



Instituto Superior de Engenharia de Lisboa

## **Mineração de Dados em Larga Escala**

Mestrado Engenharia Informática e Multimédia &  
Mestrado Engenharia Informática e Computadores

# **Parte 1**

## **Objectivos de Aprendizagem - Fase 1**

Semestre de Verão, 2022/2023

Nome: Gonçalo Fonseca | Número: A50185  
Nome: Pedro Diogo | Número: A47573

# Índice

1	Introdução	1
	Referências	2

# 1 Introdução

O conjunto de dados "*Influenza Outbreak Twitter Data*" contém uma coleção de tweets relacionados a surtos de influenza. O problema concreto de mineração de dados a ser resolvido com este conjunto de dados poderia ser identificar padrões nos dados que possam ser úteis para prever ou rastrear surtos de influenza.

Em geral, a tarefa de identificar padrões em dados está inserida no contexto mais amplo das tarefas de mineração de dados. Mineração de dados é o processo de descobrir informações úteis a partir de grandes quantidades de dados. Isso envolve a aplicação de várias técnicas, como agrupamento, classificação, mineração de regras de associação e detecção de anomalias para extrair *insights* e conhecimentos dos dados.

No contexto do "*Influenza Outbreak Twitter Data*", a tarefa de mineração de dados pode envolver a análise do texto dos tweets para identificar temas comuns, como os sintomas associados à influenza, as regiões onde ocorrem surtos e a gravidade dos surtos. Os resultados da análise poderiam então ser usados para desenvolver modelos preditivos que ajudassem os funcionários de saúde pública a se preparar e responder melhor a surtos de influenza.

## Referências

- [1] sparklyr - [Em linha] [Consult. 19 mar. 2023]. Disponível em WWW:<https://spark.rstudio.com/packages/sparklyr/latest/reference/#spark-machine-learning>.