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

Various analyses showed that there are millions of job vacancies in cybersecurity. The job market is monitored in the USA by the Cyberseek project. There are similar initiatives in Europe as well, but not as detailed and well-defined. The goal of this project is to collect cybersecurity job ads from the LinkedIn platform and analyze the dataset with different text mining approaches.

2 Progress

As a first step txt files which the initial data was represented was converted to pandas dataframe. Features like country name and job title were extracted from the text title and text description itself. Some of the text files has been modified for easier txt to dataframe transformation.

2.1 Text representation

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.

After loading of the text to dataframe the text was cleaned from punctuations and stopwords.

3 Future Plans

As next step in the research I would like to implement clustering 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.

Sample Heading (Third Level) Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Table 1. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 1 gives a summary of all heading levels.
Displayed equations are centered and set on a separate line.

$$x + y = z \tag{1}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

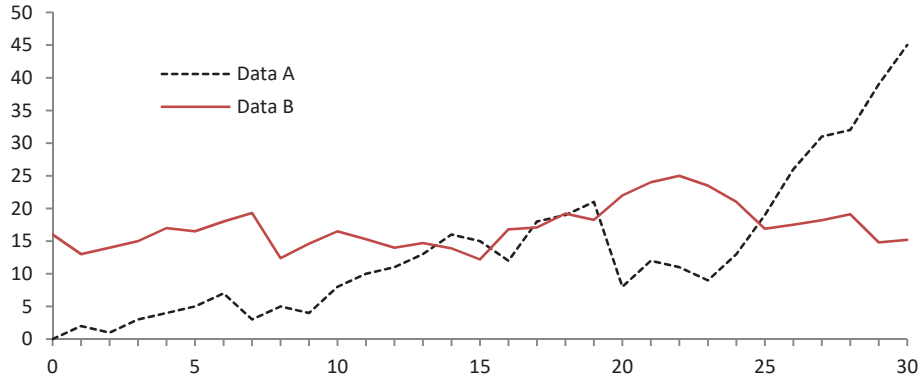

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Proof. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], and a homepage [5]. Multiple citations are grouped [1–3], [1, 3–5].

Acknowledgements Please place your acknowledgments at the end of the paper, preceded by an unnumbered run-in heading (i.e. 3rd-level heading).

References

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, http://www.springer.com/lncs. Last accessed 4 Oct 2017