EPFL

CHILI
computer-human interaction
in learning and instruction

LEARN
Center for Learning Sciences

Spring Semester 2021

Research Project: /

Skills extraction from job ads

Nicolas Kieffer

Supervisor: Richard Davis

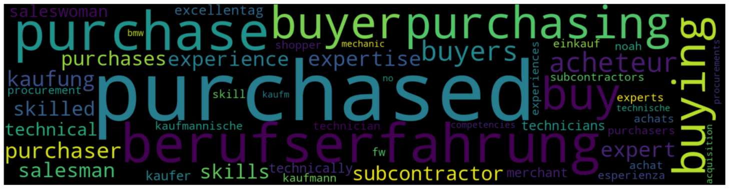
MOTIVATION

Imagine a recruitment platform where you could simply enter your skills and it would propose you all the jobs corresponding to these skills. To this purpose, the platform would have first to extract skills found in each job advertisement. This semester project aims at developing, validating, and comparing methods for extracting skills from job-ad data. Supervised methods are mostly used for related work with large amounts of labeled data. But in our case, we dispose of unlabeled data and the time and money spent on labeling it would be too much for too less. We need unsupervised methods that could be capable of extract skills from any new dataset.

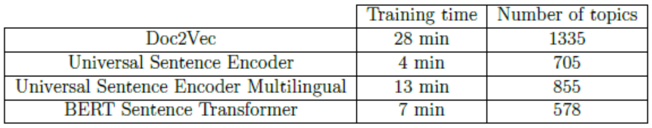
METHODS

Different methods were tried and compared to each other in order to determine which one is the most efficient and accurate. These methods are LDA with Gensim, Top2Vec using different embedding models, LDA2Vec and Contextual Topic Identification. Finally and concretely, 6 methods have been compared: LDA with 500 topics, LDA with 1000 topics, Top2Vec with Doc2Vec, Top2Vec with Universal Sentence Encoder, Top2Vec with Universal Sentence Encoder Multilingual and Top2Vec with BERT Sentence Transformer. To compare these methods, skills have been manually extracted from 10 random job ads and a F1-score has been calculated to compare accuracy between the methods by comparing the ground truth to the terms topics the different models returned. Some wordclouds of topics obtained with Top2Vec can be seen below.

logisticer
planning
punctuality
accountancy
forecasting
forecasted
statistics
timeline
schedule
gestione
prioritization
logistical
logistik
deadline
feasibility
logistic
logistique
termining
logistics
inventory
aufgaben
gestion
accounting
prioritisation
traceability
anticipating
deadlines
expectations
forecasts
planung
planned
timelines
ability
logisticsian
logistiker
logistician
logist
tasks
statistical
logisticien
predictive
managerial
scheduling
forecast
contingency
relational
schedules
distribution
logisticians
timeliness



RESULTS



	Training time	Number of topics
Doc2Vec	28 min	1335
Universal Sentence Encoder	4 min	705
Universal Sentence Encoder Multilingual	13 min	855
BERT Sentence Transformer	7 min	578

F1-score of different models:

- LDA with 500 topics: 0.138
- LDA with 1000 topics: 0.131
- Top2Vec with Doc2Vec: 0.36
- Top2Vec with USE: 0.218
- Top2Vec with USEM: 0.298
- Top2Vec with BERT: 0.313

Link to the code repo and report: (insert link here)