



# 自然语言处理与 网络舆情监控

Natural Language Processing and  
Public Opinion Monitoring

#1. 课程背景及简介



在当今互连的世界中，人们的思想，思想，观点和活动转移到数字世界中，从而产生大量数据。自然语言处理揭示了隐藏在这些单词流中的见解，可以以创新的方式来使用这些见解来改善业务，服务社会。

自然语言处理（Natural Language Processing , NLP）是一门通过建立形式化的计算模型来分析、理解和处理自然语言的学科，也是一门横跨语言学、计算机科学、数学等领域的交叉学科。具体表现形式包括机器翻译、文本摘要、文本分类、文本校对、信息抽取、语音合成、语音识别等。人类一直在探寻如何通过机器学习理解文本语义，推理文本间的关系，甚至加入联想、创新推理，目前技术主要停留在机器学习理解文本的阶段，也有相关研究人员在研究“事理图谱”等推理领域，但也是刚刚开始。自然语言处理并不是一般地研究自然语言，而在于研制能有效地实现自然语言通信的计算机系统，特别是其中的软件系统。舆情监控，整合互联网信息采集技术及信息智能处理技术通过对互联网海量信息自动抓取、自动分类聚类、主题检测、专题聚焦，实现用户的网络舆情监测和新闻专题追踪等信息需求，形成简报、报告、图表等分析结果，为客户全面掌握群众思想动态，做出正确舆论引导，提供分析依据。

本课程介绍 NLP 中的基本概念，例如文本处理、形态分析、句法分析以及词法、关系和主题语义分析等内容。本课程还涵盖了一些实际示例，来演示如何在实际生活中应用所学知识技术，以及如何使用 NLP 的民意监测系统的整个体系结构。

#2. 学习目标



本课程将解决许多挑战，如：

- ★ 如何在 Python 中处理文本和 pdf 文件
- ★ 如何使用正则表达式在 Python 中进行模式搜索
- ★ 如何利用 Spacy 进行超快速的标记化
- ★ 如何在 NLP 中进行词干和词形化
- ★ 如何利用 Spacy 进行词汇匹配
- ★ 如何用部分语音标记来处理原始文本文件
- ★ 如何使用命名实体识别
- ★ 如何利用 NLTK 进行文本情感分析
- ★ 如何利用基本的自然语言处理解决现实世界中的问题
- ★ 如何利用 Numpy 和 Pandas 数据库进行数据分析
- ★ 如何设计舆论监督系统

#3. 任课教师信息



**Dr. P K**  
老师目前是俄亥俄州立大学计算机科学与工程学院高级讲师，同时也是 Nationwide 保险公司企业首席技术官与 IT 体系架构师。他是 Open Group 认证的大师级架构师和 TOGAF 认证专家，以及俄亥俄州建筑师协会企业建筑师协会 (AEA) 与商业建筑协会会员。曾任富兰克林大学计算机科学系兼职教授，加拿大渥太华卡尔顿大学助教。

#4. 课程设置



| 周期                     | 时间             | 课程设置内容  | 课时 |
|------------------------|----------------|---|----|
| 第一周<br>学习指南<br>教授及助教辅导 | 7 月 18 日<br>周一 | 什么是 PBL 教学方法  | 1  |
|                        | 7 月 19 日<br>周二 | PBL 教学的常见形式   | 1  |
|                        | 7 月 20 日<br>周三 | 教授课-1<br>交叉学科 PBL 课程设计及知识点学习<br>学习目标：Python 语言回顾<br>描述：本课程涵盖的主题包括 python 语言的复习回顾，包括语法、文本和 pdf 的解析和处理，正则表达式的使用等。 | 3  |
|                        | 7 月 22 日<br>周五 | 助教课-1<br>知识点查漏补缺  | 2  |
|                        | 7 月 23 日<br>周六 | 教授课-2<br>制定小组项目方向<br>学习目标：自然语言处理<br>描述：本课程涵盖的主题包括对 NLTK 和 Spacy 库以及 NLP 技术的介绍，例如词干，词根化，停用词，词组匹配和标记化等。           | 3  |
|                        | 7 月 25 日<br>周一 | 助教课-2<br>知识点查漏补缺  | 2  |
|                        | 7 月 26 日<br>周二 | 教授课-3<br>交叉学科课程知识点学习<br>学习目标：文本分类与语义和情感分析   | 3  |

|                        |                |  |     |
|------------------------|----------------|--|-----|
| 第二周<br>教授及助教辅导         |                | 描述：本课程涵盖的主题包括分类指标，混淆矩阵，Scikit-Learn使用情况，以及 Spacy 的语义和单词向量，NLTK 的情感分析等。   |     |
|                        | 7 月 27 日<br>周三 | 助教课-3<br>知识点查漏补缺&<br>跟进小组项目调研进度  | 2   |
|                        | 7 月 29 日<br>周五 | 教授课-4<br>互动与项目设计跟进答疑   | 1.5 |
|                        | 7 月 30 日<br>周六 | 助教课-4<br>跟进小组项目调研进度  | 2   |
|                        | 7 月 31 日<br>周日 | 教授课-5<br>交叉学科课程知识点学习<br>学习目标：深度学习与数据分析<br>描述：本课程涵盖的主题包括深度学习基础知识的介绍，例如激活和成本函数,使用 CNN 和 RNN 进行文本分类，以及使用 Numpy 和 Pandas 库进行数据分析等。 | 2   |
| 第三周<br>教授及助教辅导<br>未来展望 | 8 月 2 日<br>周二  | 助教课-5<br>跟进小组项目调研进度  | 2   |
|                        | 8 月 3 日<br>周三  | 教授课-6<br>交叉学科课程知识点学习<br>学习目标：舆情监控<br>描述：通过本模块，学生将了解了解舆情监控的基础知识。本课程涵盖的主题包括舆情监控系统的介绍与知识，以及高水平的设计和限制等。                            | 2   |
|                        | 8 月 5 日<br>周五  | 助教课-6<br>知识点查漏补缺&<br>指导小组项目成果展示  | 2   |
|                        | 8 月 6 日<br>周六  | 教授课-7<br>教授点评小组项目成果  | 1.5 |
|                        | 8 月 7 日<br>周日  | 升学与就业方向展望  | 1   |
|                        |                | 个人规划及发展建议  | 1   |
| 总课时                    | 32             |  |     |

#5. 阅读材料



- ★ Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. Author: Sowmya Vajjala, Bodhisattwa Majumdar, Anuj Gupta, Harshit Surana. O'Reilly Publications, 2020.
- ★ Natural Language Processing in Action: Understanding, analyzing, and generating text with Python. Author: Hobson Lane, Hannes Hupke, Cole Howard. Manning Publications, First edition, 2019.
- ★ Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit (NLTK). Author: Steven Bird, Ewan Klein, and Edward Cooper. O'Reilly Publications, 2009.

#6. 项目主题



本课程使用 PBL 教学法，PBL 即项目式学习，是一种以学生为中心的教学方法，教师提供关键素材构建学习环境，学生组建团队通过在此环境里解决一个开放式项目的经历来学习。以下为本课程可选的项目主题：

- 设计 Twitter 情绪分析
- 设计亚马逊评论分类
- 设计自动文本摘要
- 使用 Spacy 进行简历和简历解析
- 自选：学生选择的主题

英文版教学大纲



|                   |   |
|-------------------|---|
| Course Title      | Natural Language Processing and Public Opinion Monitoring   |
| Credit Hours      | 32 (one credit hour is 45 minutes)  |
| Course Objectives | At the end of this class, the students will learn: <ul style="list-style-type: none"><li>★ How to work with text and pdf files in Python</li><li>★ How to leverage regular expressions for pattern searching in Python</li><li>★ How to leverage Spacy for ultra-fast tokenization</li><li>★ How to leverage stemming and</li></ul> |



|                    |   |
|--------------------|---|
|                    | <p>lemmatization in NLP</p> <ul style="list-style-type: none"><li>★ How to leverage Spacy for vocabulary matching</li><li>★ How to leverage part of speech tagging to process raw text files</li><li>★ How to leverage named entity recognition</li><li>★ How to design a system for public opinion monitoring</li><li>★ How to leverage NLTK to conduct sentiment analysis of text</li><li>★ How to work leverage basic NLP to solve real world problems</li><li>★ How to leverage Numpy and Pandas libraries for data analysis</li></ul>  |
| Course Description | <p>This course is an introduction to natural language processing with Python and will cover all the basics to set you on a path to become a world-class practitioner. We'll start off with a refresher of Python, learning how to open and work with text and PDF files, as well as learning how to use regular expressions to search for custom patterns inside of text files. Afterwards we will begin with the basics of Natural Language Processing, utilizing the Natural Language Toolkit library for Python, as well as the state-of-the-art Spacy library. We'll understand fundamental NLP concepts such as stemming, lemmatization, stop words, phrase matching, tokenization and more! We will then learn part of speech tagging, text classification, and semantic &amp; sentiment analysis. We will then apply the concepts learned to solve real world problems including spam detection, restaurant review classification, and IMDB reviews classification. We will also cover deep learning basics and introduce data</p> |

|  |  |
|--|--|
|  | analysis using Numpy and Pandas libraries. We will then cover how these techniques can be applied for developing a public opinion monitoring system. |
|--|--|

**Brief introduction of the course**

In today’s interconnected world, people’s ideas, thoughts, opinions, and activities transfer into the digital world generating large amounts of data. Natural Language Processing uncovers the insights hidden in these word streams which can be used in innovative ways to improve businesses, services, governance, and the society. Natural Language Processing is a branch of computer science and artificial intelligence that helps computers interpret the natural languages spoken by humans. This course introduces the fundamental concepts in NLP such as text processing, morphological analysis, syntactical analysis, and analysis of lexicological, relational, and topical semantics. The course also covers a few real-world examples to demonstrate how to apply the techniques learned. Finally, the architecture of a public opinion monitoring system using NLP is discussed.

|                 |   |
|-----------------|---|
|                 | <b>Topics</b>   |
| <b>Module 1</b> | Objective: Course Introduction<br>Description: Topics covered include course overview, logistics, instructor introduction, software installations and setup   |
| <b>Module 2</b> | Objective: Python Refresher<br>Description: Topics covered include review of python syntax, parsing and processing text and pdf, regular expressions  |
| <b>Module 3</b> | Objective: Natural Language Processing Fundamentals<br>Description: Topics covered include introduction to NLTK and Spacy libraries and NLP techniques such as stemming, lemmatization, stop words, phrase matching, tokenization |
| <b>Module 4</b> | Objective: Part of Speech Tagging and Named Entity Recognition<br>Description: Topics covered include part of speech tagging to automatically process raw text files, and understanding & visualizing named entity recognition    |
| <b>Module 5</b> | Objective: Text Classification<br>Description: Topics covered include classification metrics, confusion matrix, Scikit-Learn usage etc  |
| <b>Module 6</b> | Objective: Semantic and Sentiment Analysis  |

|           |  |
|-----------|--|
|           | Description: Topics covered include semantics and word vectors with Spacy, sentiment analysis with NLTK  |
| Module 7  | Objective: Real World Applications<br>Description: Working examples of three mini projects:<br>(a) spam message classification,<br>(b) restaurant review prediction, and<br>(c) IMDB review classification   |
| Module 8  | Objective: Deep Learning Basics<br>Description: Topics covered include introduction to deep learning basics such as activation and cost functions, text classification with CNN and RNN  |
| Module 9  | Objective: Data Analysis<br>Description: Topics covered include data analysis with Numpy and Pandas libraries  |
| Module 10 | Objective: Public opinion monitoring & Course Projects<br>Description: Topics covered include<br>(a)introduction to public opinion monitoring system, high level design & constraints<br>(b)Overview of course projects, business problems, available datasets, and tips |

Required Readings

- 1.Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. Author: Sowmya Vajjala, Bodhisattwa Majumdar, Anuj Gupta, Harshit Surana. O'Reilly Publications, 2020.
- 2.Natural Language Processing in Action: Understanding, analyzing, and generating text with Python. Author: Hobson Lane, Hannes Hupke, Cole Howard. Manning Publications, First edition, 2019.
- 3.Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit (NLTK). Author: Steven Bird, Ewan Klein, and Edward Cooper. O'Reilly Publications, 2009.

Suggested list of the topics for the final project

- 1.Twitter Sentiment Analysis
- 2.Amazon Review Classification
- 3.Automated Text Summarization
- 4.Resume and CV Parsing using Spacy
- 5.Custom: Student selected topic



**Criteria**

A successfully working prototype

**Class Expectation**

This class is an introduction to Natural Language Processing and no previous knowledge is expected. The only pre-requisite is basic working knowledge of Python. A Python refresher is added to ground all students with basics. The material is very hands on, with plenty of working examples. A high-level design of a public opinion monitoring system is discussed, but no implementation is included, however students can use course project to build one using the knowledge gained in this course.