Analysis of current roles in data science using Natural Language Processing (NLP) techniques

Data science is large topic, and it is often accompanied by other words such as "machine learning", "big data", "artificial intelligence", "business analytics" or "data mining". During the last years the world of data science has evolved very rapidly, and has come to develop abilities that few years ago were thought to be almost impossible. These latest technical achievements in the field of data science have created high expectations both in academics and industry, which have created lots of new opportunities around data science, and new jobs as well.

As with any new and evolving field, data science has taken in professionals from very wide and different backgrounds, such as computer science, mathematics, statistics, electronics and industrial engineering. Back in the early days of data science, the education programs were hardly updated, and few of the people working under the name "data scientists" had been properly trained for it.

As the field has progressed, so have done the job positions. Nowadays, there are lots of different roles and the naming is usually confusing. What is a data scientist? What are its responsibilities on a data science project? What skills are necessary for the job? But also, what is the difference between a data scientist and a data analyst? Or what about a data engineer, a machine learning engineer, a business intelligence engineer or a business analyst? As you can see, the landscape of data science roles is very wide and complex.

In this project, we would like to analyze the current roles in data science by using a data driven approach.

Objectives

- Gather and explore data to analyze the current professional profiles in data science.
- Based on the data, define core data science competencies and skills for each professional profile.
- Study the consistency of professional profiles among different domains: educational programs, industry job positions and other literature sources.
- Optionally, design and develop a platform to visualize the results of the project and help fellow data science professionals understand possible career paths and the skills necessary for each job profile.

Outline

1. Review the literature on data science profiles and competences. (recommended starting source: EDISON H2020 EU project)
2. Find/download/scrape a dataset with job posts.
3. Review natural language processing (NLP) techniques for text analysis and classification.
4. Analyze the obtained datasets and extract insights about the different professional profiles.

For students that enjoy the field of machine learning, big data and data analytics or that are interested on starting a career on data science, this project will provide a general overview of the field and help better understanding the standard skills requirements for some job positions. For this project, it is not necessary to have any prior experience with NLP techniques, but a programming basis is recommended.

Design and development of a survey to study the usage of data science methodologies in industry projects

Data science professionals solve problems and answer questions through data analysis. They build and train models to predict outcomes or to discover underlying patterns, with the utmost objective of gaining valuable insights for businesses. In this sense, the tools and technologies used in data analysis are evolving rapidly, enhancing data scientists abilities to reach their goal.

Despite the recent increase in computing power and access to data over the last couple of decades, the ability to use the data within the decision making process is not being maximized at all. Very often data scientists don't have a solid understanding of the questions being asked and how to apply the data correctly to the problem at hand.

Therefore, while it is true that data analysis techniques are improving day by day, data science projects are not meeting the expectations. The rate of success of these projects is very far away from the tolerable, and this is having serious consequences on the time and resources invested by companies and research centers in data analytics projects.

We believe that a critical inhibitor of success for data science projects in the lack of a well established methodology. Like traditional scientists, data scientists need a foundational methodology that serves as a guiding strategy for solving problems. This methodology, which is independent of particular technologies or tools, should provide a framework for proceeding with the methods and processes that will be used to obtain answers and results.

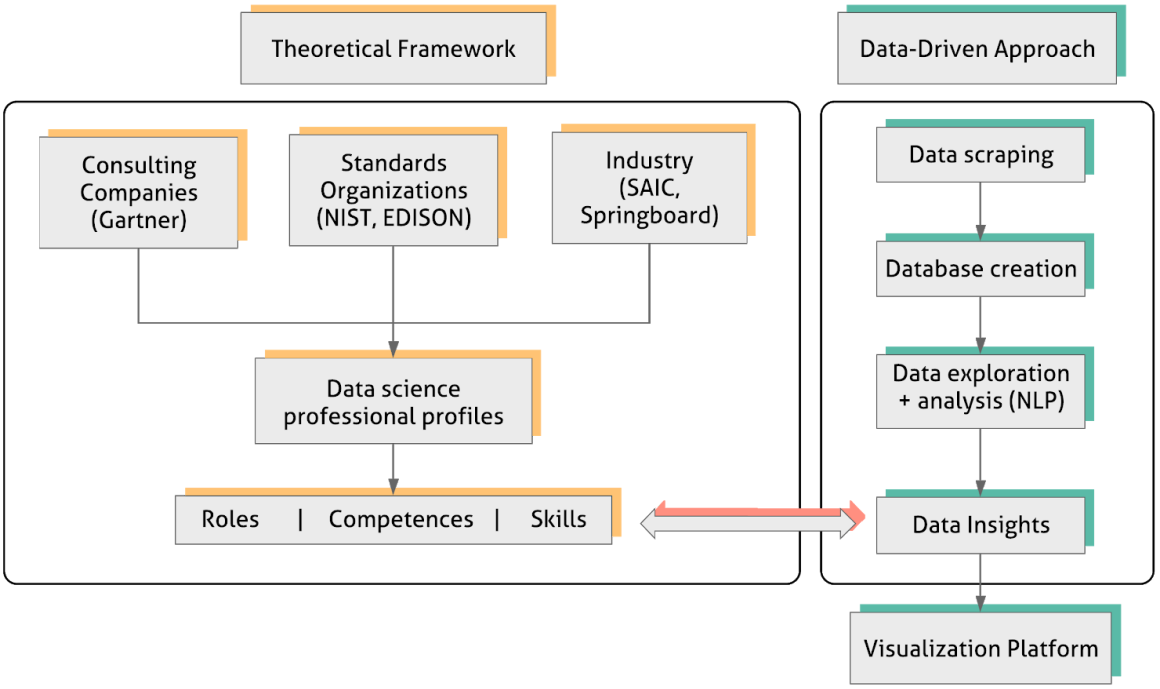
In this project, we would like to investigate the way data science teams work on their projects and the presence of data science methodologies in the industry. The main outcome of this project will be the design of a survey to gather data about the state of the field with relation to data science methodologies.

Objectives

The main objective is to design and develop a survey:

- Frame the target of the survey: business analyst, data scientists, head of departments, etc.
- Define questions to study the:
    a) major steps involved in practicing data science
    b) main organizational and technical challenges found during data science projects
    c) presence and use of data science methodologies among teams
    d) awareness of some methodologies of reference (KDD, CRISP-DM, Agile)
    e) necessary foundation for a complete methodology
- Analyze the gathered survey answers:
    a) Statistics: number of responses, average answer metrics, survey target distribution
    b) Extract insights from the survey

Oriented for students that enjoy the field of machine learning, big data and data analytics and are interested in learning about real-life data science problems. This project will provide you with a general overview of the field and the main project methodologies that are available for developing data science, which is a competence in increasing demand.



Project Timeline