## **任务一**

#### 目标：

挑选三家新闻媒体，每家新闻媒体爬取100条最新的新闻。

**要点：**

* 如若所编写爬虫可以解决js渲染问题，可额外加分。
* 如若所编写爬虫可以分版块归档新闻，可额外加分。
* 如若所编写爬虫可完整保存图片信息，可额外加分。

#### 作业格式：

新闻以txt文件保存，按文件夹归档，压缩为rar提交。

简述所用技术和思路。

## **任务二**

#### 目标：

将所爬取得新闻预处理，保留新闻主干，去掉非新闻主体部分。并进行分词和去停留词操作。

**要点：**

* 如若三个新闻网站可以使用一套通用方案可以额外加分。

#### 作业格式：

新闻处理结果以txt文件保存，文件名为新闻名称，压缩为rar提交。

简述所用技术和思路。

## **任务三**

#### 目标：

选取分析模型，对预处理之后的信息进行分析，输出每个网站（版块）的核心热词。

**要点：**

* 如若可使用多种模型，并能进行效果比较可额外加分。
* 如若可以对结果进行分类，可额外加分。
* 如若可以在此基础上有更全面分析，可额外加分。

#### 作业格式：

分析结果和分析报告。