# Cyber Security Job Vacancies Text Analysis Report 1

**Student**: Elshan Gadimov **Supervisors**: Lendák Imre, Tsegaye Misikir

[1] Eötvös Loránd University, Budapest, Pázmány Péter stny. 1/C, 1117
[2] elshanqadimov7@gmail.com

## 1 Problem Description

Research shows that the cybersecurity job market is huge and some people in this industry have skill gaps. Cyberseek is a project in USA that provides detailed, actionable data about supply and demand in the cybersecurity job market. The similar projects in Europe are not as detailed as their US counter partner.In this project I will collect job ads and perform different text mining methods on them.

## 2 Progress

### 2.1 Data

The data is given in a txt format. There are 112 text files and their titles consist of the country code, date and the position name. There are different job vacancies from EU countries like Austria, Hungary, Czech Republic, Serbia, etc.

### 2.2 Text Mining and Representation

After reading text files to pandas dataframe another features like more detailed and cleaner job title, the country name were extracted. Later I cleaned job descriptions from LinkedIn links and removed the stop words so that we could focus more on the important information. Text Representation is about representing text documents with numbers to make them computable. Bags of Words, Bags of n-grams, TF-IDF vectors and Word embeddings are text representation methods.

As a next step after lemmatization, I was able to perform TF-IDF which is used for information retrieval. For the future I will analyze TF-IDF results based on country and industry.

As a last step I used SpaCy's medium sized English model which returns vectors for the given text and compared them with cosine similarity. Based on those results the sparse matrix correlation was built.

## 3 Future Plans

As next step in this process I would like to deep dive more into existing literature as well implement clustering and some other text mining techniques on the data. Another important goal would be to increase the amount of data either manually or with the usage of APIs or scrapping.

## References

1. Cybersecurity Professionals Focus on Developing New Skills as Workforce Gap Widens https://www.austcyber.com/tools-and-resources/cybersecurity-professionals-focus (ISC)² CYBERSECURITY WORKFORCE STUDY, 2018
2. Giacomo Domeniconi, Gianluca Moro and colleagues: Job Recommendation from Semantic Similarity of LinkedIn Users' Skills. ICPRAM 2016 - International Conference on Pattern Recognition Applications and Methods https://www.scitepress.org/Papers/2016/57023/57023.pdf
3. Representing text in natural language processing https://towardsdatascience.com/representing-text-in-natural-language-processing-1eead30e57d8
4. Text Similarity Measures https://machinelearninggeek.com/text-similarity-measures/ Jun 2021
5. Ultimate guide to deal with Text Data (using Python) – for Data Scientists and Engineers, https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/. Oct 2017