* 题目：“网易首页（或搜狐）两周新闻地点的频次分析与可视化”
* 小组成员： 谢松成（硕）、李理（硕）、王怡洁（大一）、朱文昕（大一）
* 内容：

1. 获取两周的网易首页出现的所有新闻。
2. 对新闻提到的省级和市级（不统计直辖市以及特别行政区下属区县）进行频次统计。
3. 将统计到的频次用matplotlib，basemap, geoplotlib，pandas等进行可视化处理。
4. 结论总结。

* 技术路线：

1. 数据来源与获取方法

（1）利用requests, 和BeautifulSoup实现爬取网易新闻。

（2）网上下载到一份中国省市的xls文件，复制粘贴到txt文件当中，利用对txt文件的处理，得到省市列表。

1. 初步分析

（1）通过对省市名称进行分类处理，得到省市名称的简称。

（2）将省市简称保存到省市字典其中的键为省份名称，键的值为省份所包含市简称的列表。

1. 统计与展示

（1）遍历字典的键值对，遍历值列表的所以市简称，判断市简称是否在新闻中，统计每一个市的频次，再统计每一个省的频次。得到省市频次统计信息。

（2）利用pandas、matplolib和basemap生成中国地图，其中包括中国的省级行政单位。利用Matplotlib对地图进行颜色填充。

（3）做一些条形图，或者是其他的可视化方法，展示我们的分析结果。

4、可能遇到的问题以及解决方法

（1）网页中的新闻更新速度太快并且更新时间不定—以上午9:00为标准时间进行新闻抓取作为代表时间。

（2）对于工具pandas、matplolib以及basemap的使用尚在学习阶段，可能会在项目进行过程中出现问题，以及亟待解决的如何给地图添加颜色。

* 进度安排

1. 选取立项附近的两周做为爬取网站信息的时间，并完成省事词频统计。
2. 在接下来的一周到两周的时间内完成数据分析以及图表，并且进行数据的汇总，完成报告。
3. 成果展示。