复盘 3 | 自然语言处理及文本处理核心技术模块

2018-06-03 洪亮劼

AI技术内参 进入课程 >

到目前为止，我们讲完了人工智能核心技术的第三个模块——**自然语言处理及文本处理核心技术**。

整个模块**共 18 期**，**6 大主题**，希望通过这些内容，能让你对自然语言处理及文本处理核心技术有一个全面系统的认识和理解，为自己进一步学习和提升打下基础。今天我准备了 18 张知识卡，和你一起来对这一模块的内容做一个复盘。

点击知识卡跳转到你最想看的那篇文章，温故而知新。如不能正常跳转，请先将 App 更新到最新版本。

LDA 模型

LDA模型的前世今生

LDA模型，开启了“主题模型”这个领域。LDA是一个典型的产生式模型。

判别式模型天生就成为构建分类器或者回归分析的有利工具。而产生式模型更适合做无标签的数据分析，比如聚类。

LDA变种模型

LDA 扩展思路：把扩展的变量设置在上游或者是下游，从而能够对主题信息产生影响或者是受到主题信息的影响；或者把文档放到时间的尺度上，希望去分析和了解文档在时间轴上的变化。

LDA算法优化

LDA训练方法有两个，一个是基于吉布斯采样的随机方法，一个是基于变分推断的确定性方法。

加速吉布斯采样的思路：试图把原始的吉布斯采样公式拆分成好几个组成部分，并且每一个部分所代表数据的变化率是不一样的。

在优化变分推断方面，其实就是希望能够推演出一种类似SGD的变分方法。

基础文本分析

隐语义分析

隐语义分析就是利用矩阵分解的概念对“词–文档矩阵”进行分解。

隐语义分析的核心就是用无监督的方法从文本中提取特性。

在很长的一段时间里，基于SVD（奇异值分解）的隐语义模型都是标准的无监督文本挖掘的核心算法。

概率隐语义分析（PLSA）

PLSA是对文档和里面单词的联合分布进行建模。

从模型的根本特征上来说，PLSA和LDA都是对文档单词分布的一种分解。在建模的核心思想上，PLSA和LDA是一样的。

EM算法

EM算法深深根植于一种更加传统的统计参数方法，最大似然估计（MLE）。绝大多数的机器学习都可以表达成为某种概率模型的MLE求解过程。

EM算法的提出就是为了简化那些求解相对比较困难模型的MLE解。

Word2Vec

Word2Vec的基本形式

Word2Vec的核心思想是，当前的单词是从周边单词的隐含表达，或者说是词向量中产生的。

Word2Vec有两个不太一样的模型，SG和CBOW。本质上，都是希望能够利用文章的上下文信息学习到连续空间的词表达。

Word2Vec的扩展模型

Word2Vec的三个扩展：

- 把学习词向量的工作推广到句子和文章里；
- 把Word2Vec的思想扩展到另外一种离散数据图的表达上；
- 尝试在查询关键词和用户点击的网页之间建立上下文关系，使得Word2Vec模型可以学习到查询关键词以及网页的隐含向量。

Word2Vec在自然语言处理领域的应用

Word2Vec的主要应用：

- 计算词与词之间的相似度；
- 词语的类比；
- 尝试在查询关键词和用户点击的网页之间建立上自然语言处理中依靠“词包”这个输入来执行的任务，都可以相对比较容易地用“词向量”来替代。

基于深度学习的语言序列模型

RNN基础架构

对文字的深层次的理解一定是建立在对序列、对上下文的建模之中。

传统的机器学习序列模型，最经典是“隐马尔科夫模型”。

RNN的整个框架可以看作是一个加码解码的过程。

LSTM与GRU模型

RNN 的基本架构存在一个叫作“梯度爆炸”或者“梯度消失”的问题。解决这个问题的一个途径就是尝试在框架里设计“门机制”。

LSTM的思路是把隐含状态分为两个部分。一部分用来当作“存储单元”，另外一部分当作“工作单元”。

GRU模型的核心思想其实就是利用两套门机制来决定隐含单元的变化。

RNN在自然语言处理中的应用

RNN在自然语言处理中的应用场景：

- 可以利用RNN对句子层级的分类任务进行处理；
- 可以把RNN当作普遍使用的特性提取器来进行分类任务的训练，比如POS标签任务。

经典的对话模型

从方法上，对话系统可以分为“基于规则的系统”和“基于机器学习的系统”。从应用场景上，对话系统也可以分为“基于任务的对话系统”和“非任务的对话系统”。

对话系统的基本架构：自动语音识别器、自然语言理解器、对话管理器、任务管理器、NLG和TTS。

任务型对话系统技术要点

NLU这个组件的目的是把用户输入的文字信息给转换成为任务型对话系统可以理解的内部的表征形式，比如判断意图，其实就是一个多类的分类问题。

现在很多的系统中，DM和TM都是结合在一起进行构建的。在此之上往往有一个叫作“协议学习”的步骤，比较热门的方法是利用深度强化学习来对DM和TM进行统一管理。

聊天机器人技术要点

针对当前的输入，利用之前已经有过的对话进行回馈，这就是基于信息检索技术的对话系统的核心假设。

基于深度学习的对话系统逐渐成为了对话系统建模的主流，其中一个经典模型就是“序列到序列”模型。

文档情感分类

基于监督学习的文档情感分类，往往有两种形式的学习任务：二分或者多类分类问题和次序回归问题。

经实践验证，效果较好的监督学习特性：词频、词类、情感词汇。

基于非监督学习的文档情感分类，其思想核心就是设计一套“打分机制”。

情感“实体”和“方面”的提取

对于文本情感分析而言，“实体”和“方面”是两个非常重要的概念。

常用的提取技术和思路：基于“频率”的提取；利用句子中的一些特殊结构；把信息提取转换成为监督学习任务。

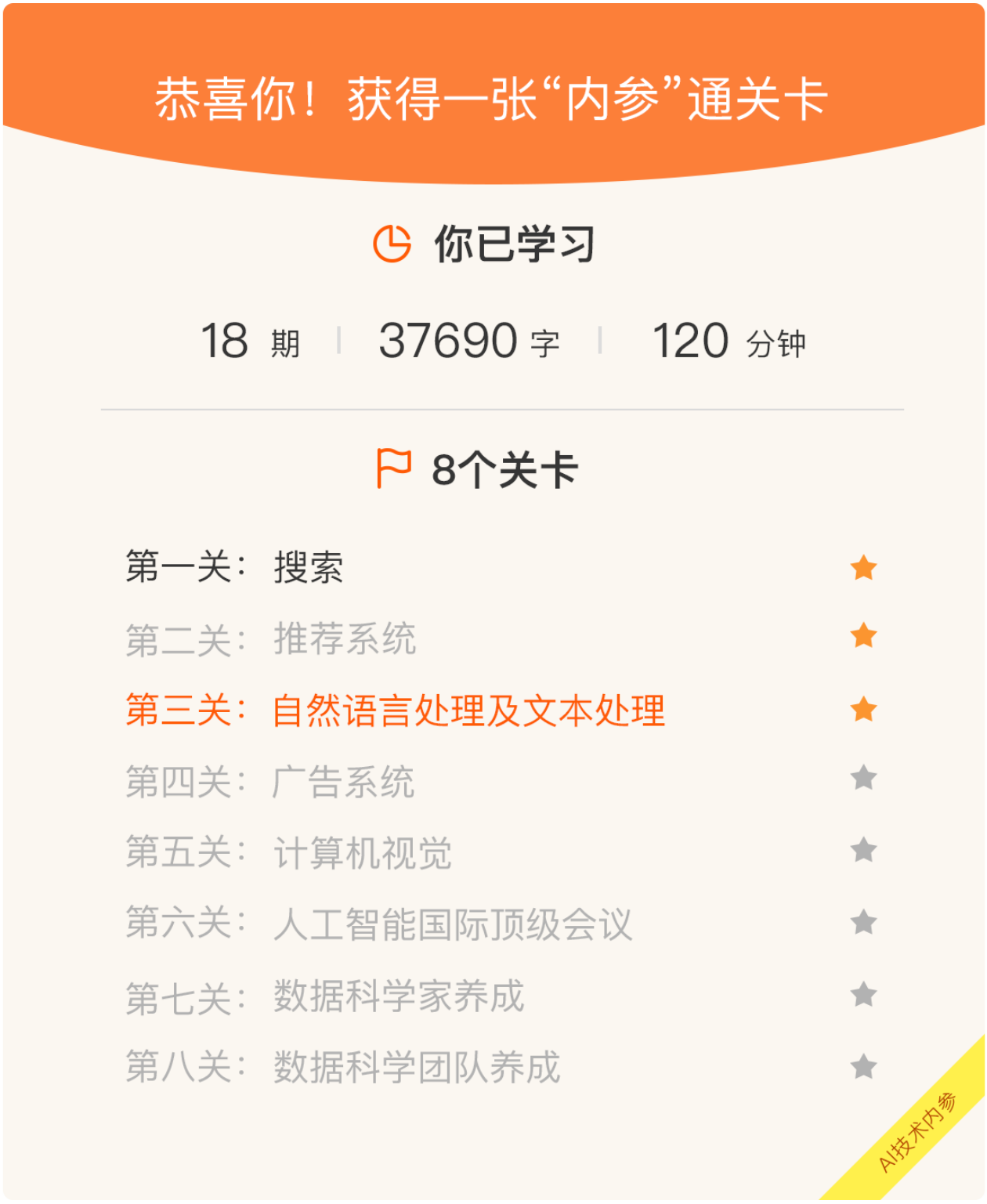
意见总结和意见搜索

基于方面的意见总结有两个特点：

- 这样的总结主要是针对物体的实体以及对应的方面来进行的；
- 意见总结需要提供数量化的总结。

意见搜索的难点在于针对意见信息的索引和检索。

积跬步以至千里



感谢你在专栏里的每一个留言，给了我很多思考和启发。期待能够听到你更多的声音，我们一起交流讨论。

上一篇　104 | 如何快速学习国际顶级学术会议的内容?

下一篇　105 | 广告系统概述

精选留言 (1)

写留言

May
2018-06-03

3

这样的复习很好!

展开