

用 Python 语言做社交媒体上的情感分析

课题大纲

01

课程背景及简介

适合人群：本课程属于通识课程，适合对计算机科学、机器学习感兴趣的同学。

联合国胜任力：技术认识

情感分析在商业洞察(客户反馈、市场研究、客户服务、风险管理等)、政治竞选(公众意见和反应、公众形象管理等)和医疗保健领域变得越来越重要。人工智能的应用数量，特别是机器学习，正在以非常快的速度增长。社交媒体上的情感分析就是其中一种应用。机器学习技术用于在标记数据集上训练模型，其中文本的情绪(积极，消极或中立)已经已知。经过训练后，这些模型可以用来预测新的、未见过的文本的情绪。

在本课程中，学生将学习机器学习的基础知识，编程语言 Python，并使用 Python 创建实现机器学习的情感分析模型。

成功完成本课程的学生将对机器学习，特别是监督学习有扎实的理解，将精通 Python，将能够创建监督学习模型(人工神经网络，逻辑回归和线性回归)。学生们还将能够应用这些概念来创建社交媒体等领域的情感分析模型。

02

学习目标

本课程将涉及：

1. 监督式学习
2. 编程语言 Python
3. 使用 Python 的机器学习来创建情感分析模型

03

导师信息

Prof. Goldsztein

教授来自阿根廷的布宜诺斯艾利斯。1992 年，他获得布宜诺斯艾利斯大学数学学士学位，1997 年获得麻省理工学院数学博士学位。在接下来的三年中(1997-2000)，他在加州理工学院担任博士后学者和应用数学讲师。自 2000 年以来，他一直是佐治亚理工学院数学学院的教员，目前是该校的正教授。Goldsztein 教授喜欢将数学应用于其他科学领域，如计算生物学、机器学习以及数学与物理的交叉，机器学习亦是他的专业领域之一。

04

课程设置

模块 1：社交媒体中的情感分析和机器学习的基础知识

学习目标：数据科学、机器学习和人工智能

简介：数据科学、机器学习、人工智能和监督学习的导论；如何将数据组织为一组示例，以及示例特征和标签；数值变量 vs 分类变量；模型在集合上的误差、均方误差、二值交叉熵误差和多分类交叉熵误差。

模块 2：介绍 Python 和 Python 在现实生活中的应用

学习目标：模型参数和训练样例集

简介：通过最小化模型在训练集上的误差来选择参数，包括线性回归、逻辑回归、神经网络和谷歌合作实验室。Python 入门：包括整数、浮点数、字符串和布尔值、变量、循环等概念。

模块 3：Python 机器学习库，包括 Numpy 和 Pandas，机器学习库在情感分析中的作用

学习目标：科学计算 Python 库

简介：用 Numpy 和 Pandas 来读取互联网上的数据；使用软件 Keras 来解决机器学习问题的案例；清理和预处理数据。

模块 4：真实世界数据集上的监督学习实例

学习目标：一个连续统问题，一个二元分类问题和一个多重分类问题

模块 5：使用机器学习的自然语言处理和情感分析

学习目标：自然语言处理导论。对推文进行情感分析。

05

延伸阅读

O Theoblad, Machine Learning For Absolute Beginners: A Plain English Introduction (Third Edition) (AI, Data Science, Python & Statistics for Beginners). 2021

(非必读)