Introduction to Big Data analysis

(FIS/07)

Laurea Magistrale in Marketing (MKT Inglese)

6 ECTS

Nicola Amoroso, PhD

Interuniversity Department of Physics

nicola.amoroso@uniba.it

Prerequisites/Requirements

The course is addressed to all students willing to get acquaintance with Big Data analysis and, specifically, with R programming language. A basic knowledge of programming and R is a plus, though no specific prerequisite is needed to successfully attend the course.

Expected learning outcomes

The course aims to provide the essential tools for Big Data analysis using the R programming language. The volume of data that enterprises acquire every day is increasing exponentially. The challenge these organizations now face is what to do with all this data and how to get insights from it. Thus, R comes into picture. R is a very amazing tool to run advanced statistical models on data. One key drawback of R, though, is that it is not very scalable. The core R engine can process and work on very limited amount of data. As Hadoop is very popular for Big Data processing, corresponding R with Hadoop for scalability is the next logical step this course will be aiming at. In practical terms, we will address the following questions:

- What are the basic strategies for exploratory data analysis?
- How can these strategies be applied to Big Data?
- What role for machine learning techniques? Can Artificial Intelligence provide useful knowledge for business and strategical stakeholders?

The objectives of this class are: (1) to be able to thoughtfully read R code and write it properly according to the users' needs; (2) to acquire a framework for analyzing big data and extracting knowledge from it; (3) to apply these concepts to real cases; (4) to develop a set of skills important to successfully manage a data analysis project.

Class meetings will revolve around lectures and in-class activities, such as exercises and programming session. In order to get the most from this course, it would be important that you can use you own laptop for exercise and practicing. The achievement of the objectives depends on the student; therefore, it is encouraged an "intelligent" and active participation to lectures and programming discussions. Students are also encouraged to ask questions, share their views and comments on the topics of the lectures and make examples of their own experiences. All this considering the needs and the harmony of the classroom.

A substantial component of this course will be a team-based throughout the whole semester project analyzing a real data analysis problem and focusing on R strategies to address it. Students will have the necessary written and oral communication skills to discuss their project and discuss of this course will be a team-based project that involves a real data analysis case. This project will run it from a critical perspective.

In order to succeed in this course:

1) Read assigned material and submit the assignments the day of the class in which it will be discussed.
2) Attend class regularly. Class discussions will give you opportunity to apply concepts discussed in the text. Class activities will give you a deeper understanding of the material and sharpen your analytical, problem solving, and reasoning skills.
3) Ask questions and participate in class discussions and activities.

Assessment methods

Attending students:
Written exam 40%
Project work (Group) 35%
Class participation 25%

Non-attending students:
Written exam 50%
Oral exam 50%

Exams

The exam will consist in an oral discussion about the semester programming project and a critical discussion of the concepts addressed during the class. These questions will be application oriented and will not typically test your recall of facts but will assess your understanding of concepts and their applications.

Class participation

Some cases will be discussed in class. The time is strictly limited so it is necessary to use it in order to be effective. The quality and frequency of comments are more important than the duration of each note. The capacity and the ability to speak and not just in front of a group is a critical business skill. The criticisms to the ideas of others, if constructive and made in a polite and respectful, are encouraged. Respect is a prerequisite for sharing and learning correctly.

Do not be upset if I challenge something you say, as we learn most when we have to defend our positions. Also, sometimes we will have to stop discussion and move on to the next topic before hearing from everyone, as there is limited time in each session and we want to use it wisely. Please don't take it personally if there isn't time to call on you.

There are ten points associated with class participation. I will monitor your participation over the semester and assign points accordingly. To get all ten points you need to participate actively, frequently, and meaningfully.

To help me know you as quickly as possible, you have to turn in a completed Personal Profile sheet by the second day of class. You could attach a recent, clear, color photo of yourself to this sheet to help me learn your name, or add a name card on your desk.

Group Project

A substantial component of this course will be a team-based project that involves a real data analysis. This project will run throughout the whole semester, and will conclude with your presentation of the project.

You will have a chance to apply the material discussed in class to a real-world project, thus experiencing what it actually means to work as a data analyst. By the end of the semester, you will be proficient in the R language, and will have developed a unique portfolio of analytical skills. The project you develop during this course will also serve as a tremendous asset to reference during your job interviews, to demonstrate analytical, organizational and creative skills, as well as the ability to work in a team.

Groups will be formed on the first days of class. Try to choose your group not on a relational or geographic proximity but on competences and abilities as it would happen in a company. This project will be ending with a presentation including.

Reading Material

Big Data Analytics with R - S. Walkowiak - PACKT Publishing

CLASS ETIQUETTE

Please turn off or silence mobile phones before class. Students involved in ongoing conversations, texting and/or answering cell phones during class time will be asked to leave without warning!

CHEATING AND PLAGIARISM

Cheating and plagiarism are serious offenses. Cheating is the attempted or unauthorized use of materials, information, notes, study aids, devices or communication during an academic exercise. Examples include but are not limited to:

- Unauthorized collaborating on a take home assignment or examination

- Using unauthorized electronic devices during an examination

- Taking an examination for another student or asking or allowing another student to take an examination for you

- Allowing others to research and write assigned papers including the use of commercial term paper services

- Signing for another student on the attendance sheet

Plagiarism is the act of presenting another person's ideas, research or writing as your own:

- Copying another person's actual words without the use of quotation marks and footnotes (a functional limit is four or more words taken from the work of another).

- Presenting another person's ideas or theories in your own words without acknowledging them.

- Using information that is not considered common knowledge without acknowledging the source.

- Failure to acknowledge collaborators on homework and laboratory assignment

My policy is to give a failing grade to any assignment that has been plagiarized or an exam in which you have cheated. If I catch a person cheating on the attendance sheet my policy is to give a zero to the cheating student for his/her attendance grade.