

PyXinwen 的功能与实现

于是 2017011414

PyXinwen 的功能与实现

- 功能概览
- 性能信息
- 算法简述
 - 查询算法
 - 推荐算法
- 其他说明

功能概览

PyXinwen 是一个使用 Django 搭建的浏览、检索近期新闻的网站。新闻是从新华网上抓取的，共1987条。

在本地运行服务器之后，输入 `127.0.0.1/news/` 进入。首页显示第1至10条新闻。每条新闻显示其标题、发布时间、摘要，如图：



页面上方的输入框可按关键词、发布日期检索新闻。下方有导航栏。

2018-09-10 07:34:08

“以房养老”是否划算？“以房养老”试点遇冷为何还要在全国开展？“以房养老”能走多远？——日前，银保监会发布了《关于扩大老年人住房反向抵押养老保险开展范围的通知》，将“以房养老”保险由原来的试点城市，扩

药品降价大势所趋 降低抗癌药价还需加把劲

2018-09-10 07:36:18

多省份调降进口药及抗癌药价格，但部分药价降幅不大——降低抗癌药价还需加把劲 徐 骏作（新华社发） 为了让更多患者顺利利用上抗癌药物，国家相关部门正不断出台新措施、新政策加快推进进口药及抗癌药降价工作

第1页，共199页

首页

»

末页

让我们来检索一条新闻看看。

北京

从

06/08/2018

到

06/30/2018

搜索

这是跳转到的检索结果的界面：

搜索结果

127.0.0.1:8000/news/search/?csrfmiddlewaretoken=h4eZ6tl5S0pcoPpnAygB2xlp3C800f1...

«回到首页

「北京」的搜索结果

共132条，用时0.00099945068359375秒

北京7月1日起将严查电动车违规销售

2018-06-27 06:58:54

不合格电动自行车、电动摩托车、电动三轮及四轮车、老年代步车.....从7月1日起，北京市将对这些违规销售电动车商户依法进行严厉查处，线上线下禁售规则相同。昨天下午，北京市工商局再次召集

北京城市副中心将新增住房18万套 通勤不超半小时

2018-06-26 07:10:32

北京城市副中心将建36个美丽家园 规划新增住房约18万套；绿色出行比例达到80%，内部公共交通通勤时间不

北京城市副中心控规草案出炉 专家：有望实现六元平衡

2018-06-22 09:45:59

据中国之声《新闻纵横》报道，昨天（21日），北京市规划国土委、通州区政府共同发布通知，《北京城市副中心控制性详细规划（街区层面）》草案已经编制完成，

北京确认1685名见义勇为者 符合条件者积分落户增20分

2018-06-21 20:21:21

见义勇为是中华民族的传统美德，截至目前，北京市共确认见义勇为人员1685人。北京每三年表彰一次相关人员，第十二届评选表彰工作将于今年底启动筹备工

上方显示检索出的条数，还有检索耗时。下方呈现检索结果，标题和正文中的关键字会被高亮。页面下方也有导航栏。

北京最大立体公交停车楼落户副中心 明年底主体完工

2018-06-15 07:58:21

原标题：北京最大立体公交停车楼落户副中心 东小营中心站可同时停放26

第1页，共14页

首页

»

末页

127.0.0.1:8000/news/detail/798/

点进一条新闻，进入详情页面：

新闻

127.0.0.1:8000/news/detail/849/

☆ 微信 钉钉 微博 头条 知乎 豆瓣 小红书 哔哩哔哩 快手 抖音 快手 抖音 快手 抖音

[«回到首页](#)

北京停车场将推广“不停车”收费 支付平台自动扣费

2018-06-12 08:44:27

原标题：北京停车场将推广“不停车”收费 在高速公路每万次不停车支付可省314升燃油；市民可通过安装ETC速通卡或手机下载APP方式实现 2017年7月28日，在北京沃尔玛山姆会员店，一位车主缴费时无需停车便驶出停简单智慧停车场。今后，北京将推广“不停车”收费，省油减排。据测算，在高速公路每万次不停车支付可节约314升燃油，减排二氧化碳670千克。

昨日（11日），2018年全国节能宣传周和全国低碳日暨北京市节能宣传周低碳日启动仪式在京举行。北京市发改委相关负责人表示，目前北京市节能减排工作进入新阶段，急需进一步加强智慧化、精细化管理。对此，全市将开展不停车收费推广活动，北京速通科技有限公司、ETCP、“停简单”等企业，将分别开展线上、线下的多种优惠活动，推动更多市民车主了解和使用不停车收费。

目前，三家企业已与市内一些停车场展开合作，其中“停简单”约有1200家，ETCP到年底将达到约2000家，ETC在首都国际机场三个航站楼、西单大悦城、国贸中心等70余家停车场相继开通了速通卡代付停车费业务。释疑1“不停车”收费能省多少能源？北京市新一轮细颗粒物(PM2.5)来源解析研究表明，市内全年PM2.5主要来源中，本地排放占三分之二。而在本地源中，移动源占比达45%，为所有污染源之首。

治理机动车污染排放成为北京市应对大气环境污染，减少温室气体排放的重要环节。据了解，机动车在排队交费、寻找车位等环节常处于怠速状态，能耗和污染物排放均高于正常行驶，同时也增加了交通拥堵。不停车收费模式可有效缓解上述问题，实现节能减排缓堵的综合效益。

从节能减排效果看，综合考虑小汽车通行时间、怠速时间、排队长度等各种因素，对比人工收费和不停车收费两种模式，据研究，在高速公路收费站，每1万次不停车支付可节约314升燃油，减排二氧化碳670千克；在停车场收费站，每1万次不停车支付可节约306升燃油，减排二氧化碳653千克。“停简单”董事长柳文超介绍，据内部测算，凭借车辆无感支付离场这一点，“停简单”每天可为全市节

拉到最下方，可以见到相关新闻推荐：

安装了ETC速通卡的车主实现不停车收费又是通过另外一种路径，车主的设置感知部件，车主进场、出场时感知部件会与车上的电子标签联动得接连通速通卡实现扣费。据了解，ETC最初应用于高速公路，在经过一定李洁祎) +1 不停车收费停车场支付宝北京 【纠错】 责任编辑：聂晨

北京新建公共停车场须配建充电设施

北京停车场将推广“不停车”收费

北京南站首设网约车专属车位 地铁末班车延长

京津冀高速路出行信息将共享发布

京津冀高速路出行信息将共享发布 三地服务区和收费站将统一标准

这次我们检索“北京 停车”，试一试多关键字检索：



性能信息

共有新闻 1987 条，查询耗时在 $ms \sim 10^{-1}ms$ 数量级。

算法简述

查询算法

构建了三个模型，如下：

```
from django.db import models
```

```

class NewsPiece(models.Model):
    news_title = models.CharField(max_length=200)
    news_content = models.CharField(max_length=10000)
    pub_date = models.CharField(max_length=200)
    news_abstract = models.CharField(max_length=200)
    def __str__(self):
        return self.news_title

class Tag(models.Model):
    newspiece = models.ForeignKey(
        NewsPiece,
        on_delete=models.CASCADE,
    )
    name = models.CharField(max_length=255)
    def __str__(self):
        return self.name

class ContentTag(models.Model):
    newspiece = models.ForeignKey(
        NewsPiece,
        on_delete=models.CASCADE,
    )
    name = models.CharField(max_length=255)
    def __str__(self):
        return self.name

```

`NewsPiece` 是新闻模型，`Tag` 是标题中分出的词的模型，`ContentTag` 是正文中分出的词的模型。新闻模型与分词模型是「一对多」的外键关系。

查询时，使用 `NewsPiece.objects.filter(tag__name='关键词')` 以及 `NewsPiece.objects.filter(contenttag__name='关键词')` 进行检索。得到查询结果后，从标题和正文中找到关键词，插入 `` 标签，再传递给模板。

推荐算法

出于尽量模拟真实情况的考量，决定采用「在线」的算法。算法流程如下：

- 点开一篇文章时，对于文章的分词集中的每一个词，计算其 `tf-idf` 值。`tf-idf` 值是一种综合考量词频、词的独特性的值，值越高表现一个词对该文章的代表性越高。
- 取 `tf-idf` 值最高的两个词，检索包含这两个词的所有文章。
- 对于上步得到的每一篇文章，计算其与本文的 *jaccard* 相似度。选取相似度最高的5篇显示。

本算法是综合性能和效果考量的。理论上，直接计算全部文章与本文的 *jaccard* 相似度，便能够得到最精确的推荐，但这样算法耗时过长，极度影响用户体验。因此，在算法中添加了第二步，即依据 `tf-idf` 值先对文章进行筛选，从而减少需要进行 *jaccard* 相似度计算的文章数量。实测效果较好，是好的折衷算法。

其他说明

`polls` 文件夹是当时按照官方文档进行练习时建立的 App，与本项目无关。

爬虫的代码在 `/news/views.py` 中的 `init_data` 函数中。进入 `127.0.0.1/news/work/` 运行爬虫。