web数据管理

30*1 填空 PPT原句

5*8 名词解释

简答 3*10

没有编程题

一 不考

二 爬虫：6，7，app爬虫分类不考，nutch考

三 与代码有关的不考代码，考技术点知识点。一二三要求高，四要求低一点。第四部分：不考4.7，4.8；

四 考思想和用到的工具

五 第五部分不考，包装器例子不考、1.1，1.2不考，

六 算法的细节不考，规则、观察不考，理解思想；没有公式，不考（自适应算法）

七 重点结构化数据两个例子的细节不考（逻辑回归和决策树等四个工具ctr不考）知道非结构化数据的两种形式（离散化连续化）

八 全看

九 三部分神经网络中，预训练语言模型不考，NNLM只知道是什么，起到什么作用。不考具体模型结构和损失函数。word2vec重点

十 全看

十一 全看 两个问题（**pagerank**、textrank）要记公式

十二

图像检索去掉

图片有哪些特征 不同层次 颜色纹理低级 形状高级

cnn没讲

不同颜色空间转换公式不用记

要记直方图和颜色矩

纹理主要讲的两种：全局（信号处理，主要是思想，滤波器 卷积操作，需要人工设置滤波器，粗略的讲解）和局部（lbp，要求掌握详细/形状特征hog定义思想实现过程（不用公式）整个作用和优缺点,sift没讲完）

题目猜测：

大题：各说两种前端和后端的反爬策略，以及爬虫测略

填空题：

1.DFS比BFS好是因为：

爬虫的BFS比DFS好是因为_____
2.礼貌性是

- Web服务器有显式或隐式的策略控制爬虫的访问

3.网页基于承载内容的分类（数据型）和（文档型）
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。

文档解析、词条化、词项归一化、次干还原、词型归并

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

名词解释：

1.web数据抽取

2.pagerank算法

3.**textRank**算法 https://zhuanlan.zhihu.com/p/126733456

4.比较**fasttext**和word2vec

5.HOG,SIFT

6.TF/IDF

7.RE

8,统计语言模型

9.word2vec的优缺点，以及2大模型特点

10.文档哈希的思想，shingle算法，LSH，minihash

11.LDA

12.

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

4.CBIR如何将局部特征转为全局特征

5.网页排序算法PageRank、HITS、HillTop

6.比较nutch,bs4,

7.pageRank textrank