Skill Scanner

An AI-based Recommendation System for Employers, Job Seekers and Educational Institutions

https://doi.org/10.3991/ijac.v16i1.34779

Tim Schlippe✉, Koen Bothmer
IU International University of Applied Sciences, Germany
tim.schlippe@iu.org

Abstract—Skills are the common ground between employers, job seekers and educational institutions which can be analyzed with the help of artificial intelligence (AI), specifically natural language processing (NLP) techniques. In this paper we explore a state-of-the-art pipeline that extracts, vectorizes, clusters, and compares skills to provide recommendations for all three players—thereby bridging the gap between employers, job seekers and educational institutions. As companies hiring data scientists report that it is increasingly difficult to find a so-called "unicorn data scientist" [1], we conduct our experiments and analysis using companies' job postings for a data scientist position, job seekers' CVs for that position, and a curriculum from a master's program in data science. However, our investigated methods and our final recommendation system can be applied to other job positions as well. Our best system combines Sentence-BERT [2], UMAP [3], DBSCAN [4], and K-means clustering [5]. To also evaluate feedback from potential users, we conducted a survey, in which the majority of employers', job seekers' and educational institutions' representatives state that with the help of our automatic recommendations, processes related to skills are more effective, faster, fairer, more explainable, more autonomous and more supported.

Keywords—AI in education, recommender system, recommendation system, up-skilling, natural language processing

1 Introduction

Access to education is one of people's most important assets and ensuring inclusive and equitable quality education is goal 4 of United Nations' Sustainable Development Goals [6]. This goal should not only refer to general education in the private environment, but also to specific education in the professional environment. If people have the right education for the professional environment, they have a better chance to get jobs that allow them to have a good life. Unfortunately, there are often still gaps between the skills that are needed in the job market, the skills that job seekers have and the skills that are taught in educational institutions [7].

To solve this problem, all three players—employers, job seekers, and educational institutions—need to be aligned. Since today's natural language processing (NLP) methods are good at extracting information from text, there are already NLP approaches to extract either text data from job seekers' CVs, employers' job postings or educational institutions' learning curricula and give recommendations to one of these players. However, this way employers, job seekers and educational institutions usually use AI in isolation from one another. For example, [8] present a Word2Vec-based [9] recommendation system which informs employers how well job seekers' CVs fit job postings. LinkedIn has a system that recommends jobs to job seekers based on their CVs [10]. [11] investigate how AI-based recommendations help job seekers find study programs based on their personal profile. [12] use a combination of knowledge graph and BERT for finding suitable candidates in a corpus of CVs.

Connecting and supporting each of these three players—employers, job seekers[1], and educational institutions—will provide the most value as demonstrated in Figure 1: (1) Employers want to automatically check which of their required skills are covered by applicants' CVs (*Find and Select*) and know which courses their employees can take to acquire missing skills (*Upskill Workforce*). (2) Job seekers want to know which skills from job postings are missing in their CV (*Fit to Demand*), and which study programs they can take to acquire missing skills (*Find Program*). (3) In addition, educational institutions want to make sure that skills required in job postings are covered in their curricula (*Fit to Demand*) and they want to recommend study programs (*Advise*).

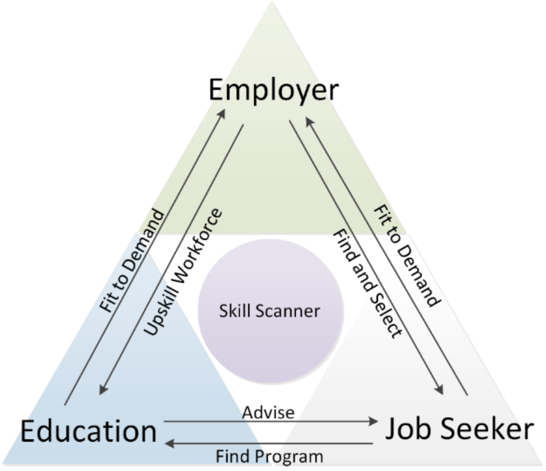


Fig. 1. Connecting and supporting employers, job seekers, and educational institutions.

Since skills are the common ground between these three players which can be analyzed with the help of AI, we investigate several NLP techniques to extract, vectorize,

[1] The term "job seeker" refers to current applicants and individuals who wish to advance towards a position.

cluster and compare skills. Then we combine the optimal methods in a pipeline which serves as the basis for our online service *Skill Scanner*[2] that outputs statistics and recommendations for all three players. Our goal was to help employers, job seekers and educational institutions adapt to the job market's needs. Consequently, we used job postings, which represent the job market's needs as reference. Our recommendation system determines which skills in the job postings are covered and which skills are missing. These representative skills, which we draw from a large set of job postings, are referred to as "*market skills*" in this paper.

As companies hiring data scientists find that it is difficult to find a so-called "unicorn data scientist" [1], we conducted our experiments and analysis using companies' job postings for a data scientist position, job seekers' CVs for that position, and a curriculum from a master's program in data science. However, our investigated methods and our final recommendation system can be applied to other job positions as well.

In the next section, we will present the latest approaches of recommendation systems for employers, job seekers, and educational institutions, as well as techniques for clustering and comparing text data. Section 3 describes our experiments to extract, vectorize, cluster and compare skills. We conclude our work in Section 4 and suggest further steps.

2 Related Work

Automatically ranking CVs is a valuable tool for employers. For example, [12] rank candidates for a job based on semantic matching of skills from LinkedIn profiles and skills from their job description, relying on a taxonomy of skills. Recent advancements in NLP offer opportunities to improve these methods: Particularly word embeddings, i.e. vector representations of words in a semantic vector space, are able to deal with similar words and synonyms since in this representation word embeddings of words with similar context are nearby in vector space. [8] use word embeddings from Word2Vec [13] to match CVs to jobs. [11] combine a knowledge graph and BERT for finding suitable candidates in a corpus of CVs: First, BERT is used for named entity recognition which labels the CV's keywords with classes. Then the knowledge graph is used to deal with the keywords' relationships, association, and aggregation.

Our best system also works with embeddings—however with sentence embeddings—to vectorize the skills. In addition, we use a cluster approach to find synonymous skills in job postings and CVs. The benefit of our clustering approach is that we do not rely on an additional language-dependent and manually created component like a knowledge graph.

Recommendation systems for job seekers have been investigated by [14,15,16]. As in the systems for employers, text data from social media profiles such as

[2] https://github.com/KoenBothmer/SkillScanner

LinkedIn or Facebook is usually processed [9,17]. Researchers at LinkedIn [18] have built a taxonomy of 35k standardized skills and use semantic matching to measure the similarity in job descriptions and job seekers' profiles.

Our technique is similar to [23], but instead of a taxonomy, we use a clustering approach. The benefit is that our model can pick up new skills without the need to update a taxonomy.

[10] give a systematic review of recent publications on course recommendation. Most related work focuses on recommending courses to potential students. They report a growing popularity of data mining techniques in those systems. To cope with the challenges of different levels of abstraction and synonyms in the course materials and students' documents, they first cluster the content, which they can then compare. K-means is usually used for clustering. To help employers recommend appropriate courses for their employees, [19] suggest a framework called "Demand-aware Collaborative Bayesian Variational Network (DCBVN)".

Compared to the related work, we propose courses for students and employees based on k-means clustering extended with additional steps to detect outliers in the clusters. While the job market is not considered in the recommendation process of other approaches, we use information from employers' job postings—denoted as *market skills* in this paper—as valuable information. We believe that this feature helps educational institutions design study programs that better prepare their students for the job market.

3 NLP to Extract, Vectorize, Cluster and Compare Skills

While in [20] we only briefly explained our NLP pipeline for extracting, vectorizing, clustering and comparing skills, in this paper we will further elaborate on the technical details. Our goal was to help employers, job seekers, and educational institutions adapt to the *market skills*. Our recommendation system *Skill Scanner* determines which skills in the job postings are covered and which skills are missing. In this section we will describe our NLP pipeline to process the skills.

3.1 A Pipeline to Extract, Vectorize, Cluster and Compare Skills

For a certain job position, (1) *Skill Scanner* takes a CV, a job posting or a learning curriculum as input, (2) extracts the skills of the provided document, (3) compares the document's extracted skills to a skill set which represents the job market's needs (*market skills*) and (4) returns information of which *market skills* are covered or missing in the provided document compared to the job market's needs [8].

To be able to compare the skills in the provided document to the *market skills*, we need to cope with the challenges of different levels of abstraction and synonyms among the skills in the uploaded document and the *market skills*. Consequently, we apply the following 4 steps when we gather the *market skills* and when we upload a document to be analyzed which are visualized in Figure 2:

1. *Retrieving skill sets*: Extract skill requirements.
2. *Pre-processing skill sets*: Map skill requirements to a semantic vector space.
3. *Removing outliers*: Remove outliers from skill requirements.
4. *Clustering skill sets*: Cluster skill requirements to cope with the challenges of different levels of abstraction and synonyms.
5. *Skill Scanner*: Provide a user interface, compute intersections among the skill sets of employers, job seekers, and educational institutions in relationship to the *market skills* and visualize recommendations.

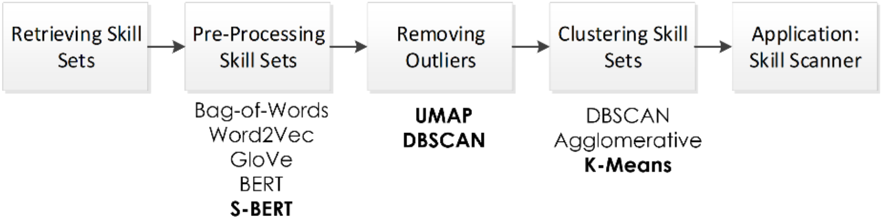
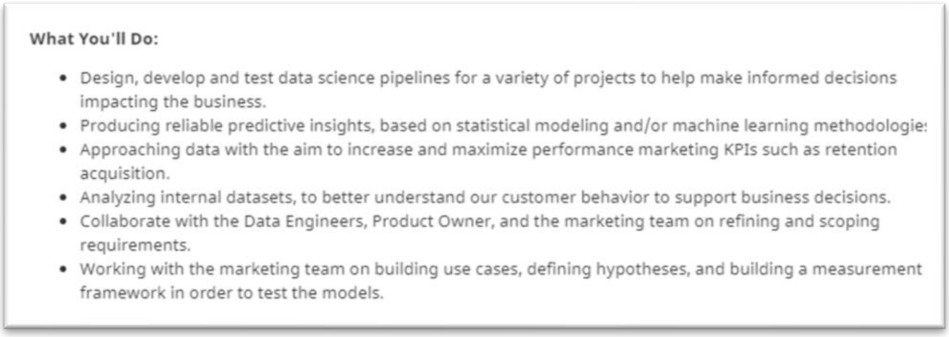


Fig. 2. Pipeline to Extract, Vectorize, Cluster, and Compare Skills.

3.2 Retrieving Skills

Whereas the job market is not considered in other approaches—especially not in the recommendation systems for educational institutions [7]—we use information from employers' job postings as valuable information. We believe that this feature helps educational institutions design study programs that better prepare their students for the job market. To retrieve job postings, we developed a web scraper based on the *BeautifulSoup* package [21].



What You'll Do:

- Design, develop and test data science pipelines for a variety of projects to help make informed decisions impacting the business.
- Producing reliable predictive insights, based on statistical modeling and/or machine learning methodologie
- Approaching data with the aim to increase and maximize performance marketing KPIs such as retention acquisition.
- Analyzing internal datasets, to better understand our customer behavior to support business decisions.
- Collaborate with the Data Engineers, Product Owner, and the marketing team on refining and scoping requirements.
- Working with the marketing team on building use cases, defining hypotheses, and building a measurement framework in order to test the models.

Fig 3. Required skills in a job posting of a data scientist position.

Fig. 3 demonstrates the tendency of employers to express their skill requirements in bullet points. We use this tendency to extract skills from job postings by first extracting all bullet points. Since at that step we naively appoint all bullet points as skill requirements, we analyzed methods to deal with outliers that are not skill requirements as described in Section 3.4.

With our web scraper, we retrieved 21.5k bullet points in 2,633 job postings to build a set of representative skills for the position of a data scientist in English from Indeed.com and Kaggle.com. In this work, we refer to this representative set of skills as the *market skills*.

3.3 Vectorizing Skills

To compute distances between skills we mapped the skills to a semantic vector space. Like [8,11], we experimented with word embeddings to vectorize the skills. To represent the skills which usually consist of several words, we investigated stacking and averaging the word embeddings in a skill after they were produced with Word2Vec [13] and GloVe [22]. In addition, we explored sentence embeddings. As Bidirectional Encoder Representations from Transformers (BERT) [24] models are successful in NLP tasks, we also experimented with Sentence-BERT [2], a modification of the pre-trained BERT transformers.

Sentence-BERT (44.2%) outperformed word embedding like GloVe (39.5%) by 12% in Silhouette score [28] at the end of our pipeline.

3.4 Removing Outliers

To remove outliers in the vectorized skills and allow our clustering techniques to perform better, we reduce the dimensionality of the semantic vector space created by Sentence-BERT. For that we experimented with combinations of PCA [25], UMAP [3], and DBSCAN [4]. Using (1) UMAP to reduce the vectorized skills to two dimensions and (2) DBSCAN to remove outliers in the 2-dimensional space gave the best results.

As shown in Fig. 4, DBSCAN finds clusters in dense areas and assigns a cluster ID to all points belonging to the same cluster. Outliers—i.e. in our case clusters which do not represent skills or are not part of a dense area—are assigned the cluster ID -1. Let’s take a closer look at the two dots highlighted in the figure: The blue dot which belongs to cluster 0 represents “Strong SQL development skills”. The purple dot is one of the outliers which is distributed in the space and does not belong to a dense area. Therefore, it is categorized at cluster -1. It represents COVID-19 information and is no skill required on the job market. Sometimes, DBSCAN returns positive cluster IDs for dense areas containing only outliers. In those cases, we discarded them manually. Please refer to the interactive version of Fig. 4 to gain insight in the content of each cluster.[3]

[3] https://storage.googleapis.com/public-hosting-paper/skill_model/umap_dbscan.html

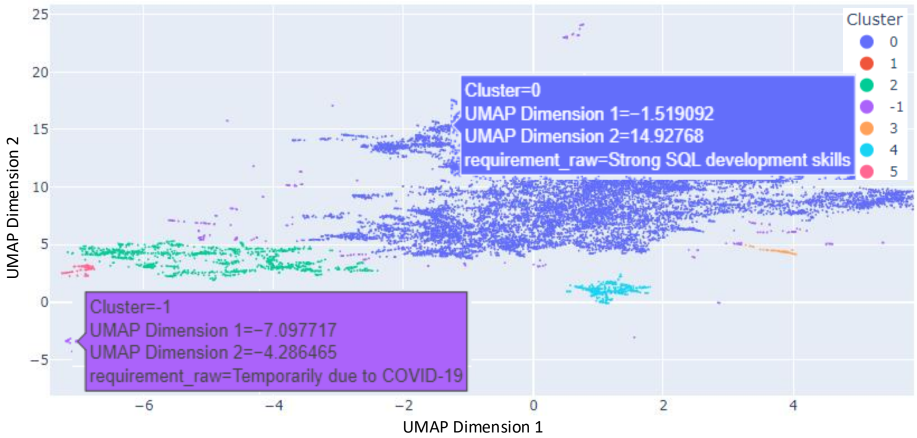


Fig 4. DBSCAN clustering of skills vectorized by Sentence-BERT and dim-reduced by UMAP.

With those steps we reduced the 21.5k potential skills retrieved with our web scraper to 18.8k skills.

However, since the 2D vectors did not contain enough information for further analysis of the skill dataset, we applied another clustering to the original 768-dimensional vectors that remained after removing outliers in the 2-dimensional space.

3.5 Clustering Skills

Similar to [19], we use a clustering approach to find synonymous skills in job postings, CVs, and learning curricula. But instead of a taxonomy, we use a clustering approach. The benefit is that our model can pick up new skills without the need to update a taxonomy. While hierarchical clustering approaches have not proven to be robust against outliers [26], K-means clustering has been successfully used in clustering word embeddings [27] and is adaptable and scalable [5]. Consequently, we used K-means to cluster our 768-dimensional vectors with the cosine distance as the distance metric. K was chosen as 31 with the highest Silhouette score of 44%.

Fig. 5 shows the resulting skill clusters of the data scientist position marked in different colors after applying K-means. Please refer to the interactive version of Fig. 5 to gain insight in the content of this cluster and the rest of the model.[4]

[4] https://storage.googleapis.com/public-hosting-paper/skill_model/cluster_k31.html

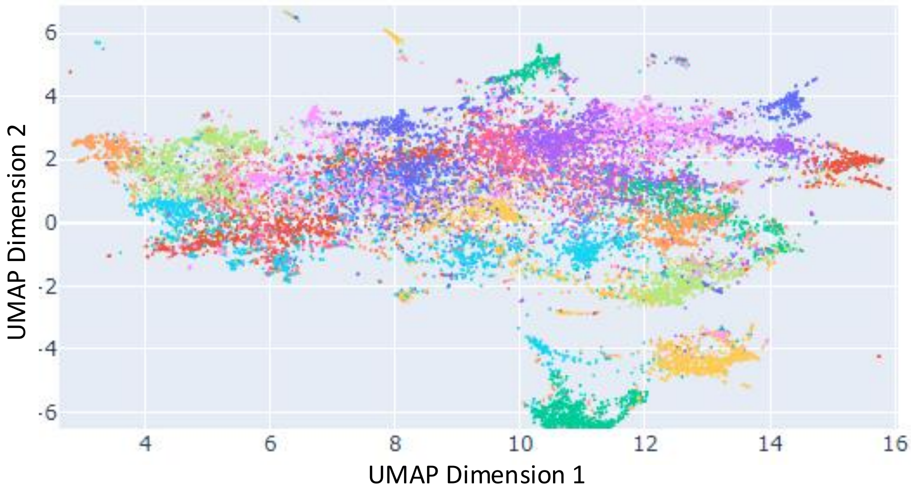


Fig. 5. UMAP-projection of skills vectorized by Sentence-BERT and clustered by K-means.

3.6 Skill Scanner: Comparison and Analysis

After retrieving clusters and vectors representing the skill of each cluster, we perform mathematical operations to find covered and missing *market skills*. For example, Figure 6 shows a section of a report in *Skill Scanner* where overlaps between skills in job postings and learning curricula were calculated. There, using the visualizations, educational institutions are shown the importance of certain skills in data scientist profiles along with the lack of coverage of those skills in their curricula.

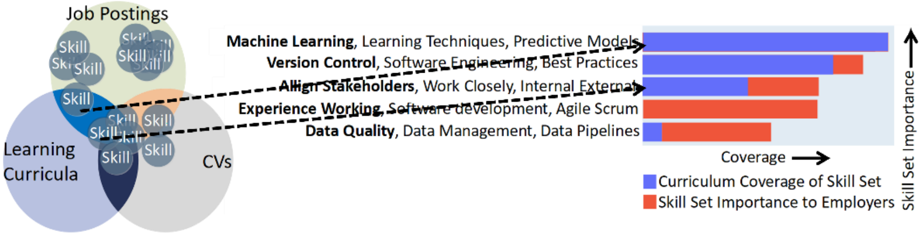

Job Postings
Skill
Learning Curricula
CVs
Machine Learning, Learning Techniques, Predictive Models
Version Control, Software Engineering, Best Practices
Allign Stakeholders, Work Closely, Internal External
Experience Working, Software development, Agile Scrum
Data Quality, Data Management, Data Pipelines
Coverage
Skill Set Importance
Curriculum Coverage of Skill Set
Skill Set Importance to Employers

Fig. 6. Skill Comparison and Analysis between *Market Skills* and Learning Curricula.

In case of a single provided job posting, the bars show exclusively the skills specified in this provided job posting. Otherwise, the bars show all *market skills*. In both cases the skills are sorted by importance in the *market skills* which is represented by the bar lengths. Each skill is described by the 3 most frequent bigrams used in the *market skills*. The blue part in each bar demonstrates how well the skills in the provided learning curriculum match, whereas the red part shows how much is missing indicating the room for improvement. This representation of the skills with the bigrams, the coverage and the room for improvement is used consistently in all reports.

Technically, the bigrams at each bar on the y-axis are the most common bigrams that are located in a *market skill* cluster gained in step 4 *Clustering skill sets* of *Skill Scanner*'s pipeline (see Figure 2). How well a skill in the provided curriculum matches a skill in the job posting or in the *market skills* is determined by the distance of the skill's vector specified in the curriculum to the centroid of the cluster.

Reports as in Fig. 6 were shown to 108 representatives of our 3 players in a survey [29]. The majority finds that with our system, processes related to skills are more effective, faster, fairer, more explainable, more autonomous and more supported. 89% of all participants are not averse to apply our recommendation system. 67% of job seekers would certainly use it.

4 Conclusion and Future Work

In this paper we have presented a state-of-the-art pipeline that extracts, vectorizes, clusters, and compares skills to provide recommendations for employers, job seekers and educational institutions. With the help of NLP techniques our system processes skills which are the common ground between all 3 players. The job market dictates what job seekers should learn, and educational institutions should teach. Therefore, our system processes skills in job postings, CVs, and learning curricula and outputs recommendations for employers, job seekers, and educational institutions based on present and missing skills and their importance to employers. With our clustering approach we do not have to update a taxonomy as skill requirements change. We provided detailed explanations of our clustering techniques together with screenshots and online access to our interactive scatter plots.

Future work may be to apply our pipeline to other job positions and expand it to other domains. Furthermore, as we used the pre-trained Sentence-BERT it may be analyzed if a fine-tuned Sentence-BERT leads to further improvement.

5 References

[1] Baškarada, S., Koronios, A.: Unicorn Data Scientist: The Rarest of Breeds, Program: Electronic Library and Information Systems, Vol. 51 No. 1, pp. 65–74. (2017)

[2] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, EMNLP-IJCNLP (2019)

[3] McInnes, L., Healy J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv, abs/1802.03426 (2018)

[4] Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD. AAAI Press, 226–231 (1996)

[5] Lloyd, S.P.: Least Squares Quantization in PCM. Techn. Report RR-5497, Bell Lab (1957)

[6] United Nations: Sustainable Development Goals: 17 Goals to Transform our World (2021), https://www.un.org/sustainabledevelopment/sustainable-development-goals

[7] Palmer, R.: Jobs and Skills Mismatch in the Informal Economy. 978-92-2-131613-8 (2017)

[8] Fernández-Reyes, F.C., Shinde, S.: CV Retrieval System Based on Job Description Matching Using Hybrid Word Embeddings, Computer Speech & Language, vol 56 (2019)

[9] Geyik, S.C., Guo, Q., Hu, B., Ozcaglar, C., Thakkar, K., Wu, X., Kenthapadi, K.: Talent Search and Recommendation Systems at LinkedIn: Practical Challenges and Lessons Learned. SIGIR (2018)

[10] Guruge, D.B., Kadel, R., Halder, S.J.: The State of the Art in Methodologies of Course Recommender Systems—A Review of Recent Research Data, 6(2), 18 (2021)

[11] Wang, Y., Allouache, Y., Joubert, C.: Analysing CV Corpus for Finding Suitable Candidates using Knowledge Graph and BERT. DBKDA (2021)

[12] Faliagka, E., Iliadis, L. Karydis, I., Rigou, M., Sioutas, S., Tsakalidis, A. Tzimas, G.: On-line Consistent Ranking on E-Recruitment: Seeking the Truth Behind a Well-Formed CV. Artif Intell Rev 42, pp. 515–528 (2014)

[13] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. ICLR (Workshop Poster) (2013)

[14] Siting, Z., Wenxing, H., Ning, Z., Fan, Yang: Job Recommender Systems: A Survey. ICCSE (2012)

[15] Hong, W., Zheng, S., Wang, H., & Shi, J. (2013). A Job Recommender System Based on User Clustering. J. Comput., 8, 1960-1967.

[16] Alotaibi, S: A Survey of Job Recommender Systems. Int. J. Phys. Sci. (2012)

[17] Diaby, M., Viennet, E., Launay, T.: Toward the Next Generation of Recruitment Tools: An Online Social Network-Based Job Recommender System. ASONAM (2013)

[18] Li, J., Arya, D., Ha-Thuc, V., Sinha, S.: How to Get Them a Dream Job? Entity-Aware Features for Personalized Job Search Ranking. SIGKDD (2016)

[19] Wang, C., Zhu, H., Wang, P., Zhu, C., Zhang, X., Chen, E., Xiong, H.: Personalized and Explainable Employee Training Course Recommendations: A Bayesian Variational Approach. ACM Trans. Inf. Syst. (2021)

[20] Bothmer, K., Schlippe, T.: Investigating Natural Language Processing Techniques for a Recommendation System to Support Employers, Job Seekers and Educational Institutions. The 23rd International Conference on Artificial Intelligence in Education (AIED) (2022).

[21] Hajba, G.L.: Using Beautiful Soup. In: Website Scraping with Python. Apress (2018)

[22] Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. EMNLP (2014)

[23] Le, Q., Mikolov, T.: 2014. Distributed Representations of Sentences and Documents. ICML.

[24] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019)

[25] Pearson, K.: On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine. 2 (11): 559–572 (1901)

[26] Rani, Y., Rohil, H.: A Study of Hierarchical Clustering Algorithm. International Journal of Information and Computation Technology (Vol. 3, Issue 10) (2013)

[27] Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G.; Does Deep Learning Help Topic Extraction? A Kernel K-Means Clustering Method with Word Embedding. Journal of Informetrics, 12 (4), 1099–1117 (2018)

[28] Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. 20: 53–65. (1987)

[29] Bothmer, K., Schlippe, T.: Skill Scanner: Connecting and Supporting Employers, Job Seekers and Educational Institutions with an AI-based Recommendation System. The Learning Ideas Conference 2022 (15th annual conference), New York, New York (2022)

6 Authors

Prof. Dr. Tim Schlippe is a professor of Artificial Intelligence at IU International University of Applied Sciences and CEO of the company Silicon Surfer. (email: tim.schlippe@iu.org)

Koen Bothmer is a data scientist who works as a digital R&D specialist at FrieslandCampina. (email: koenbothmer@gmail.com).