Currently, a large portion of the data available for analysis consists of unstructured information in the form of natural language texts. This information includes web pages (Wikipedia, digital newspapers, blogs) or social media (Facebook, Twitter). Being able to analyse these texts using Natural Language Processing (NLP) algorithms is highly beneficial for organizations to make better decisions.

Machine learning algorithms are not capable of understanding text or characters, which is why NLP performs all the necessary pre-processing to convert this text data into a machine-readable format (numbers) and enable various types of subsequent analysis to extract relevant information. Some common applications of NLP include text classification, information retrieval and extraction, summarization, machine translation, and automated question-answering systems, among others.

This is a theoretical and practical course that includes theory sessions, practical sessions, and mixed sessions. The theory classes will be taught in Spanish.
PREVIOUS KNOWLEDGE

Relationship to other subjects of the same degree
There are no specified enrollment restrictions with other subjects of the curriculum.

Other requirements
The recommended prerequisite is to have passed the Machine Learning course (first semester).

OUTCOMES

2262 - M.U. en Ciencia de Datos
- Students should communicate conclusions and underlying knowledge clearly and unambiguously to both specialized and non-specialized audiences.
- Students should demonstrate self-directed learning skills for continued academic growth.
- Students should possess and understand foundational knowledge that enables original thinking and research in the field.
- Be able to assess the need to complete their technical, scientific, language, computer, literary, ethical, social and human education, and to organise their own learning with a high degree of autonomy.
- Capacidad para trabajar en equipo para llegar a soluciones de problemas interdisciplinarios usando técnicas de análisis de datos.
- Ser capaces de acceder a herramientas de información (bibliográficas y de empleo) y utilizarlas apropiadamente.
- Ser capaces de asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio, aplicando los conocimientos adquiridos en la identificación de salidas profesionales y yacimientos de empleo.
- Extraer conocimiento de conjuntos de datos en diferentes formatos.
- Modelar la dependencia entre una variable respuesta y varias variables explicativas, en conjuntos de datos complejos, mediante técnicas de aprendizaje máquina, interpretando los resultados obtenidos.
- Usar las técnicas de procesado de lenguaje natural para analizar textos extrayendo conocimiento útil de ellos.

LEARNING OUTCOMES

Learn the different techniques for extracting and processing information from unstructured text data.

Know how to transform text into numerical vectors (embeddings) for use in machine learning algorithms.
Understand and apply DL models on text, especially attention-based models (transformers).

Learn about the various applications of NLP in text classification, knowledge extraction, and text generation.

**DESCRIPTION OF CONTENTS**

1. Introduction to the Natural Language Processing
   1.1. What is NLP?
   1.2. The importance of text
   1.3. Historical approaches to NLP
   1.4. Applications and workflow

2. Use of text in Python
   2.1. Text strings in Python
   2.2. Regular expressions
   2.3. Text loading
   2.4. Web content scraping
   2.5. NLP libraries in Python

3. Text pre-processing
   3.1. Text division
   3.2. Text cleaning and normalization
   3.3. Morphological analysis
   3.4. Semantic analysis
   3.5. Grammatical analysis

4. Features extraction
   4.1. Simple features
   4.2. Bag of Words model
   4.3. TF-IDF model
   4.4. Word vectors (word embeddings)
   4.5. Document vectors
5. NLP applications

5.1. Classification
5.2. Information extraction
5.3. Text mining
5.4. Information retrieval
5.5. Sequential models

6. Deep learning in NLP

6.1 Recurrent Neural Networks in NLP
6.2 Attention in deep learning
6.3 Transformers
6.4 Applications

WORKLOAD

<table>
<thead>
<tr>
<th>ACTIVITY</th>
<th>Hours</th>
<th>% To be attended</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory practices</td>
<td>28,00</td>
<td>100</td>
</tr>
<tr>
<td>Theory classes</td>
<td>28,00</td>
<td>100</td>
</tr>
<tr>
<td>Theoretical and practical classes</td>
<td>4,00</td>
<td>100</td>
</tr>
<tr>
<td>Attendance at events and external activities</td>
<td>2,00</td>
<td>0</td>
</tr>
<tr>
<td>Development of group work</td>
<td>15,00</td>
<td>0</td>
</tr>
<tr>
<td>Development of individual work</td>
<td>3,00</td>
<td>0</td>
</tr>
<tr>
<td>Study and independent work</td>
<td>15,00</td>
<td>0</td>
</tr>
<tr>
<td>Readings supplementary material</td>
<td>5,00</td>
<td>0</td>
</tr>
<tr>
<td>Preparation of evaluation activities</td>
<td>10,00</td>
<td>0</td>
</tr>
<tr>
<td>Preparing lectures</td>
<td>10,00</td>
<td>0</td>
</tr>
<tr>
<td>Preparation of practical classes and problem</td>
<td>10,00</td>
<td>0</td>
</tr>
<tr>
<td>Resolution of case studies</td>
<td>15,00</td>
<td>0</td>
</tr>
<tr>
<td>Resolution of online questionnaires</td>
<td>5,00</td>
<td>0</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>150,00</strong></td>
<td></td>
</tr>
</tbody>
</table>

TEACHING METHODOLOGY

Theoretical activities.
- Expository development of the subject with student participation in resolving specific questions.
- Completion of individual evaluation questionnaires.

Practical activities.

Learning through problem-solving, exercises, and case studies through which competencies on different aspects of the subject are acquired.

Laboratory and/or computer classroom work.

Learning through the completion of activities carried out individually or in small groups, conducted in computer classrooms.

---

**EVALUATION**

Objective test, consisting of one or several exams that will include both theoretical-practical questions and problems (between 20% and 80%).

Evaluation of practical activities based on the development of papers/reports, oral presentations, and e-learning tools of the University (between 20% and 80%).

Evaluation based on the participation and level of involvement of the student in the teaching-learning process, taking into account regular attendance to the scheduled face-to-face activities and the periodic resolution of proposed questions and problems (between 0% and 20%).

---

**REFERENCES**

**Basic**
- Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python. O'Really Media, 2009

**Additional**
- Jacob Eisentein. Natural Language Processing. 2018 (disponible bajo licencia CC-BY-NC-ND)
- Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana. Practical Natural Language Processing. OReally Media, 2020