信息内容安全

课程考核

累加式考核

- 考试成绩：60％
- 实验成绩：20％
- 平时作业：20%

卷面：选择 10个=20分

简答 5道 = 20分 （绪论，捕包，管控，P2P）简单明了

计算 6道 = 50 分 （字符串匹配，分类算法）

设计 1 道 = 10分 （综合设计一个系统）

（设计思路，结构，主要技术）

第一章　信息内容安全概述

（10分）

- 什么是网络空间?网络空间的四个基本要素包括哪些?
- 什么是网络空间安全?网络空间安全学科的研究方向有哪些?
- 《网络安全法》中的主体、客体主要有哪些?试列举各主体的基本责任和义务。
- 什么是网络空间主权?基本原则是什么?(独立平等自主管辖)
- 什么是信息内容安全?

主要涉及对传播信息的有效审查监管，剔除非授权信息（非法信息、泄密信息、垃圾邮件等），保护授权信息

- 信息内容安全技术主要包括哪些?（获取、识别、分析、管控）
- 信息内容安全技术面临的挑战是什么?(数据量大，计算复杂度高，网络技术新，社会矛盾深)

第二章　网络信息获取

（5分）

- 1 网络信息被动获取
  - 网卡的四种接收模式，旁路数据获取需要网卡在混杂模式。
  - 串行和旁路数据获取的区别
  - BPF的原理（在协议栈处理之前拷贝，应用tcpdump）
  - IP头、TCP头、UDP头的关键字段有哪些
  - Libpcap或winpcap捕包的基本流程

第二章　网络信息获取

（5分）

- 高性能捕包
  - 网络数据包由网卡到用户空间进行了几次拷贝。
  - 操作系统消除拷贝的方式（DMA方式，mmp共享内存的原理，ebpf）
  - 网卡设备厂商零拷贝的技术（DPDK）
  - ebpf xdp和DPDK实现零拷贝的区别

第二章　网络信息获取

（10分）

- 3 网络信息主动获取
  - 网络信息搜索系统的一般结构（四个部分）
  - 通用爬虫的一般框架（队列）
  - 单机爬虫抓取算法，多机抓取算法
  - PageRanks算法及本思想，会计算（M矩阵的构建，PR值的推导
  - dead end问题和spider traps问题如何修正
  - 网络信息的主动获取和被动获取的区别

第二章　网络信息获取

（5分）

- 4 社交网络和P2P信息获取
  - P2P 系统结构的分类
  - 各种结构的P2P系统内容发布和检索的方式
  - 结构化P2P的分布式哈希表结构DHT的原理
  - 给定节点数如何构造每个节点的路由表，保证O(n)的时间复杂度可以找到任何一个节点
  - KAD网络节点路由表K桶的构造，节点查询的原理

第三章　字符串匹配

（25-30）

- 1 模式串匹配算法的分类
- 2 单模式匹配算法
- 3 BF算法，了解基本思想，知道时间复杂度
- 4 KMP算法
  - 算法思想，如何控制不回溯
  - next函数， 和文本串无关，寻找最长前缀后缀
  - 给定模式串如何求next数组
  - kmp算法手动推导
  - 如何求nextval数组

第三章　字符串匹配

■5 BM算法

- 算法基本思想，由右向左匹配
- 坏字符原则和好后缀原则
- 坏字符bmbc[]数组的构造
- 好后缀bmgs[]数组的构造
  - ✓ 先计算suffixes数组，找到不同位置i能和后缀匹配上的最大长度
  - ✓ 根据suffixes求好后缀，三种情况
  - ✓ 给定文本串和模式串能计算两个数组，能进行字符串匹配的手动推导，参见作业。

■ 6 单模式匹配算法时间复杂度的比较

第三章　字符串匹配

- 7 多模式匹配算法AC和WM
- 8 了解Trie树结构，AC算法的基础
- 9 AC算法
  - 转向函数g，失效函数f，输出函数output的构造，能手动计算推导。
  - AC算法的时间复杂度分析，初始化时间只和模式集字符数有关，匹配时间复杂度只和文本串的字符数有关。和模式串个数和长度无关。
  - AC算法的优化，内存占用问题
  - 了解行压缩和位图方法的思想
  - 掌握双数组方法，能够进行推导计算。

第三章　字符串匹配

- 10 WM算法
  - 基本思想和 BM算法相似
  - 算法关键的数据结构，SHIFT表，HASH表， PREFIX表，PAT_PTR表的构造
  - 要对模式集按最短的模式截断
  - 选择合适的HASH函数的重要性
  - 能够对WM算法手动计算和推导
- 11 AC 和 WM的性能比较，各自适用的场景
- 12 AC算法并行化处理，如何切分文本
- 13 基于AC双数组的IP地址多模式匹配，能推导
- 14 最大公共子串和最大公共子序列的求解算法，了解算法原理。
- 15 正则表达式匹配算法的思想

第四章 信息内容分析与挖掘

（20-25分）

- 文本分类与文本聚类的区别
- 2 文本表示重点掌握 TF-IDF模型
- 3 jieba分词的原理，了解能自己表述出来
- 4 基于决策树的文本分类方法，能够计算 类别信息熵、属性的信息熵、信息增益、属性分类信息度量、信息增益率。计算能推导一级分支即可。
- 5 基于贝叶斯的分类，能计算朴素贝叶斯分类，掌握多项式模型，能根据多项式模型进行文本分类。

第四章 信息内容分析与挖掘

- 6 支持向量机分类svm，掌握基本思想和原理就可以，知道核函数的作用。
- 7 KNN分类算法，掌握基本思想和原理就可以
- 8 k-means聚类算法 能计算，推理分类
- 9 基于密度的聚类DBSCAN，了解什么是核心点，边界点和噪音点，知道两个关键参数的含义，能计算推理和分类
- 10 基于层次的聚类，两种模型，能够计算和聚类

第五章 信息内容安全管理

（5-10分）

- 信息内容安全管理的目标是什么：剔除非授权信息，保护授权信息。
- TCP重置攻击，RST攻击的原理，seq、ack和滑动窗口的作用，能够计算rst报文的seq值
- P2P网络的管控方法
  - 索引污染和索引毒害的区别
  - 资源占用攻击的原理
  - 数据欺骗攻击的原理
  - 数据块污染攻击的原理
  - eclipse攻击的原理

第五章信息内容安全管理

- 隐私保护技术（5-10分）
  - 隐私数据不同属性的分类：显示标识符、准标识符、敏感属性
  - 基于K匿名的隐私度量方法，掌握k-anonymity、l-diversity和t-closeness的基本概念，各度量能解决的问题，了解即可。
  - 基于差分隐私的度量方法，能解决的问题，了解即可。
  - 常用的隐私保护技术的分类，熟练掌握四种类别，各种隐私保护方法的分类归属。
  - 同态加密技术的原理，了解即可
  - 联邦学习的原理，横向、纵向适用的场合。
  - 基于位置的隐私保护，攻击模型和保护方法了解即可

- 答疑时间地点：

  周一，周三上午，周二，周四，周五下午 格物楼二楼办公室

- 其他时间可以QQ提前联系
- 考试之前作业和实验报告必须提交