PL01 –Processamento de Linguagem Natural

MIA – Mestrado em Inteligência Artificial

Plano de Aula - PL01

Apresentação da UC

Processamento de Linguagem Natural

Tarefas de Pré-Processamento

Exemplo Prático

Ficha Exercícios (fe01)

Apresentação da UC

Apresentação da UC

Equipa Docente:

José Manuel Ferreira Machado
jmac@di.uminho.pt
Tel: 253604438
Gab. 3.10

Hugo Peixoto
hpeixoto@di.uminho.pt
Gab. 1.17

Apresentação da UC

Objetivos:

O conteúdo programático proposto foi desenhado para satisfazer o grande objetivo da UC: uniformizar as competências dos alunos, em particular em conceitos gerais e proficiências de introdução à Inteligência Artificial e ao Processamento de Linguagem Natural.

A sequência definida para a apresentação e discussão dos diversos tópicos do programa foi definida como uma introdução completa e atualizada dos conceitos-chave em IA, apresentando gradualmente e naturalmente as diferentes abordagens, nomeadamente as suas especificidades e limitações.

Apresentação da UC

Resultados da Aprendizagem:

- Uniformizar as competências dos alunos, em particular em conceitos gerais e proficiências em Inteligência Artificial e em Processamento de Linguagem Natural;

- Apresentar uma visão geral da Inteligência Artificial, com ênfase na utilidade e aplicação das diferentes abordagens de resolução de problemas;

Apresentação da UC

Resultados da Aprendizagem:

- Compreender o modelo lógico de representação de conhecimento, o desenvolvimento de mecanismos de raciocínio para a resolução de problemas e a seleção e implementação de modelos de representação de conhecimento e raciocínio mais adequados à resolução de problemas reais;

- Estudar conceitos básicos de Língua Natural e seus principais elementos, conceber, desenvolver e testar sistemas de língua natural dentro de uma variedade de aplicações diferentes e o desenvolvimento de sistemas com a capacidade de extrair conhecimento a partir de textos.

Apresentação da UC

Programa:

1. Inteligência Artificial
2. Métodos de Resolução de Problemas e de Procura
3. Conhecimento, Raciocínio e Planeamento
4. Processamento de Linguagem Natural
5. Conceitos Avançados
6. Text Mining
7. Aplicações

Apresentação da UC

Bibliografia:

- Artificial Intelligence: A Modern Approach, Stuart Russell and Peter Norvig, (3rd Edition), ISBN 978-9332543515, 2015.
- Blueprints for Text Analytics Using Python, Albrecht, J., Ramachandran, S., Winkler, C.. O'Reilly Media., 2020.
- Natural Language Processing with Python, Bird, S., O'Reilly Media. 2016.
- An Introduction to Corpus Linguistics, Kennedy, G. Taylor & Francis., 2014.
- Artificial Intelligence: Foundations of Computational Agents, Poole and Mackworth, 2nd ed., ISBN 978-1107195394, 2017.

Métodos de Ensino:

As aulas teóricas decorrem com a exposição e a discussão dos diversos tópicos considerados no programa da unidade curricular, com recurso a situações práticas de aplicação real.

As aulas práticas-laboratoriais decorrem com a análise e a resolução de exercícios práticos, e a demonstração de aplicações práticas, acompanhando os conteúdos lecionados na componente teórica.

A presença nas aulas teóricas e práticas será controlada.

Métodos de Avaliação:

A avaliação da aprendizagem envolve dois instrumentos de avaliação:

a) trabalhos práticos de desenvolvimento (com peso de 50%) e

b) uma prova escrita (com peso de 50%).

A classificação final é calculada pela ponderação dos diversos instrumentos de avaliação. É considerado aprovado o aluno cuja classificação final seja superior ou igual a 10 (dez) valores.

Para o cálculo da classificação final estabelecem-se notas mínimas. Estas notas mínimas não podem ser inferiores a 10 (dez) valores.

TOMORROW NEEDS ENGINEERING

Processamento de Linguagem Natural

Processamento de Linguagem Natural

O que é Processamento de Linguagem Natural?

O Processamento de Linguagem Natural, ou PLN, é uma subárea da inteligência artificial (IA) que visa à criação de sistemas capazes de entender, interpretar e gerar linguagem humana de forma eficaz e natural.

O objetivo principal do PLN é construir tecnologias que permitam aos computadores comunicar-se com os humanos na sua própria língua, além de automatizar a interpretação de texto e fala para várias aplicações práticas.

Processamento de Linguagem Natural

Importância do Processamento de Linguagem Natural?

- **Interseção Tecnológica**: O PLN é o ponto de interseção entre a linguística computacional, a inteligência artificial e a ciência da computação, abrindo portas para inovações tecnológicas significativas que transformam a forma como interagimos com as máquinas.
- **Era Digital**: Na era digital, a quantidade de dados gerados em linguagem natural é monumental. O PLN permite a análise e o processamento desses dados em grande escala, transformando-os em insights valiosos e ações estratégicas para as empresas.
- **Automatização e Eficiência**: Ferramentas e sistemas baseados em PLN automatizam tarefas que envolvem linguagem, como tradução automática, resumo de textos, assistentes virtuais e atendimento ao cliente, aumentando a eficiência e a produtividade.
- **Acessibilidade e Inclusão**: O PLN também desempenha um papel crucial na criação de tecnologias acessíveis, ajudando pessoas com deficiências através de sistemas de reconhecimento e síntese de fala, por exemplo.

Processamento de Linguagem Natural

TOMORROW NEEDS ENGINEERING

Exemplos de PLN:

- **Análise de Sentimento em Feedback de Clientes**: Utilização do PLN para avaliar automaticamente o sentimento e a emoção em comentários de clientes, proporcionando uma compreensão mais profunda da satisfação do cliente e áreas para melhoria.

- **Assistência Virtual Inteligente**: Desenvolvimento de *chatbots* e assistentes virtuais que entendem consultas em linguagem natural, oferecendo suporte ao cliente eficiente e personalizado, além de reduzir a carga sobre equipas de atendimento.

- **Resumo Automático de Documentos**: Implementação de soluções de PLN para gerar resumos concisos de documentos extensos, relatórios e artigos, economizando tempo e melhorando a eficiência operacional.

Processamento de Linguagem Natural

Benefícios de PLN:

- **Marketing**: Melhoria na segmentação de público e personalização de campanhas através da análise de tendências de consumo e feedback de clientes, aumentando a eficácia das estratégias de marketing.

- **Recursos Humanos (RH)**: Otimização de processos de recrutamento e seleção através da análise automática de currículos e descrições de vagas, identificando rapidamente os candidatos mais adequados.

Processamento de Linguagem Natural

Benefícios de PLN:

- **Call Center**: Melhoria da experiência do cliente com respostas mais rápidas e precisas, além da análise de transcrições de chamadas para identificar padrões, problemas recorrentes e oportunidades de treino.

- **Comunicação Estratégica**: Análise de tendências de redes sociais, media e opinião pública para orientar decisões estratégicas de comunicação e gestão de crise, assegurando que as mensagens da empresa são efetivas e bem recebidas.

Desafios da implementação do PLN

- Necessidade de grandes volumes de dados de treino de qualidade;
- Complexidade de interpretar nuances e contextos específicos;
- Questões de privacidade e ética na análise de dados exigem atenção e cuidados rigorosos.

Processamento de Linguagem Natural

Desafios da implementação do PLN

Processamento de Linguagem Natural em Português: Lidar com as especificidades e nuances do português, como a riqueza de verbos e a variação sintática, requer modelos e algoritmos ajustados para capturar efetivamente a essência da língua.

Ambiguidade da Linguagem: A linguagem humana é repleta de ambiguidades, sejam elas léxicas, sintáticas ou semânticas. Técnicas avançadas de PLN e modelos de IA, como redes neurais profundas, estão a ser utilizadas para superar estes desafios.

Técnicas e Ferramentas do PLN

- **Análise Léxica**: Processo de conversão do texto em palavras ou *tokens*, preparando-os para análise mais profunda.
- **Análise Sintática**: Determinação da estrutura gramatical do texto, identificando relações entre palavras e frases.
- **Análise Semântica**: Extração do significado e interpretação do contexto das palavras e frases no texto.
- **Machine Learning e Deep Learning**: Técnicas que permitem aos sistemas de PLN aprender e melhorar a partir de grandes volumes de dados de texto.

Processamento de Linguagem Natural

Futuro do PLN - Desafios e Oportunidades:

- **Avanços Tecnológicos**: A contínua evolução em IA e machine learning promete avanços significativos no PLN, tornando as interações com máquinas ainda mais naturais e intuitivas.
- **Superação de Barreiras Linguísticas**: Uma das grandes promessas do PLN é a capacidade de quebrar barreiras linguísticas, facilitando uma comunicação global sem precedentes.
- **Compreensão Contextual Profunda**: O desenvolvimento de modelos mais sofisticados que compreendem nuances culturais e contextuais representa um desafio contínuo e uma oportunidade para tornar o PLN verdadeiramente universal.

Tarefas de Pré-Processamento

Tarefas de Pré-Processamento

Lematização – Definição

A lematização é o processo de reduzir uma palavra à sua forma base ou raiz, conforme encontrada no dicionário. A lematização tem em conta o contexto e a morfologia da palavra para garantir que o resultado seja um termo válido.

Esta técnica é fundamental para o PLN, pois ajuda a padronizar palavras que aparecem em diferentes formas gramaticais. Isto melhora a recuperação de informações, a análise de sentimentos e a classificação de textos, tornando os dados textuais mais estruturados e comparáveis.

Tarefas de Pré-Processamento

Lematização - Exemplos

Dado um conjunto de palavras:

"corro", "correndo", "correu", "corrida"

Ao aplicar a lematização, todas estas formas podem ser reduzidas à raiz **"correr"**, garantindo uma representação unificada do termo.

Tarefas de Pré-Processamento

Lematização – Aplicação Prática

A **lematização** é amplamente utilizada em motores de pesquisa e sistemas de análise de texto. Por exemplo, ao pesquisar “corredor” num motor de pesquisa que utiliza lematização, o sistema pode devolver resultados relacionados a "correr", "correu" e "corrida", e proporciona assim respostas mais abrangentes e precisas.

Tarefas de Pré-Processamento

Stemming – Definição

O **stemming** é um processo de normalização de texto que reduz palavras à sua raiz, removendo sufixos e afixos. Diferente da lematização, o stemming não considera o significado da palavra, apenas aplica regras para cortar partes do termo. Desta forma pode gerar resultados que não são palavras reais.

Esta técnica é útil para recuperação de informações e indexação de documentos, pois reduz a variação de palavras semelhantes, melhorando a eficiência de pesquisas e análise textual.

Tarefas de Pré-Processamento

Stemming - Exemplos

Dado um conjunto de palavras:
"correr", "corrida", "corredor", "correu"

Após a aplicação do stemming, todas podem ser reduzidas à raiz **"corr"**, que não é uma palavra válida, mas mantém a essência do termo original.

Tarefas de Pré-Processamento

Stemming – Aplicação Prática

O stemming é amplamente utilizado em motores de pesquisa e sistemas de mineração de textos. Permite que documentos com palavras derivadas de um mesmo radical sejam encontrados de maneira mais eficiente, reduzindo o processamento necessário para análise de grandes volumes de texto.

Tarefas de Pré-Processamento

Stop Words – Definição

As **stopwords** são palavras comuns que geralmente não carregam significado relevante num determinado texto e, por isso, costumam ser removidas no pré-processamento de dados em PLN. Estas palavras incluem artigos, preposições, pronomes e conjunções, como "o", "é", "e", "mas", "de", entre outras.

A remoção de stopwords ajuda a reduzir o ruído nos dados e melhora a eficiência de algoritmos de pesquisa, análise de sentimentos e modelação de tópicos, pois permite que o foco fique nas palavras mais informativas.

Tarefas de Pré-Processamento

Stop Words – Exemplos de remoção

Texto original:

"O cão e o gato correram pelo parque."

Após a remoção de stopwords:

"cão gato correram parque."

Tarefas de Pré-Processamento

Stop Words – Aplicação Prática

Em motores de pesquisa, a remoção de **stopwords** evita que termos irrelevantes influenciem os resultados.

Por exemplo, uma pesquisa sobre **”os melhores livros de ficção"**, palavras como **”os"** ou **”de"** são descartadas, garantindo uma busca mais eficiente e precisa.

Tarefas de Pré-Processamento

Contagem de Valores – Definição

A **contagem de valores** é um método de análise de dados que identifica a frequência de ocorrência de cada valor único dentro de um conjunto de dados. Essa técnica é amplamente usada para entender a distribuição de categorias em bases de texto, tabelas ou bases de dados estruturados.

No contexto do PLN, a contagem de valores pode ser aplicada para identificar padrões de palavras, analisar categorias de sentimentos ou classificar entidades em textos.

Tarefas de Pré-Processamento

Contagem de Valores – Exemplos

Texto de entrada:

"O gato é um animal de estimação popular. O cão também é um animal de estimação. Muitas pessoas têm gato ou cão."

Contagem de valores de palavras:

Palavra	Frequência
gato	2
cão	2
animal	2
estimação	2
muitas	1
pessoas	1

Tarefas de Pré-Processamento

Contagem de Valores – Aplicação Prática

A contagem de valores é útil para **análise de frequência de palavras, categorização de documentos e deteção de tendências em grandes volumes de texto**, sendo aplicada em mineração de textos, *chatbots* e motores de recomendação.

Tarefas de Pré-Processamento

TF–IDF – Definição

- TF-IDF (*Term Frequency – Inverse Document Frequency*) é uma forma de **atribuir peso** às palavras de um documento:

- **TF** (frequência no documento): quanto a palavra aparece **neste** texto.

- **IDF** (inverso da frequência nos documentos): quão **rara** é a palavra **no corpus todo**. Palavras muito frequentes no documento **e** raras no corpus recebem **peso alto**; palavras comuns em todo lado (ex.: “hoje”, “ano”) recebem **peso baixo**.

Tarefas de Pré-Processamento

TF–IDF – Exemplos

Corpus com 3 documentos (já sem stopwords):

D1: “governo anuncia medidas hoje”

D2: “oposição critica medidas hoje”

D3: “hoje há jogos futebol”

Cálculos (log natural):

N = 3 documentos

df(hoje)=3 ⇒ **IDF(hoje)=ln(3/3)=0**

df(governo)=1 ⇒ **IDF(governo)=ln(3/1)≈1.10**

df(medidas)=2 ⇒ **IDF(medidas)=ln(3/2)≈0.405**

Tarefas de Pré-Processamento

TF–IDF – Aplicação Prática

Keywords por documento (top termos com maior TF-IDF).

Pesquisa/IR: ranking de documentos mais relevantes para uma query.

Classificação/Clustering: usar a matriz TF-IDF como *features*.

Filtrar ruído: reduzir influência de termos muito genéricos do domínio.

TOMORROW NEEDS ENGINEERING

Exemplo Prático – Google Colab

Exemplo prático

1. Aceder ao Google Colab

2. Fazer o upload do documento fornecido n

A contagem de valores é útil para **análise de frequência de palavras, categorização de documentos e deteção de tendências em grandes volumes de texto**, sendo aplicada em *text mining*, *chatbots* e motores de recomendação.

TOMORROW NEEDS ENGINEERING

Ficha de Exercícios 01

PL01 –Processamento de Linguagem Natural

MIA – Mestrado em Inteligência Artificial