BIT

# 爬虫

第四组：吴志伟，陈曦，单则安，刘杰龙，陈思益，倪俊峰

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

德以明理　学以精工

Contents

# 结构大纲

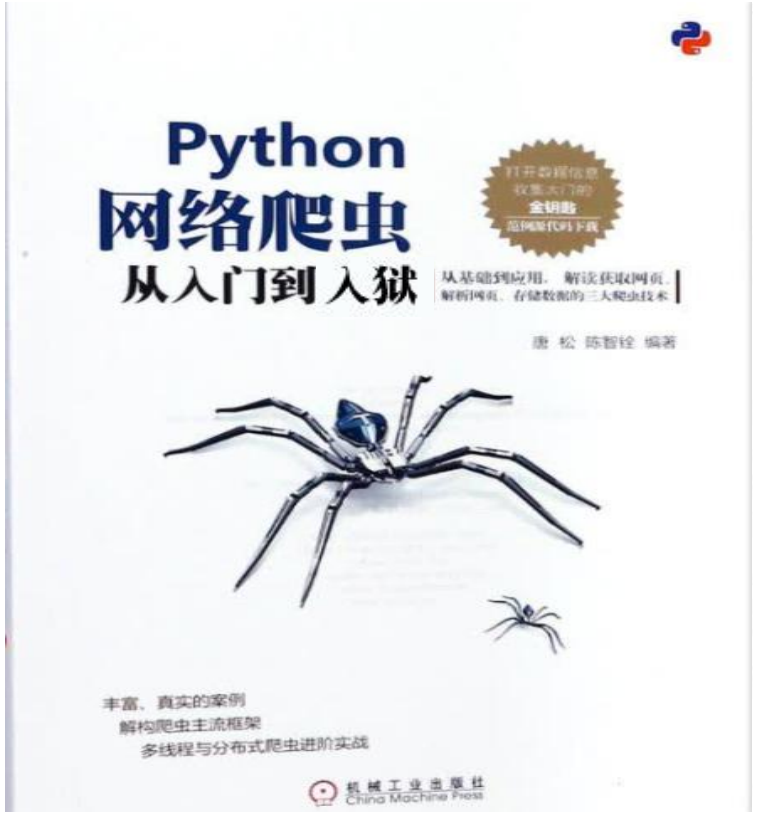
01

# 爬虫简介

主讲人：吴志伟

# 1.1 引言


爬虫学的好，
监狱进的早。



# 1.2 爬虫的定义

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 何为爬虫？

**网络爬虫**，是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本。

VS

## 为何叫爬虫？

由于专门用于检索信息的“机器人"程序像蜘蛛一样在网络间爬来爬去，因此，搜索引擎的“机器人”程序就被称为“蜘蛛"程序，即爬虫。

# 1.3 发展历史

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 1989年

Tim Berners-Lee发明的万维网，引入三个重要技术。

**统一资源定位器(URL)**，我们通过它来访问我们想看的网站；

**内嵌的超链接**，让我们可以在网页之间导航，例如产品详情页；

网页不仅包含**文本**，还包括**图像、音频、视频**和**软件组件**。

## 1990年

第一个**网络浏览器**由Tim Berners-Lee发明，被称为**WorldWide网页(无空间)**，以WWW项目命名。在网络出现一年后，人们有了一条途径去浏览它并与之互动。

1991年第一个**网页服务器**和第一个http:// 网页页面。

## 1993年

6月第一台网页机器人——**万维网漫游器**，用来测量网页的大小。

12月首个基于爬虫的网络搜索引擎—**JumpStation**。**JumpStation**带来了新的飞跃。它是第一个依靠网络机器人的**WWW搜索引擎**。

# 1.3 发展历史

## 2000年

（**API**表示应用程序编程接口）2000年，Salesforce和eBay推出了自己的API，程序员可以用它访问并下载一些公开数据。网页API通过收集网站提供的数据，为开发人员提供了一种更友好的网络爬虫方式。

## 2004 年

2004年，**Beautiful Soup**发布。它被认为是用于网络爬虫的最复杂和最先进的库，也是当今常见和流行的方法之一。

## 2005-2006年

网络抓取软件的可视化，2006年，Stefan Andresen和他的Kapow软件发布了**网页集成平台6.0版本**，它允许用户轻松简单的选择网页内容，并将这些数据构造成可用的excel文件或数据库。

# 爬取过程：

（1）**分析目标网站。**明晰目标网站结构，理清关键数据位置。

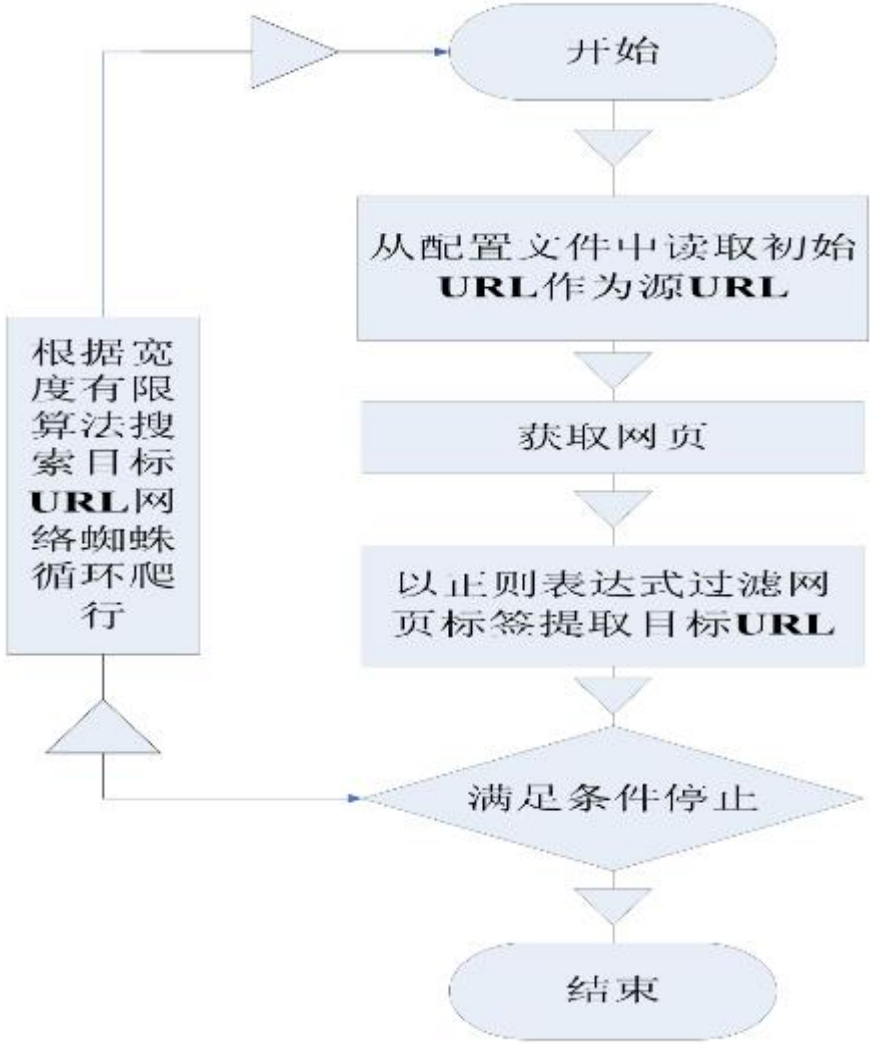
（2）**发起请求**：使用http库或浏览器模拟工具向目标站点发起请求，即发送一个Request。

（3）**获取响应内容。**如果得到了一个Response。说明浏览器能正常响应。

（4）**解析内容**：解析html数据：正则表达式（RE模块），第三方解析库如Beautifulsoup，lxml等。解析json数据：json模块。

从而获得想要的关键数据信息，或者是下一个待爬取的URL地址。

（5）**保存数据。**



# 1.5 爬虫的种类

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

通用网络爬虫

deep web 爬虫

聚焦网络爬虫

增量式爬虫

02

# 爬虫种类介绍

主讲人：刘杰龙
陈思益

# 通用网络爬虫

通用网络爬虫又称**全网爬虫**。爬行对象从一些种子URL扩充到整个Web，主要为门户站点搜索引擎和大型 Web 服务提供商采集数据。 这类网络爬虫的爬行范围和数量巨大，对于爬行速度和存储空间要求较高，对于爬行页面的顺序要求相对较低，同时由于待刷新的页面太多，通常采用并行工作方式，但需要较长时间才能刷新一次页面。虽然存在一定缺陷，通用网络爬虫适用于为搜索引擎**搜索广泛的主题**，有较强的应用价值。

# 2.1.2 通用网络爬虫

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

**广度优先策略**是按照网页内容目录层次深浅来爬行页面，处于较浅目录层次的页面首先被爬行。当同一层次中的页面爬行完毕后，爬虫再深入下一层继续爬行。

## 优势

- 能够有效控制页面的爬行深度，避免遇到一个无穷深层分支时无法结束爬行的问题
- 实现方便
- 需存储大量中间节点

## 劣势

- 需较长时间才能爬行到目录层次较深的页面

# 2.1.2 通用网络爬虫

**深度优先策略**是按照深度由低到高的顺序，依次访问下一级网页链接，直到不能再深入为止。 爬虫在完成一个爬行分支后返回到上一链接节点进一步搜索其它链接。 当所有链接遍历完后，爬行任务结束。

## 优势

- 适合垂直搜索或站内搜索

## 劣势

- 爬行页面内容层次较深的站点时会造成资源的巨大浪费

# 2.2.1 Deep web 爬虫

## Deep web

Web 页面按存在方式可以分为**表层网页**（Surface Web）和**深层网页**（Deep Web，也称 Invisible Web Pages 或 Hidden Web）。表层网页是指传统搜索引擎可以索引的页面，以超链接可以到达的静态网页为主构成的 Web 页面。Deep Web 是那些大部分内容不能通过静态链接获取的、隐藏在搜索表单后的，只有用户提交一些关键词才能获得的 Web 页面。

# 2.2.1 Deep web 爬虫

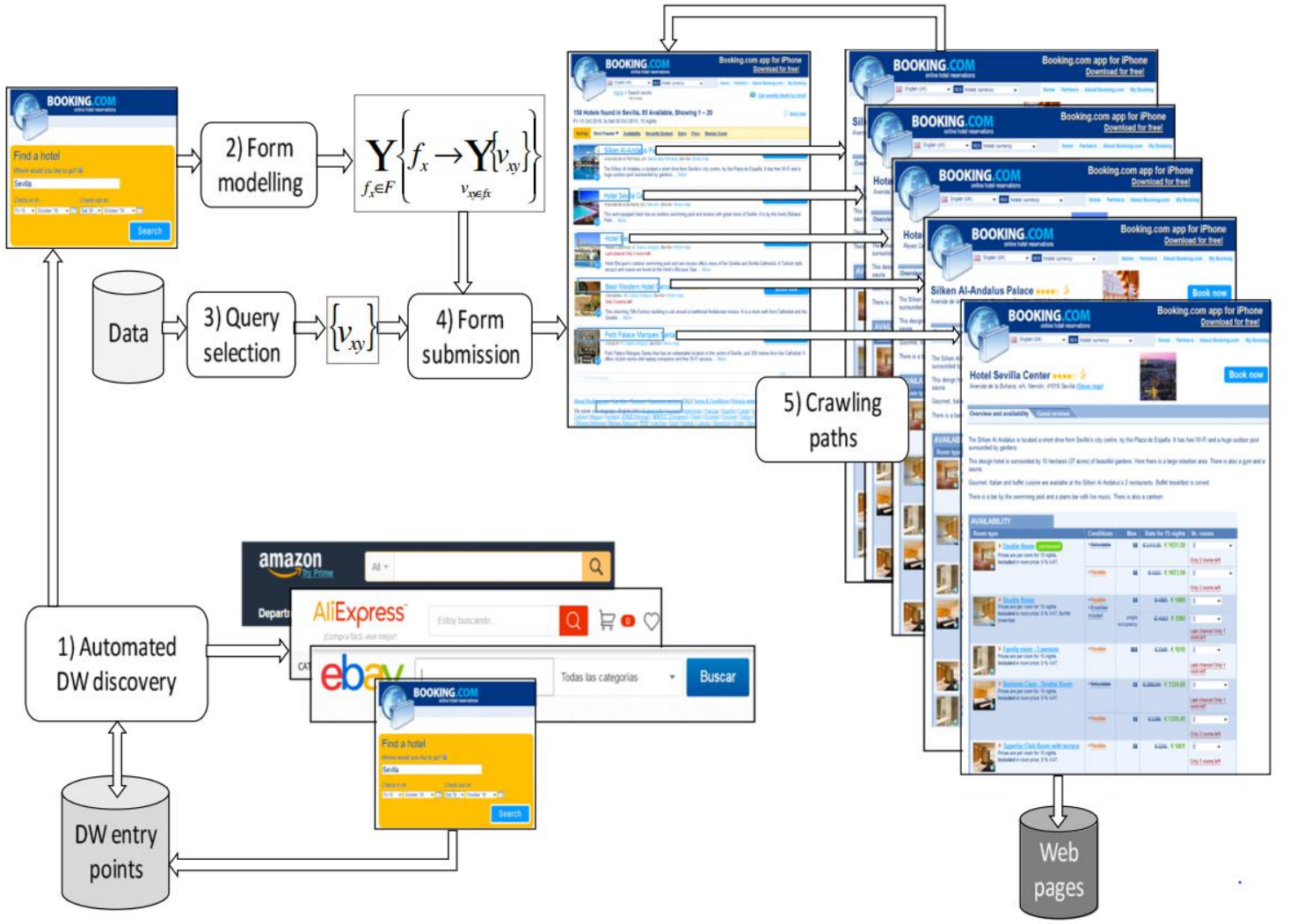
# 2.2.2 Deep web 爬虫

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

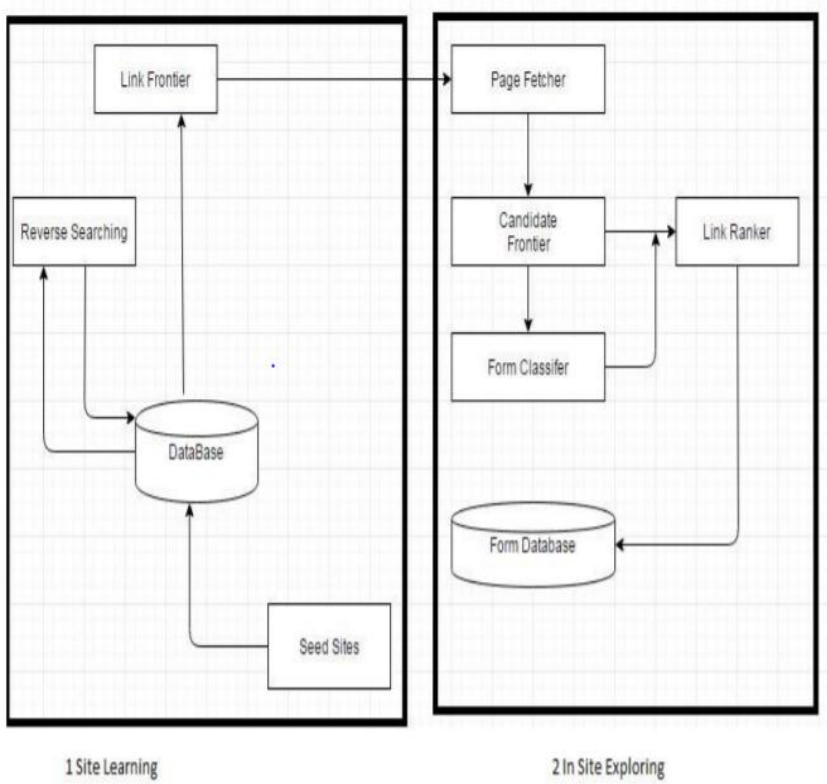
## Deep web 爬虫关键步骤

- 自动查找深层Web入口点。
- Form 建模。
- 查询选择。
- 表单提交。
- 学习爬行路径。

# 2.2.2 Deep web 爬虫



# 2.2.2 Deep web 爬虫




## 自动查找深层Web入口点

- 步骤1: 收集与关键字相关的种子站点。
- 步骤2: 访问种子站点并收集该页面上的所有链接。
- 步骤3: 访问每一个链接，收集所有关键词。
- 步骤4: 计算页面上的外部链接。
- 步骤5:根据关键词、链接数量计算每个链接的页面排名。

# 2.2.2 Deep web 爬虫

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## Form modelling

自动填充表单的一个主要步骤是将为人类设计的搜索表单转换为可以被机器理解的形式,可以自动处理表单并不是一件容易的事情,因为每个搜索表单都是专门为特定网站创建的，没有标准、指南或建议来指导表单设计者。

## Query selection

当对表单建模后，可以通过实例化模型并为其字段选择有效的值组合来自动填充表单。每个值的组合都涉及一个新的表单提交，该表单提交将由Web站点服务器接收。服务器使用表单值组成在其数据库上执行的查询，并生成带有该查询结果的响应页面。

# 2.2.2 Deep web 爬虫

# 2.2.2 Deep web 爬虫



## Blind crawlers

从网站上下载尽可能多的页面;它们从种子页面开始，迭代地跟踪它提供的每个链接，直到下载了从种子页面可以访问的每个页面。

## Focused crawlers

只关注那些可能指向包含某一主题信息的页面的链接通常对下载的页面进行分类，以检查它们是否属于该主题。

## Ad-hoc crawlers

只跟随那些可能导致页面包含与特定用户相关的信息的链接，而不一定属于同一主题。

# 2.3.1 聚焦网络爬虫

**聚焦网络爬虫**（Focused Crawler），又称**主题网络爬虫**（Topical Crawler），是指选择性地爬行那些跟主题有相关性内容的网络爬虫。和通用爬虫相比，聚焦爬虫只需要爬行与主题相关的页面，极大地节省了资源，还可以很好地满足一些特定人群对特定领域信息的需求。

# 2.3.2 聚焦网络爬虫优势

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 优势1

相比通用爬虫只能提供粗略的信息，主题爬虫主题明确且系统能够精准地获取有效信息。

VS

## 优势2

主题爬虫在存储网页URL时需要判断该URL与主题的相关性，尽可能地筛选出与主题相关的页面。

# 2.3.3 聚焦爬虫的系统结构

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 网页获取

模拟客户端发送HTTP请求，获取服务器端的响应后下载网页，完成爬虫系统爬取工作。

## 网页过滤

筛选与主题有关的URL，通过筛选抓取与主题相关的页面，确保主题爬虫系统的准确率。

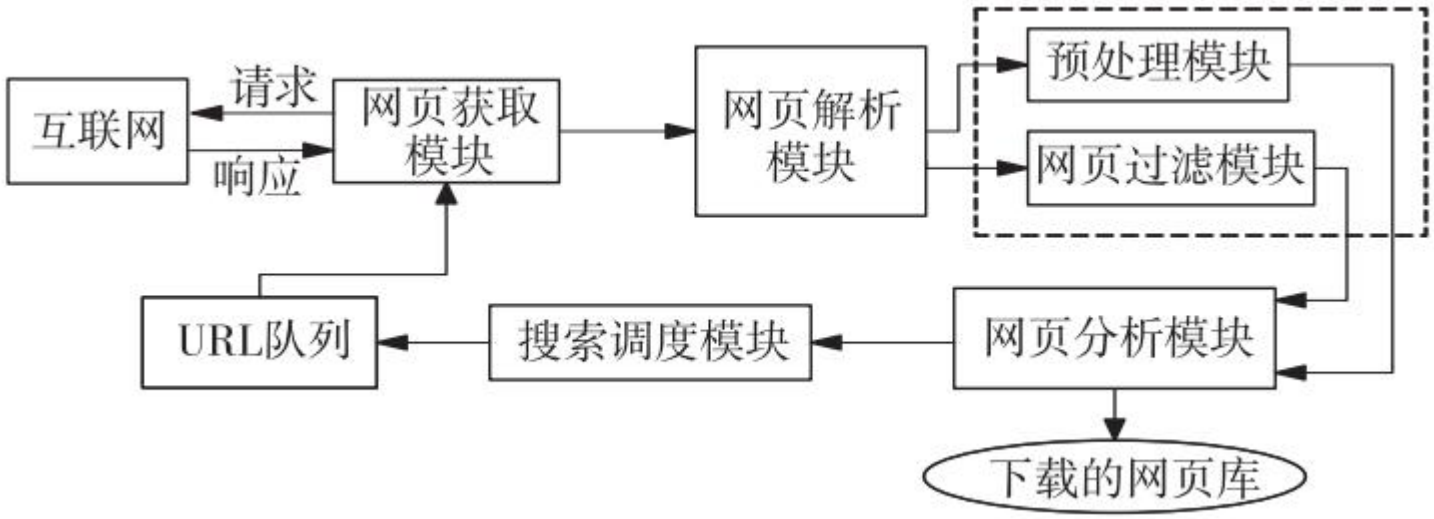
## 网页存储

将网页解析模块解析出来的数据通过文件或数据库的形式存储起来。

## 网页分析

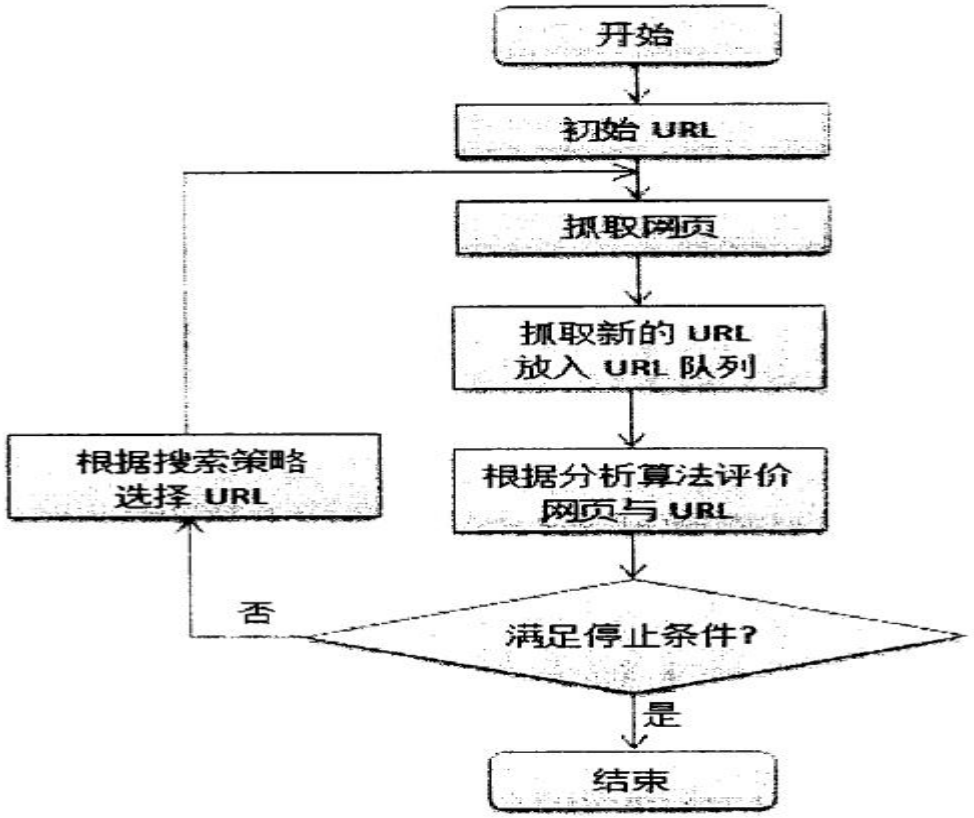
第一部分是主题相关度判断，第二部分是主题相关度预测。

# 2.3.4 聚焦爬虫模块




聚焦爬虫模块图

# 2.3.5 聚焦爬虫过程




聚焦爬虫过程图

# 2.3.6 聚焦爬虫相似度判别算法

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 向量空间模型

将文本处理转换为在向量空间上的向量运算，将每一篇文档表示为向量空间上的某一维度，通过计算向量在空间的相似度来衡量文档之间的相似度。

## 语义相似度

文本中能够观察到的量只有词频和文档频率，文本语义的分析方法是一种对以这两个量为主要思想的计算基础，使得计算机能够“懂”得人类的语言。

# 2.3.7 聚焦爬虫搜索策略

- **静态搜索策略**依照确定的规则进行搜索，搜索策略的规则不会因为网页结构、文本信息的改变而改变。例如**广度优先搜索、深度优先搜索、最佳优先搜索。**
- **动态搜索策略**以**高效、快速**完成爬取任务为第一宗旨，实时调整搜索路线。动态搜索策略会实时根据URL的主题相关度而进行调整。例如**基于文本内容的Fish-Seach、Shark-Search，以及基于链接分析的PageRank、HITS、HillTop。**

# 2.3.7 Fish-Seach



算法描述如下：

- 首先取与主题相关的种子链接页面放入待爬URL队列中。
- 若当前页面与主题相关，将该页面的前a*width个链接放入到URL 队列顶部，增加爬行的宽度，其中参数a和width都是给定的初始值。
- 若当前页面与主题无关，将该页面的前width个链接放入到URL队 列中部，即与主题相关链接的后面 。
- 若是其他情况，将该网页的子链接放入到URL队列的尾部，当有充 足的时间时才对这些链接进行爬取。
- 对于相关性的描述，可以定义一个变量potential_score，当网页与主题相关时，potential_score设为1，当网页与主题无关时，potential_scire设为0.5，其他情况potential_score设为0，运用potential_score的值对待爬行URL队列进行排序。

# 2.3.8 聚焦网络爬虫研究方向

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 基于网页内容

| 方法 | 文献 | 查准率 | 召回率 | F值 |
|---|---|---|---|---|
| 基于改进 PageRank 算法 | [4] | 0.7 | \ | \ |
| 基于 KNN 分类算法的主题网络爬虫 | [5] | 0.75 | \ | \ |
| 基于 ODP 主题描述和 VSM 主题相关度改进 | [6] | 0.64 | 0.24 | 0.24 |
| 基于词向量语义模型构建主题爬虫 | [9] | 0.46 | 0.69 | 0.44 |
| 基于 SVM 分类器的支持向量构建 KNN 分类器 | [8] | 0.80 | \ | \ |
| 基于关键词和 SVM 的动态主题爬虫 | [11] | 0.92 | \ | \ |
| 基于 URL 和锚文本语义特征改进 | [12] | 0.69 | \ | \ |

## 基于链接分析

| 方法 | 文献 | 查准率 | 召回率 | F值 |
|---|---|---|---|---|
| 基于 URL 模式集的主题爬虫 | [14] | 0.69 | 0.52 | 0.61 |
| 基于页面子链接分析的链接排序算法 | [15] | 0.55 | \ | \ |
| VIPS 分析网页深度 + 多粒度鲨鱼搜索算法 | [17] | 0.66 | \ | \ |
| 分类器引导的主题爬虫且链接上下文 | [20] | \ | 0.61 | \ |
| 基于 Best-First 算法 + HITS 算法 | [19] | 0.61 | 0.75 | \ |
| 基于内容分块-选择性链接上下文的聚焦爬虫 | [23] | \ | 0.80 | \ |
| HTML 分析 + 文本密度分析 + 多因子相似度 | [24] | 0.67 | 0.48 | \ |

# 2.3.9 聚焦网络爬虫研究方向

## 基于网页内容

目前，基于网页内容的主题爬虫计算文本相似度的判断方法大致分为两类：a）**基于字词统计模型**，如向量空间模型；b）**基于语义理解模型**。研究人员希望使用语义相关性使网络爬虫可以获得更精确的结果。在整个主题相似度判别过程中，首先确定主题爬虫的主题，再根据网页内容、结构信息计算网页主题相关度和抓取 URL 的相关度，依据网页主题相关度判断待抓取链接和抓取链接的优先级。此类爬虫通常能获得较高的准确率。

## 基于链接分析

互联网中数十亿的网页通过万维网上的超链接进行链接，研究人员试图通过有效的方式获取链接上下文的含义，从而对链接上下文进行解析和提取，或者基于网页内容对传统链接选择算法改进 ，使网络爬虫采集过程中的准确度提升。该类算法通过分析网页链接判断网页的重要性，强调了页面链接的权威性对用户的需求是有意义的，同时从网页正文、链接锚文本以及锚文本上下文的网页内容分析和链接分析结合解决了 “主题漂移”问题，提高主题爬取的**准确性**。

# 2.3.10 聚焦网络的实战

- 为了弥补通用搜索引擎的不足，实现对特定主题信息的检索，出现了垂直搜索引擎，它检出的结果更准确，挖掘信息的层次更深，无效信息更少，更能适应**垂直领域**的服务。
- 垂直搜索引擎是面向特定领域为特定用户服务的一种搜索引擎，是对专业领域信息的深层次挖掘，它将信息过滤、筛选、梳理后集成在一起，为用户提供了面向专业知识的检索。
- 垂直搜索引擎与全文搜索引擎工作原理类似，区别在于抓取模块中的爬虫程序与主题词库。

# 2.4.1 增量式网络爬虫简介

某些网站会定时在原有网页数据的基础上更新一批数据。在遇到这样的场景时，便可以采用增量式爬虫。增量爬虫技术就是通过爬虫程序监测某网站数据更新的情况，以便可以爬取到该网站更新后的新数据。实现增量式爬虫的核心是**去重**。

# 2.4.2 增量式爬虫的思路

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

在发送请求之前判断这个 URL 是否曾爬取过（适合不断有新页面的网站）

在解析内容后判断这部分内容是否曾爬取过（适合页面内容定时更新的网站）

写入存储介质时判断内容是否已存在于介质中（最大限度达到去重的目的）

# 2.4.3 增量式爬虫去重

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 方法一

对爬取过程中产生的 URL 进行存储，存储在 Redis 的 set 中。当下次进行数据爬取时，首先在存储 URL 的 set 中对即将发起的请求所对应的 URL 进行判断，如果存在则不进行请求，否则进行请求。

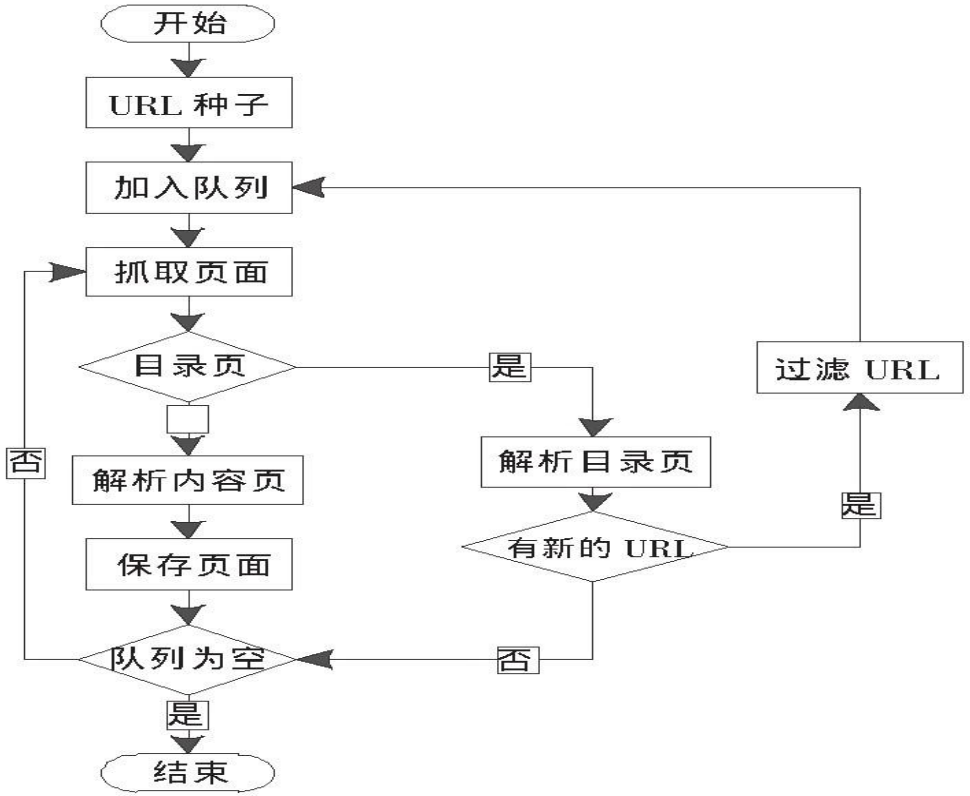
## 方法二

对爬取到的网页内容进行唯一标识的制定（数据指纹），然后将该唯一标识存储至 Redis 的 set 中。当下次爬取到网页数据的时候，在进行持久化存储之前，可以先判断该数据的唯一标识在 Redis 的 set 中是否存在，从而决定是否进行持久化存储。

# 2.4.3 Bloomfilter

基于Redis的Bloomfilter去重，既发挥了Bloomfilter的海量去重能力，又发挥了Redis的可持久化能力。Bloomfilter是一个很长的二进制向量和一系列随机映射Hash函数。通常辨别某个元素是否在集合中的常用方法是用已知元素和集合中的元素进行对比。Bloomfilter能够在较短时间内检查某一元素是否在集合内。

# 2.4.4 增量式爬虫过程




增量式爬虫过程图

03

# 开发库及框架

主讲人：倪俊峰

# 3.1 常用库

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 页面爬取

实现Http请求操作

- **urllib**
- **requests**
- selenium
- aiohttp

## 页面分析

从网页中提取信息

- **lxml**
- **beautifulsoup**
- pyquery

## 存储库

与数据库交互

- pymysql
- pymongo

# 3.1.1 urllib

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## request

HTTP请求模块，可以用来模拟发送请求。只需要给库方法传入URL以及额外的参数，就可以模拟在浏览器里输入网址然后回车一样。

## error

异常处理模块，如果出现请求错误，我们可以捕获这些异常，然后进行重试或其他操作以保证程序不会意外终止。

## parse

工具模块，提供了许多URL处理方法，比如拆分、解析、合并等。

## robotparser

网站识别模块，主要是用来识别网站的robots.txt文件，然后判断哪些网站可以爬，哪些网站不能爬。

# 3.1.1 urllib

main.py

```python
import urllib.request

response = urllib.request.urlopen("https://www.python.org")
html = response.read()
print(html.decode('utf-8'))
```

```
<!doctype html>
<!--[if lt IE 7]>   <html class="no-js ie6 lt-ie7 lt-ie8 lt-ie9">   <![endif]-->
<!--[if IE 7]>      <html class="no-js ie7 lt-ie8 lt-ie9">          <![endif]-->
<!--[if IE 8]>      <html class="no-js ie8 lt-ie9">                 <![endif]-->
<!--[if gt IE 8]><!--><html class="no-js" lang="en" dir="ltr">  <!--<![endif]-->

<head>
    <meta charset="utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">

    <link rel="prefetch" href="//ajax.googleapis.com/ajax/libs/jquery/1.8.2/jquery
    <link rel="prefetch" href="//ajax.googleapis.com/ajax/libs/jqueryui/1.12.1/jqu

    <meta name="application-name" content="Python.org">
    <meta name="msapplication-tooltip" content="The official home of the Python Pr
    <meta name="apple-mobile-web-app-title" content="Python.org">
    <meta name="apple-mobile-web-app-capable" content="yes">
    <meta name="apple-mobile-web-app-status-bar-style" content="black">

    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <meta name="HandheldFriendly" content="True">
    <meta name="format-detection" content="telephone=no">
    <meta http-equiv="cleartype" content="on">
    <meta http-equiv="imagetoolbar" content="false">
```

# 3.1.2 requests

**requests**是python的第三方库，它是对urllib的进一步封装，因此在使用上显得更加的便捷。


## 功能特性

Requests 完全满足今日 web 的需求。

- Keep-Alive & 连接池
- 国际化域名和 URL
- 带持久 Cookie 的会话
- 浏览器式的 SSL 认证
- 自动内容解码
- 基本/摘要式的身份认证
- 优雅的 key/value Cookie
- 自动解压
- Unicode 响应体
- HTTP(S) 代理支持
- 文件分块上传
- 流下载
- 连接超时
- 分块请求
- 支持 `.netrc`

Requests 支持 Python 2.6—2.7以及3.3—3.7，而且能在 PyPy 下完美运行。

# 3.1.2 requests

```python
import requests

response1 = requests.get(url='http://httpbin.org/get')
headers = {
    'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, '
                  'like Gecko) Chrome/91.0.4472.114 Safari/537.36 '
}
response2 = requests.get(url='http://httpbin.org/get',headers=headers)
print(response1.text)
print(response2.text)
```


```
{
  "args": {},
  "headers": {
    "Accept": "*/*",
    "Accept-Encoding": "gzip, deflate",
    "Host": "httpbin.org",
    "User-Agent": "python-requests/2.26.0",
    "X-Amzn-Trace-Id": "Root=1-61495393-33dd7ff814c081f56a26d447"
  },
  "origin": "114.246.201.140",
  "url": "http://httpbin.org/get"
}

{
  "args": {},
  "headers": {
    "Accept": "*/*",
    "Accept-Encoding": "gzip, deflate",
    "Host": "httpbin.org",
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (
    "X-Amzn-Trace-Id": "Root=1-61495393-557da9dd57064a2674e15806"
  },
  "origin": "114.246.201.140",
  "url": "http://httpbin.org/get"
}
```

# 3.1.3 lxml

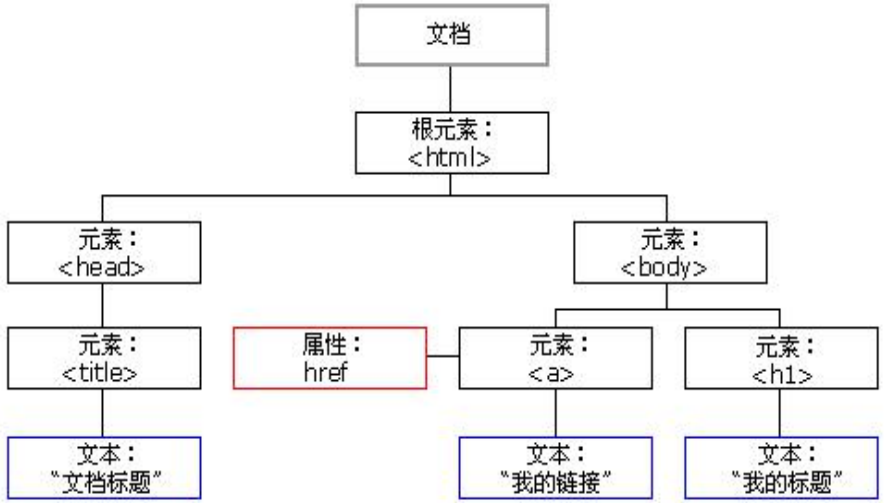
## lxml - 使用 Python 的 XML 和 HTML

lxml 是 Python 语言中用于处理 XML 和 HTML 的功能最丰富且易于使用的库。

### » 介绍

lxml XML 工具包是 C 库**libxml2**和**libxslt**的 Pythonic 绑定 。它的独特之处在于它将这些库的速度和 XML 功能完整性与本机 Python API 的简单性相结合，大部分兼容但优于众所周知的 **ElementTree** API。最新版本适用于从 2.7 到 3.9 的所有 CPython 版本。有关lxml 项目的背景和目标的更多信息，请参阅**介绍**。一些常见问题的回答中的**常见问题**。

**HTML DOM 树**



# 3.1.3 lxml

XPath，全称 XML Path Language，即 XML 路径语言，最初是用来搜寻 XML 文档的，但同样适用于 HTML 文档的搜索。所以在做爬虫时完全可以使用 XPath 做相应的信息抽取。

## 选取节点

XPath 使用路径表达式在 XML 文档中选取节点。节点是通过沿着路径或者 step 来选取的。 下面列出了最有用的路径表达式：

| 表达式 | 描述 |
| --- | --- |
| nodename | 选取此节点的所有子节点。 |
| / | 从根节点选取（取子节点）。 |
| // | 从匹配选择的当前节点选择文档中的节点，而不考虑它们的位置（取子孙节点）。 |
| . | 选取当前节点。 |
| .. | 选取当前节点的父节点。 |
| @ | 选取属性。 |

# 3.1.3 lxml

```
html = """
<html>
    <head>
        <title>The Dormouse's story</title>
        <meta name="viewport" content="width=device-width, initial-scale=1.0">
    </head>
    <body>
        <p class="title" id="story">The Dormouse's story</p>
        <p class="story">Once upon a time there were three little sisters and their names were
            <a href="https://example.com/elsie" class="sister" id="link1">Elsie</a>,
            <a href="https://example.com/lacie" class="sister" id="link2">Lacie</a> and
            <a href="https://example.com/tillie" class="sister" id="link3">Tillie</a>;
        </p>
        <p class="story">and they lived at the bottom of a well.</p>
        <p class="story">...</p>
    </body>
</html>
"""
```

```
html_test = etree.HTML(html)
html_title = html_test.xpath('//body/p[@class="title"]')
html_story1 = html_test.xpath('//body/p[@class="story"][1]')
html_sister = html_test.xpath('//body/p/a')
html_story_left = html_test.xpath('//body/p[position()>2]')
print(html_title[0].text.strip())
print(html_story1[0].text.strip())
for sister in html_sister:
    print(sister.text.strip(), end=" ")
print()
for left in html_story_left:
    print(left.text.strip())
```

```
The Dormouse's story
Once upon a time there were three little sisters and their names were
Elsie Lacie Tillie
and they lived at the bottom of a well.
...

Process finished with exit code 0
```

# 3.1.4 Beautiful Soup

## Beautiful Soup 4.4.0 文档

Beautiful Soup 是一个可以从HTML或XML文件中提取数据的Python库.它能够通过你喜欢的转换器实现惯用的文档导航,查找,修改文档的方式.Beautiful Soup会帮你节省数小时甚至数天的工作时间.

这篇文档介绍了BeautifulSoup4中所有主要特性,并且有小例子.让我来向你展示它适合做什么,如何工作,怎样使用,如何达到你想要的效果,和处理异常情况.

文档中出现的例子在Python2.7和Python3.2中的执行结果相同

你可能在寻找 Beautiful Soup3 的文档,Beautiful Soup 3 目前已经停止开发,我们推荐在现在的项目中使用Beautiful Soup 4, 移植到BS4

**Beautiful Soup**将复杂HTML或XML文档转换成一个复杂的树形结构（文档树），树上每个节点都是一个Python对象。所有对象可以归纳为4种类型:**Tag**，**NavigableString**，**BeautifulSoup** ,**Comment** 。

# 3.1.4 Beautiful Soup

```python
from bs4 import BeautifulSoup

html = """
<html>
    <head>
        <title>The Dormouse's story</title>
        <meta name="viewport" content="width=device-width, initial-scale=1.0">
    </head>
    <body>
        <p class="title" id="story">The Dormouse's story</p>
        <p class="story">Once upon a time there were three little sisters and their names were
            <a href="https://example.com/elsie" class="sister" id="link1">Elsie</a>,
            <a href="https://example.com/lacie" class="sister" id="link2">Lacie</a> and
            <a href="https://example.com/tillie" class="sister" id="link3">Tillie</a>;
        </p>
        <p class="story">and they lived at the bottom of a well.</p>
        <p class="story">...</p>
    </body>
</html>
"""
bs = BeautifulSoup(html, 'lxml')
print(bs.title)
p_tag = bs.find('p')
print(p_tag)
print(p_tag.name, p_tag['class'], p_tag['id'])
print(p_tag.string)
```

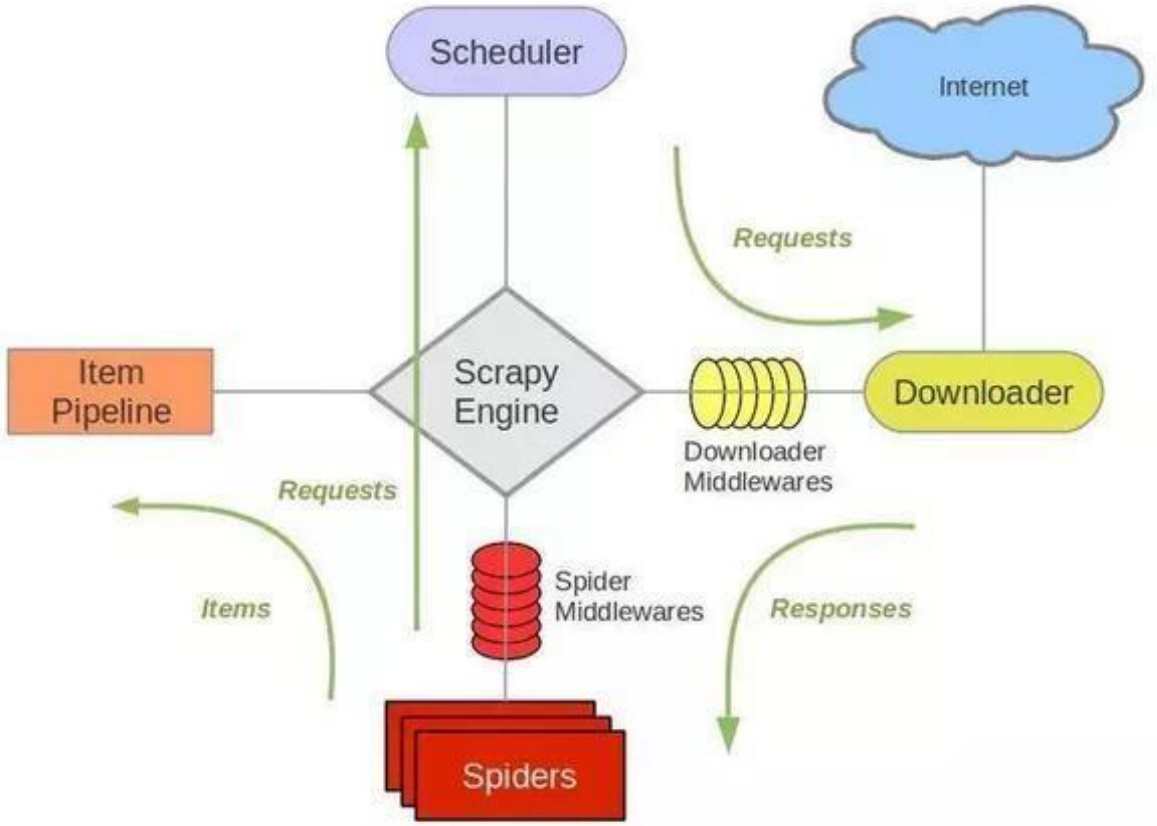
```
<title>The Dormouse's story</title>
<p class="title" id="story"><b>The Dormouse's story</b></p>
p ['title'] story
The Dormouse's story

Process finished with exit code 0
```


# Scrapy框架

- 是一个快速、高层次的屏幕抓取和web抓取框架，用于抓取web站点并从页面中提取结构化的数据。
- Scrapy吸引人的地方在于它是一个框架，任何人都可以根据需求方便的修改。它也提供了多种类型爬虫的基类，如BaseSpider、sitemap爬虫等。

# 3.2.1 Scrapy框架




- **Scrapy引擎(Scrapy Engine)**：是整个框架的核心，用来处理整个系统的数据流和触发事务。
- **调度器(Scheduler)**：接收引擎发过来的请求并将其加入队列中，在引擎再次请求的时候将请求提供给引擎。
- **下载器(Downloader)**：是所有组件中负担最大的，它用于高速地下载网络上的资源。
- **爬虫(Spider)**：帮助用户定制自己的爬虫(通过定制正则表达式等语法)，用于从特定的网页中提取自己需要的信息。
- **实体管道(Item Pipeline)**：用于处理爬虫spider提取的实体。主要功能是持久化实体、验证实体的有效性，以及清除不需要的信息。

# 3.2.1 Scrapy框架

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 1、定义Item

```python
import scrapy


class TxmoviesItem(scrapy.Item):
    # define the fields for your item here like:
    name = scrapy.Field()
    description = scrapy.Field()
```

## 2、写爬虫程序

```python
import scrapy
from ..items import TxmoviesItem


class TxmsSpider(scrapy.Spider):
    name = 'txms'
    allowed_domains = ['v.qq.com']
    start_urls = ['https://v.qq.com/x/bu/pagesheet/list?append=1&channel=cartoon&iarea=1&list'
                  'page=2&offset=0&pagesize=30']
    offset = 0

    def parse(self, response):
        items = TxmoviesItem()
        lists = response.xpath('//div[@class="list_item"]')
        for i in lists:
            items['name'] = i.xpath('./a/@title').get()
            items['description'] = i.xpath('./div/div/@title').get()
            yield items

        if self.offset < 120:
            self.offset += 30
            url = 'https://v.qq.com/x/bu/pagesheet/list?append=1&channel=cartoon&iarea=1&listpage=2&offset={}&page
                .format(str(self.offset))
            yield scrapy.Request(url=url, callback=self.parse)
```

# 3.2.1 Scrapy框架

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 3、管道输出

```python
class TxmoviesPipeline:
    def process_item(self, item, spider):
        print(item)
        return item
```

## 4、执行

```python
from scrapy import cmdline

cmdline.execute('scrapy crawl txms -o txms.json'.split())
```

txms.json


```
[
{"name": "斗罗大陆", "description": "此生不悔入唐门"},
{"name": "开心锤锤", "description": "普通锤锤的爆笑日常"},
{"name": "星辰变", "description": "穿星辰沧海 赴羽立之约"},
{"name": "武庚纪", "description": "神力觉醒，三界大战！"},
{"name": "完美世界", "description": "岁月淹埋，休想把我沉浮！"},
{"name": "武神主宰", "description": "武神跌落，浴火少年再起"},
{"name": "斗破苍穹 第4季", "description": "少年不屈 异火不熄"},
{"name": "绝世武魂", "description": "吞噬龙血，少年逆天崛起"},
{"name": "灵剑尊", "description": "天地三界，我为至尊！"},
{"name": "无上神帝", "description": "仙王觉醒，重归万界巅峰"},
{"name": "万界仙踪", "description": "仙魔一念，人间千载"},
{"name": "小品一家人", "description": "小品的搞笑温馨日常"},
{"name": "万界神主", "description": "陨落古神，遨游苍蓝"},
{"name": "魔道祖师", "description": "魔道祖师完结篇更新中"},
{"name": "万界独尊", "description": "涅槃重生，逆天改命！"},
{"name": "2021腾讯视频动漫年度发布", "description": "七大类型百部国漫登场"},
{"name": "狂神魔尊", "description": "玄京废少，搞怪逆袭"},
{"name": "我气哭了百万修炼者", "description": "大孝子气人逆天之路"},
{"name": "独步逍遥", "description": "少年热血闯红尘"},
{"name": "迷你小洞", "description": "迷你世界爆笑同人动画"},
{"name": "元气食堂", "description": "吃货熊大卫的美食日常"},
{"name": "飞狗MOCO", "description": "柯基与主人的搞笑日常"},
{"name": "妖神记", "description": "踏足武道巅峰"},
{"name": "首席御灵师", "description": "殿前比试，踢到钢板了！"},
{"name": "游侠战纪", "description": "逆天改命※血脉觉醒"},
{"name": "眷思量", "description": "超高颜值古风3D动画"},
{"name": "猪屁登", "description": "和猪屁登一起传递正能量"},
{"name": "全职法师 第5季", "description": "超点！莫凡火雷双系狂揍审判员"},
{"name": "逆天至尊", "description": "鸿蒙一念间，以血祭陈渊。"},
{"name": "面膜妈妈养娃-小视频特别版", "description": "暖心家庭的温馨生活"}
]
```

# 3.2.2 Pyspider框架

Pyspider是一个带有强大的WebUI、脚本编辑器、任务监控器、项目管理器以及结果处理器的框架。它支持多种数据库后端、多种消息队列和Javascript渲染页面采集。

# 3.2.2 PySpider框架

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 后起之秀

### PySider

- 提供了webui，爬虫编写和调试都是在WebUi 中进行
- 内置了pyquery 作为选择器
- 支持PhantomJS来进行Javascript渲染页面的采集

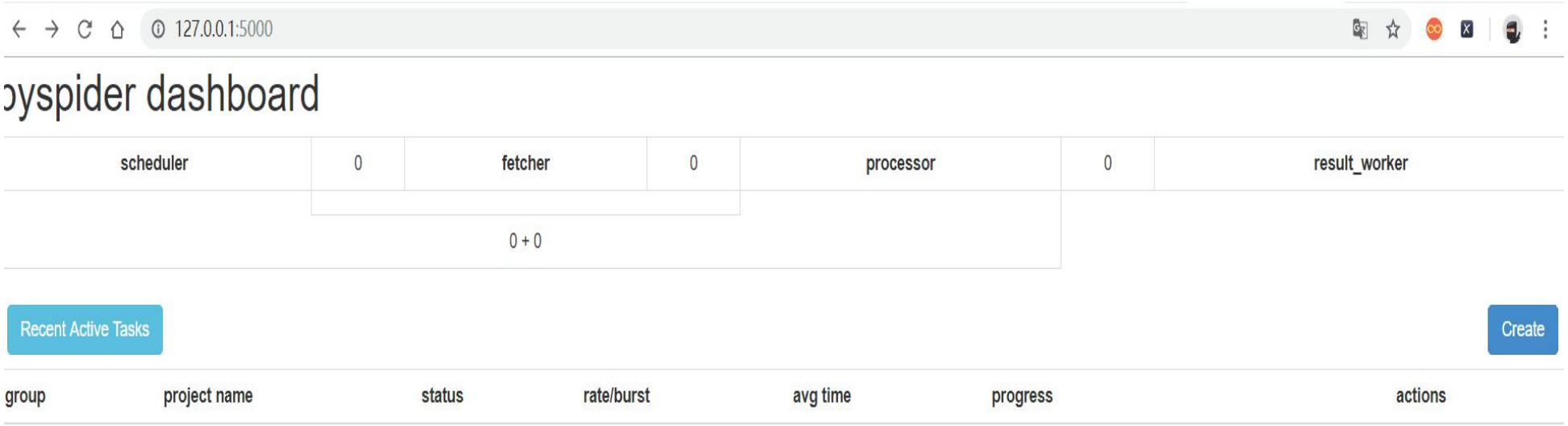
VS

## 老当益壮

### Scrapy

- 采用代码和命令行的操作模式
- 对接了 Xpath / CSS 选择器和正则表达式
- 可以对接Scrapy-Splash组件实现Javascript 渲染页面的采集，不过需要额外的配置

# 3.2.2 PySpider框架



# 总结

04

# demo展示

主讲人：陈曦

# 4.1 主题——外交部例行记者会数据的爬取与分析

# 4.1 主题——外交部例行记者会数据的爬取与分析

首页 外交部长 外交部 外交动态 政府信息公开 驻外机构 国家和组织 资料 服务 移动客户端

首页 > 发言人表态 > 例行记者会

## 发言人表态

- 发言人简历
- 例行记者会
- 发言人表态和电话答问

## 例行记者会

- 2021年9月23日外交部发言人赵立坚主持例行记者会(2021-09-23)
- 2021年9月22日外交部发言人赵立坚主持例行记者会(2021-09-22)
- 2021年9月17日外交部发言人赵立坚主持例行记者会(2021-09-17)
- 2021年9月16日外交部发言人赵立坚主持例行记者会(2021-09-16)
- 2021年9月15日外交部发言人赵立坚主持例行记者会(2021-09-15)
- 2021年9月14日外交部发言人赵立坚主持例行记者会(2021-09-14)
- 2021年9月13日外交部发言人赵立坚主持例行记者会(2021-09-13)

# 4.1 主题——外交部例行记者会数据的爬取与分析

**总台央视记者：21日，习近平主席出席第76届联大一般性辩论并发表重要讲话，提出“全球发展倡议”。这是中方首次使用“全球发展倡议”这一提法。你能否进一步介绍有关情况？**

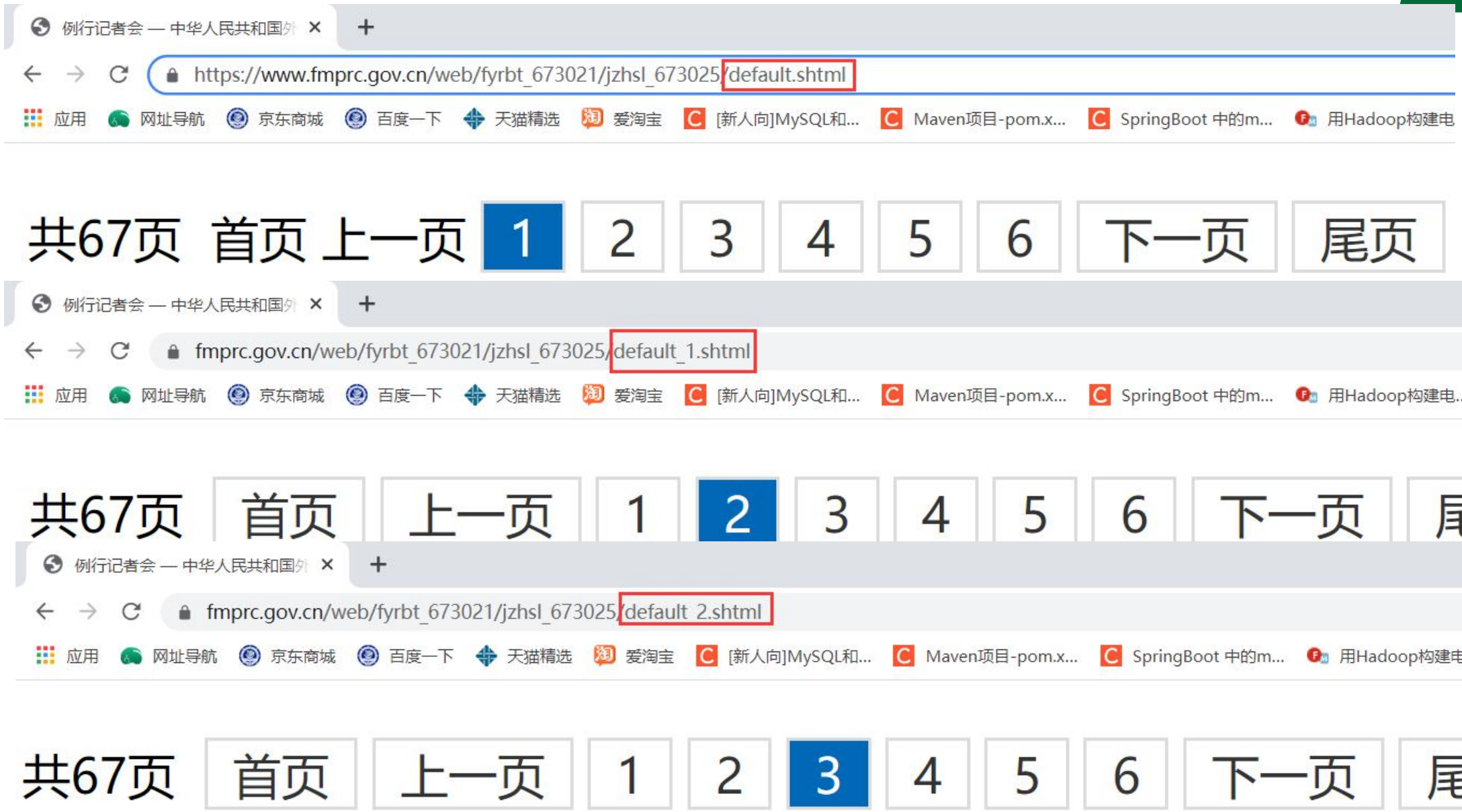
赵立坚：昨天，习近平主席以视频方式出席第76届联大一般性辩论并发表重要讲话。习主席在讲话中提出“全球发展倡议”，呼吁国际社会加快落实2030年可持续发展议程，推动实现更加强劲、绿色、健康的全球发展。习主席提出这一重大倡议，为因应世界变局擘画了蓝图，为全球共同发展指明了方向。

和平与发展是当今时代的主题。当前，南北差距、“发展鸿沟”不断加大，发展中国家长期陷于发展滞后的困境。特别是新冠肺炎疫情吞噬了过去10年全球减贫成果，发展中国家遭受重创，经济增长下滑，粮食安全问题突出，气候变化挑战增多，获得疫苗困难重重。发展中国家面临政治、经济、社会和民生等多重危机，落实2030年议程、实现可持续发展面临严峻挑战。与此同时，当前新工业革命浪潮方兴未艾，数字经济、绿色发展和疫情催生的新业态、新模式为发展中国家实现跨越式发展带来新机遇。

# 4.2.1 准备工作


2021年9月17日外交部发言人赵
fmprc.gov.cn/web/fyrbt_673021/jzhsl_673025/t1907872.shtml
应用 网址导航 京东商城 百度一下 天猫精选 爱淘宝 [新人向]MySQL和... Maven项目-pom.x... SpringBoot 中的m... 用Hadoop构建电...
2021年9月17日外交部发言人赵立坚主持例行记者会
2021年9月16日外交部发言人赵
fmprc.gov.cn/web/fyrbt_673021/jzhsl_673025/t1907464.shtml
应用 网址导航 京东商城 百度一下 天猫精选 爱淘宝 [新人向]MySQL和... Maven项目-pom.x... SpringBoot 中的m... 用Hadoop构建电...
2021年9月16日外交部发言人赵立坚主持例行记者会
2021年9月15日外交部发言人赵
fmprc.gov.cn/web/fyrbt_673021/jzhsl_673025/t1907205.shtml
应用 网址导航 京东商城 百度一下 天猫精选 爱淘宝 [新人向]MySQL和... Maven项目-pom.x... SpringBoot 中的m... 用Hadoop构建电...
2021年9月15日外交部发言人赵立坚主持例行记者会

# 4.2.1 准备工作


例行记者会 — 中华人民共和国外
https://www.fmprc.gov.cn/web/fyrbt_673021/jzhsl_673025/default.shtml
应用 网址导航 京东商城 百度一下 天猫精选 爱淘宝 [新人向]MySQL和... Maven项目-pom.x... SpringBoot 中的m... 用Hadoop构建电
共67页 首页 上一页 1 2 3 4 5 6 下一页 尾页
例行记者会 — 中华人民共和国外
fmprc.gov.cn/web/fyrbt_673021/jzhsl_673025/default_1.shtml
应用 网址导航 京东商城 百度一下 天猫精选 爱淘宝 [新人向]MySQL和... Maven项目-pom.x... SpringBoot 中的m... 用Hadoop构建电... 简 基于物
共67页 首页 上一页 1 2 3 4 5 6 下一页 尾页
例行记者会 — 中华人民共和国外
fmprc.gov.cn/web/fyrbt_673021/jzhsl_673025/default_2.shtml
应用 网址导航 京东商城 百度一下 天猫精选 爱淘宝 [新人向]MySQL和... Maven项目-pom.x... SpringBoot 中的m... 用Hadoop构建电... 简 基于
共67页 首页 上一页 1 2 3 4 5 6 下一页 尾页

# 4.2.1 准备工作

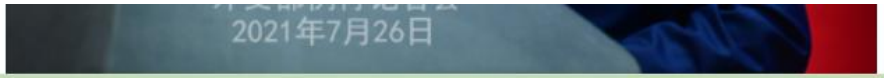
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 2021年7月16日外交部发言人赵立坚主持例行记者会(2021-07-16)
- 2021年7月15日外交部发言人赵立坚主持例行记者会(2021-07-15)
- 日外交部发言人赵立坚主持例行记者会(2021-07-14)
- 2021年7月13日外交部发言人赵立坚主持例行记者会(2021-07-13)

a 346.58 × 20

```
<li>…</li>
<li>…</li>
<li>…</li>
<li>…</li>
<li> == $0
  <a href="./t1891716.shtml" target="_blank">2021年7月13日外交部发言人赵立坚主持例行记者会</a>
  "(2021-07-13)"
</li>
<li>…</li>
```

p 872 × 84




**总台央视记者：日前，王毅国务委员兼外长与巴基斯坦外长库雷希在四川成都举行第三次中巴外长战略对话。发言人能否介绍有关情况?**

赵立坚：7月24日，王毅国务委员兼外长在成都与巴基斯坦外长库雷希举行了第三次中巴外长战略对话。双方就共同关心的国际地区问题深入交换意见。

王毅国务委员指出，中巴建交70年来，两国携手同心，克服无数艰难险阻，战胜诸多风险挑战，结下了“铁杆”情谊，建成了全天候战略合作伙伴关系。高度互信、倾力相助、共谋和平、共促发展是中巴关系最鲜明的特征，也成为双方携手前进的最大底气。中方愿同巴方以建交70周年为契机，加快构建新时代更加紧密的中巴命运共同体，为两国人民创造更多福祉，为地区稳定繁荣作出更大贡献。

```
<div class="info">…</div>
<div id="News_Body_Txt_A" class="content"> == $0
  <p align="justify"> </p>
  <p align="center">…</p>
  <p align="justify">…</p>
  <p align="justify"> 赵立坚：7月24日，王毅国务委员兼外长在成都与巴基斯坦外长库雷希举行了第三次中巴外长战略对话。双方就共同关心的国际地区问题深入交换意见。 </p>
  <p align="justify">…</p>
  <p align="justify">…</p>
  <p align="justify">…</p>
  <p align="justify">…</p>
  <p align="justify"> 赵立坚：一段时间以来，国际社会批评美方将溯源问题政治化，要求调查德特里克堡生物实验室的理性声音不断增多。
  </p>
  <p align="justify">…</p>
  <p align="justify">…</p>
  <p align="justify">…</p>
  <p align="justify"> </p>
  <p align="center">…</p>
  <p align="justify">…</p>
```

# 4.2.2 代码

```
58  #for循环得到全部的目录链接
59  url="https://www.fmprc.gov.cn/web/fyrbt_673021/jzhsl_673025/"
60  catalog_list=[]#目录链接
61  catalog_list.append(url)
62  for i in range(1,67):
63      catalog_list.append(url + "default_" + str(i) + ".shtml")
```

# 4.2.2 代码

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

```
66  #由目录链接得到所有的发言链接及发言内容
67  item_list=[]#格式化存储所有发言数据
68  for catalog in catalog_list:
69      r = requests.get(catalog)#使用requests库打开每一页目录
70      r.encoding = 'utf-8'
71      page=BeautifulSoup(r.text,features="html.parser")#BeautifulSoup库解析html元素
72      div=page.find('div',class_='rebox_news')
73      li_list=div.findAll('li')
74      for li in li_list:#对于每一li（列表项目）标签
75          item={}
76          a=li.find('a')#BeautifulSoup库解析<a></a>元素
77          item['href']=a['href']#<a></a>元素里href属性指向例行记者会网页链接
```

# 4.2.2 代码

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

```
82      #打开例行记者会链接
83      detail_url=url+item['href']
84      detail_request=requests.get(detail_url)#requests库打开每一记者会链接
85      detail_request.encoding='utf-8'
86      detail_page=BeautifulSoup(detail_request.text,features="html.parser")#BS库解析html
87      p_list=detail_page.findAll('p')#<p></p>标签里存储了所有的发言内容
88      question_answer_list=getDailyQuestionAnswer(p_list)#格式化每日记者会的数据并返回
89      item['question_answer_list']=question_answer_list
90      item_list.append(item)
```

# 4.2.3 结果



JSON 原始数据 头

保存 复制 全部折叠 全部展开 (慢) 过滤 JSON


answer: " 答：我们关注到有关报道。首先我想说，中国同东盟国家共同维护南海地区和平稳定的努力有目共睹，中国为推动和平解决朝鲜半岛核问题所付出的巨大和不懈努力也是有目共睹的。 近日，美澳就南海问题发表了一些言论，中方已表明了有关立场。我想进一步强调的是，在地区国家的共同努力下，当前本地区保持了和平稳定的发展势头，这一局面值得珍惜。南海问题已回归由直接当事国通过谈判磋商解决争议的轨道。中国和菲律宾就妥善处理南海问题 会议。同时，中国和东盟国家还达成"南海行为准则"框架，为下一步"准则"磋商奠定了坚实基础。中国和东盟国家也在积极开展海上务实合作，落实一批"早期收获"项目。希望有关方充分尊重和支持地区 不是相反。"

999:

href: "./t1467387.shtml"

title: "2017年6月2日外交部发言人华春莹主持例行记者会"

date: "2017-06-02"

spokesman: "华春莹"

question_answer_list:

0:

question: "问：6月1日，美国总统特朗普在白宫宣布，美将退出《巴黎协定》并立即停止执行协定的所有减排标准。中方对此有何评论？"

journalist: ""

answer: " 答：中方密切关注美国宣布退出《巴黎协定》问题。我们认为，《巴黎协定》凝聚了国际社会应对气候变化的最广泛的共识，各方应共同珍惜和维护这一来之不易的成果。 中国政府高度重 发展理念，采取切实政策行动积极应对气候变化，取得有目共睹的成效。这既是中国作为发展中大国承担的国际责任，也是中国可持续发展的内在要求。 未来，中国将继续做好应对气候变化各项工作 。中方愿与有关各方加强合作，共同推动《巴黎协定》实施细则的后续谈判和有效落实，推动全球绿色、低碳、可持续发展。 "

1:

question: "问：今天，威海"5·09客车放火案"新闻发布会称，调查结果显示事故原因是校车司机纵火，这是一起事关韩国人的事故。中方对此有何评论？"

journalist: ""

answer: " 答：今天上午，威海有关部门召开了"5·09客车放火案"侦办和善后处置工作情况新闻发布会并介绍了案情。经公安机关侦查认定，发生在山东威海陶家疙瘩隧道的"5·09案件"是一起人为实施的放火案件，事发车辆驾驶员实施了这起个人极端严重暴力犯罪并在作案中死亡。这起不幸事件造成包括中韩儿童在内的12名无辜人员遇难，我们对此深感震惊和悲痛。中方再次对此次案件中的中韩无辜遇难者表示深切哀悼，对他们的家属表示诚挚慰问。 中方将继续全力做好善后工作。希望媒体客观公正报道，充分理解体谅这起不幸事件带给各方特别是家属的悲痛与沉重。 "

2:

question: "问：菲律宾马尼拉一处度假村发生火灾，菲律宾政府方面称是抢劫导致的，在30多名遇难者中是否有中国公民？"

journalist: ""

answer: " 答：中方注意到有关报道。我们对事件中的无辜遇难者表示哀悼，对受伤人员表示慰问。相信菲方能妥善处理这一事件。关于在这次不幸事件中是否有中国公民伤亡，目前仍在核实，中方正就此同菲方保持密切联系。 "

3:

question: "问：据报道，6月1日，联合国的外交官称，美国于当天向安理会提交了对朝鲜施加新制裁的决议草案，预计安理会将于6月2日就议案进行投票表决。你能否证实上述消息？中方对此持何立场？"

journalist: ""

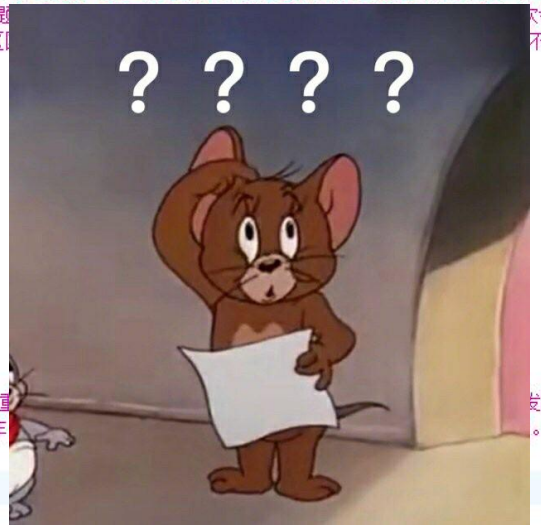
answer: " 答：在半岛问题上，中方始终坚持实现半岛无核化、坚持维护半岛和平稳定、坚持通过对话协商解决问题。对朝鲜试射弹道导弹问题，联合国安理会决议有明确规定。安理会已就近期朝鲜使用弹道导弹技术发射行为发表主席新闻谈话，表明反对立场。 同时，我们近期也多次强调，半岛正面临能否开启对话的关键时期，半岛问题有关各方都应保持克制，不做相互刺激，加剧地区局势紧张的事，共同维护本地区和平稳定。安理会的讨论和行动应有助于实现这一目标。中方本着这一精神和原则参与安理会有关讨论。"

4:

question: "问：关于美国退出《巴黎协定》，许多官员和专家认为美方退出是不负责任的行为。中方怎么看？是否也认为这是不负责任的行为？"

journalist: ""

answer: " 答：中方密切关注美国宣布退出《巴黎协定》问题。我们注意到《联合国气候变化框架公约》秘书处以及多国领导人已就此发表了声明。我们认为，《巴黎协定》凝聚了国际社会应对气候变化问题的最广泛的共识，各方应努力共同珍惜和维护这一来之不易的成果 "

# 4.3 分析


在过去的1000场记者会中，
耿爽回答了 **2822**个问题，
华春莹回答了 **2444**个问题，
赵立坚回答了 **2036**个问题，
陆慷回答了 **1320**个问题，
汪文斌回答了 **1171**个问题。

发言人风格不同

回答每个问题，
汪文斌平均用字最多为 **294字**，
耿爽用词最少，为 **212字**。

# 4.3 分析


**问：能否确认一下，中巴经济走廊框架下合作项目总投资是190亿美元吗？是否是迄今为止的总投资额？**

答：是的。

# 4.3 分析


2021年2月4日，汪文斌驳斥**《环球时报》**关于“**新疆存在针对妇女的系统性性侵与虐待”**的提问，用了**1742字**，有理有据驳斥了涉疆谣言。

**《环球时报》记者：美国务院发言人在回答提问时表示，美国对新疆“再教育营”存在的针对妇女的系统性性侵与虐待深感不安，中国政府应允许国际观察员立即对这些令人震惊的指控进行独立调查。我们也注意到，BBC报道称其获得最新“详细证词”显示新疆存在针对妇女的系统性性侵与虐待。反华学者郑国恩（Adrian Zenz）称BBC搜集到的证词提供了有关性虐待和酷刑的权威且详细的证据。中方对此有何评论?**

汪文斌：昨天我已经就BBC不实报道作出回应。我愿再次强调几点：

第一，新疆根本不存在所谓“再教育营”。新疆依法设立的职业技能教育培训中心属于学校性质，与英国设立的“转化和脱离项目”、法国设立的去极端化中心没有本质区别，都是预防性反恐和去极端化的有益尝试和积极探索，目的是为了从源头上消除恐怖主义、宗教极端主义，完全符合《联合国全球反恐战略》《联合国反暴力极端主义行动计划》等一系列反恐决议的原则和精神。关于教培中心问题，我们已多次介绍，大家可以读一下2019年8月中国国务院新闻办公室发布的《新疆的职业技能教育培训工作》白皮书。

第二，根本不存在所谓“针对女性的系统性性侵和虐待”。中国是法治国家，尊重和保障人权是中国宪法规定的基本原则，并在中国的各项法律制度和中国政府开展的各项工作中得到充分体现。新中国成立以来中国妇女解放和发展事业取得了空前的成就，各族妇女依法享有政治权利、文化教育权利、劳动与社会保障权利、财产权利、人身权利和婚姻家庭权利。

# 4.3 分析

教培中心严格贯彻落实宪法和法律规定，保障参与培训学员的基本权利不受侵犯，严禁以任何方式对学员进行人格侮辱和虐待。在2月1日新疆维吾尔自治区在北京举办的第三场涉疆问题新闻发布会，以及2月3日中国常驻日内瓦代表团和新疆维吾尔自治区共同举办的“新疆是个好地方”视频会议上，都有教培中心结业女学员讲述她们在那里的生活经历，讲述她们是如何摆脱极端思想、掌握劳动技能、过上正常生活的体会和感受，大家可以去看一下有关报道。

第三，关于新疆经济社会发展情况，中方已发布8本涉疆白皮书，新疆维吾尔自治区政府举办了25场新闻发布会，邀请了来自100多个国家的1200多名外交官、记者和宗教团体代表等赴新疆参访。在事实和真相面前，国际上一些反华势力炮制的各种谎言和虚假信息不攻自破。

希望美国政府有关人士正视新疆稳定发展的事实，倾听新疆2500万各族人民的呼声，采取基于事实和负责任的态度，不要被个别媒体的假新闻所误导。我们坚决反对任何外部势力借涉疆问题干涉中国内政，必将继续坚定维护国家主权安全发展利益。

近一段时期我们看到了太多针对新疆、针对中国的虚假信息和抹黑之辞。在这些虚假信息的背后，我们常常看到一些熟悉的名字，比如BBC。

2020年7月17日，BBC“新闻之夜”栏目采访了一名名叫早木热·达吾提（Zumrat Dawut）的维吾尔族女子，此人爆料其各种所谓“证词”。但事实是，这个人编造了太多谎言。她自称曾被“关押在‘再教育营’”，但事实上，她从未在教培中心学习过。她自称曾被“强制绝育”、“摘除子宫”，但事实上，她于2013年3月在乌鲁木齐市妇幼保健院妇产医院生第三个孩子时，自己在分娩志愿同意书上签字，表示同意剖宫产、要结扎，随后在医院做了剖宫产和结扎手术，根本没有被绝育，更没有被摘除子宫。她还声称其年迈的父亲数次遭到新疆当局的“拘押和调查”，不久前去世，死因不明。但事实上，她的父亲一直同其子女正常生活，从未被调查或拘押，于2019年10月12日因心脏病去世。她的三哥和五哥都对这些情况进行了澄清。

# 4.3 分析

这个叫早木热·达吾提的人，已经成为反华势力攻击炒作新疆的演员和工具。

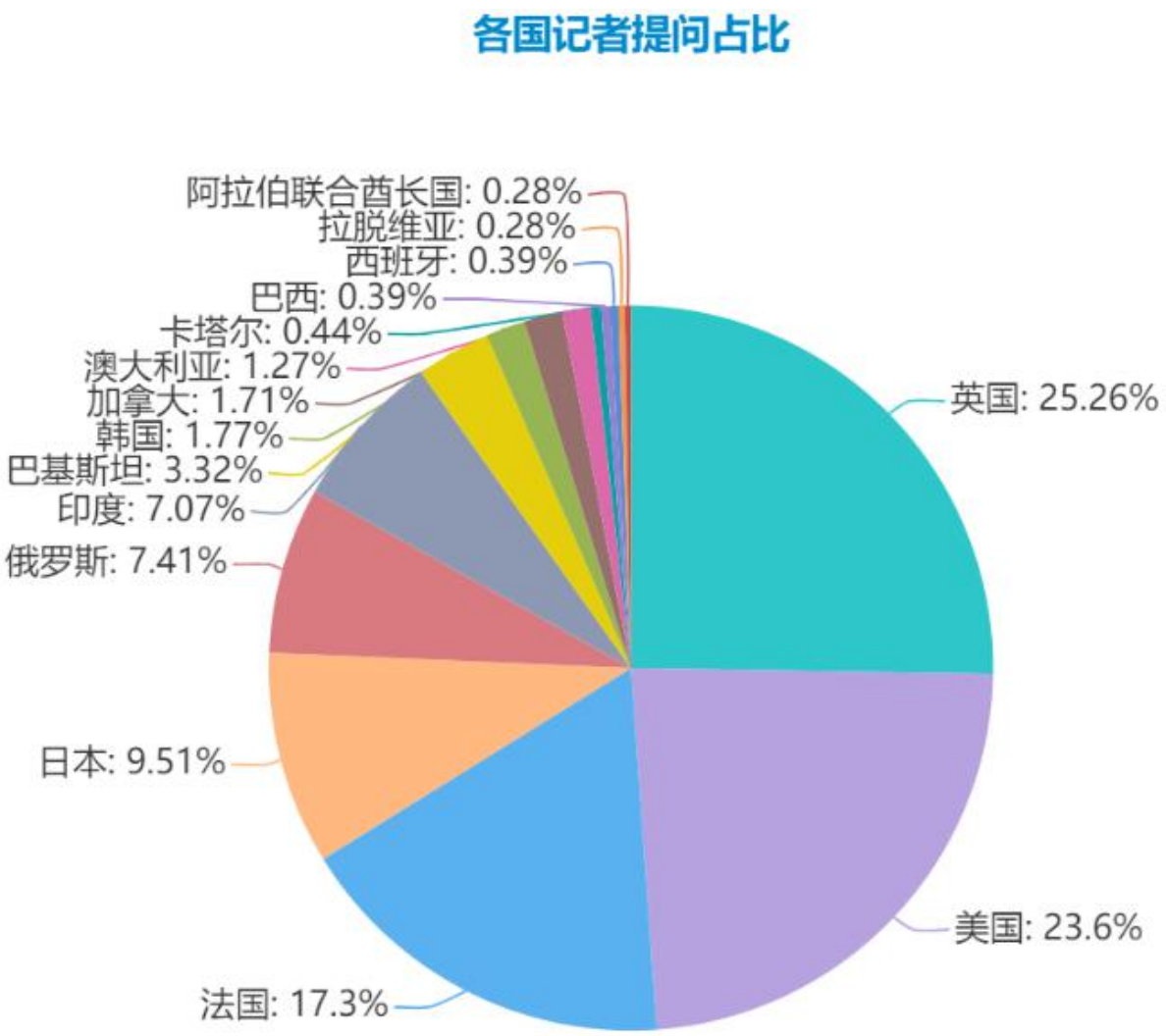
还有一个我们熟悉的名字叫郑国恩，我想大家都不陌生。郑国恩是美国于1983年成立的极右翼组织——“共产主义受害者纪念基金会”的成员。这个人热衷于炮制涉疆谣言，诽谤中国。他发表的有关报告和言论，早已被事实证明是虚假信息。他在报告当中捏造了一个将违法生育者送入教培中心的所谓“墨玉名单”。但事实是，这个名单上的绝大多数人都是墨玉县当地街道的居民，过着正常的生活。只有极个别受到宗教极端思想蛊惑、有轻微违法犯罪行为的人，才依法接受职业技能教育培训。郑国恩还在他的报告中，把新疆正常招录民警，猜测是为所谓“拘留运动”做准备；把深受新疆各族群众欢迎的“访惠聚”工作，想象成是“拘留运动”的“决策基础”；把充分保障儿童上学的寄宿制学校和学前教育，臆想成是“拘留运动”的“兜底保障”；把少数民族群众自主自愿到外地就业，无端地猜测为“强迫劳动”。他的这种生拉硬扯、荒诞不经的“联想”，活脱脱就是一个“不怕不敢想，就怕想不到”的痴人说梦的心态。

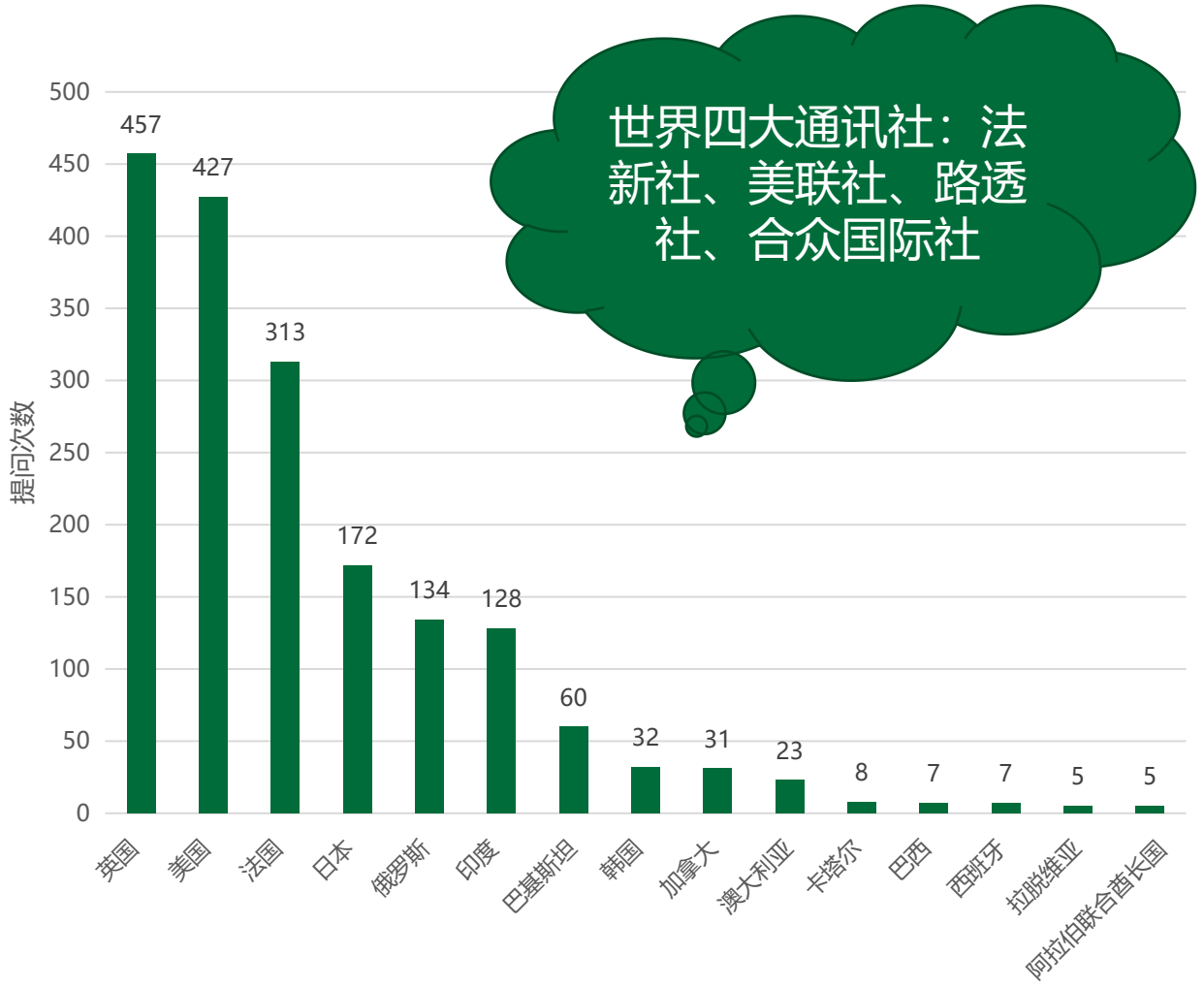
当大家了解到这些事实真相之后，再听到、看到BBC、郑国恩关于涉疆的报道或者报告的时候，大家的脑海里可能会画一个大大的问号：这是不是又一个关于新疆的谎言呢？

# 4.3 分析——热点地区词汇


深圳
中印关系
德国
伊朗
南海
加拿大
武汉
埃及
香港特区政府
俄罗斯
俄方
钓鱼岛
阿富汗
朝鲜
英方
韩方
西方
中澳
香港
津巴布韦
涉疆
印尼
印方
东盟
意大利
印度
华盛顿
东亚
半岛
西班牙
土耳其
巴西
日方
澳方
中日关系
朝方
澳大利亚
香港特区
斯里兰卡
朝鲜半岛
新加坡
马来西亚
巴基斯坦
朝美
加方
东盟国家
尼泊尔
新疆
柬埔寨
美国
金砖
日本政府
巴方
委内瑞拉
伊核
法国
非洲
中美关系
洞朗
湖北
越南
美朝
马尔代夫
欧洲
东盟
巴黎
中巴
中日
美方
菲律宾
缅甸
亚洲
西藏
欧盟
韩国
中俄关系
安理会
中欧
中日韩
联合国安理会
新西兰
福岛
台湾地区
联合国
上海
伊朗核
中朝
巴勒斯坦
中美
美国政府
苏丹
英国
中非
伊拉克
沙特
台湾
涉港
叙利亚

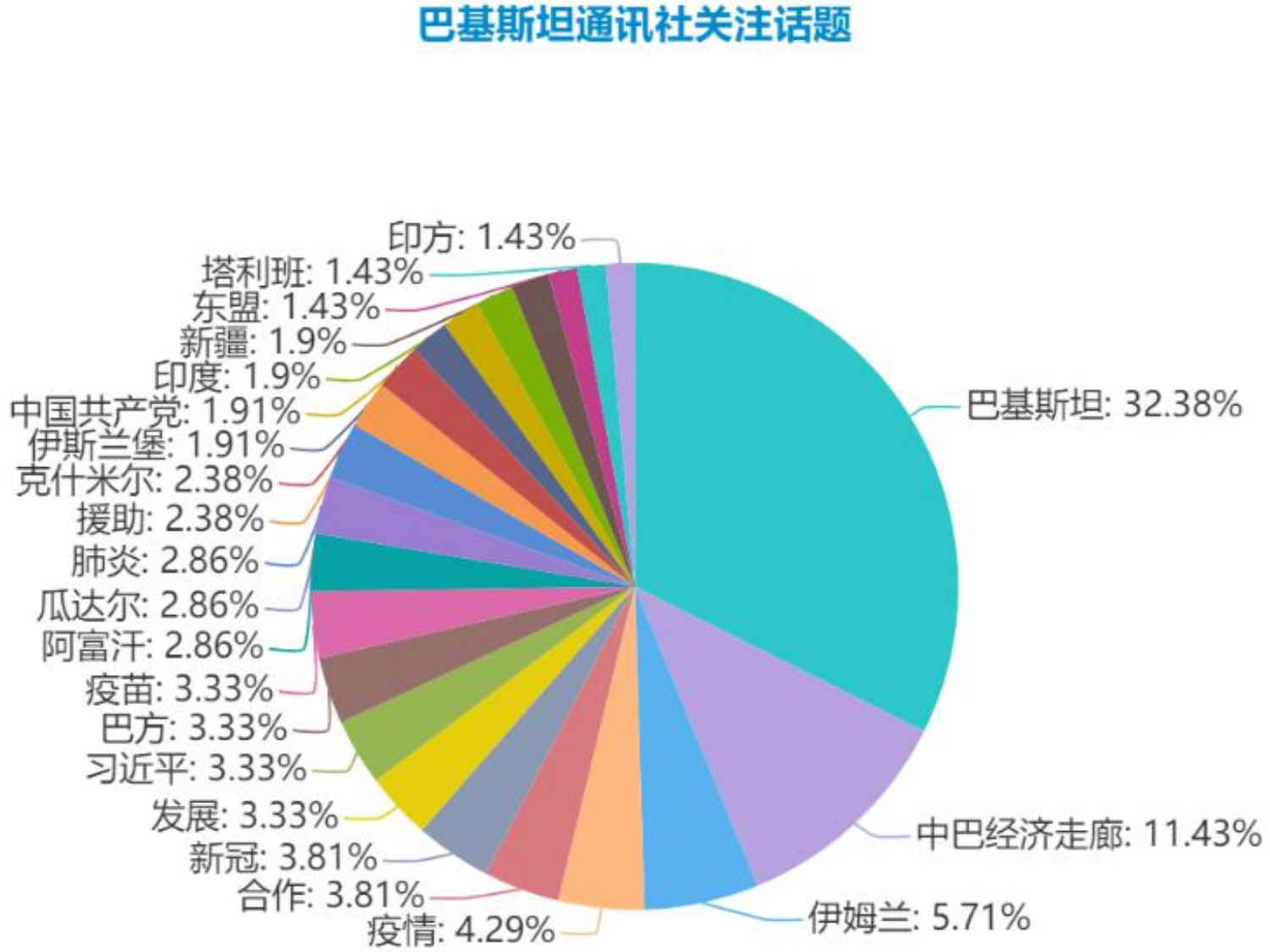
# 4.3 分析——哪国记者提问最多



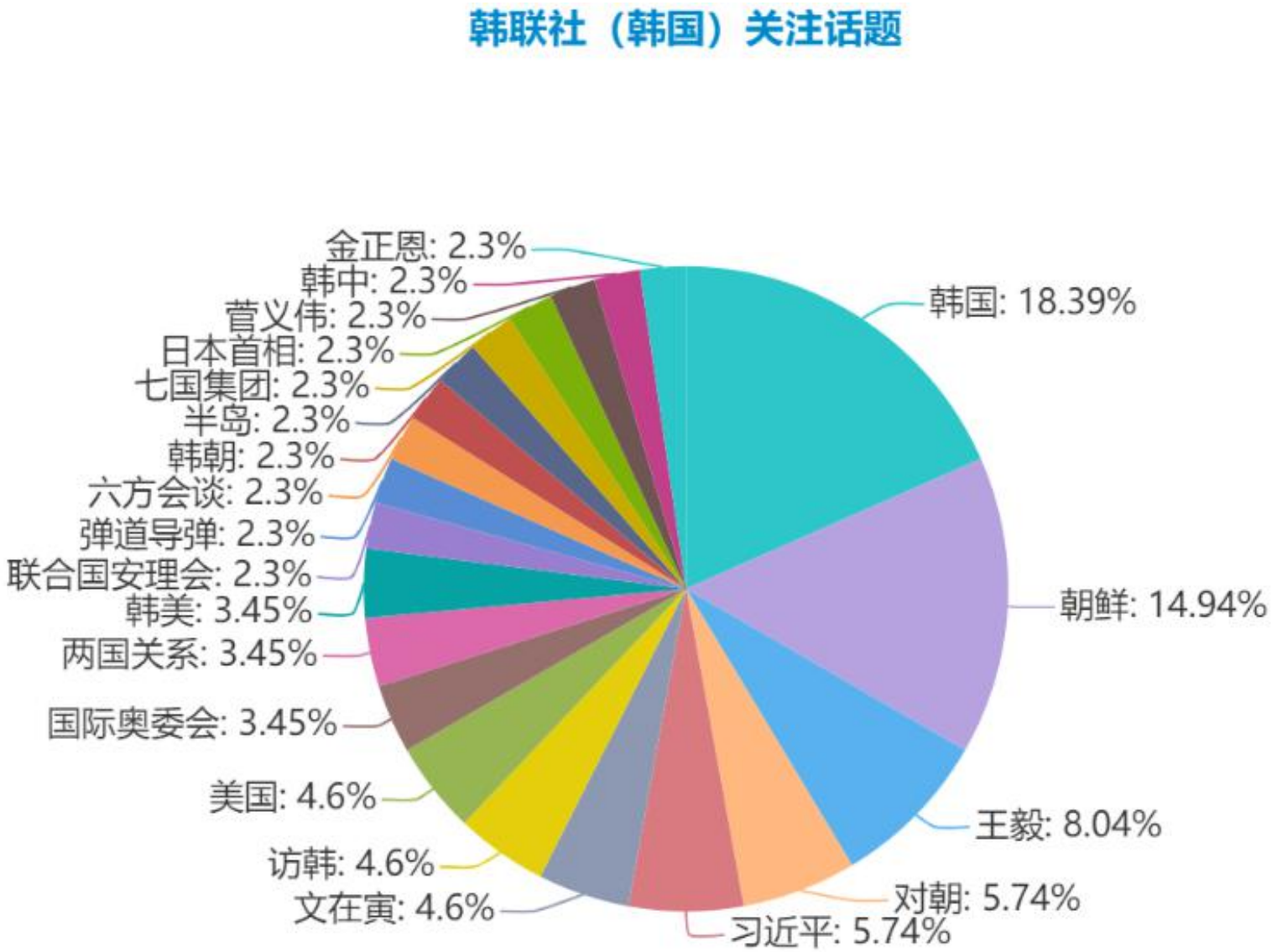




世界四大通讯社：法新社、美联社、路透社、合众国际社

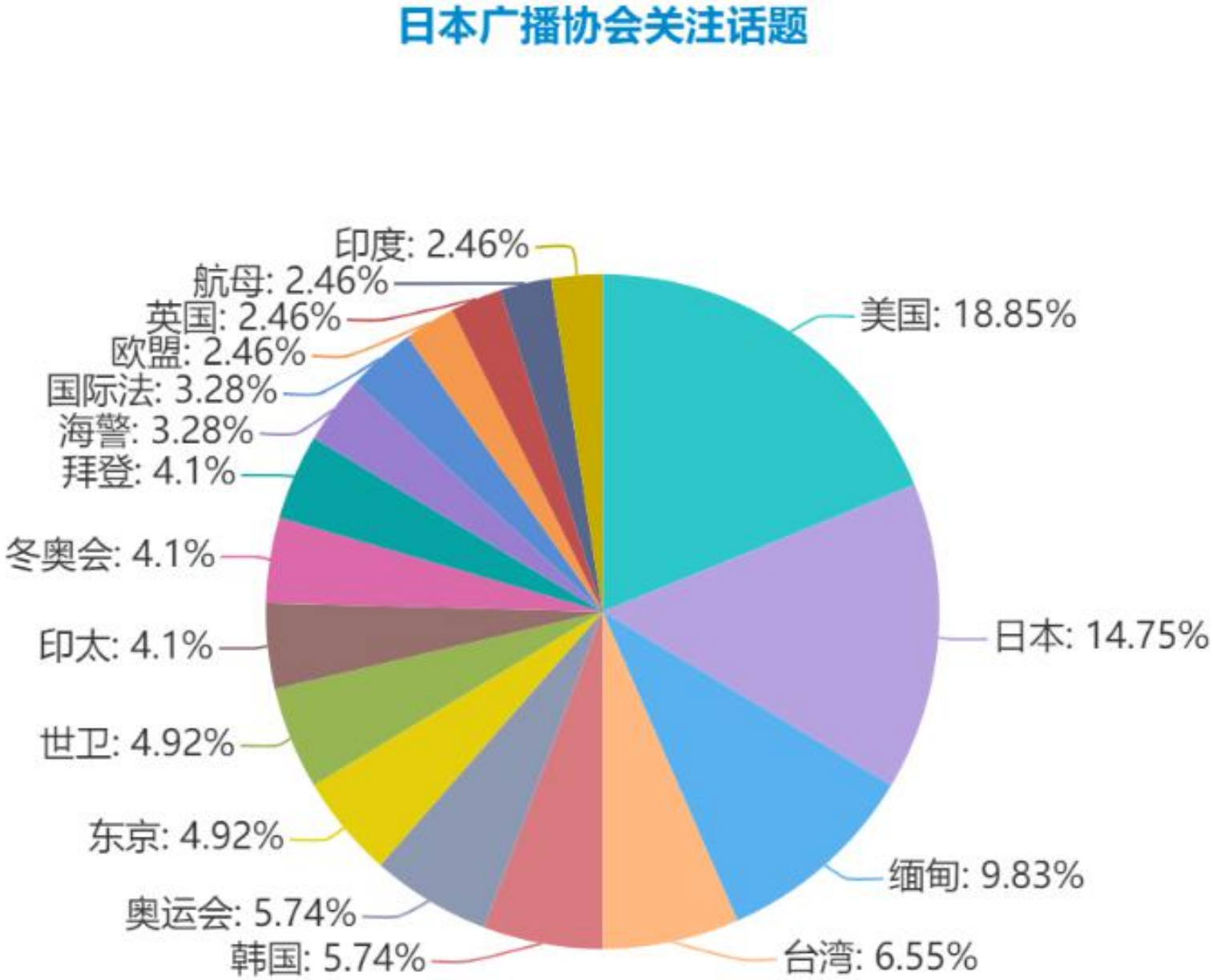
# 4.3 分析——各国记者提问高频词



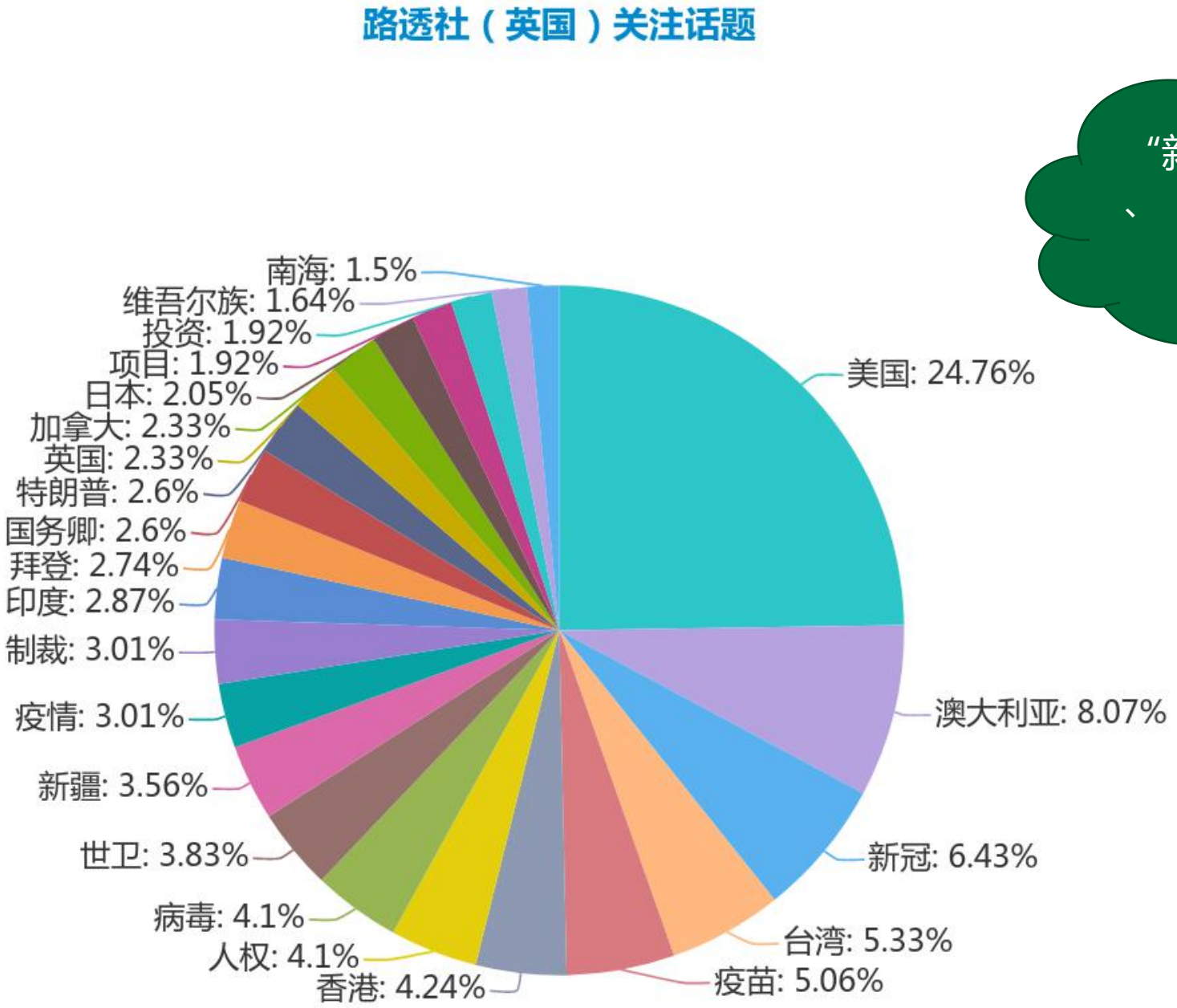
# 4.3 分析——各国记者提问高频词



# 4.3 分析——各国记者提问高频词

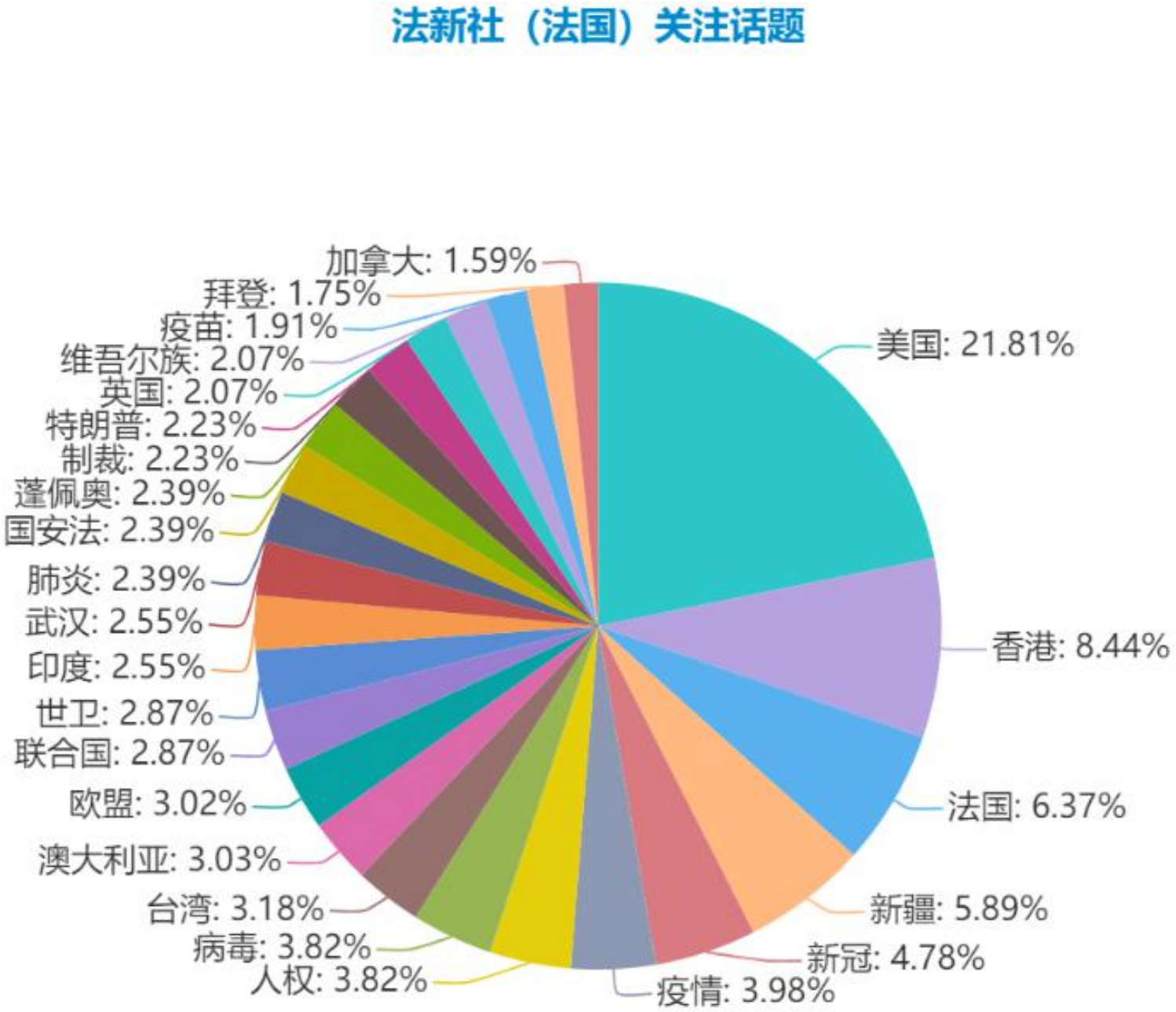


# 4.3 分析——各国记者提问高频词

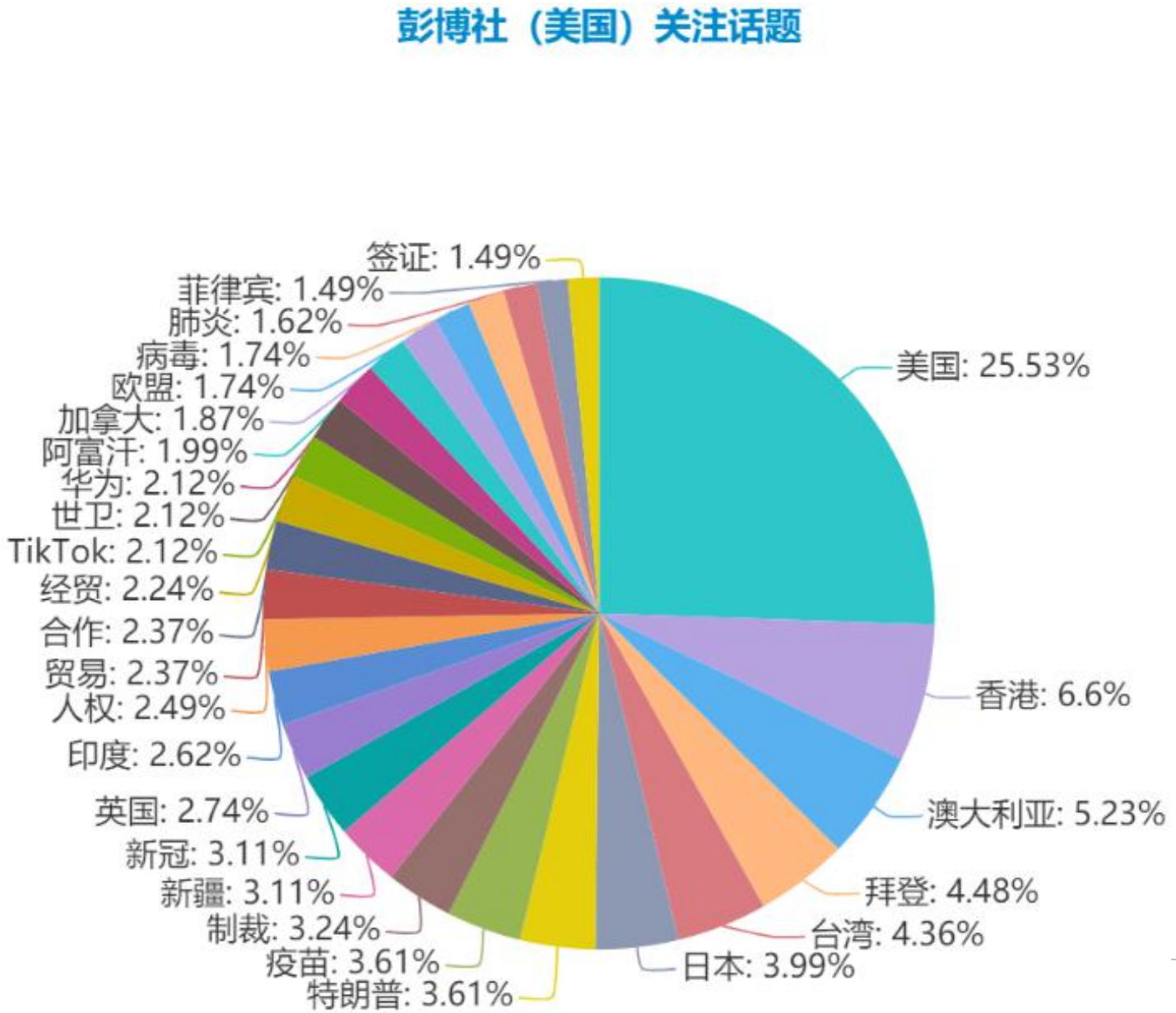



“新疆”、“香港”、“人权”等高频词开始出现

# 4.3 分析——各国记者提问高频词



# 4.3 分析——各国记者提问高频词



05

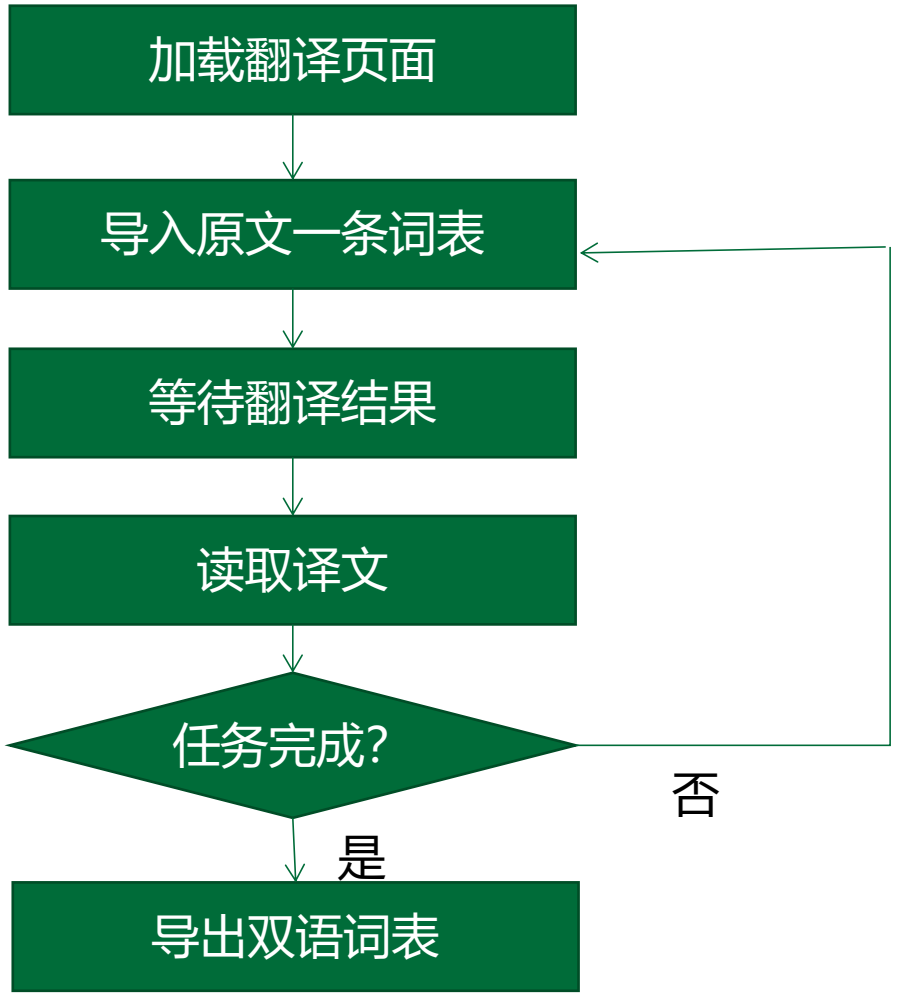
# 爬虫技术前沿

主讲人：单则安

## 基于Gecko浏览器内核的谷歌翻译爬虫

此方法模拟浏览器加载网页，完成用户输入，触发执行脚本，最终获得目标数据。应用上述方法，设计并实现了面向谷歌翻译的专用爬虫，能够采用“多次少取”的方式解决大规模语料的自动翻译问题。

# 5.1 谷歌翻译爬虫




## 网络舆情监测

采用网络爬虫技术从百度指数获取某一“热门事件”的数据，并对这些数据进行预处理，进而建立网络舆情的 Logistic 微分方程模型。结合已有数据，采用智能算法确定微分方程解中的 3 个关键参数；最后应用于网络舆情预测。

## 开展政务信息公开的审计

按传统模式，按5%的比例进行抽查审计，无法实现全覆盖，且需要频繁的人工点击网站、核对数据，效果不好。以完全自主研发的分布式云计算平台为核心，以自定义方式灵活应对不同格式的网页数据源，从而解决各地预决算公开审计难题。

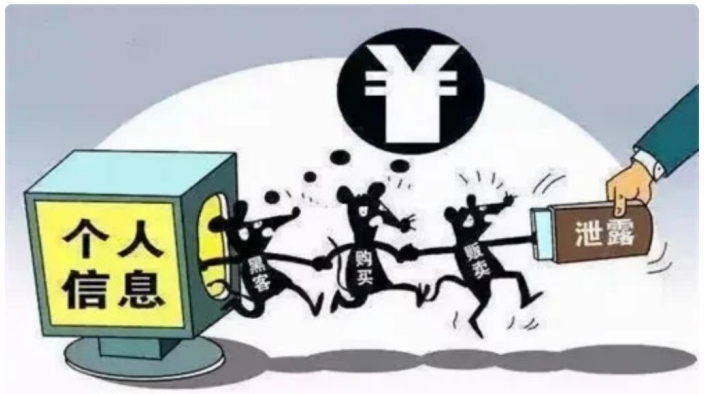
# 5.4.1 反爬虫

# 为什么要反爬?

**警方案例：利用“爬虫”非法获取信息？抓！**

二三里客户端
9月10日 22:02 二三里客户端官方账号 ＋关注

2021年9月份，泰安警方在工作中发现，犯罪嫌疑人刘某等人通过网络“爬虫”等软件非法获取他人信息并将信息出售进行获利。经过缜密调查，警方将犯罪嫌疑人刘某等人抓获，刘某等人对其犯罪事实供认不讳。目前，刘某等人已被刑事拘留。




**北京市海淀区人民法院**

**刑事判决书**

（2017）京 0108 刑初 2384 号

公诉机关北京市海淀区人民检察院。

被告单位上海晟品网络科技有限公司，统一社会信用代码×××，住所地×××，法定代表人张某某。

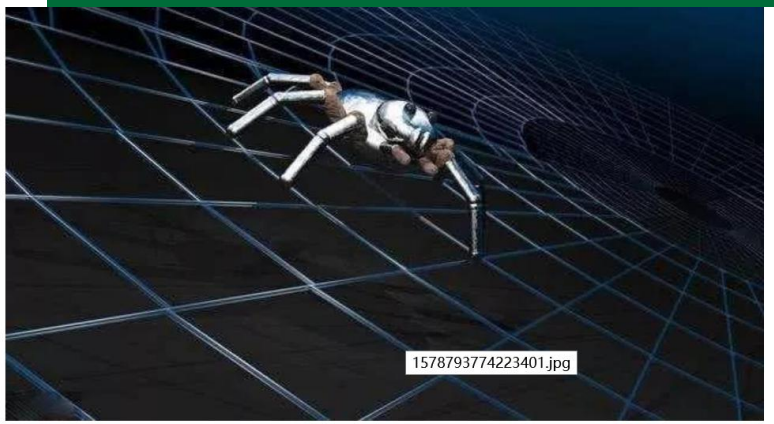
诉讼代表人潘娟，上海晟品网络科技有限公司行政主管

辩护人李智慧，北京市冠衡律师事务所律师。

被告人张某某，男，1982 年 3 月 5 日出生于黑龙江省五常县，公民身份号码×××，汉族，大学肄业，上海晟品网络科技有限公司法定代表人，户籍所在地黑龙江省五常市，因涉嫌犯非法获取计算机信息系统数据罪，于 2017 年 3 月 4 日被羁押，2017 年 4 月 7 日被取保候审。

辩护人王珺，北京市兰台律师事务所律师。

辩护人陈怡，上海宏翰律师事务所律师。




案情简介：上海某网络科技有限公司经营技术开发、技术服务等业务。该公司主要负责人员张某等人经共谋，于2016年至2017年间采用“爬虫”技术非法抓取北京某网络技术有限公司服务器中存储的视频数据。法院以非法获取计算机信息系统数据罪分别判处被告单位罚金20万元，判处被告人张某等四人一年至九个月不等的有期徒刑，并处罚金。据悉，该案系全国首例利用“爬虫”技术非法入侵其他公司服务器抓取数据，进而实施复制被害单位视频资源的案件。

# 怎样反爬?

**1.User-Agent控制请求**

User-Agent中可以携带一串用户设备信息的字符串，包括浏览器、操作系统等信息。我们可以通过在服务器设置user-agent白名单，只有符合条件的user-agent才能访问服务器。它的缺点就是很容易被爬虫程序伪造头部信息，进而被破解掉。

**2.session访问限制**

session是用户请求服务器的凭证，网络爬虫往往通过携带正常用户session信息的方式，模拟正常用户请求服务器。因此，我们同样可以根据短时间内的访问量的大小判断是否为爬虫程序，将疑似爬虫程序的用户的session加入黑名单。

**3.蜘蛛陷阱**

蜘蛛陷阱通过引导爬虫程序陷入无限循环的陷阱，消耗爬虫程序的资源，导致其崩溃而无法继续爬取数据。此方法的缺点就是会新增许多浪费资源的文件和目录，而且对正常网站排名有影响，会造成搜索引擎的爬虫程序也无法爬取信息，进而导致在搜索引擎的网站排名靠后。

# 怎样反爬?

**4.IP限制**

我们可以在服务器设置一个阈值，将短时间内访问量大的IP地址加入黑名单，禁止其访问，以达到反爬虫的目的。

**5.验证码**

在用户登录或访问某些重要信息时可以使用验证码来阻挡爬虫程序。验证码分为图片验证码、短信验证码、数值计算验证码、滑动验证码、图案标记验证码等。

**6.数据加密**

前端请求服务器前，将请求参数、user-agent、cookie等参数进行加密，用加密后的数据请求服务器，这样的话网络爬虫程序不知道我们的加密规则，就无法进行模拟请求我们的服务器。但是，这种方式的加密算法是写在js代码里的，很容易被用户找到并且破解。

**7.对 Cookie 进行限制**

用户向访问网站发送 Request 时,数 据中会包含特定的 Cookie 数据, 网站将会通过对Cookie值的验证来判断该用户操作是爬虫脚本还是真实的用户,当用户第二次及第 三次打开网页访问无 Cookie数据时,则说明该操作为爬虫脚本。

# 5.4.3 反爬虫



**爬虫技术发展** >> **反爬虫保护隐私** >> **针对反爬的爬虫技术**

爬虫这种自动化技术的确为人类互联网带来了许多好处,但是同样的,滥用爬虫技术也会有很多坏处,水能载舟亦能覆舟,因此,我们要在法律规制和自身行为规范下学会正确使用这种技术,才能最大化的发挥其优势,避免造成对互联网环境的危害。

# 感谢聆听