人工智能之最新自然语言处理技术与实战

- **课程介绍：**

  自然语言处理(简称 NLP)是计算机科学和人工智能研究的一个重要方向，研究计算机和理解和运用人类语言进行交互的问题，它是集语言学、机器学习、统计学、大数据于一体的综合学科。
  本课程主要介绍了 NLP 中的常用知识点：分词、词法分析、句法分析、向量化方法、经典的 NLP 机器学习算法，还重点介绍了 NLP 中最近两年来基于大规模语料预训练的词嵌入模型及应用。同时本课程偏重于实战，不仅系统地介绍了 NLP 的知识点，还讲解如何实际应用和开发，每章节都有相应的实战代码。

- **课程时间：4 天**

- **学习对象**

  1.希望从事 NLP 工作的 IT 技术人员、开发人员等。
  2.高校、科研院涉及 NLP 工作的学生和研究人员。

- **学习目标：**

  1.掌握 NLP 基础；
  2.分词；词法、句法分析
  3.文本向量化
  4.HMM 与 CRF
  5.基于深度学习 NLP 算法；
  6.神经语言模型
  7.词嵌入方法
  8.基于大规模语料预训练的词嵌入

- **课程大纲**

  第一天：传统的 NLP
  一、NLP 基础知识
  1、自然语言处理简介
  2、中文 NLP 的主要任务
  3、常见的 NLP 系统
  4、NLP 的研究机构与资源

二、中文分词
1、基于字符串匹配的分词
2、统计分词法与分词中的消歧
3、命名实体识别
4、常用分词工具：JIEBA
三、文本的相似性
1、VSM
2、TF-IDF
3、初步情感分析
四、隐马尔科夫模型
1、形式化定义
2、三个问题
3、评估问题与向前向后算法
4、解码问题：维特比算法
5、学习问题：Baum-Welch 算法
五、条件随机场
1、最大熵原理
2、无向图模型
3、最大团上的势函数
4、工具：CRF++

第二天：从传统到现代
一、从 LSA 到 LDA
1、LSA 与 SVD 分解
2、pLSA
3、LDA
二、神经网络语言模型
1、维数的诅咒
2、n-gram 语言模型
3、NNLM 的具体实现
4、改进的思路
三、word2vec
1、one-hot 与 Distributed
2、CBOW
3、skip-gram
4、Hierachical Softmax
5、Negative Sampling
四、循环神经网络（RNN）
1、RNN 的基础架构
2、RNN 的示例
3、LSTM
4、GRU

第三天：预训练模型之一（变形金刚、芝麻街、独角兽及其他）

一、GloVe
1、与 word2vec 的区别
2、统计共现矩阵
3、用 GloVe 训练词向量
二、Transformer
1、所有你需要的仅仅是“注意力”
2、Transformer 中的 block
3、自注意力与多头注意力
4、位置编码（为什么可以抛弃 RNN）
三、三大特征抽取器的比较
1、CNN、RNN 与 Transformer 的比较
2、融合各种模型
四、Elmo
1、双向语言模型
2、工作原理
3、Elmo 的应用场景
五、GPT
1、“一定会有人用它干坏事”
2、GPT 的内部架构
3、Transformer 的演示
4、自注意力机制的改进
5、GPT 的应用场景

第四天：预训练模型之二（站上 BERT 的肩头）
一、BERT 的前世今生
1、之前介绍的模型回顾
2、现代 NLP 的最新应用场景
3、条条大路通 BERT
二、BERT 详解
1、原理与方法
2、BERT 的应用场景
3、BERT 源码简介
三、站在 BERT 肩膀上的新秀们
1、ERNIE
2、XLnet