## 2021

2021山东大学软件学院web数据管理期末考试回忆

一、填空题（顺序记不住了，部分题）

1.DFS比BFS是（）

2.礼貌性是（）

3.网页基于承载内容的分类（）和（）

4.颜色矩是计算颜色通道的（）、（）和（）

5.词项词表的处理步骤 文档解析、（）、（）和（）去除停用词表（个人感觉少了一空 但试卷上确实是3空）去除停用词的方法有（）和（）

6.统计语言模型是（），应用是（）

7.LBP的全称是（），它是（）特征描述子，记录像素点和（）

8.中文分词的方法有 基于NLP的，基于（）的和基于（）的

二、名词解释

1.正则表达式

2.web数据抽取

3.倒排索引

4.TF/IDF

5.BOF

三、问答题

1.反爬虫的策略和爬虫怎么应对

2.网页排序算法有哪些，简述思想

3.HMM模型是什么，基于HMM模型的中文分词思想

4.什么是PLSA，LSA和PLSA的异同

复习的话主要就是结合老师给的范围好好看PPT，基本都是需要背诵记忆的，考前别吃麻辣烫，会拉肚子，很惨，溜了~

————————————————

版权声明：本文为CSDN博主「AlAuAu」的原创文章，遵循CC 4.0 BY-SA版权协议，转载请附上原文出处链接及本声明。

原文链接：<https://blog.csdn.net/AlAuAu/article/details/118184324>

## 2020

一、 填空（部分）：

1.网页预处理可以采取\_\_\_\_\_\_和\_\_\_\_\_\_

2.统计语言模型是\_\_\_\_\_\_\_\_\_，应用有\_\_\_\_\_\_\_\_\_\_

3.爬虫的BFS比DFS好是因为\_\_\_\_\_\_\_\_\_\_

4.CBIR技术的核心是\_\_\_\_\_\_\_和\_\_\_\_\_\_\_\_

5.web爬虫可以用\_\_\_\_\_\_\_判重

二、名词解释

1.web数据抽取

2.HMM

3.web spam

4.布尔检索模型

5.HOG

三、简答：

1.简述反爬虫与爬虫的博弈

2. web爬虫爬取的数据有哪些存储格式？选取的原则？

3.比较bs4和scrapy

3.向量空间模型特点

4.CBIR如何将局部特征转为全局特征

都是老师最后一节课划得考点，没有超纲的内容。填空只要大概能记住整个知识体系就没问题，没有很偏的问题，注意一些概念的定义非常可能出填空题。

————————————————

版权声明：本文为CSDN博主「llwswyn」的原创文章，遵循CC 4.0 BY-SA版权协议，转载请附上原文出处链接及本声明。

原文链接：<https://blog.csdn.net/Departuresna/article/details/108394207>

## 2019

比较坑的填空题:

1.网页预处理可以采取\_\_\_\_\_\_和\_\_\_\_\_\_

2.底层特征和高级语义之间的差距被形象的称作\_\_\_\_\_\_\_\_\_\_

3.Tamura纹理中,\_\_\_\_\_\_\_\_和\_\_\_\_\_\_\_\_和方向性(directionality)对于图像检索尤为重要。

名词解释:

1.web数据抽取

2.HMM

3.LBP

4.链码

5.HOG

简答题:

1.简述反爬虫与爬虫的博弈

2.信息检索技术都有哪些,描述流程

3.向量空间模型特点

3.颜色特征都有哪些?

4.图像检索算法，编码本训练的方法

————————————————

版权声明：本文为CSDN博主「单身狗尊享」的原创文章，遵循CC 4.0 BY-SA版权协议，转载请附上原文出处链接及本声明。

原文链接：<https://blog.csdn.net/qq_44220394/article/details/118182977>