified (frequency) and the number of unique n-grams that led to it (diversity). If a concept is directly available in the ontology, its score is set to the maximum score.

Concept selection Once the relevance score is identified for all the generated topic. The topic are plotted distributionally and the elbow method is implemented in order to implement the top relevant topic which could be help full in order to do the matching. [6]

Combined generated graph The obtained topics from the both the semantic and syntactic modules are combined together. It then explored the topics by inferring all their direct super topics, exploiting the superTopicOf relationship with in ontology. For instance, when the classifier extracts the topic "machine learning", it will infer also "artificial intelligence". All the identified returned by both the module are stored in the dictionary . and it is further converted into graph with the help of networkx library.

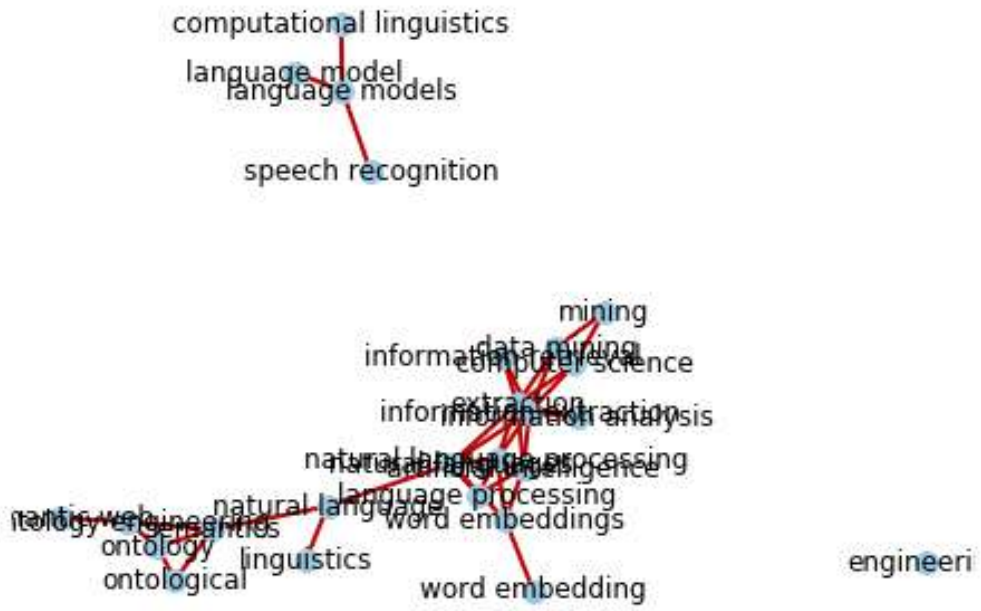


Figure 16: Generated Skill graph from job post/CV

8.2.2 CULTURAL MATCH

As a reminder, to describe or understand the cultural profile we have Organizational Culture Dimensions that have two antonym concepts which have several descriptors and each those have terms. So, in order to do the matching a profile of the CV or job posting is done with the help of the cultural graph, so each text will have a cultural graph profile. And then this profiles will be compared.

In order to better understand, we will explain the procedure by steps:

- Calculate the cosine similarities between the descriptors' terms and the text (this is done by using the embedded model)
  - *ex: decentralized organization = 0.62, centralized organization = 0.52*
- The mean of the similarities values of the descriptors' terms will be stored in each of the antonym concepts, this way, we'll know the belonging of each of the antonym concepts to the text, or, in other words, we will have the text profiled in the cultural graph.

  - *ex: small power distance = 0.86, large power distance = 0.56*

- So, we can obtain several cultural graphs that profile different CVs or job postings
- To compare an euclidean distance between both profiles would be measured.

This is how the similarity between a CV (or several CVs) and job posting can be done.

8.2.3 EDUCATION MATCH

For the education match between the job and CVs. The sovren parser parses the degree and the name of school from the job post and cv's. Lookup dictionary is created with all the equivalent degree related to particular degree. for eg MSc, master, BAC+5 belongs to same category. This match is done by keyword match in the dictionary where the required degree from the job post is searched in the degree obtained in candidate Cv.

if candidate required degree is inferior to the degree required in job post. The candidate gets rejected and is not processed for further stages.

8.2.4 REQUIRED SKILL MATCH

All the required skill parsed from the Job post are collected and is matched to the Skill/Concept found in the candidate skill graph. Based on the number of skill from the job post matched in obtained skill from candidate CV scored is assigned. For eg if out of 4 skill 3 skill is matched in candiate CV, then the calculated score is 0.75.

8.3 MULTIPLE CVS TO JOB POST MATCHING

One task is to match a CV with a job post, where the thing to do is to somehow compare different points of view of the texts such as culture, domain, education, and others. And another task to accomplish is to compare several CVs to a job posting, task that actually would be the the most common used by recruiters in order to understand the match of a candidate's profile with the job post a recruiter is promoting. In order to do this we will serve ourselves of two steps process: *filtering and matching*.

8.3.1 FILTERING

In order to accomplish this task, we will use the theory seen in HAY criteria, where there are some requirements that must be met and other that could be not mandatory.

As one of the proposals, the user would be able to "tune" the mandatory fields in order to implement by himself this filtering process. This could be done as a second iteration over the proposed solution since for now, the proposed solution has limited analysis axes and so, the development of this part would be excessive comparing to the actual functionality. For the moment, as a first approach, some filtering concepts have been imposed and CVs would be filtered taking them into account.

So, as per this iteration, simply the education axis will be taken into account. So, a comparison of required education level against the actual education level will me made and just the ones that meet or exceeds the requirements will be passed.

For this end, sovern software used helps a lot by telling us already the required skills that a job posting has.

8.3.2 SORTING

For the matching part, the help of an multi criteria algorithm is used in order to accomplish the task. When having several CVs, even hundreds, the help of a sorting algorithm may be very cherished. So, how to sort? As recruiters may have it's own idea in mind of the aspects they want to favor depending on the situation, the sector and the needs of the company, we propose a user friendly sorting by letting the recruiters to "tune" the parameters.

The parameters could be highly tuned in different aspects if the user needed it. Meaning that for each main axis of study, different sub parameters could be adjusted according to needs. For example, if there was a team that needed someone with a soft and friendly character because it's full of strong and difficult characters (true story), the recruiter should be able to tune this part of the sorting algorithm. But, mainly, the user should be able to tune five or maximum seven dimensions since, as per research and recommendation, those are the maximum quantity of dimensions that a person can handle. So, even if the tuning could be expanded, as a first proposal, seven is our maximum and the inputs will be directly asked to the user.

The sorting dimensions to take into account will be integers between zero and three included with the meanings: { 0-not interested, 1-poorly interested, 2-interested, 3-very interested }. In this way, the user can express his interest in each of the axis in a scale from zero to three. The algorithm used is the multi-criteria majority-rule sorting algorithm [20].

Algorithm 1: Multi-Criteria Majority-Rule Sorting algorithm adapted to our solution

input : values:[[CVs dimensions valuations]], weights, min_max (to normalize)

output: Sorted CVs according to input criteria and CVs evaluation against a job posting

handle not expected inputs

nomalized_weights $\leftarrow normalizeWeights(**)$

majority_rule_values $\leftarrow$ [ ]

for $cv\_scores \in values$ **do**

  majority_rule_value_for_cv $\leftarrow 0$

  for $(dimension, score) \in cv\_scores$ **do**

    min, max, norm_weight $\leftarrow getDimensionMinMaxNormalizedWeightValues(**)$

    normalized_score $\leftarrow norm\_weight * score$

    majority_rule_value_for_cv $\leftarrow majority\_rule\_value\_for\_cv + normalized\_score$

  $AppendToMajorityRuleValues(majority\_rule\_values, majority\_rule\_value\_for\_cv)$

$AddColumnToValues(values, title = "MRValues", data = majority\_rule\_values)$

return $SortValuesOfColumn(values, column = "MRValues")$

This way the user is able to tune the dimensions indicating which interest him the most, having $n^{n-1} * (n-1)$ different combinations to adjust, being n the number of dimensions to study. In our cases, we're proposing for the moment, and as part of the first iteration four dimensions skills, domain skills, culture, required skills .

9 EVALUATION

In order to evaluate our system, we have implemented two different use cases. In the first use case we test the function **ManyToOne matching** that allows to filter CVs and give an score to the selected ones. The second use case evaluates the function **OneToOne matching** that allows to see the level of correspondence between one CV and one Job post.

The CV's used to evaluate this tool were downloaded from the Naukri data base [17] using the following filters: "Search by Keywords: data", "Total Experience: 0 to 2 years" and "Candidate Age: 20 to 30 years". The job posts used were obtained from Linkedin. The search of this job posts were focus on a data science internship with 0 to 2 years of experience. Both, the CV's and Job post were parsed using the demo version of the Sovren parser tool [24]. We collected 120 CV's and 11 job posts. This data is stored in our Mongo database.

For this evaluation, we have selected 5 different CVs which include different profiles. We can define them as technical and business oriented profiles. We have selected also two different job posts, technical and a business oriented one aswell. The test dataset was the following:

ID	MongoID
CV1	5e60f5895a90883323e38bcc
CV2	5e60f58a5a90883323e38bdc
CV3	5e60f58a5a90883323e38bcf
CV4	5e60f58b5a90883323e38bf1
CV5	5e60f58a5a90883323e38bdb

ID	MongoID	Job Offer
BusinessOrientedJob	5e60f5895a90883323e38bcc	Intern Data Science - Product Analytics at Criteo
TechnicalOrientedJob	5e64cbef837ba015d90abc79	Intern Data Science at Multivac

9.1 USE CASE 1 - ONETOMANY MATCHING

9.1.1 RESULT FROM TEST 1: 5 CVS AND BUSINESS ORIENTED JOB POST

- **Input weights to define the priority of sections:** DomainSkillsMatch: 2, SkillsMatch: 2, CultureMatch: 2

ID	DomainSkillsMatch	SkillsMatch	CultureMatch	MRValues (Score)
CV1	1.428	1.420	0.944	0.743
CV2	1.415	1.418	0.913	0.739
CV3	1.420	1.417	0.889	0.736
CV4	1.427	1.431	0.845	0.730
CV5	1.414	1.419	0.828	0.728

- **Input weights to define the priority of sections:** DomainSkillsMatch: 3, SkillsMatch: 3, CultureMatch: 1

ID	DomainSkillsMatch	SkillsMatch	CultureMatch	MRValues (Score)
CV1	1.428	1.420	0.944	0.774
CV2	1.415	1.418	0.913	0.772
CV3	1.420	1.417	0.889	0.771
CV4	1.427	1.431	0.845	0.769
CV5	1.414	1.419	0.828	0.768

9.1.2 RESULT FROM TEST 2: 5 CVS AND TECHNICAL ORIENTED JOB POST

- **Input weights to define the priority of sections:** DomainSkillsMatch: 2, SkillsMatch: 2, CultureMatch: 2

ID	DomainSkillsMatch	SkillsMatch	CultureMatch	MRValues (Score)
CV3	1.418	1.432	0.940	0.743
CV4	1.427	1.426	0.928	0.741
CV5	1.414	1.420	0.924	0.740
CV2	1.414	1.417	0.891	0.736
CV1	1.429	1.435	0.849	0.731

- **Input weights to define the priority of sections:** DomainSkillsMatch: 3, SkillsMatch: 3, CultureMatch: 1

ID	DomainSkillsMatch	SkillsMatch	CultureMatch	MRValues (Score)
CV3	1.418	1.436	0.940	0.774
CV4	1.429	1.434	0.928	0.774
CV5	1.414	1.420	0.924	0.773
CV2	1.414	1.417	0.891	0.772
CV1	1.432	1.443	0.849	0.769

From these results we can observe how CV3 CV4 and CV5 are more likely to be selected for a technical job offer and CV1 and CV2 are more likely to be selected for a business oriented job.

We also observe that changing the weight to evaluate each section, is not changing the sorting of the CVs, however the MR score does change.

9.2 USE CASE 2 - ONETOONE MATCHING

Since CV1 and CV3 were the CV's with great score in both cases. Let's analyze why did they have this result.

9.2.1 RESULT FROM TEST 3: CV1 AND BUSINESS ORIENTED JOB POST

```
========= RESULT ==========

Culture match has been done with Word2Vec approach, so values mean how related the cultural terms are to CV ones.
* For definitions, please, look up on article
(<CV_to_Culture_similarity>, <Job_Posting_to_Culture_similarity>) <- <Cultural_aspect_evaluated>
(74.44, 75.8) <- small power distance
(63.99, 63.63) <- large power distance
(79.38, 79.33) <- collectivist culture
(76.22, 75.81) <- indivualist culture
(63.06, 64.09) <- feminine culture
(57.95, 57.16) <- masculine culture
(73.49, 74.92) <- weak uncertainty avoidance
(76.73, 77.04) <- strong uncertainty avoidance
(82.34, 82.37) <- long term orientation
(84.86, 80.7) <- short term orientation
(71.92, 73.71) <- indulgence
(62.83, 60.87) <- restraint

Overall Culture Matching: 0.9449211317955092 %
The candidate degree overqualified matches to the job profile
Domain_Skills Matching: 1.4224166673391714 %
OverAllSkills Matching: 1.42090911604134 %
Required Skill Matching: 0.0 %
execution stored
```

Figure 17: System output explaining the correlation between CV1 and a job post

9.2.2 RESULT FROM TEST 4: CV3 AND TECHNICAL ORIENTED JOB POST

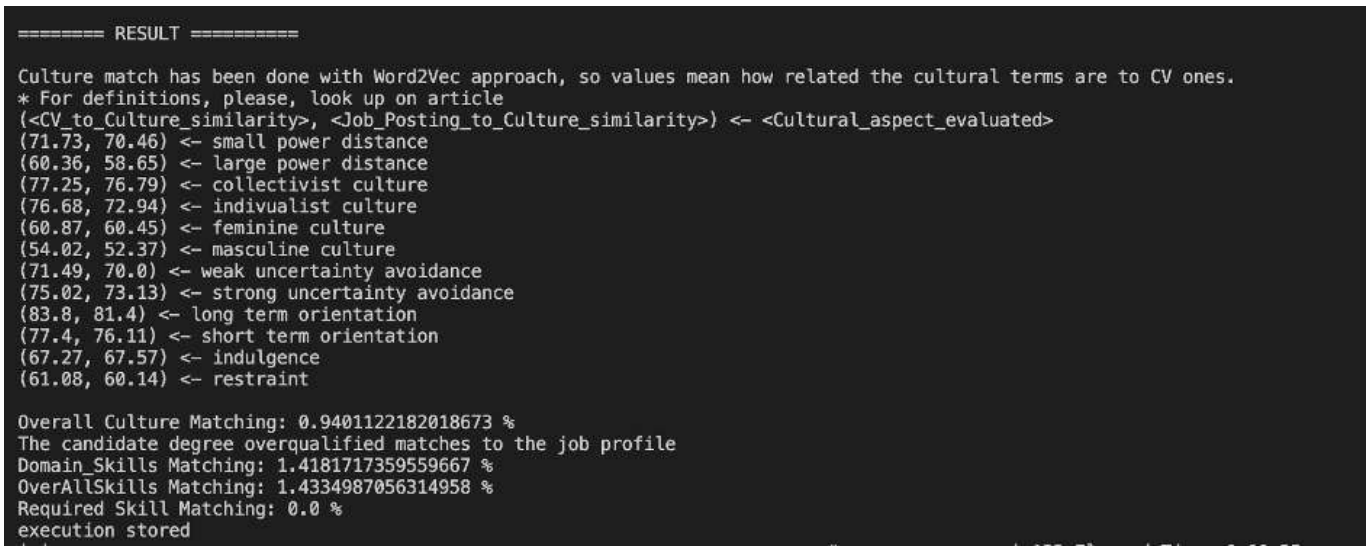


Figure 18: System output explaining the correlation between CV3 and a job post

10 CONCLUSIONS

About the process, a clear algorithm for recommendation could be implemented. Possible models for the CV parsing and recommendation processes are variate, not a complete approach can be found in research. We implemented a divide and conquer methodology for the model. We can approach each problem and solve each one with the best tool such as ontologies, embeddings, direct match, expert evaluation, machine learning. Develop an algorithm for the whole process according to the existence or not of data. About the product, it would reduce time in the recruiting process, save money, invest recruiters in more productive activities in order to increase retention and productivity of the team decrease recruiter's bias. Besides, it would encourage best candidates' fitting on the organization, which would increase the company value as a consequence.

11 FUTURE PROSPECTIVE

The algorithm clearly shows a winning result as per the given time frame to complete the project. However there are many improvement which are possible in order to improve the result. In the scope of this project we explored mainly the domain of skill graph matching. The other important domain which left unattended is the Job titles. Adding the dimension of job title in current skill ontology's can help to leverage deep match between candidate to the job positions. Moreover creating the word2vec model on the domain related data set will greatly help to explore closely the different concepts in the particular domain and to get the idea about their similarity and dissimilarity amongst them. In the scope of this project we did not consider the work experience of the candidate as we only focused to match candidate at a beginner level. However we wish to improve it further in order to adopt it for the experience professional. Most importantly in order to deal with the cultural prospective of an organization we lacked the data from the organization as we believed that CV's of candidate accepted within the organization best describe the organization cultural values. By training the model on such data we could have achieved better result in cultural prospective.

References

[1] Va. ALEXANDRIA. *Average Cost-per-Hire for Companies Is $4,129, SHRM Survey Finds*. accessed march 2020. 2016. URL: https://www.shrm.org/about-shrm/press-room/press-releases/pages/human-capital-benchmarking-report.aspx.

[2] S. Amdouni and W. Ben abdessalem Karaa. "Web-based recruiting: Framework for CV structuring". In: (2010), pp. 1–7. DOI: 10.1109/AICCSA.2010.5587018.

[3] AndreasFische. "Improved quadratic time approximation of graph edit distance by combining Hausdorff matching and greedy assignment". In: ().

[4] P Cappelli. "Your approach to hiring is all wrong". In: *Harvard Business Review* 97.3 (2019), pp. 48–58.

[5] Jennifer A. Chatman. "Matching People and Organizations: Selection and Socialization in Public Accounting Firm". In: *Administrative Science Quarterly* 36.3 (1991), pp. 459–84. URL: http://faculty.haas.berkeley.edu/chatman/papers/34_MatchingPeopleOrgs.pdf.

[6] "Finding a Kneedle in a Haystack: Detecting Knee Points in System Behavior". In: ().

[7] Enrico Motta KMi Francesco Osborne. "Klink-2: Integrating Multiple Web Sources to Generate Semantic Topic Networks". In: ().

[8] Asunción Gómez-pérez, Jaime Ramírez, and Boris Villazón-terrazas. "Reusing Human Resources Management Standards for Employment Services". In: (2007).

[9] *Hay Guide Chart*. accessed march 2020. URL: https://en.wikipedia.org/wiki/Hay_Guide_Chart.

[10] *Hay Job Evaluation Methodology: An Overview*. accessed march 2020. 2016. URL: https://peoplecentre.wordpress.com/2016/07/05/hay-job-evaluation-methodology-an-overview/.

[11] *HR-XML - Human Resources XML*. accessed march 2020. URL: https://schemas.liquid-technologies.com/HR-XML/2007-04-15/.

[12] Christopher D. Manning Jeffrey Pennington Richard Socher. "Global Vectors for Word Representation". In: *Computer Science Department, Stanford University, Stanford, CA 94305* (2014). URL: https://nlp.stanford.edu/projects/glove/.

[13] Cath Sleeman Jyldyz Djumalieva. "An open and data-driven taxonomy of skills extracted from online job adverts". In: (2018).

[14] Tianya LI. "Organizational Culutre Employee Behavior". Thesis. Lahti University of Applied Sciences, Degree programme in Business Information Technology, 2015.

[15] Martin Luenendonk. *What is Recruitment? Definition, Recruitment Process, Best Practices*. CLEVERISM. URL: https://www.cleverism.com/what-is-recruitment/.

[16] Kate Reilly Maria Ignatova. *https://business.linkedin.com/talent-solutions/blog/trends-and-research/2018/4-trends-shaping-the-future-of-hiring*. accessed march 2020. 2018. URL: https://business.linkedin.com/talent-solutions/blog/trends-and-research/2018/4-trends-shaping-the-future-of-hiring.

[17] *Naukri resume database*. https://freesearch.naukri.com. Accessed: 2019-02-18.

[18] *O*NET Resource Center*. accessed march 2020. URL: https://www.onetcenter.org/overview.html.

[19] Georgia Parmelee. *Georgia Tech Guest Post: Meet Your Newest Job Recruiter, the Algorithm*. (accessed march 2020). 2019. URL: https://www.americaninno.com/atlanta/from-the-community-atlanta/georgia-tech-guest-post-meet-your-newest-job-recruiter-the-algorithm/.

[20] Alexandru-Liviu Olteanu Patrick Meyera. "Handling imprecise and missing evaluations in multi-criteria majority-rule sorting". In: *Computers Operations Research* 110 (2019). DOI: https://doi.org/10.1016/j.cor.2019.05.027.

[21] Radim ?eh??ek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50. URL: https://radimrehurek.com/gensim/index.html.

[22] Angelo Salatino. "The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles". In: ().

[23] Tracy Hammond Shiqiang Guo Folami Alamudun. "RésuMatcher: A personalized résumé-job matching system". In: *Expert Systems with Applications* 60 (2016), pp. 169–182. DOI: https://doi.org/10.1016/j.eswa.2016.04.013.

[24] *Sovren Parsing Tool*. sovren.com. Accessed: 2019-02-18.

[25] *The talent challenge: Harnessing the power of human skills in the machine age*. pwc. URL: https://www.pwc.com/gx/en/ceo-survey/2017/deep-dives/ceo-survey-global-talent.pdf.

[26] Z. Wang, X. Tang, and D. Chen. "A Resume Recommendation Model for Online Recruitment". In: (2015), pp. 256–259. DOI: 10.1109/SKG.2015.31.