## 基于语义解析的网页自动摘要提取研究

1. **开发环境：**
   1. Windows 10 64 位系统 / macOS High Sierra 10.13.4
   2. Python 3.6.4 (64 位)，至少需要安装下列依赖库

可以直接 pip install xxx，附上开发时用的版本

* + 1. Django – 2.0.2，项目 Web 框架
    2. PyMySQL – 0.8.0，数据库接口
    3. BeautifulSoup4 - 4.6.0，解析网页代码，提取内容
    4. Requests – 2.18.4，发送网页请求，获取内容
    5. SnowNLP – 0.12.3，中文自然语言处理库
  1. MySQL 5.7.x，数据库，安装完毕后需要配置好，设置好中文编码！
  2. PyCharm 2017.3.2，Professional 版本，学生可以免费使用，或者也可以破解。可直接创建 Django 项目，免去许多配置

为保证正确运行，请尽量使用相同的开发环境。 涉及到python 的路径设置，请在 Pycharm中调试时，选择正确的运行路径。

注意：涉及到中文文本的处理，请务必使用 UTF-8 编码！

1. **项目结构：（基于 Django 开发的Web界面）**
2. newsex，管理整个 Web 的前后端交互、请求、渲染等；
3. newsCrawler，网页爬虫，获取新闻数据，同时根据新闻文本生成关键词和自动摘要，并写入数据库中
4. NewsAbExtraction 文件夹，Django 项目的设置
5. Static 文件夹，网页端使用到的 css 样式和一些 js 代码，用于渲染网页
6. Templates 文件夹，html 网页代码，通过 Django 注入数据，作为前端 UI
7. Sql 文件是数据库表的建立和数据插入
8. **主要代码文件说明**

本项目是 Django 框架开发的，可以不使用网页 UI 直接运行代码，命令行输出。

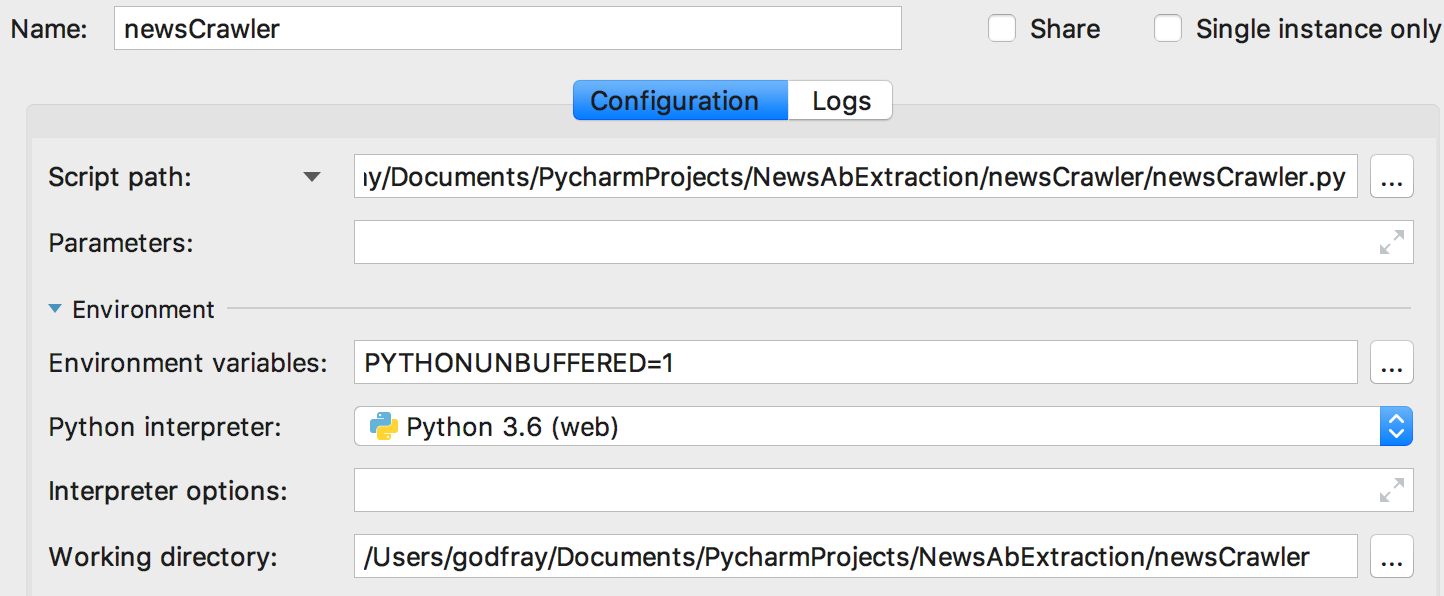
* 1. 新闻网页内容的提取，主要在 newsCrawler 文件夹下
* databaseUtil.py，提供了一系列数据库中新闻实体的插入、按要求查询等接口
* newsCrawler.py，是通过爬虫获取网页信息、解析、计算后，生成结构化化的数据表示，可供数据库读取写入。这个文件的最后是 main 函数，即运行爬虫的入口
  1. Django 框架，网页 UI
* 主要是newsex下的 views.py，里面提供了各种请求的处理和前端渲染
* 在项目中的 manage.py 文件是运行网页的入口，包括 newsex 文件夹中的大多数文件在内都是创建项目时生成的

1. **程序运行方式**

可以通过 Pycharm 导入打开，支持命令行运行和 Web UI 运行。

*注意：由于 python 环境和项目设置的路径不同，需要根据实际情况修改。直接运行可能会出错。有的时候需要重新选择一下 Python interpreter 就没有问题了。*

1. 运行爬虫程序，手动修改爬取新闻的页码范围，运行 newsCrawler.py，即在合法目录下，输入 “python newsCrawler.py”，等待结果输出，写入数据库。 Pycharm 中的设置主要是 Script path 和 Working directory



B) Web UI 演示运行，命令行中，在第一层文件夹下，使用 “python manage.py runserver 8000” 运行，当有如下的提示后，可以访问 <http://127.0.0.1:8000> ，看到项目的 UI 界面，下面的操作请根据演示视频直接在网页上操作即可。

System check identified no issues (0 silenced).

April 06, 2018 - 12:56:27

Django version 2.0.2, using settings 'NewsAbExtraction.settings'

Starting development server at http://127.0.0.1:8000/

Quit the server with CONTROL-C.

推荐在 Pycharm 中调试与运行，但是需要 Professional 版本，保持默认配置即可

