Project for Data Engineering Candidates

This project is about extracting textual data from HTML files and using the data for categorizing companies using [NAICS (North American Industry Classification System)](https://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2017).

In more detail, we would like to assign to every company one or more NAICS codes based on publicly available information about the company and on descriptions of NAICS codes. Descriptions of NAICS subsector codes are provided in the HTML files in the zipped file. Snippets of text about 4 companies are provided as separate text files. All text starting with “=====” should be ignored, as it’s included for reference only.

You are expected to write a Python program that will be given a file with a company’s description (similar to the examples given) as input and will output NAICS codes that match the description (best matching codes first).

We suggest that you use ElasticSearch as the storage backend for this project. Extra points are given for creating a website (or API) that will accept companies’ descriptions as input and will output corresponding NAICS codes. Submit all your code (parsing HTML, loading text data, matching company description to NAICS codes).

We don’t expect this project to take longer than 6 hours.

Questions? Please ask: Yakov Keselman, yakov@suplari.com