ACADEMIC PROGRAM

INFORMATION RETRIEVAL AND ANALYSIS

B.F.A. IN **COMPUTER SCIENCE**

MODALITY: ON CAMPUS

ACADEMIC YEAR: 2023-2024

Name of the course:	Information Retrieval and Analysis
Degree :	Computer Science
Location:	Centro Universitario de Tecnología y Arte Digital
Area:	Data Engineering
Year:	3º
Teaching period:	2
Type:	OBM
ECTS credits:	6
Teaching modality:	On campus
Language:	English
Lecturer / Email	-
Web page:	http://www.u-tad.com/

SUBJECT DESCRIPTION

Area description

The contents of the subject allow students to understand the flow of searching, ingesting, storing, processing and analyzing data information and brings students closer to the techniques and technologies necessary for managing large amounts of data.

Subject description

The objective of this subject is to proficiently handle the most modern techniques for capturing information from the Internet (known as “webscrapping”) in addition to the different APIs offered by many Internet services for capturing information available on the Internet. It is an essentially practical discipline with countless applications in many areas of Data Science, and for which we will use opensource environments for which there are a multitude of libraries for the analytical objectives proposed in the subject: R and python.

But there is no doubt, as always in Data Science, data preparation (cleaning, consolidation), as well as basic text management techniques (search, regular expressions) will be the first step to be able to address real datasets of considerable size.

Once the data is obtained, this subject goes into detail into two areas of application of data analysis and science techniques that are highly specialized due to the type of data on which they are based:

Text Mining or text mining, for which it will be necessary to know the mathematics underlying text processing in statistical languages, such as R (also python and others). Below we will see two libraries or methodologies of the R environment for text mining: quanteda and the tidyverse approach. And with them we will see how to view large amounts of texts.

Finally, we will see two current text mining analytics techniques: finding topics or themes in text sets (Topic Mining) and sentiment analysis. To address this last part of the subject we will have datasets resulting from the processing of all the previous phases.

And, secondly: Social Network Analysis (SNA) that uses all the available information from nodes connected to each other (social network, or connected graph). This is a very powerful methodology in this era of hyperconnectivity. As an important part of SNA techniques we will see visualization, with the help of a fascinating software: gephi.

Finally we will see applications of Deep Learning models in NLP, specifically the so-called Large Language Models (LLM), some of which (GPT-2 in particular but some more and I am sure that they will be released as the course develops) are available in the frameworks more modern DL (tensorflow and pytorch

COMPETENCIES AND LEARNING OUTCOMES

Competencies

BASIC AND GENERAL SKILLS

CG1 - Ability to understand, schedule and solve problems trough software development

CG3 - Knowledge of the scientific fundamentals applicable to the resolution of computer problems

CG4 - Ability to simplify and optimize computer systems by understanding their complexity

CG9 - Ability to learn, modify and develop new software solutions

CG10 -Use of creative techniques to carry out computer projects

CG11 - Ability to search, analyze and manage information for insights capture

BC1: Students should demonstrate knowledge in an area of study that builds upon the foundation of general secondary education and goest beyond at a level that, while supported by advanced textbooks, also encompasses certain aspects derived from the cutting edge of their field of study.

BC2: Students should be able to apply their knowledge to their work or vocation in a professional manner, and they should possess the competencies typically demonstrated through the development and defence of arguments as well as problem-solving within their field of study.

BC3: Students must possess the ability to gather and interpret relevant data (usually within their field of study) in order to make judgments that involve reflection on socially, scientifically, or ethically significant issues.

BC4: Students should be capable of conveying information, ideas, problems, and solutions to both specialized and non-specialized audiences.

BC5: Students should have developed the learning skills necessary to pursue further studies with a high degree of autonomy.

TRANSVERSAL SKILLS

CT1 - Knowledge of the definition, scope and implementation of the fundamentals of project management methodologies for technology projects

CT2 - Knowledge of the main sectorial players and the life cycle of a digital content development and commercialization project

CT4 -Ability to update the knowledge acquired in the management of digital tools and technologies according to the current state of affairs of the sector and the technological solution

CT5 -Development of the basic skills for digital entrepreneurship.

SPECIFIC SKILLS

CE3 - Knowledge of relational algebra and querying in procedural languages for the design of standardized database schemas based on entity-relationship models

CE10 - Ability to work with a release manager and generate application documentation automatically.

Learning outcomes

Upon completion of the degree, the graduate will be able to:

- To know and develop storage procedures and data management in distributed environments.

- To know and apply supervised, unsupervised and semisupervised learning techniques.

- To know and apply Deep Learning techniques

- To be able to retrieve information using web scraping or standard APIs

- To know and understand Natural Language Processing techniques

- To be able to analyze social networks contents.

- To understand the nature and representation of digital images.

- To know the applications of neural networks to the analysis and generation of sound, static images and video.

- To develop software solutions for computer vision.

- To develop a fully-fledged data project applying iterative methodology from design to delivery.

CONTENTS

Textual information. Relevance and similarity models

Textual and non-textual information retrieval.

Web scrapping

Social neteworks analysis

SUBJECT SYLLABUS

1. The R ecosystem

1.1. R programming environments (IDE): RStudio

1.2. The R language

1.3. R libraries: installation, maintenance

1.4. The Tidyverse approach to data science.

2. Text Mining with R.

2.1. Basics of text in R

- Character encoding

- Regular expressions

- The stringr library

2.2. tm and quanteda bookstores

23. The tidyverse approach to text mining.

2.4. Text display: wordclouds.

3. Text type data sources

3.1. Twitter and other social networks. API's to these sources

3.2. Web scrapping.

3.3. Other data sources

4. Modeling and analysis in text mining

4.1. Finding the topics of a text: topicmodels.

4.2. Sentiment analysis.

5. Social Network Analysis (Social Network Analysis -SNA).

5.1. Theoretical aspects of SNA

5.2. The enron case

5.3. The igraph R bookstore

5.4. Visualize networks: gephi.

5.5. SNA applications

6. Applications of Deep Learning in NLP: LLMs

6.1. Transformers

6.2. NLP in keras /TF and pytorch.org environments

6.3. LLMs in TF/Pytorch APIS

6.4. Huggingface (https://huggingface.co/)

TRAINING ACTIVITIES AND TEACHING METHODOLOGIES

TRAINING ACTIVITIES

LEARNING ACTIVITIES	Total hours	Hours of presence
Theoretical / Expository classes	29,38	29,38
Practical classes	23,25	23,25
Tutorials	4,00	2,00
Independent study and autonomous work of the student	50,00	0,00
Elaboration of work (group or individual)	31,88	0,00
Evaluation Activities	5,25	5,25
Project Follow-Up	6,25	6,25
TOTAL	150	66,13

Teaching methodologies

Expository method or master lesson

Case learning

Learning based on problem solving

Project based learning

Cooperative or collaborative learning

inquiry learning

Flipped classroom methodology

Gamification

Just in time Teaching (JITT) or classroom on time

Expository method or master lesson

Case method

Learning based on problem solving

Project based learning

Cooperative or collaborative learning

inquiry learning

Flipped classroom methodology

Gamification

TEMPORAL DEVELOPMENT

DIDACTIC UNITS / TOPICS TIME PERIOD

1. R - environment and libraries

2. Text Mining with R

- quanteda

- Text Mining with tidyverse February 2024

3. Sources of text data.

- Twitter etc. API's to these sources.

- Web scrapping.

- Other sources. March 4-7, 2024

4. Text Mining analyzes

4.1. Topicmodels

4.2. Sentiment Analysis

March 11-21, 2024

5. SNA

-The Enron case

- Gephi

- SNA igraph

- Cascades April 4-15, 2024

6. Deep Learning and LLMs

April 18 – End of classes 2024

Final project - Individual presentation (mandatory) May 13-16, 2024

EVALUATION SYSTEM

ASSESSMENT SYSTEM	MINIMUM SCORE RESPECT TO THE FINAL ASSESSMENT (%)	MAXIMUM SCORE RESPECT TO THE FINAL ASSESSMENT (%)
Assessment of participation in class, exercises or projects of the course	10	30
Assessment of assignments, projects, reports, memos	40	80
Objective test	10	60

GRADING CRITERIA

ASSESSMENT SYSTEM	ORDINARY EVALUATION	EXTRAORDINARY EVALUATION
Assessment of participation in class, exercises or projects of the course	10	10
Assessment of assignments, projects, reports, memos	80	80
Objective test	10	10

General comments on the evaluations/assessments

● The evaluation of participation in class, in practices or in projects of the subject will be carried out based on attendance and active participation in class and in the rest of the activities developed during the course. This aspect will represent 10% of the final grade for the subject in the ordinary call.

● In the last two weeks of the course, an objective test (partial) will be taken that will account for 10% of the grade. If this objective test is passed (>= 5), it will not be necessary to take the exam that will take place on the ordinary call date.

● Throughout the course, 2 associated activities or exercises will be proposed and within the time frame of each of the following topics:

o Text Mining

or SNA

These must be delivered before the proposed deadline (always before May 3, 2024) and uploaded to the virtual platform (BlackBoard). Each of these activities will be graded strictly independently and that grade will be 45% of the final grade for the subject.

● In order to pass the subject, it will be an essential requirement to have at least a grade of 5 (out of 10) *in each and every one of the proposed activities*. In case of not exceeding 5 in all the deliveries of the "partial"

tasks or activities, the recoveries must be made that will be proposed in the month of May 2024 with a delivery date that cannot be postponed on the ordinary call date for the subject.

● As a final project, a work on the application of any of the technologies included in the subject is proposed, and it must be presented publicly (in class) according to the rules that we will publish, and which must be delivered on Blackboard and presented, necessarily, in one of the the last two days of the subject this year 2024: May 13 or 16.

o The final work and its presentation are *strictly individual*.

o Your grade will account for 40% of the total score for the subject, and for this grade to be averaged with the rest of the grades it must exceed 5 (out of 10).

o There will be great freedom regarding the topic of the work or project, as well as its presentation format.

● To pass the subject in the ordinary call, the arithmetic mean of the average of partial works (always greater than or equal to 5) and of the final work (also, greater than or equal to 5). This formula summarizes these requirements:

min(partial_work_scores) = 5

partial_work_grades = sum(partial_work_scores) / 2

min(grade_final_work) = 5

min(exam_or_objective_test) = 5

qualifies_BAIN_2024 = ([(qualifies_partial_work + qualifies_final_paper) / 2] * 0.9) +

(0.1 * exam_or_objective_test)

qualifies_final_BAIN_2024 = (qualifies_BAIN_2024 * 0.9) + attendance_participation_etc

● The final qualify_BAIN_2024 score must be equal to or greater than 5 (always out of 10) in order to consider the subject passed in the ordinary call.

● If any of the minimum score requirements for each of the partial assignments, the final assignment, or the objective test are not met, the student must present/repeat all those components that are below the minimum. required and if any of these parts is suspended, it must be submitted to the ordinary May call with the recoveries that will be published during the month of May 2024. Likewise, if the final work is suspended, a new one must be made and presented on the call date ordinary.

● If the student does not obtain approval in the ordinary call in May, the student may take the extraordinary call in June/July 2024, keeping those parts approved in the ordinary call, and must retake the parts that have been failed according to the rules that we will publish for the extraordinary call.

● The grade for class participation in the ordinary session will be maintained in the extraordinary session.

● No grades of any kind will be kept between different academic years.

General considerations about the development of classes:

● It is not allowed to consume drinks or food in the classroom. The presence of any type of drink on the tables is also not permitted, even in closed containers.

● Active participation will be required from the student, necessary for the development of the classes.

● The student will be required to behave well at all times during classes. Bad behavior that prevents the normal development of the class may lead to expulsion from the classroom for a period of time to be determined by the teacher.

LIST OF REFERENCES (BOOKS, PUBLICATIONS, WEBSITES):

Basic bibliography (Text Mining)

R for data professionals: an introduction

https://www.datanalytics.com/libro_r/

Text Mining with R- A Tidy Approach

https://www.tidytextmining.com/

Additional bibliography (Text Mining)

There are literally hundreds of tutorials online for each of the topics we are going to cover. These are the most essential, in my opinion:

https://bookauthority.org/books/best-text-mining-books

https://en.wikibooks.org/wiki/R_Programming/Text_Processing

https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

https://quanteda.io/

https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf

https://rstudio-pubs-static.s3.amazonaws.com/266565_171416f6c4be464fb11f7d8200c0b8f7.html

https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html

Gayo-Abello, D. (2023): Information retrieval and text mining. Available online: https://danigayo.prof/teaching/RI-SIW-WebSem/

Basic bibliography (SNA)

igraph tutorial:

https://kateto.net/networks-r-igraph

https://kateto.net/tutorials/

Robert A. Hanneman and Mark Riddle (2005): Introduction to social network methods https://faculty.ucr.edu/~hanneman/nettext/

gephi tutorials

https://gephi.org/users/

Additional bibliography (SNA)

There cannot be a more complete list of SNA bibliography than that of the Lada Adamic course:

https://github.com/ladamalina/coursera-sna/blob/master/Syllabus.pdf

And about Spain, Twitter, the works of Mari Luz Congosto (who was also a professor at UTAD) are great:

https://www.researchgate.net/profile/Mariluz_Congosto

Deep Learning and LLMs

Jurafsky, D. ; Martin, J. H. (2023): Speech and Language Processing (3rd draft ed.). Available online: https://web.stanford.edu/~jurafsky/slp3/

https://keras.io/guides/keras_nlp/getting_started/

https://keras.io/guides/keras_nlp/transformer_pretraining/

https://keras.io/examples/nlp/

https://pytorch.org/tutorials/index.html

https://keras.io/api/keras_nlp/

REQUIRED MATERIALS, SOFTWARE AND TOOLS

Type of classroom
Theory classroom

Board and projection system

Materials:
Personal Computer

Software:
R última versión (https://cran.r-project.org/), en el momento de escribir esta guía,v. 4.3.2 (en general será necesaria versión >= 4.3.x)

y RStudio Desktop:(https://posit.co/download/rstudio-desktop/)

+ librerías stringr, tm, quanteda, tidyverse, tidyte