PROJECT DESIGN PHASE 1

PROPOSED SOLUTION:

DATE	1 OCT 2022
TEAM ID	PNT2022TMID47296
TITLE	Skill and Job Recommender Application
TEAM LEAD	JAYA LAKSHMAN M
TEAM MEMBERS	NAVEEN K SABARINATHAN S SANTHOSH P
MAXIMUM MARKS	2

Solution:

A. Data Collection

We automatically collected a set of job vacancies/offers from the Brazilian recruitment site called Cath and a set of Brazilian professional profiles from the well-known LinkedIn. We make available these datasets in a public repository with personal data anonymized. It is important to note that we collected more data from similar sites but, due to the validation issues, we only managed to work with these two sources in our framework. To perform job offers scraping, we created a list of keywords from the IT industry and used them as search terms. For each keyword, we search all the related job offers using Catho's search engine and save the retrieved results in our database; thus, the content's quality is highly related to the quality of Catho's search engine.

Additionally, the scraper is set up to avoid duplicate job offers, thus all the job offers are unique. On the other hand, to perform professional profiles scrapping, we created a list of areas of professional practice from the IT industry and, from

that, we search among the professional contacts of first, second, and third degree of our research group using Linkedin's search engine and save the retrieved results in our database; thus, all the professional profiles also are unique. We use text mining approaches to process both profiles and job offers data. Therefore, we selected the work experience, education, and competencies/skills from the profiles and, the description and title from the job offers. Finally, we concatenate these fields into a new one and discard the original fields, thus we end up with a document-like representation for each job offer and professional profile.

B. Data preparation :

Although we retrieved data from job search sites using only IT keywords, there were still some job offers that do not correspond to this field, then, the first step in this phase is filtering out job offers that do not belong to the IT field. To achieve this, we use a dictionary of weighted IT terms to match each job order in its document-like format. This way, we calculate the weighted sum of each word of the job offer in the dictionary and divide it by the appearances of the rest of the words in the document (job offer). Finally, we get a score with a value from 0 to 1, where a higher value indicates that the offer contains many relevant words on IT and it is very likely that corresponds to this field. Subsequently, we select only those job offers with a value of this score greater than 0.5. This setback only happens with the job offers since profiles were collected only into an IT professionals network. Once job offers and profiles are filtered, the second step is text pre processing. In this task, we perform stop words removal, tokenization, and lemmatization for the Portuguese language. The third step, feature representation, aims to represent these documents (job offers and profiles) as vector space models. For this purpose, we adopted two approaches: word embeddings and TF-IDF. Unlike the former, the latter technique does not require so much effort to be implemented From the variety of word embedding representations, we selected Word2Vec, which has different variants. We explore the two model architectures CBOW and Skip-Gram, and also the use of n-grams (bigrams and trigrams) to find the variation that best fit our problem. This way, we tested 5 different representations, TF-IDF, Word2Vec using CBOW, Word2Vec using Skip-Gram, Word2Vec using CBOW with n-grams, and Word2Vec using Skip-Gram with n-grams. For the Word2vec models, a vector space size of 200 was selected after some initial experimentation. For both word embedding and TF-IDF representation, we only used the corpus composed of the job offers. Although we lose some data, it was necessary since we realized that job seeker profiles added some noise because of the existence of professionals with very different background and skill set looking for a job in IT, which could foster spurious relations among skills. Finally, we transform both job offers and profiles into these 5 new representations and then proceed to use them in the recommendation

phase. In Table 1, we can see the description of the corpora used for our word embeddings