Topic - Job Advertisement Deduplication

Deduplication is a basic condition to produce high quality statistics from online job advertisements as companies often publish job advertisements on different web portals. Posting advertising of the same jobs must be identified and removed using automatic and robust solutions that allow the treatment of big amounts of data in an efficient manner to avoid double counting.

That's why Eurostat released the "deduplication" challenge. The competition dataset contained 112 000 multilingual online job advertisements, retrieved from around 400 websites active in the European Union.

We've achieved the 3rd award in this challenge (0.82 F1 score) by identifying duplicated job advertisements (full, semantic, temporal or partial duplicates). Our approach based on lightweight LLM and semantic relations between adverts. Focus on scalability secures also the 3rd award in Reproducibility category – the most reproducible and scalable solutions for regular production.

Jakub Żerebecki, Mikołaj Tym - we are 5th year Computer Science and Econometrics students at Poznan University of Economics and Business. We are interested in the development of AI, we wrote bachelor's thesis on this topic. We have a few years of commercial experience as Data Scientists.