# 浙江工业大学之江学院

# 期终试卷样卷

二级学院：　　理学院　　专业名称：信计专业

课程名称：爬虫与 Web 数据挖掘　课程代码：30800500

主讲教师：　　宋丛威

| 题序 | 一 | 二 | 三 | 四 | 总分 |
|---|---|---|---|---|---|
| 计分 | | | | | |

一、填空题 (每空 2 分, 共 20 分):

1) 向网站发出请求的方法有_____、_____;

2) 写出匹配国内 13 位电话号码的正则表达式_____;

3) 写出至少两种常见 HTML 标签名_____、_____;

4) 用于存储网页数据的数据库有_____、_____;

5) BeautifulSoup 常用 HTML 标签搜索方法有_____、_____;

6) 统一资源定位符简称_____;

二、判断题 (每空 2 分, 共 10 分):

1) Python 有大量优秀的第三方库进行网络编程; (　　)

2) Python 只适合数据分析，不适合设计网络爬虫; (　　)

3) requests 与 beautifulSoup 的组合经常与用于 Python 的网络爬虫设计; (　　)

4) 下载电影用迅雷比用 Python 写爬虫更方便; (　　)

5) beautifulSoup 其实是一个 HTML 语言的解析器，将 HTML 翻译成 Python 的数据结构; (　　)

三、选择题 (每空 2 分, 共 10 分):

1) 号称 Python 网络爬虫编程绝配是一组合? (　　)
(A) 美女与野兽 (B) 牛奶与巧克力 (C) requests 与 bs4 (D) urllib 与 re

2) 下面哪一种是常用的反爬策略? （　　）
(A) 更换电池 (B) 更换 WiFi (C) 增加头信息 (D) 增加元信息

3) 下面哪个是 Python 的网络爬虫框架; （　　）
(A) beautifulSoup (B) requests (C) re (D) scrapy

4) 下面哪个网络状态码表示请求成功; （　　）
(A) 400 (B) 404 (C) 501 (D) 200

5) 下面哪一件事情是 Python 网络编程无法直接实现的? （　　）
(A) 下载一部小说 (B) 瘫痪金融网络 (C) 入侵 FBI (D) 获得友情或爱情

四、计算题 (共 60 分):

1) (10 分) 下载指定网站中的一部网络小说.

2) (10 分) 下载新闻，并统计其中的文字或词语的频数.

3) (40 分) 利用爬虫框架，开发一款爬虫软件.