八斗教育-八斗大数据项目实战

八斗学院

第一阶段：开发基础

本部分为基础课程，快速进入信息流推荐领域前，打好开发基础，做充分的前期准备，学完本章节，你可以掌握基本的Python开发能力，运用hadoop、Spark平台应对日常工作中的大数据批处理工作，同时对数据挖掘有个基本理论入门，利用传统机器学习、深度学习工具满足基本的数据训练和预估需求

1、Python语法与数据分析基础

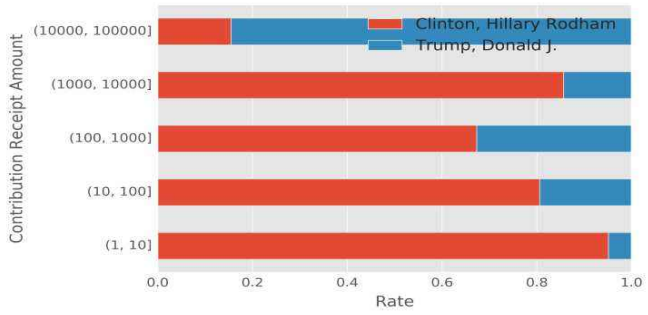
【理论部分】Python时下最流行编程语言之一，在大数据挖掘、机器学习、深度学习中，作为开发首选

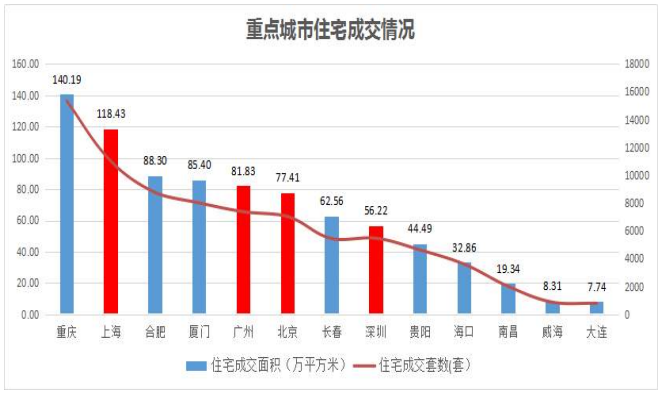
- Python语言基础
- Numpy、Pandas、Series、DataFrame数据分析
- matplotlib图像展示
- Echarts可视化图标展示

【实践部分】通过以下实践体现Python语言的魅力，帮助同学快速熟悉并无枯燥地运用Python语言

八斗学院

项目	描述
《美国大选数据分析》	基于Numpy库完成大选数据统计
《全国房产销量数据可视化》	基于PyEcharts的房产标签可视化展示





2、大数据生态基础

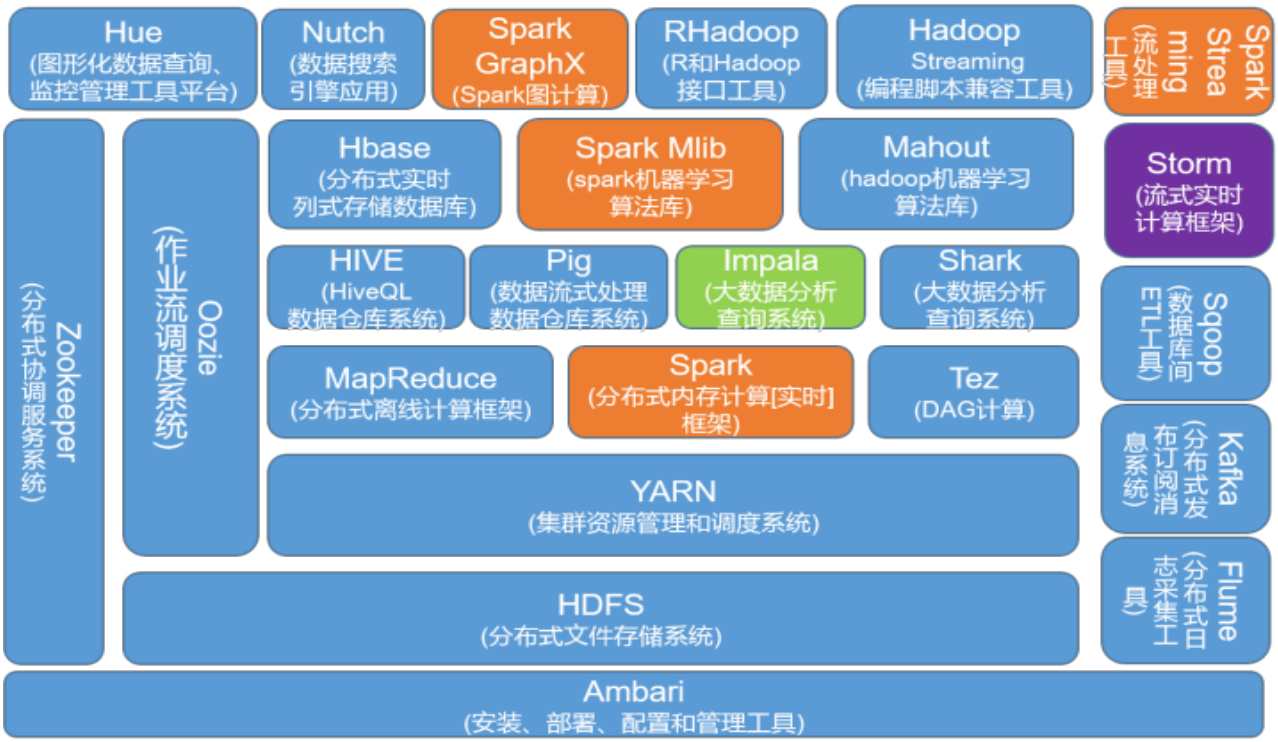
【理论部分】Hadoop及Spark以成为日常工作中存储和处理数据的标准架构

- Hadoop生态学习
- Spark框架

【实践部分】通过以下实践，熟练掌握大数据处理工具

项目	描述
《文本大数据的词频统计》	基于MapReduce、Spark集群统计大规模文本词频
《大规模日志挖掘用户IP映射》	用户IP离线映射服务的批量运算

八斗学院



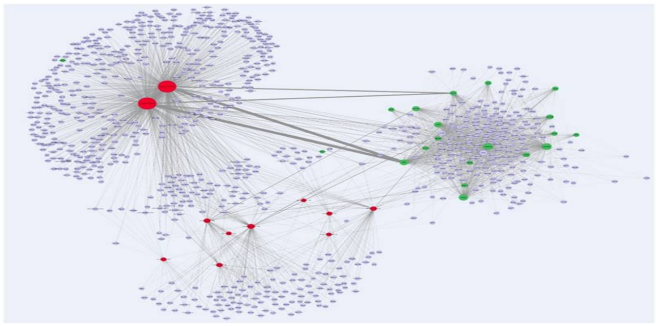
3、机器学习基础概念

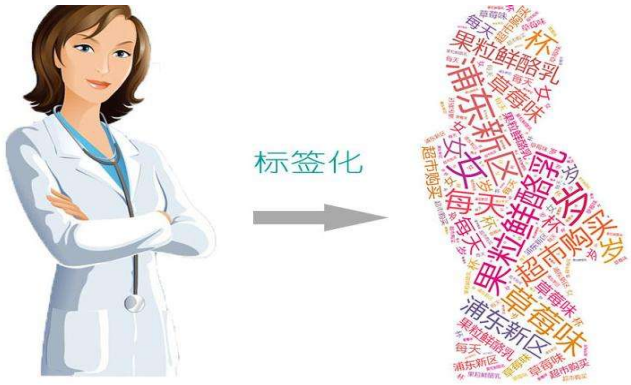
【理论部分】基于文本、视频等信息流的推荐场景中，无论从用户行为分析，还是到物品自然语言表达，机器学习无处不在，已成为该领域人才必不可少的必备技能之一

> - 分类模型(Naive Bayesian、Decision Tree)
> - 聚类模型(Kmeans)
> - Scikit-learn通用机器学习算法库

【实践部分】通过以下实践，快速使用机器学习基础工具，实现一个自己训练的模型

项目	描述
《多标签文本分类模型》	利用LightGBM工具，针对爬虫批量下载的新闻类目，进行模型分类预估
《用户聚类及群体标签化服务》	基于Kmeans算法，针对用户个体行为特征，进行画像提取，挖掘群体内在关联





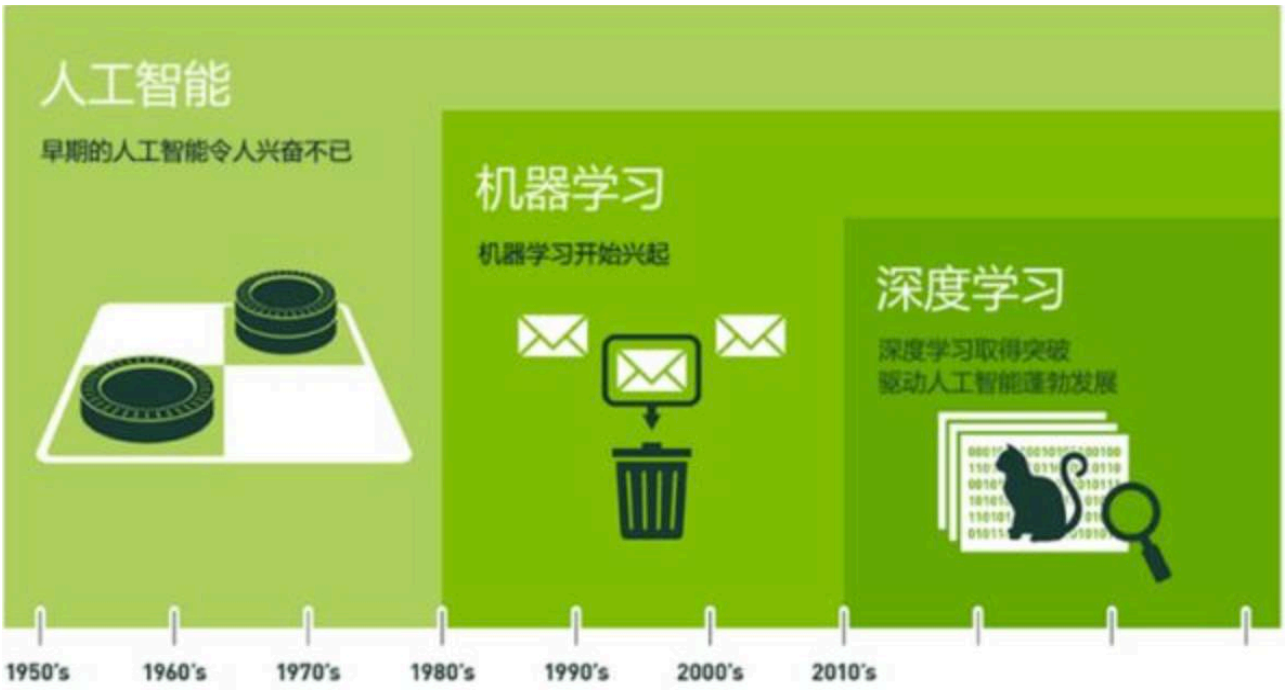
4、Pytorch深度学习基础

【理论部分】深度学习技术，是人工智能（AI）中发展最迅速的领域之一，目前在各大互联网、大数据公司中，均可见到该技术的许多落地应用和场景，日益成为大数据挖掘人才的必备技能

- Pytorch安装，Tensor基础练习
- 前馈神经网络、梯度反向传播

【实践部分】通过以下实践，快速掌握深度学习工具的使用

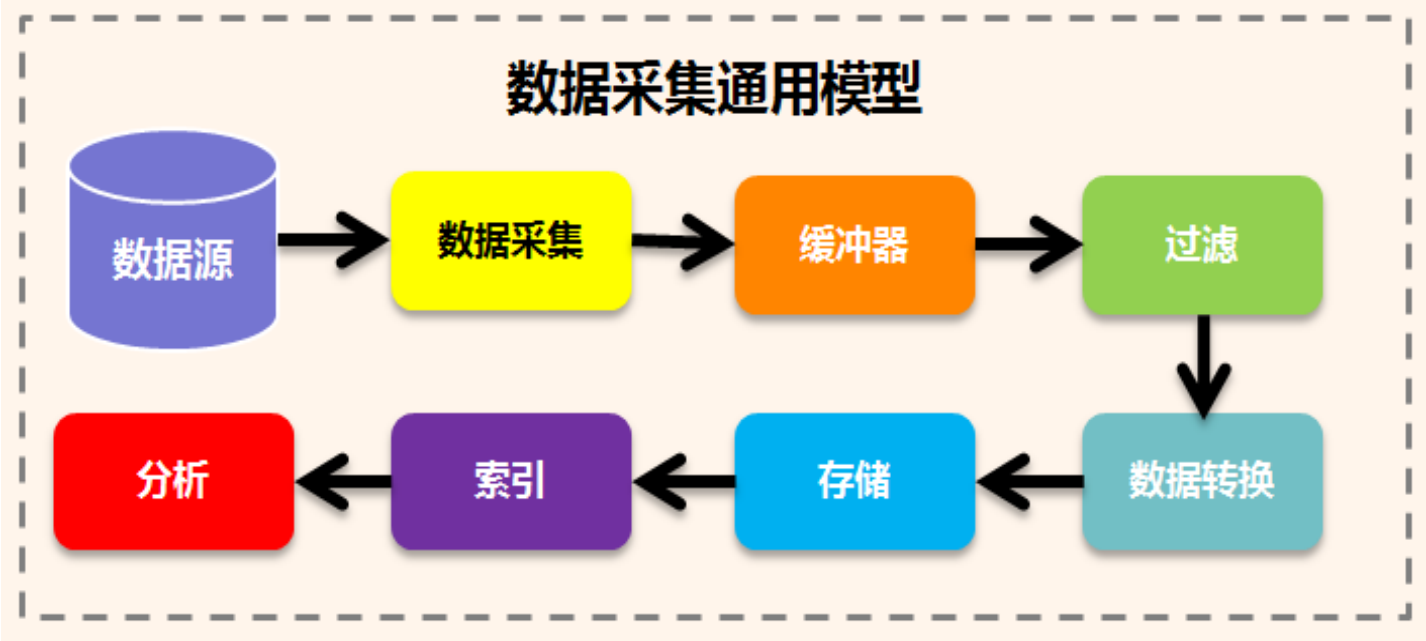
项目	描述
《基于DNN的短视频标题分类》	利用Pytorch实现的DNN分类模型，完成对短视频标题分类任务



八斗学院

第二阶段：用户日志实时收集系统项目实战

伴随着互联网及应用程序规模的不断扩大，各类日志数据呈数量级的倍数增长，同时数据是驱动业务有效实施的根本报障，对用户行为日志的收集和监控，必须采取系统性的应对方案。



1、Flume数据采集工具

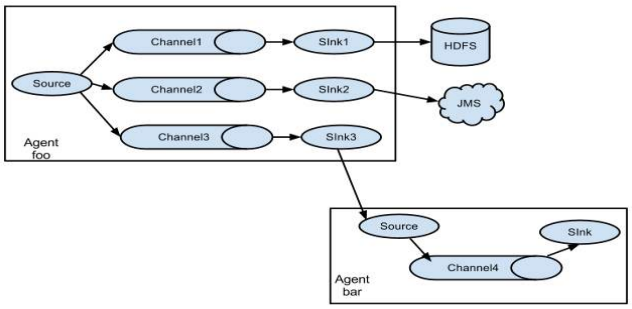
【理论部分】Flume是一个分布式、可靠、和高可用的海量日志聚合的系统，支持在系统中定制各类数据发送方，用于收集数据；同时，Flume提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力

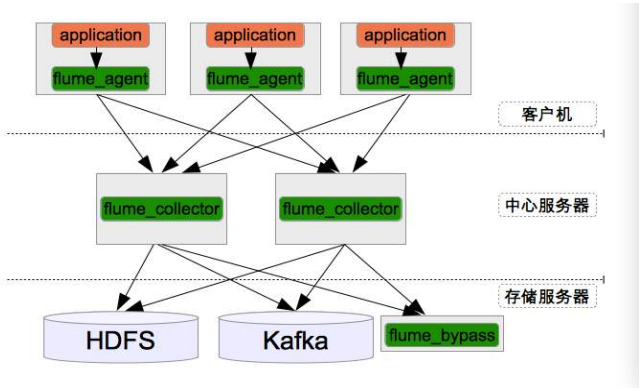
> - Flume架构、核心组件
> - Flume集群搭建

【实践部分】通过以下实践，快速搭建一套Flume分布式集群，实时采集用户行为日志

项目	描述
《基于Flume的用户观影行为日志采集系统》	利用Flume分布式日志采集解决方案，完成用户行为数据收集

八斗学院





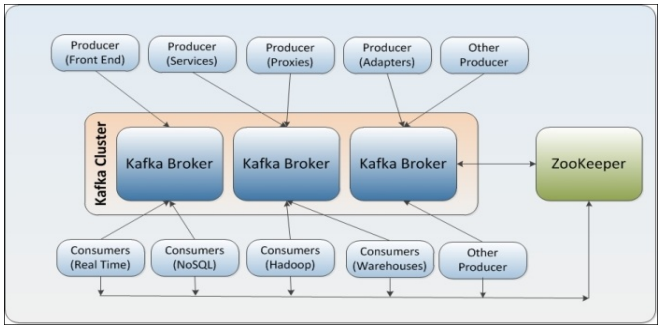
2、Kafka数据消息队列

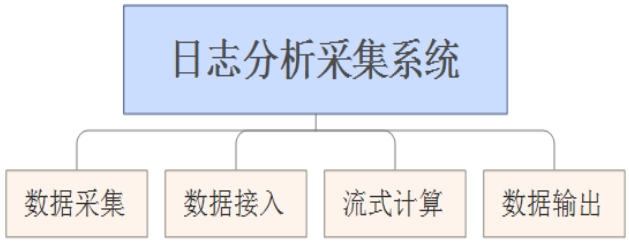
【理论部分】Kafka是一个分布式、支持分区的（partition）、多副本的（replica），基于zookeeper协调的分布式消息系统

- Kafka架构、Topic、Partition等基础概念
- Kafka集群搭建、Producer、Consumer开发

【实践部分】通过以下实践，快速搭建一套Kafka分布式集群，实现Flume+Kafka结合，完成用户行为日志采集+消息传输系统

项目	描述
《基于Flume+Kafka的用户行为日志消息系统》	利用Flume+Kafka分布式解决方案，完成用户行为数据收集





3、实时流Spark Streaming系统

【理论部分】Spark Streaming属于Spark的核心API,它支持高吞吐量、支持容错的实时流数据处理

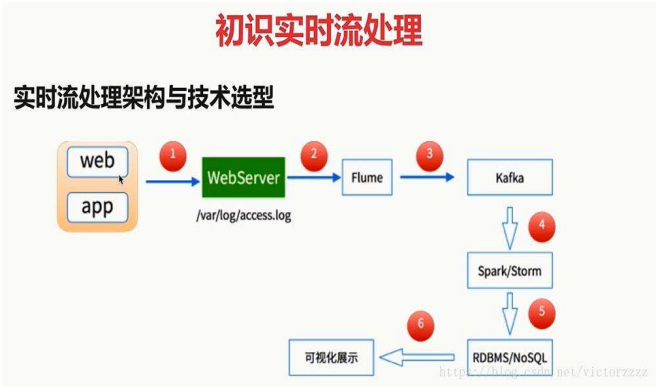
- Streaming&Hbase

八斗学院

- Streaming&SQL

【实践部分】通过以下实践，快速搭建一套Spark Streaming实时消息计算引擎

项目	描述
《基于Streaming的用户实时数据计算系统》	利用Flume+Kafka分布式解决方案，同时对接Streaming实时流计算引擎，打通完整消息流转通路



4、数据分布式批量处理

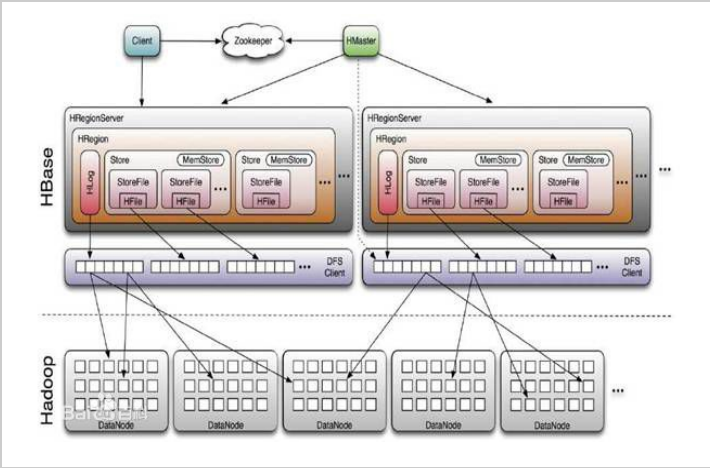
【理论部分】深入Hadoop生态内部，系统学习MapReduce、Spark、Hive、Hbase组件

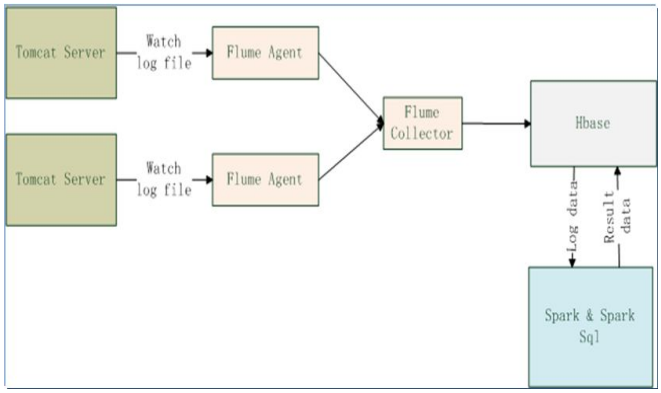
- Spark集群搭建、Spark计算引擎开发
- MapReduce、Hive数据分析
- Hbase基础概念及基本操作

【实践部分】通过以下实践，快速掌握Hbase数据存储技术

项目	描述
《基于MapReduce的视频元信息的批量建库及分析》	使用MapReduce、Spark批处理计算引擎操作Hbase，完成批量数据建库功能，同时配合Hive数据分析

八斗学院





第三阶段：用户数据分析系统项目实战

用户行为日志蕴含着大量有价值的信息，谁能够深入了解用户的行为习惯、兴趣偏好与浏览路径等特征，就能在业务竞争中占据更加有利的地位。

1、信息标签化的中文分词

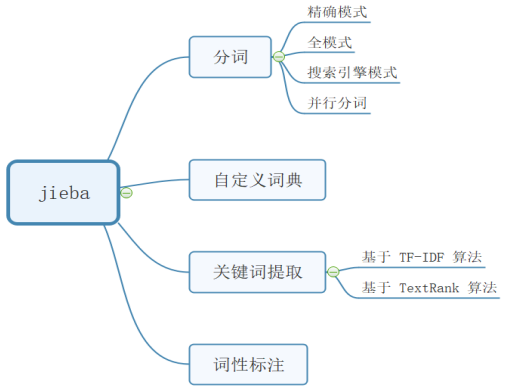
【理论部分】学习分词算法，熟练运用到项目问题中去

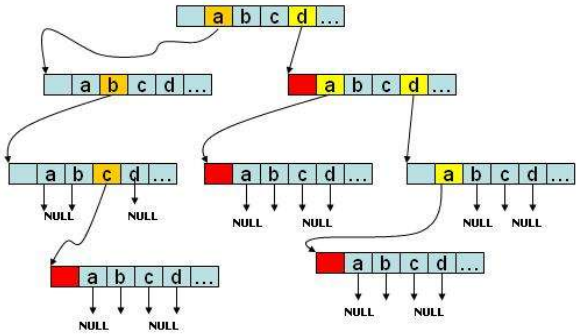
- 手写分词算法
- 第三方Jieba分词工具

【实践部分】通过以下实践，快速掌握中文分词技术

八斗学院

项目	描述
《基于中文分词的文本标签提取》	使用Jieba中文分词工具，结合MapReduce计算引擎，批量完成大规模本章标签提取服务
《基于中文分词的用户画像》	使用Jieba中文分词工具，结合MapReduce计算引擎，完成用户批量标签化功能





2、群体智慧下的人群聚类

【理论部分】学习聚类算法原理和使用，结合人群分组实际的场景，运用算法解决问题

- 基于聚类算法的人群分组
- Kmeans的原理和使用

【实践部分】通过以下实践，快速完成群体分组服务

项目	描述
《基于Kmeans的人群划分》	结合之前的用户画像标签化，得到用户向量化表达，在Kmeans算法运作下，完成群体分组

八斗学院

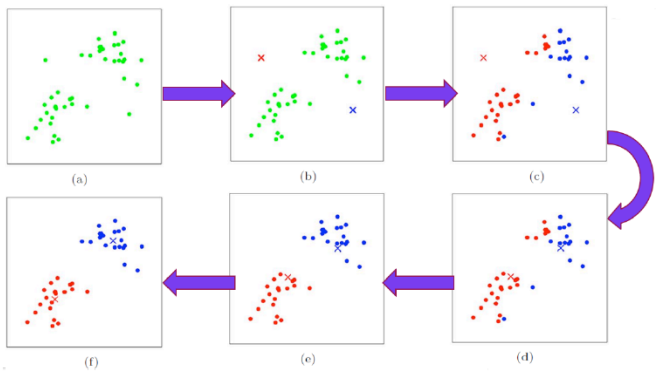
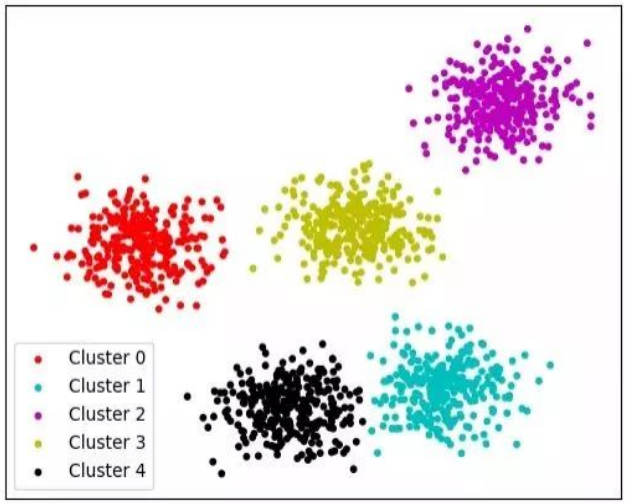


Figure 2: Procedure of K-means



第四阶段：个性化推荐系统实战

随着信息技术的快速发展，“信息过载”问题日趋严重，个性化推荐作为解决信息过载问题的重要方式一直是广大学者和业界的研究热点。

1、基础推荐系统召回算法

【理论部分】掌握基本推荐召回算法，学习ES检索系统

- 协同过滤：User Based、Item Based算法

【实践部分】通过以下实践，快速完成针对海量信息的内容推荐系统

项目	描述
《基于内容的信息推荐系统》	分析物品内容特征标签，借助表示算法完成推荐系统

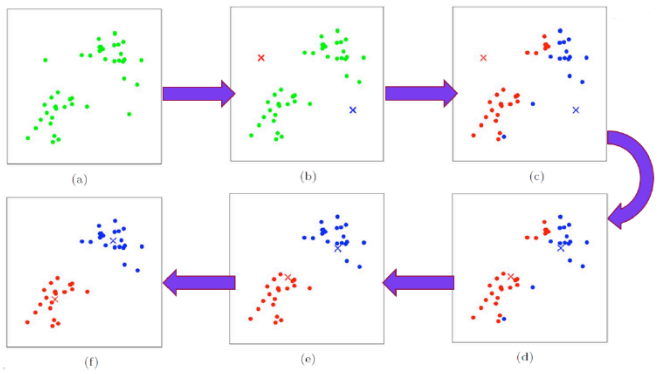
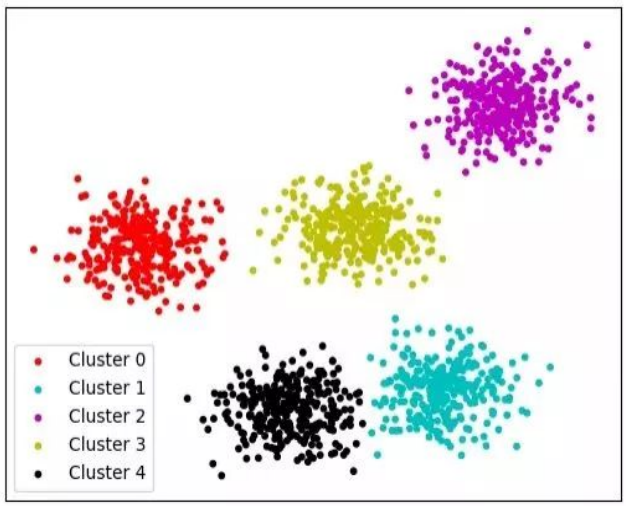


Figure 2: Procedure of K-means



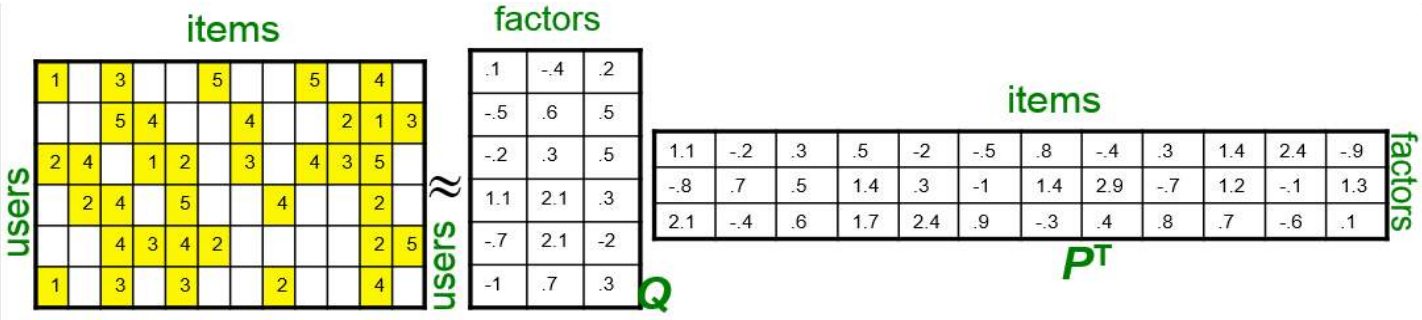
2、基于MF隐因子分解的召回方案

【理论部分】掌握基本推荐MF特征表达，和召回算法

- ALS、SVD等矩阵分解推荐方法

【实践部分】通过以下实践，快速完成基于MF的推荐系统

项目	描述
《基于MF的内容信息推荐系统》	利用MF算法，完成物品推荐系统



3、基于Nearest Neighbor Search的检索服务

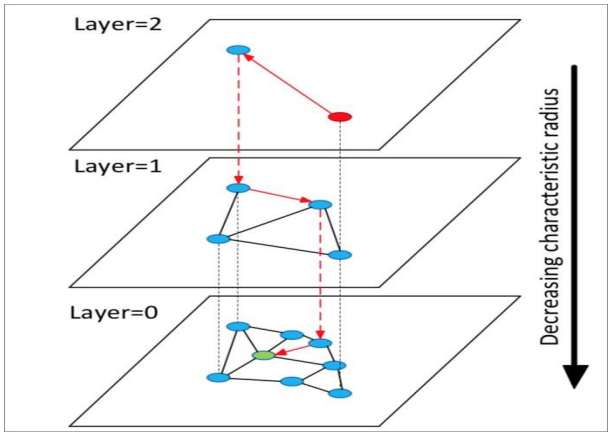
【理论部分】掌握基本ANN检索技术

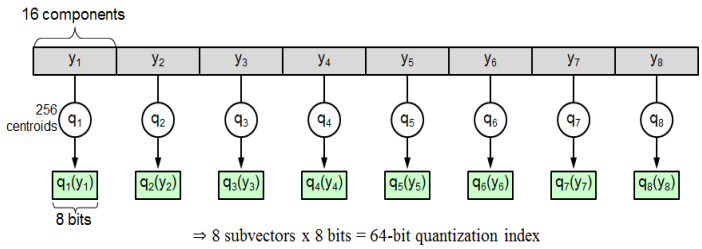
八斗学院

- Annoy近邻召回检索
- 基于HNSW的近邻召回检索

【实践部分】通过以下实践，快速完成基于ANN的检索系统

项目	描述
《基于ANN的内容信息检索系统》	利用ANN算法，完成物品在线检索





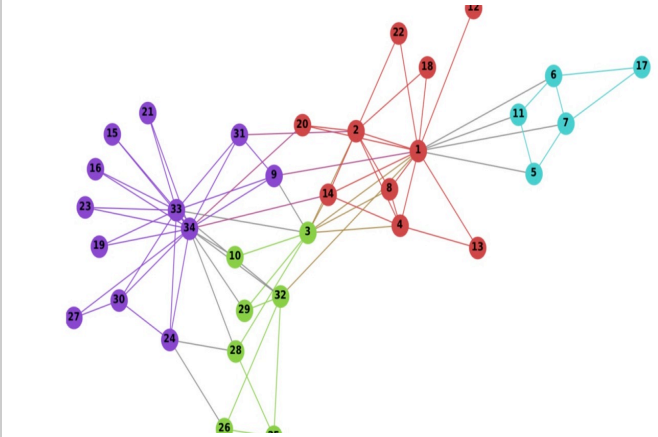
4、基于Graph Embedding的召回策略

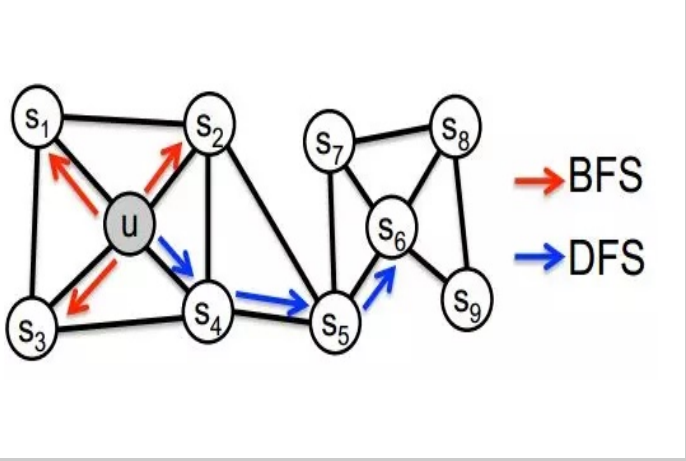
【理论部分】掌握基本图模型技术

- Graph算法原理和应用
- DeepWalk、LINE、Node2vec

【实践部分】通过以下实践，快速完成基于ANN的检索系统

项目	描述
《基于Graph Embedding的物品表示》	利用DeepWalk算法，完成物品向量化特征表示





第五阶段：点击率预估项目实战

点击率(CTR,Click-Through Rate)是用来衡量物品会被用户点击的可能性大小的指标,提高CTR预估的准确性不仅能够提高用户体验，增强物品的推广效果，同时还会带来更多的财富收益，所以如何能更准确预估CTR就成了项目中的重中之重。

1、LR、FM、FFM、DeepFM

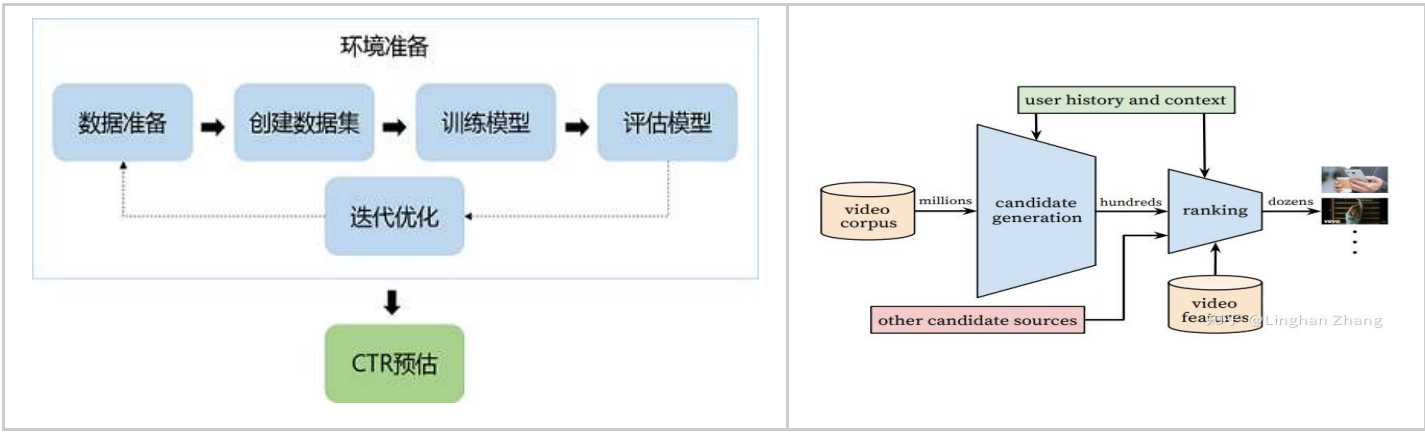
【理论部分】掌握基本排序算法，构建回归模型

- 逻辑回归（LR）、梯度下降(GD)算法原理和应用
- 稀疏数据下的特征组合问题

【实践部分】通过以下实践，快速完成基于LR的模型

八斗学院

项目	描述
《基于LR的点击率模型》	利用LR、FM算法，完成点击率模型



2、DNN、Wide&Deep推荐模型

【理论部分】掌握基于DNN深度学习的点击率模型

- 基于DNN的点击率预估模型

【实践部分】通过以下实践，快速完成基于DNN的点击率模型构建

项目	描述
《基于DNN的点击率模型》	利用Wide&Deep算法，完成点击率模型

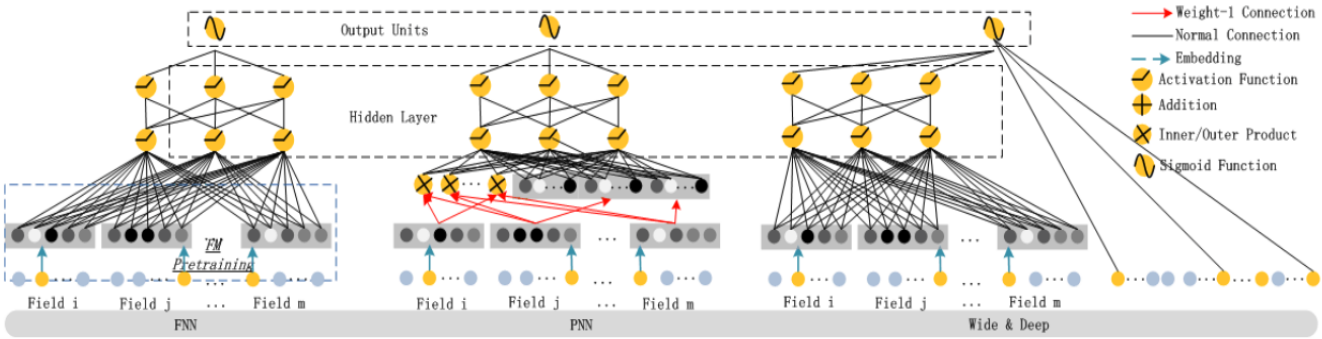


Figure 5: The architectures of existing deep models for CTR prediction: FNN, PNN, Wide & Deep Model

3、实时FTRL在线学习方法

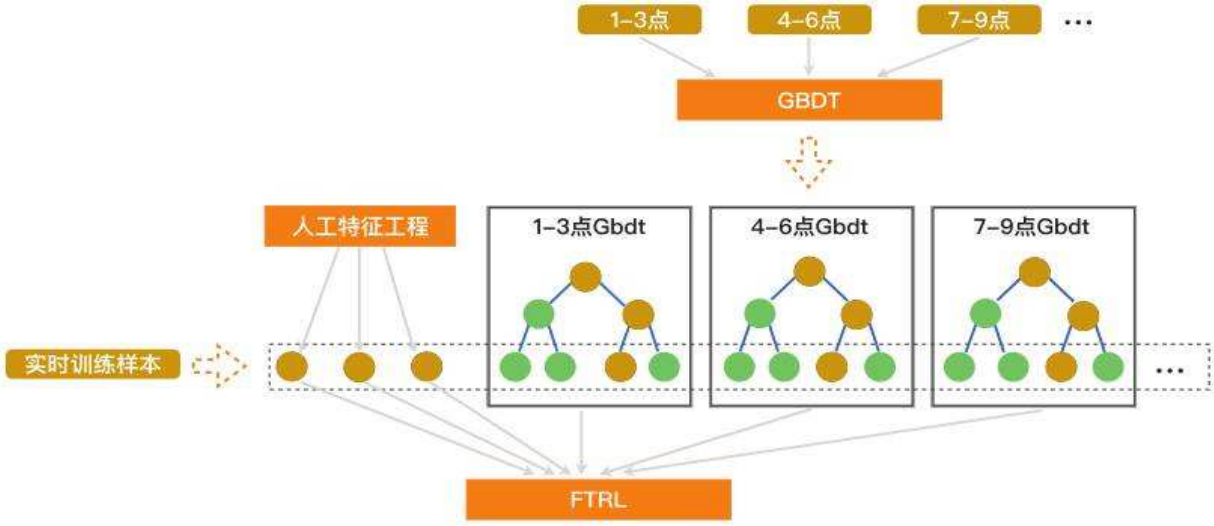
【理论部分】掌握基于FTRL在线模型学习

- 基于FTRL的在线学习模型

八斗学院

【实践部分】通过以下实践，快速完成基于FTRL的点击率模型构建

项目	描述
《基于FTRL的点击率模型》	利用FTRL算法，完成模型在线学习



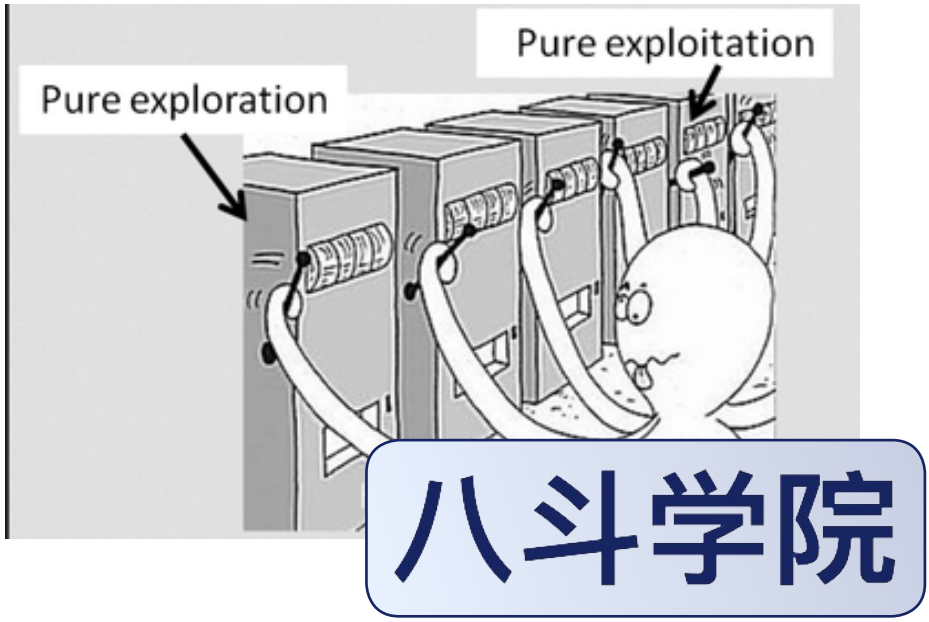
4、推荐系统中的EE和Bandit策略

【理论部分】学习Explore & Exploit算法

> - 推荐系统中的探索和利用问题

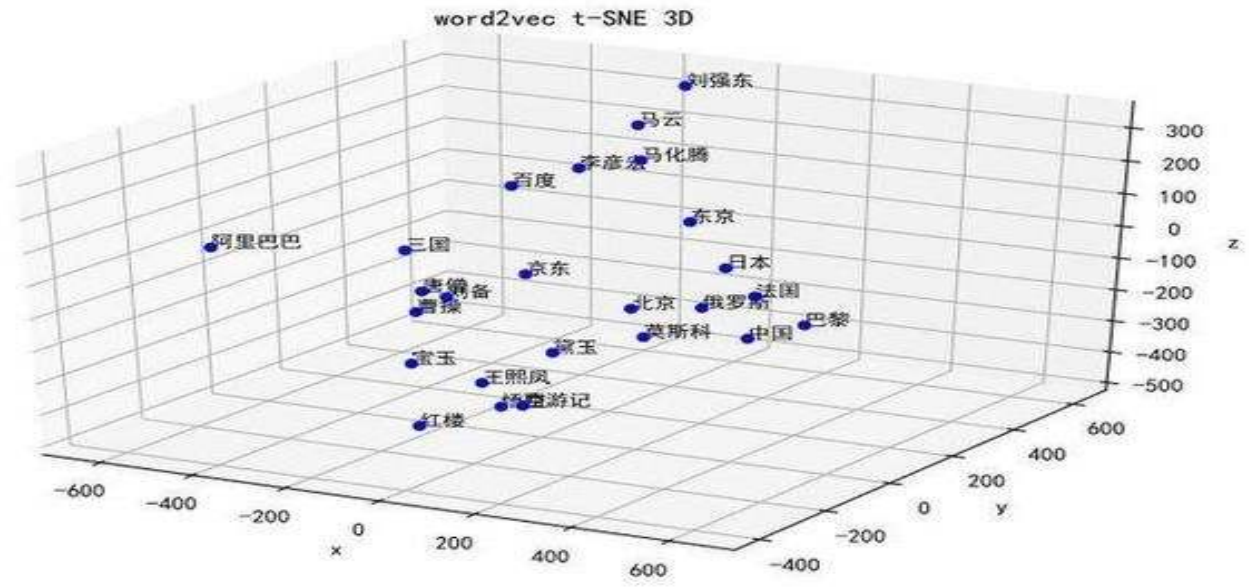
【实践部分】通过以下实践，快速完成基于EE的物品选择模型

项目	描述
《基于EE的物品选择模型》	基于EE算法的候选召回模型



第六阶段：文本表征学习项目实战

Representation Learning文本表示是自然语言处理的基础工作,是信息检索、文本分类、问答系统的关键问题。



1、基于Word2Vec的文本特征提取方案

【理论部分】学习自然语言表示模型构建与应用

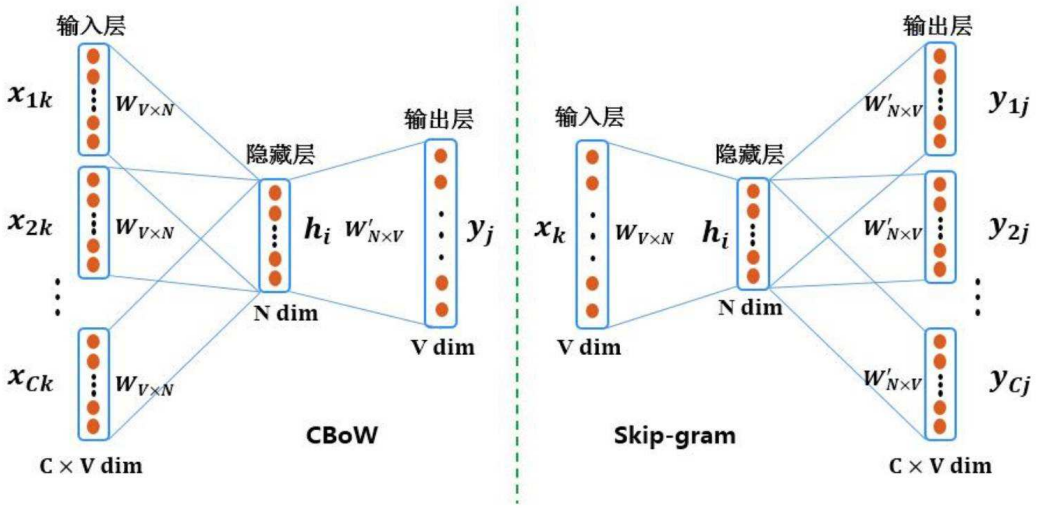
- Skip-Gram原理

【实践部分】通过以下实践，快速完成自然语言表示模型构建

项目	描述
《基于Word2Vec的文本相似度计算》	利用Word2Vec算法，完成文本相似度计算

八斗学院

图 3 CBoW 模型和 Skip-gram 模型图示



资料来源：盈灿咨询

2、基于RNN的文章自动生成系统

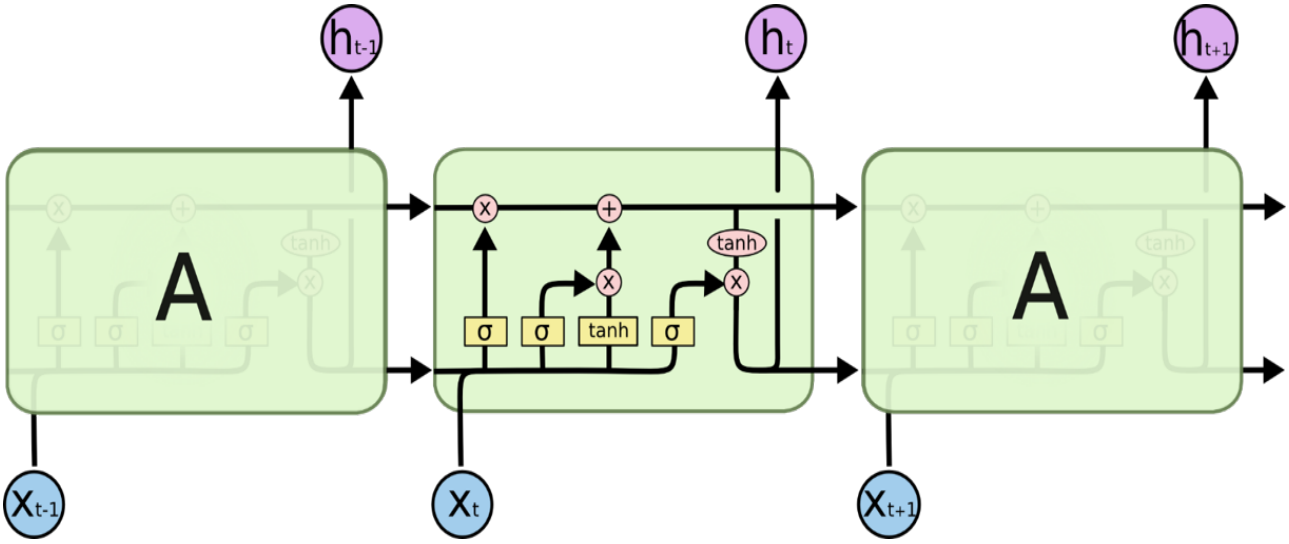
【理论部分】学习LSTM原理，并了解实际业务中的应用

八斗学院

- LSTM原理和应用

【实践部分】通过以下实践，完成基于LSTM的自动写书

项目	描述
《基于LSTM的自动文本书写功能》	使用LSTM，完成文本自动生成



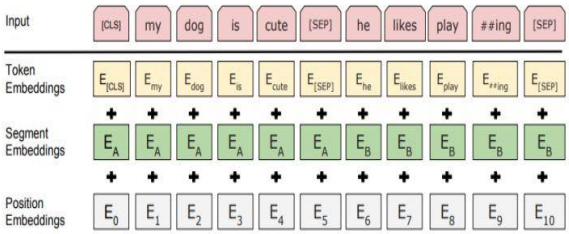
3、深度语义理解Bert模型

【理论部分】学习Transform原理，Bert技术

> - 深度语义挖掘模型

【实践部分】通过以下实践，完成基于Self-Attention技术的文本分类工作

项目	描述
《基于Bert的文本分类》	使用Bert技术，提高文本分类模型的效果



八斗学院