# 搜索引擎基本概念

王树森

http://wangshusen.github.io/


查询词（query）


深度学习|
深度学习教程
深度学习框架
深度学习算法
深度学习pytorch
深度学习的基本原理
深度学习工作站
深度学习电脑配置
深度学习公开课
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
123
空格
搜索

查询建议
(SUG)


深度学习
深度学习教程
深度学习框架
深度学习算法
深度学习pytorch
深度学习的基本原理
深度学习工作站
深度学习电脑配置
深度学习公开课
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
123
空格
搜索


深度学习|
深度学习教程
深度学习框架
深度学习算法
深度学习pytorch
深度学习的基本原理
深度学习工作站
深度学习电脑配置
深度学习公开课
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
123
空格
搜索


查询词
最新
最热
视频
文档 1
文档 2
文档 3

搜索结果页


深度学习|
深度学习教程
深度学习框架
深度学习算法
深度学习pytorch
深度学习的基本原理
深度学习工作站
深度学习电脑配置
深度学习公开课
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
123
空格
搜索


查询词
最新
最热
视频
标签/筛选项
文档 1
文档 2
文档 3


深度学习
深度学习教程
深度学习框架
深度学习算法
深度学习pytorch
深度学习的基本原理
深度学习工作站
深度学习电脑配置
深度学习公开课
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
123
空格
搜索


查询词
最新
最热
视频
文档 1
文档单列曝光
文档 2
文档 3


深度学习|
深度学习教程
深度学习框架
深度学习算法
深度学习pytorch
深度学习的基本原理
深度学习工作站
深度学习电脑配置
深度学习公开课
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
123
空格
搜索


查询词
最新
最热
视频
文档 1
文档单列曝光
文档 2
文档 3


查询词
最新
最热
视频
文档 1
文档 2
文档双列曝光
文档 3
文档 4

# 曝光和点击


# 文档点击率

- 曝光：用户在搜索结果页上看到文档，就算曝光。
- 文档点击：在曝光之后，用户点击文档，进入文档的详情页。
- 文档点击率：文档点击总次数 / 文档曝光总次数。


# 查询词点击率（有点比）

- 查询词点击：用户点击搜索结果页上任意一篇文档，就算“查询词点击”。
- 查询词点击率（有点比）：查询词点击总次数 / 搜索总次数。
- 查询词首屏点击：用户点击搜索结果页首屏的任意一篇文档，就算“查询词首屏点击”。
- 查询词首屏点击率（首屏有点比）：查询词首屏点击总次数 / 搜索总次数。


# 对比

- 文档点击率：10% 左右。
- 查询词点击率（有点比）：70% 左右。
- 查询词首屏点击率（首屏有点比）：60% 左右。
- 有点比重要性高于文档点击率（Why?）

# 垂搜 vs 通搜

# 垂直搜索

- 垂直搜索（垂搜）：针对某一个行业的搜索引擎。
- 电商搜索：Amazon、淘宝、京东、拼多多……
- 学术搜索：Google Scholar、知网……
- 本地生活搜索：Yelp、大众点评、美团、饿了么……
- 酒店机票搜索：Booking、美团、携程、东航……
- 租售房搜索：Zillow、Redfin、Airbnb、贝壳……
- 招聘搜索：LinkedIn、脉脉、Boss直聘……

# 垂直搜索

- 垂搜的文档普遍是结构化的，容易根据文档属性标签做检索筛选。
  - 电商：可以限定品牌、卖家、价格、颜色。
  - 学术：可以限定关键词、作者、期刊、年份。
  - 本地生活：可以限定类目、商圈、距离。
- 垂搜用户的意图明确。
  - 大众点评用户搜索“寿司”，目的是找寿司餐厅。
  - 淘宝用户搜索“拳击”，目的是找拳击相关的商品。

# 通用搜索

- 通用搜索（通搜）：覆盖面广，不限于一个领域。（例：谷歌、百度、必应、抖音……）
- 文档来源广，覆盖面大。（例：网页、视频、图片、商品、直播、店铺……）
- 没有结构化，检索的难度大。
- 用户使用通搜的目的各不相同，较难判断用户意图。
- 本课程主要研究通用搜索。

# 课程安排

# 课程内容

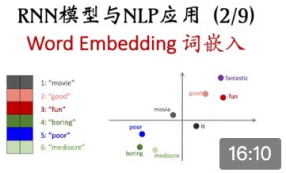
- **基础知识**：用户满意度、评价指标、搜索链路。
- **相关性**：定义与分档、评价指标、文本匹配、语义匹配。
- **查询词处理**：分词、NER、词权重、类目、意图、改写。
- **召回**：文本召回、向量召回、离线召回。
- **排序**：排序模型、训练。
- **查询词推荐**：推词场景、推词召回、推词排序。

# Prerequisite

- NLP & 深度学习（尤其是 Attention 和 BERT）。

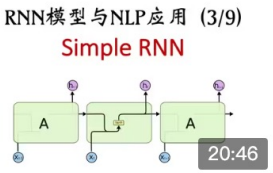

RNN模型与NLP应用(1/9)：数据处理基础

9347 2021-5-15



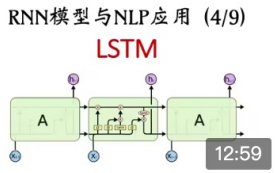

RNN模型与NLP应用(2/9)：文本处理与词嵌入

1.2万 2021-5-15



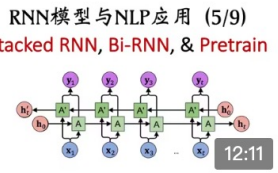

RNN模型与NLP应用(3/9)：Simple RNN模型

9663 2021-5-15



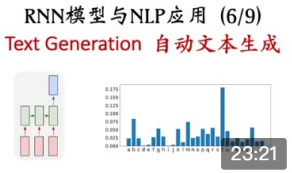

RNN模型与NLP应用(4/9)：LSTM模型

8531 2021-5-15



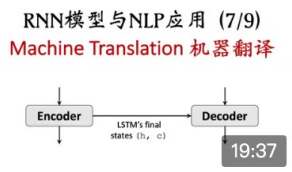

RNN模型与NLP应用(5/9)：多层RNN、双向RNN、预训练

6991 2021-5-15



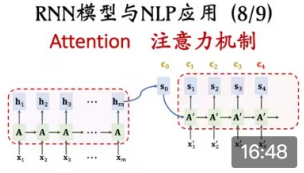

RNN模型与NLP应用(6/9)：Text Generation (自动文本生成)

8742 2021-5-15



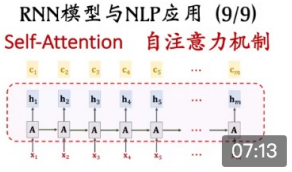

RNN模型与NLP应用(7/9)：机器翻译与Seq2Seq模型

8852 2021-5-15



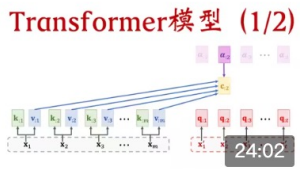

RNN模型与NLP应用(8/9)：Attention (注意力机制)

1.7万 2021-5-15



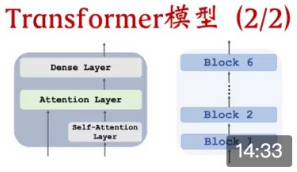

RNN模型与NLP应用(9/9)：Self-Attention (自注意力机制)

1万 2021-5-15



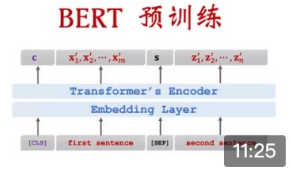

Transformer模型(1/2): 剥离RNN，保留Attention

1.9万 2021-5-13




Transformer模型(2/2): 从Attention层到Transformer网络

1.2万 2021-5-13




BERT (预训练Transformer模型)

1.1万 2021-5-13

# Prerequisite

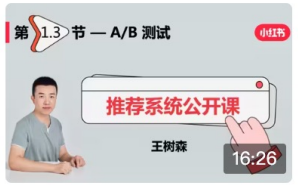
- NLP & 深度学习（尤其是 Attention 和 BERT）。
- 推荐系统的基础知识（A/B测试、向量召回、排序模型、多样性算法）。


概要01：推荐系统的基本概念

2.6万 2022-4-19


概要02：推荐系统的链路

1.2万 2022-4-19




概要03：推荐系统的AB测试

8297 2022-12-24


召回01：基于物品的协同过滤（ItemCF）

1.5万 2022-4-19


召回02：Swing 模型

9514 2022-4-21


召回03：基于用户的协同过滤（UserCF）

8890 2022-4-22


召回04：离散特征处理

8519 2022-4-24


召回05：矩阵补充、最近邻查找

9035 2022-4-26


召回06：双塔模型——模型结构、训练方法

1.6万 2022-4-28


召回07：双塔模型——正负样本

1万 2022-5-1


召回08：双塔模型——线上服务、模型更新

8669 2022-5-2


召回09：双塔模型+自监督学习

9540 2023-2-5

# Prerequisite

- NLP & 深度学习（尤其是 Attention 和 BERT）。
- 推荐系统的基础知识（A/B测试、向量召回、排序模型、多样性算法）。
- YouTube： https://www.youtube.com/c/ShusenWang
- Bilibili： https://space.bilibili.com/1369507485
- GitHub： https://github.com/wangshusen/SearchEngine

# Thank You!

http://wangshusen.github.io/