Ficha de Unidade Curricular – (Versão A3ES 2018-2023)

1. Caracterização da Unidade Curricular.

1.1. Designação da unidade curricular ***(1.000 carateres).***
Processamento de Linguagem Natural/Natural Language Processing

1.2. Sigla da área científica em que se insere ***(100 carateres).***
IC

1.3. Duração[1] ***(100 carateres).***
Semestral

1.4. Horas de trabalho[2] ***(100 carateres).***
162

1.5. Horas de contacto[3] ***(100 carateres).***
67.5H (T: 22.5H TP: 24H ; PL: 21H)

1.6. ECTS ***(100 carateres).***
6

1.7. Observações[4] ***(1.000 carateres).***
UC optativa, comum com outros ciclos de estudos

1.7. Remarks ***(1.000 characters).***
Elective, shared with other courses

2. Docente responsável e respetiva carga letiva na Unidade Curricular ***(preencher o nome completo) (1.000 carateres).***
Matilde Pós-de-Mina Pato (67.5)

3. Outros docentes e respetivas cargas letivas na unidade curricular ***(1.000 carateres).***

4. Objetivos de aprendizagem (conhecimentos, aptidões e competências a desenvolver pelos estudantes). ***(1.000 caracters).***

Os estudantes que terminam com sucesso esta unidade curricular serão capazes de:

1. Adquirir os conceitos linguísticos fundamentais relevantes para o processamento de texto em linguagem natural (PLN).
2. Entender os algoritmos e as técnicas básicas e do estado da arte para lidar com texto em linguagem natural.
3. Familiarizar-se com as ferramentas e os recursos linguísticos mais avançados em PLN.
4. Compreender e empregar métricas de avaliação para diferentes tarefas do PLN.
5. Ser capaz de formular um problema de classificação em PLN e resolvê-lo com as técnicas, algoritmos e ferramentas apropriados.
6. Ler e compreender as tarefas que são realizáveis com as tecnologias atuais.
7. Escrever relatórios técnicos e elaborar apresentações técnicas com análise comparativa e discussão de diferentes resultados.

4. Intended learning outcomes (knowledge, skills, and competencies to be developed by the students). ***(1.000 characters).***

Students who complete this course unit will be able to:

1. Acquire the fundamental linguistic concepts relevant to natural language text processing.
2. Understand the essential and state-of-the-art algorithms and techniques for dealing with natural language text.
3. Become familiar with the most advanced linguistic tools and resources in NLP.

4. Understand and employ evaluation metrics for different NLP tasks.
5. Be able to formulate a classification problem in NLP and solve it with the appropriate techniques, algorithms, and tools.
6. Read and understand current research on NLP.
7. Write technical reports and prepare presentations with comparative analysis and discussion of different solutions.

5. Conteúdos programáticos ***(1.000 carateres).***

I. Introdução ao PLN: linguagens formais e naturais, ambiguidade, variabilidade linguística e outros, processamento em cadeia.
II. Processamento básico de texto: expressões regulares, tokenização, normalização, lematização, *stemming* e segmentação.
III. Modelos probabilísticos de linguagem: *n-grams* e suavização.
IV. Classificação de texto: *bag-of-words*, Naive Bayes, engenharia de *features*; classificadores generativos e discriminativos.
V. Semântica: semântica lexical e vectorial, análise semântica, desambiguação do sentido das palavras e *entity linking*, *embeddings* de palavras.
VI. Modelos sequenciais: modelos de Markov, *conditional random fields*; marcação de classe gramatical e reconhecimento de entidades mencionadas.
VII. Aprendizagem profunda em PLN: modelos neuronais de linguagem, redes neuronais recorrentes, redes *encoder-decoder,* redes *transformer*.
VIII. Aplicações: classificação de texto, recuperação de informação, extração de informação, resposta automática a perguntas, sistemas de conversação, e outras aplicações.

5. Syllabus ***(1.000 characters).***

I. Introduction to Natural Language Processing (NLP): formal and natural languages, ambiguity, linguistic and another variability, chain processing.
II. Basic text processing: regular expressions, tokenisation, normalisation, lemmatisation, stemming, and segmentation.
III. Probabilistic language models: n-grams and smoothing.
IV. Text classification: bag-of-words, Naive Bayes, feature engineering, generative and discriminative classifiers.
V. Semantics: lexical semantics, vector semantics, semantic analysis, word sense disambiguation and entity linking, and word embeddings.
VI. Sequential models: hidden Markov models, conditional random fields, grammar class marking and named entity recognition.
VII. Deep learning in NLP: neural models of language, recurrent neural networks, encoder-decoder networks, attention, transformer networks.
VIII. Applications: text classification, information retrieval, information extraction, automatic question answering, conversational systems, and other applications.

6. Demonstração da coerência dos conteúdos programáticos com os objetivos de aprendizagem da unidade curricular ***(1.000 carateres).***

Esta UC constitui o primeiro contacto dos estudantes com a área de processamento de linguagem natural, proporcionando uma abordagem estruturada à aprendizagem das competências e conhecimentos necessários para trabalhar com dados de linguagem natural. Os tópicos (I) e (II) introduzem a cultura geral necessária para compreender o tema e aferir os objectivos de aprendizagem (1) e (2); os tópicos (III) a (VIII) permitem aferir o cumprimento dos objectivos de aprendizagem (3) a (6). Com a realização e apresentação do trabalho prático e a elaboração do respetivo relatório é possível aferir o objetivo de aprendizagem (7).

6. Evidence of the syllabus's coherence with the curricular unit's intended learning outcomes ***(1.000 characters).***

This course is the student's first contact with the field of natural language processing, providing a structured approach to learning the necessary skills and knowledge to work with natural language data. Topics (I) and (II) introduce the general culture required to understand the subject and assess learning objectives (1) and (2);

issues (III) to (VIII) enable the assessment of the achievement of learning goals (3) to (6). With the realisation and presentation of the practical work and the elaboration of the respective report, it is possible to assess the learning objective (7).

7. Metodologias de ensino (avaliação incluída) ***(1.000 carateres).***

Metodologia de ensino é baseada na abordagem *Problem-Based Learning* (PBL). Pretende-se privilegiar a autonomia do estudante no desenvolvimento de soluções para problemas complexos, adequados ao seu nível cognitivo. Incentiva-se o trabalho em grupo e a discussão/reflexão em sessões de grupo. As aulas destinam-se à apresentação dos temas e de exemplos práticos de aplicação. Os objetivos de aprendizagem de (1) a (6) são avaliados através da componente teórica (CT), constituída por avaliação presencial. Os objetivos de aprendizagem (1) a (7) são avaliados através da componente prática (CP), que consiste na realização de trabalhos práticos (TP), pedagogicamente fundamental, a escrita dos relatórios e uma apresentação em contexto de sala. A classificação final é obtida através da média aritmética simples de ambas as componentes. Para aprovação na UC, a classificação mínima da CT é de 8 valores. A avaliação em épocas especiais consiste na elaboração de 2 TPs e de 1 exame escrito.

7. Teaching methodologies (including assessment) ***(1.000 characters).***

The teaching methodology is based on the Problem-Based Learning (PBL) approach. It is intended to privilege student autonomy in developing solutions to complex problems appropriate to their cognitive level. Group work and discussion/reflection are encouraged in group sessions. Classes are designed to present the topics and practical examples of application. Learning objectives (1) to (6) are assessed using the theoretical component (CT), consisting of a face-to-face evaluation. Learning objectives (1) to (7) are evaluated through the practical part (CP), which consists of pedagogically actual practical work (TP), the writing of a report and a presentation in class. The final classification is obtained through the simple arithmetic mean of both components. For approval in the UC, the minimum mark of the CT is 8 points. Special-season evaluation consists of 2 TPs and one written exam, each component worth 50% of the final mark.

8. Demonstração da coerência das metodologias de ensino com os objetivos de aprendizagem da unidade curricular ***(3.000 carateres).***

As aulas destinam-se à apresentação das bases teóricas dos conteúdos programáticos (aulas teóricas). Nas aulas, são desenvolvidos pequenos projetos e analisados casos de estudo (aulas teórico-práticas). Privilegia-se uma forma de apresentação interativa. A componente laboratorial (aulas práticas) serve para aplicar num ambiente controlado as técnicas apresentadas. O trabalho autónomo (extra-aula) é guiado pelo trabalho prático (projeto), concebido para consolidar as competências de conceção e desenvolvimento dos conteúdos programáticos. O projeto é apresentado aos estudantes no início do semestre guiando os exemplos e tópicos lecionados. Os objetivos de aprendizagem são identificados nos guiões apresentados aos estudantes, permitindo clarificar as competências que são necessárias adquirir no desenvolvimento do projeto e nas aulas práticas.

8. Evidence of the teaching methodologies' coherence with the curricular unit's intended learning outcomes ***(3.000 characters).***

The classes aim to present the theoretical basis of the course contents (academic classes). In class, small projects are developed, and case studies are analysed (theoretical-practical classes). An interactive form of presentation is favoured. The laboratory component (practical courses) serves to apply the techniques presented in a controlled environment. Autonomous work (extra-class) is guided by experimental work (project) designed to consolidate the design and development skills of the course contents. The project is presented to students at the beginning of the semester, guiding the examples and topics taught. The learning objectives are identified in the guides given to the students, allowing clarification of the skills necessary to acquire in the development of the project and the practical classes.

9. Bibliografia de consulta/existência obrigatória ***(1.000 carateres).***

Dan Jurafsky and James H. Martin, *Speech and Language Processing*, Prentice Hall, 2023 (3rd edition draft).
Jacob Eisenstein, *Natural Language Processing,* MIT Press, October 2019. ISBN: 9780262042840.
Yoav Goldberg, *Neural network methods for natural language processing*. Morgan & Claypool Publishers, 2017.

ISBN: 9781627052986. DOI: 10.2200/S00762ED1V01Y201703HLT037

[1] Anual, semestral, trimestral, …

[2] Número total de horas de trabalho.

[3] Discriminadas por tipo de metodologia adotado (T - Ensino teórico; TP - Ensino teórico-prático; PL - Ensino prático e laboratorial; TC - Trabalho de campo; S - Seminário; E - Estágio; OT - Orientação tutorial; O - Outro).

[4] Assinalar sempre que a unidade curricular seja optativa.