Rational Search Engine

Zihan Qiu
2020011628

Zhengzhao Ma
2020012362

Zihan Zhou
2020010749

Abstract

”clickbait” and ”information cocoon” are nonnegligible issues when people more and more rely on Internet. The information we get actually affect our cognition about the world, therefore we hope to develop a tool that can help users to efficiently, and more importantly, ”rationally” make use of the vast Internet, easily getting rich desired information. In this project, we utilize NLP tools to build a readily available and practical system. Our codes are in this repository

1 Introduction

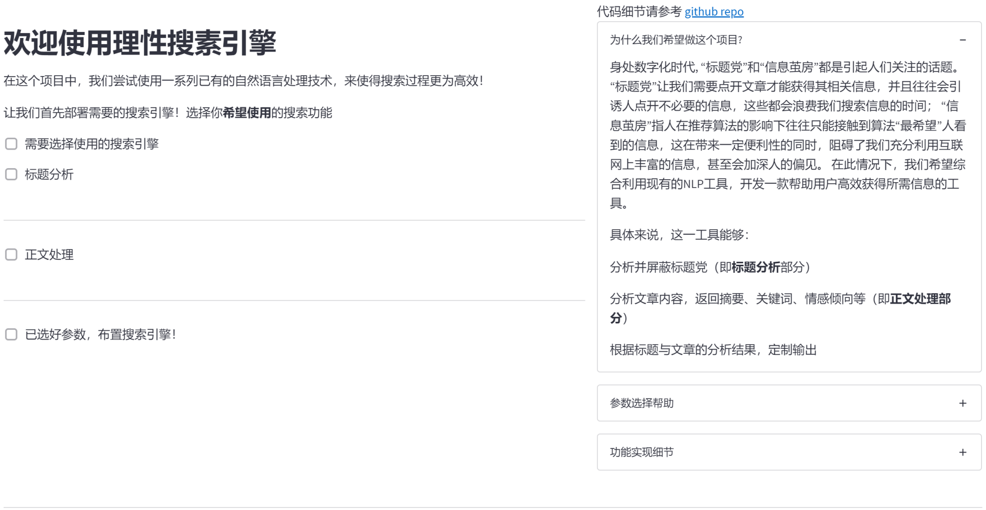
In the information age, ”clickbait” and ”information cocoon” are often topics of concern. Clickbaid is common and annoying, it makes us hard to get the articles' main idea without opening it, which waster lots of time. Although new technologies have given us the ability to access and produce massive amounts of information, it's hard to utilize diversity news with limited time and concentration; this situation becomes even worse when powerful recommendation algorithms rule our information channel. Under this circumstance, we think, if the search engine can help us process the search results, we can get diversity information more efficient. To be concrete, we hope the search engine can

for titles: analyze the title and detect clickbait, hid the title if necessary;

for contexts: avoid duplicate search result, give summaries about the articles, give rough emotional tendency and content classification;

for search results' output: enable customized rank (like according to the reliability of the title) and demonstrate (like whether showing summarise).

With these functions, we hope this search engine can help users to obtain information efficiently. Thanks to the fast development of NLP field and vast python community, we can build a comprehensive system which alleviates the problems and carries forward the advantages of internet.



欢迎使用理性搜索引擎

在这个项目中，我们尝试使用一系列已有的自然语言处理技术，来使得搜索过程更为高效！

让我们首先部署需要的搜索引擎！选择你**希望使用**的搜索功能

□ 需要选择使用的搜索引擎

□ 标题分析

□ 正文处理

□ 已选好参数，布置搜索引擎！

代码细节请参考 github repo

为什么我们希望做这个项目? −

身处数字化时代, “标题党”和“信息茧房”都是引起人们关注的话题。“标题党”让我们需要点开文章才能获得其相关信息，并且往往会引诱人点开不必要的信息，这些都会浪费我们搜索信息的时间；“信息茧房”指人在推荐算法的影响下往往只能接触到算法“最希望”人看到的信息，这在带来一定便利性的同时，阻碍了我们充分利用互联网上丰富的信息，甚至会加深人的偏见。 在此情况下，我们希望综合利用现有的NLP工具，开发一款帮助用户高效获得所需信息的工具。

具体来说，这一工具能够:

分析并屏蔽标题党（即**标题分析**部分）

分析文章内容，返回摘要、关键词、情感倾向等（即**正文处理部分**）

根据标题与文章的分析结果，定制输出

Figure 1: Initial interface of our Rational Search Engine

2 Related Work

[2] analyzes the use of NLP techniques in search engines; [3] provides a time-based and keyword-based filtering system with a stepwise filtering according to users' choice, but only depends on keywords in titles, which fails when the titles are 'messy' like clickbits; [4, 5] provides a method for using NLP techniques in news headline classification; [1] provides a reference for context classification. With reference to the evaluating article titles' process in [7], we use keywords and grammar to mark titles as well as detecting clickbit.

3 Processing Method

The system can be roughly divided into four stages, including extensive crawling, title and context processing and final result showing.

3.1 Get Content from Internet

We use python program to call given web severs for search result with given keywords.

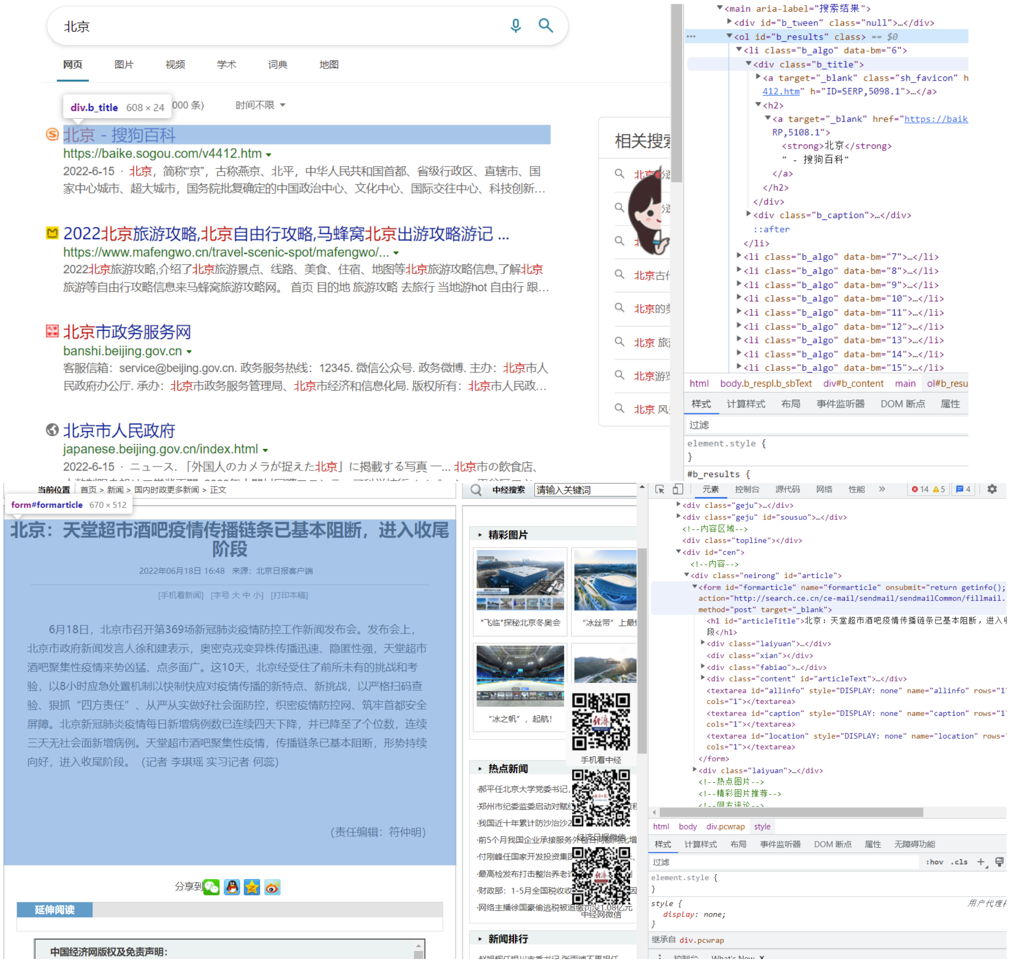
3.1.1 Package Used

Request: getting html code from web severs, do simple division and operation that allow the users to get some content.

BeautifulSoup: a package that can get data and content from html or xml files, operate and process them with less time used.

For more details, please visit document of BeautifulSoup and document of Request.

3.1.2 Content Gaining



北京 - 搜狗百科
https://baike.sogou.com/v4412.htm
2022-6-15 · 北京，简称"京"，古称燕京、北平，中华人民共和国首都、省级行政区、直辖市、国家中心城市、超大城市，国务院批复确定的中国政治中心、文化中心、国际交往中心、科技创新...

2022北京旅游攻略,北京自由行攻略,马蜂窝北京出游攻略游记 ...
https://www.mafengwo.cn/travel-scenic-spot/mafengwo/...
2022北京旅游攻略,介绍了北京旅游景点、线路、美食、住宿、地图等北京旅游攻略信息,了解北京旅游等自由行攻略信息来马蜂窝旅游攻略网。 首页 目的地 旅游攻略 去旅行 当地游hot 自由行 跟...

北京市政务服务网
banshi.beijing.gov.cn
客服信箱：service@beijing.gov.cn. 政务服务热线：12345. 微信公众号. 政务微博. 主办：北京市人民政府办公厅. 承办：北京市政务服务管理局、北京市经济和信息化局. 版权所有：北京市人民政...

北京市人民政府
japanese.beijing.gov.cn/index.html
2022-6-15 · ニュース. 「外国人のカメラが捉えた北京」に掲載する写真 —... 北京市の飲食店、

```
<main aria-label="搜索结果">
  <div id="b_tween" class="null">…</div>
  <ol id="b_results" class> == $0
    <li class="b_algo" data-bm="6">
      <div class="b_title">
        <a target="_blank" class="sh_favicon" h
        412.htm" h="ID=SERP,5098.1">…</a>
        <h2>
          <a target="_blank" href="https://baik
          RP,5108.1">
            <strong>北京</strong>
            " - 搜狗百科"
          </a>
        </h2>
      </div>
      <div class="b_caption">…</div>
      ::after
    </li>
    <li class="b_algo" data-bm="7">…</li>
    <li class="b_algo" data-bm="8">…</li>
    <li class="b_algo" data-bm="9">…</li>
    <li class="b_algo" data-bm="10">…</li>
    <li class="b_algo" data-bm="11">…</li>
    <li class="b_algo" data-bm="12">…</li>
    <li class="b_algo" data-bm="13">…</li>
    <li class="b_algo" data-bm="14">…</li>
    <li class="b_algo" data-bm="15">…</li>
```

```
element.style {
}
#b_results {
```

北京：天堂超市酒吧疫情传播链条已基本阻断，进入收尾阶段

2022年06月18日 16:48 来源：北京日报客户端

[手机看新闻] [字号 大 中 小] [打印本稿]

6月18日，北京市召开第369场新冠肺炎疫情防控工作新闻发布会。发布会上，北京市政府新闻发言人徐和建表示，奥密克戎变异株传播迅速、隐匿性强，天堂超市酒吧聚集性疫情来势凶猛、点多面广。这10天，北京经受住了前所未有的挑战和考验，以8小时应急处置机制以快制快应对疫情传播的新特点、新挑战，以严格扫码查验、狠抓"四方责任"、从严从实做好社会面防控，织密疫情防控网、筑牢首都安全屏障。北京新冠肺炎疫情每日新增病例数已连续四天下降，并已降至了个位数，连续三天无社会面新增病例。天堂超市酒吧聚集性疫情，传播链条已基本阻断，形势持续向好，进入收尾阶段。（记者 李琪瑶 实习记者 何蕊）

（责任编辑：符仲明）

·郝平任北京大学党委书记,
·郑州市纪委监委启动对赋
·我国近十年累计防沙治沙2
·前5个月我国企业承接服务外包合同额同比增
·付刚峰任国家开发投资集团
·最高检发布打击整治养老
·财政部：1-5月全国税收收
·网络主播徐国豪偷逃税被追缴罚没1.08亿元

```
<div class="geju">…</div>
<div class="geju" id="sousuo">…</div>
<!--内容区域-->
<div class="topline"></div>
<div id="cen">
  <!--内容-->
  <div class="neirong" id="article">
    <form id="formarticle" name="formarticle" onsubmit="return getinfo();
    action="http://search.ce.cn/ce-mail/sendmail/sendmailCommon/fillmail.
    method="post" target="_blank">
      <h1 id="articleTitle">北京：天堂超市酒吧疫情传播链条已基本阻断，进入收
      段</h1>
      <div class="laiyuan">…</div>
      <div class="xian"></div>
      <div class="fabiao">…</div>
      <div class="content" id="articleText">…</div>
      <textarea id="allinfo" style="DISPLAY: none" name="allinfo" rows="1"
      cols="1"></textarea>
      <textarea id="caption" style="DISPLAY: none" name="caption" rows="1"
      cols="1"></textarea>
      <textarea id="location" style="DISPLAY: none" name="location" rows="
      cols="1"></textarea>
    </form>
    <div class="laiyuan">…</div>
    <!--热点图片-->
    <!--精彩图片推荐-->
```

```
element.style {
}
style {
  display: none;
}
```

Figure 2: Top: Search Engines' code of the result pages; Down: the html code of detected pages

To get suitable format of search results for downstream processing, we execute the following steps:

Use Request package to access the search Engines' servers and gain the code of the result pages, shown in **Figure 2 Top**.

Use ETREE to get titles of search result from the specific area in the search engine's page, get the links of the news page by detecting the hyperlinks in these areas.

Use Request again to get the html code of the news pages detected, change their encoding such that we can detect Chinese characters in there pages.

Use bs4 to divide all blocks in the news page and get their content with Chinese strings, shown in **Figure 2 Down**.

Take the blocks with more than 50 Chinese characters, merge them together as the main content of the page searched. Thus, the noise in the news page, like advertisements, can be removed.

3.1.3 Search Engines Used

Baidu News: all webs found are news webs, while some of the links has no stable connection, thus the number of webs after screening is less.

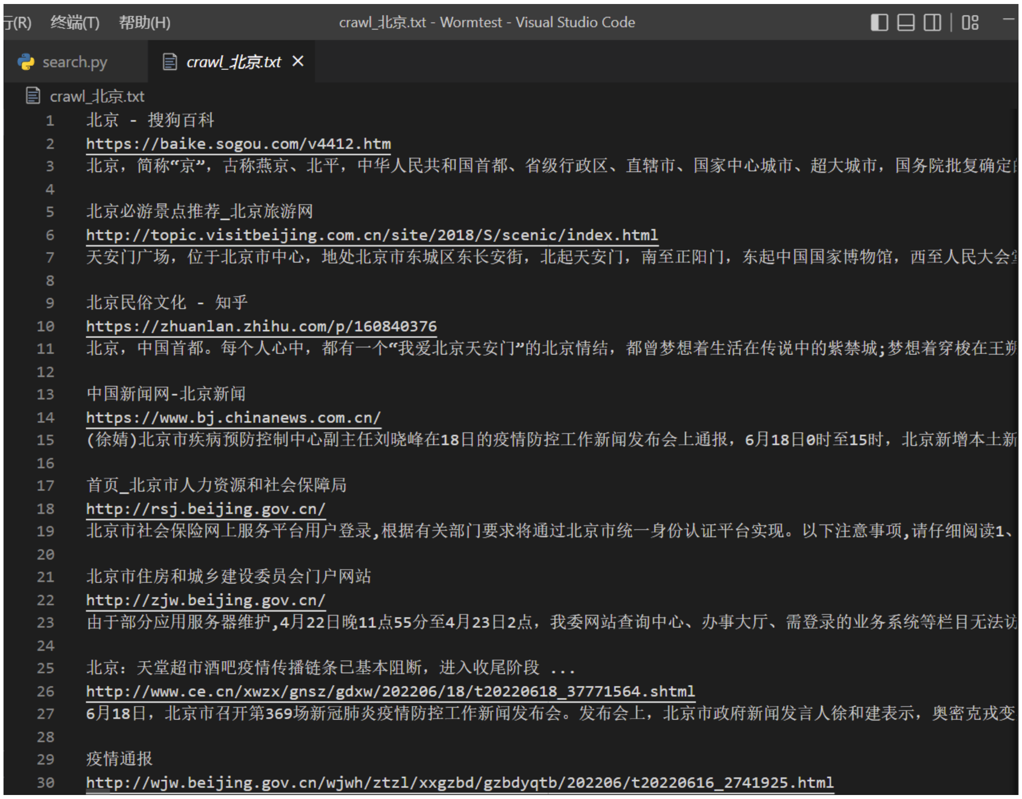
Bing: may find many different kinds of webs, while the number of webs found is larger

3.1.4 Screening Method

In order to prevent getting repeated result in one search and meaningless pages, we take a few screening method to remove some news pages in this part. In this part, all but one pages with repeated titles would be removed because they usually have the same content, and we only need one of them.

Further, pages with not enough content, i.e. with no content blocks that has more than 50 characters, would be removed from the search result. They may have linking problems or have no long paragraphs to express what happens.

3.1.5 Search Result



```
北京 - 搜狗百科
https://baike.sogou.com/v4412.htm
北京，简称“京”，古称燕京、北平，中华人民共和国首都、省级行政区、直辖市、国家中心城市、超大城市，国务院批复确定

北京必游景点推荐_北京旅游网
http://topic.visitbeijing.com.cn/site/2018/S/scenic/index.html
天安门广场，位于北京市中心，地处北京市东城区东长安街，北起天安门，南至正阳门，东起中国国家博物馆，西至人民大会

北京民俗文化 - 知乎
https://zhuanlan.zhihu.com/p/160840376
北京，中国首都。每个人心中，都有一个“我爱北京天安门”的北京情结，都曾梦想着生活在传说中的紫禁城;梦想着穿梭在王

中国新闻网-北京新闻
https://www.bj.chinanews.com.cn/
(徐婧)北京市疾病预防控制中心副主任刘晓峰在18日的疫情防控工作新闻发布会上通报，6月18日0时至15时，北京新增本土

首页_北京市人力资源和社会保障局
http://rsj.beijing.gov.cn/
北京市社会保险网上服务平台用户登录,根据有关部门要求将通过北京市统一身份认证平台实现。以下注意事项,请仔细阅读1、

北京市住房和城乡建设委员会门户网站
http://zjw.beijing.gov.cn/
由于部分应用服务器维护,4月22日晚11点55分至4月23日2点，我委网站查询中心、办事大厅、需登录的业务系统等栏目无法

北京：天堂超市酒吧疫情传播链条已基本阻断，进入收尾阶段 ...
http://www.ce.cn/xwzx/gnsz/gdxw/202206/18/t20220618_37771564.shtml
6月18日，北京市召开第369场新冠肺炎疫情防控工作新闻发布会。发布会上，北京市政府新闻发言人徐和建表示，奥密克戎变

疫情通报
http://wjw.beijing.gov.cn/wjwh/ztzl/xxgzbd/gzbdyqtb/202206/t20220616_2741925.html
```

Figure 3: Search result example

Here we have got the search result, for following operations, we need to change these result into a standard form.

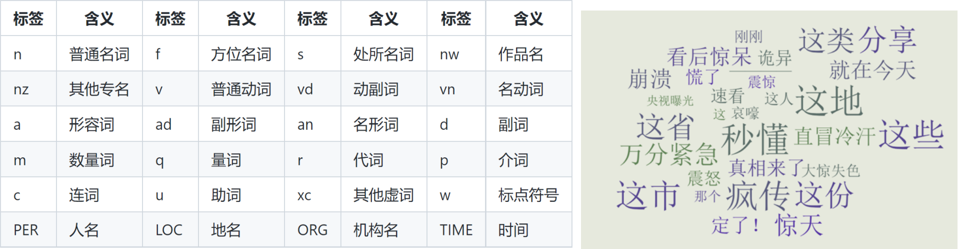
The search result can be returned in a list form, each item in the list contains the title, links and main content of one pages after the screening. We can also save the result in .txt files for further process, each page takes 3 lines, first is their title, the second is their links and the last one is their content.

For 200 raw data searched,we usually get 40-50 available pages. They'll be sent to further operations.

3.2 Title Processing

For the titles of the corresponding searching results, our processing goal is to detect clickbat. To solve this problem, we first ask: what good titles and bad titles respectively have in common? Recalling what we have learnt about the necessary elements in title in high school and the typical words in clickbit, we use grammar information to mark good titles and keywords to mark bad ones.

Additionally, we also enable classification on titles, details are shown in the context processing parts.



标签	含义	标签	含义	标签	含义	标签	含义
n	普通名词	f	方位名词	s	处所名词	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号
PER	人名	LOC	地名	ORG	机构名	TIME	时间

刚刚
这类分享
看后惊呆
诡异
就在今天
崩溃
慌了
震惊
央视曝光
速看
这人
这地
这
哀嚎
这省
秒懂
直冒冷汗
这些
万分紧急
真相来了
大惊失色
震怒
这市
那个
疯传
这份
定了！
惊天

Figure 4: left: part of speech index; right: ”bad words”

3.2.1 Grammar Based Positive Marking

To get the grammar structure of the title, we first use Lexical Analysis of Chinese to do words segmentation and return words' the part of speech (like verb, noun, shown in **Figure 4 left**). Then we compare the titles' structure with our ”good grammar dictionary” (we list possible structures with complete information like '`n-v-n, adj-n-adv-v`'), if the title contains good words' pairs, we will give it credit.

3.2.2 Keywords Based Negative Marking

We first collect common words in clickbit like in fig, then build a ”bad keywords dictionary”(shown in **Figure 4 right**) . Given a title, we first segment it into meaningful words, and give it negative score if it has ”bad keywords”.

After these two marking process (notice marking weights can be adjusted), we can: **1.**compare the final score with the user-setted threshold and hid the titles with low score; **2.** rank the search results according to the score.

3.3 Context Processing

For the passages of the corresponding searching results, we mainly focus on the following three parts: Keywords extraction, Summary generation and Text classification.

3.3.1 Keywords extraction

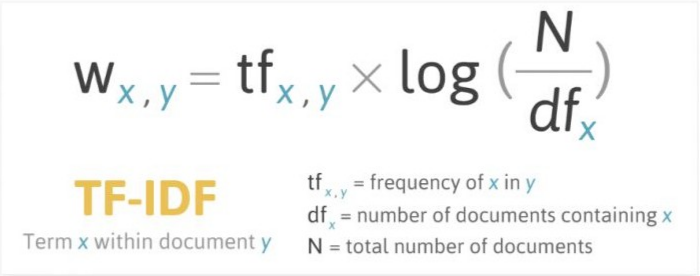

$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$
TF-IDF
Term x within document y
$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Figure 5: TF-IDF demonstrate

This part takes the reference of the project Chinese Keyphase extractor(CKPE)[1]. The algorithm is mainly based on the Term Frequency-Inverse Document Frequency(TF-IDF). In terms of the inverse document frequency part, the relevant weights of each terms have been generated via over a million news texts. The large amount of data guarantees the accuracy of elimination over those commonly-used terms. For each specific passage we use pkuseg as a word segmentation tool to extract potential keywords, calculating the term frequency over the passage. Combining the two parts together, we get the final results about whether a term is essential for the text. We provide the top 5 keyphases for users to help them get a basic insight about the context.

3.3.2 Summary generation

In version 1.0, we simply use an extraction model based on BERT to abstract essential sentence in the passage. It encodes every sentence in the passage to a vector, and then we use a clustering algorithm to extract key sentences that are closest to the centroid as the candidate summary sentences. [6] In version 1.2, we provide another summary generation models based on GPT2, which is a conditional generation model. For those two kinds of summary generation models, the conditional generation model is more powerful when the essential information is separated in different sentences. GPT2 can learn to mix them together and output more effective information. [2] However, in some cases, the conditional generation model may misunderstand the original meaning of the passage. For example, GPT2 may get confused when a passage is full of numbers or statistics. They may output the mismatching number. On the contrary, BERT ensures the accuracy of the information in the abstract.

[1] https://github.com/dongrixinyu/chinese_keyphrase_extractor

[2] https://github.com/qingkongzhiqian/GPT2-Summary

3.3.3 Text classification and emotional tendency analysis

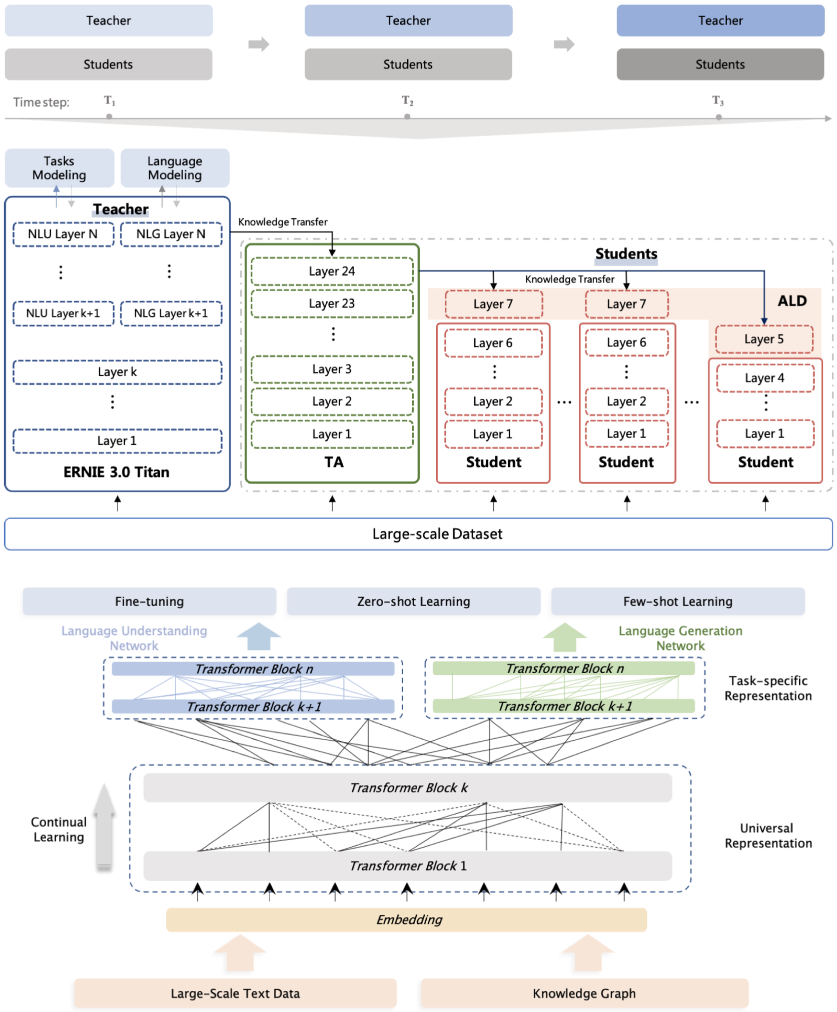
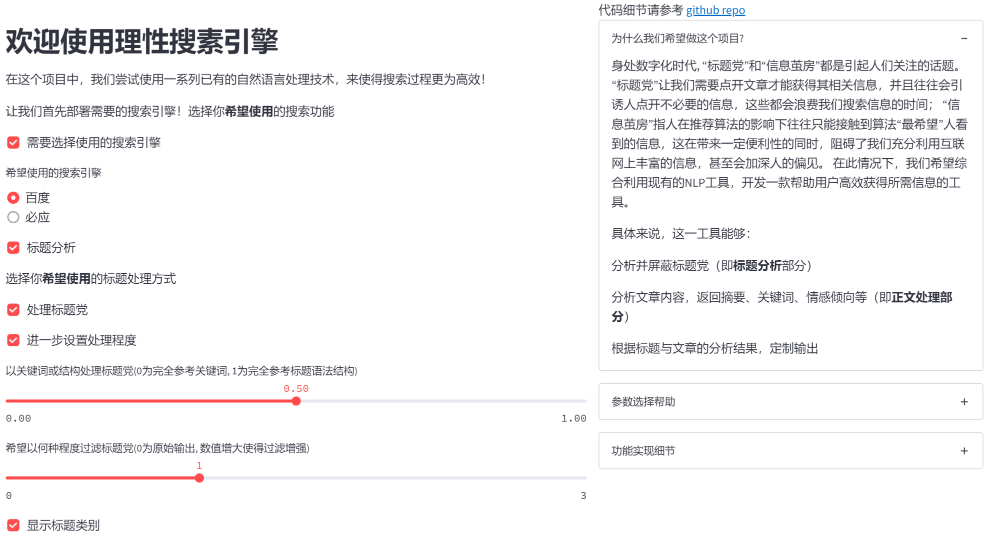


Figure 6: ERNIE3.0 model. Top: Teacher student structure; Down: using knowledge graph

In this part, we classify the passages according to their emotion and their topics. The whole part mostly based on ERNIE-titan. Except for plain texts, ERNIE3.0 has also been trained on a large scale knowledge graph. This universal pre-training enhances the ability of knowledge extraction. Besides, ERNIE3.0 combines both auto-regressive network and auto-encoding network, so that the trained model can handle both natural language understanding and generation tasks through zero-shot learning, few-shot learning or fine-tuning. [8] For ERNIE-titan, they promote two techniques called OFD and ALD to distill the large model and get great result. With ERNIE-titan, We evaluate how well these classification labels fit to their titles and determine whether we should display the passages to the users.

3.4 Web App Deploy

Informed by [3], we depoly a web app through streamlit to enable customized and interactive processing (As first shown in **Figure 1**). For the output part, we utilize the title and context processing results, support ranking final result by (positive/negative) emotional tendency and title scores. We also offer easily parameter selection interfaces and different demonstrating styles as shown in **Figure 7** and **Figure 8**.



欢迎使用理性搜索引擎

在这个项目中，我们尝试使用一系列已有的自然语言处理技术，来使得搜索过程更为高效！

让我们首先部署需要的搜索引擎！选择你**希望使用**的搜索功能

☑ 需要选择使用的搜索引擎

希望使用的搜索引擎

● 百度
○ 必应

☑ 标题分析

选择你**希望使用**的标题处理方式

☑ 处理标题党

☑ 进一步设置处理程度

以关键词或结构处理标题党(0为完全参考关键词, 1为完全参考标题语法结构)

0.50

0.00 1.00

希望以何种程度过滤标题党(0为原始输出, 数值增大使得过滤增强)

1

0 3

☑ 显示标题类别

代码细节请参考 github repo

为什么我们希望做这个项目? −

身处数字化时代, “标题党”和“信息茧房”都是引起人们关注的话题。“标题党”让我们需要点开文章才能获得其相关信息，并且往往会引诱人点开不必要的信息，这些都会浪费我们搜索信息的时间；“信息茧房”指人在推荐算法的影响下往往只能接触到算法“最希望”人看到的信息，这在带来一定便利性的同时，阻碍了我们充分利用互联网上丰富的信息，甚至会加深人的偏见。 在此情况下，我们希望综合利用现有的NLP工具，开发一款帮助用户高效获得所需信息的工具。

具体来说，这一工具能够:

分析并屏蔽标题党（即**标题分析**部分）

分析文章内容，返回摘要、关键词、情感倾向等（即**正文处理部分**）

根据标题与文章的分析结果，定制输出

参数选择帮助 +

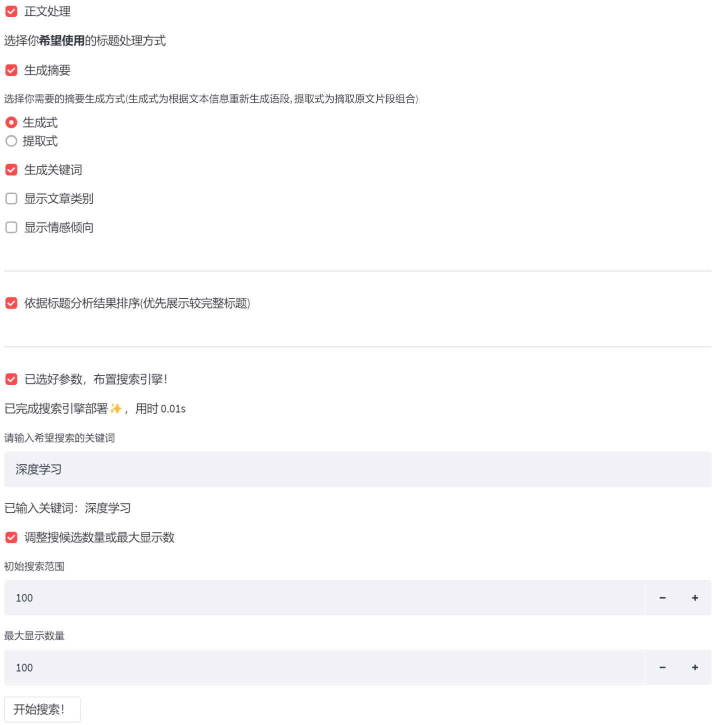
功能实现细节 +

Figure 7: Web App interface for selecting search engine and title processing

Moreover, streamlit also supports us to deploy the project on cloud server, allowing users to get access to the search engine from different devices. For more detail, please visit our github repository . We believe after further tuning and refining, this project is practical in real daily life.

4 Results

Some search results' are shown in **Figure 9, 10 and 11**. We also recommend watching the demo



☑ 正文处理

选择你**希望使用**的标题处理方式

☑ 生成摘要

选择你需要的摘要生成方式(生成式为根据文本信息重新生成语段，提取式为摘取原文片段组合)

● 生成式
○ 提取式

☑ 生成关键词

□ 显示文章类别

□ 显示情感倾向

☑ 依据标题分析结果排序(优先展示较完整标题)

☑ 已选好参数，布置搜索引擎！

已完成搜索引擎部署✨， 用时 0.01s

请输入希望搜索的关键词

深度学习

已输入关键词：深度学习

☑ 调整搜候选数量或最大显示数

初始搜索范围

100

最大显示数量

100

Figure 8: Web App interface for selecting context processing and out put method

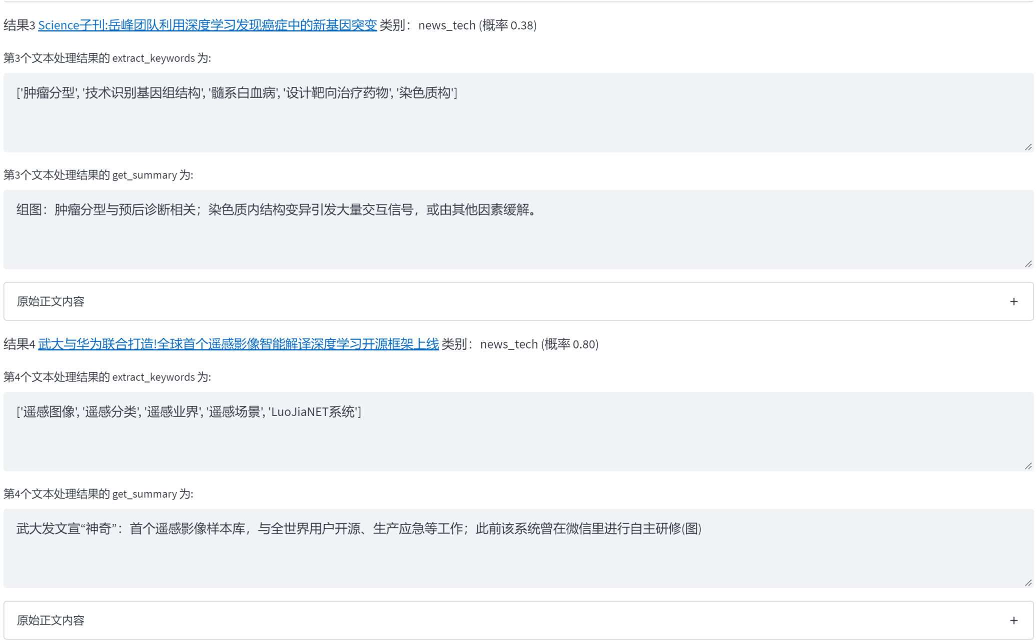
5 Discussion

In this project, we implement a practical adjusted search engine, which we hope can mitigate the concerns about ”clickbait” and ”information cocoon”. Because of time and resources limitation, we only perform five rounds of updates about functions and interfaces, we plan to enable parallel processing to accelerate the searching and processing.

6 Contribution

Report part: Zhengzhao Ma finishes Content Gaining (part 3.1) and helps checking the report; Zihan Zhou finishes Context Processing (part 3.3) and helps checking the report; Zihan Qiu finish other parts of the report and edits the Latex file.

Code part: Zhenzhao Ma contributes for the `search` parts; Zihan Zhou offers related files about context processing; Zihan Qiu finishes the codes for title processing and web app deploy, organizes all the codes.



结果3 Science子刊:岳峰团队利用深度学习发现癌症中的新基因突变 类别：news_tech (概率 0.38)

第3个文本处理结果的 extract_keywords 为:

['肿瘤分型', '技术识别基因组结构', '髓系白血病', '设计靶向治疗药物', '染色质构']

第3个文本处理结果的 get_summary 为:

组图：肿瘤分型与预后诊断相关；染色质内结构变异引发大量交互信号，或由其他因素缓解。

结果4 武大与华为联合打造!全球首个遥感影像智能解译深度学习开源框架上线 类别：news_tech (概率 0.80)

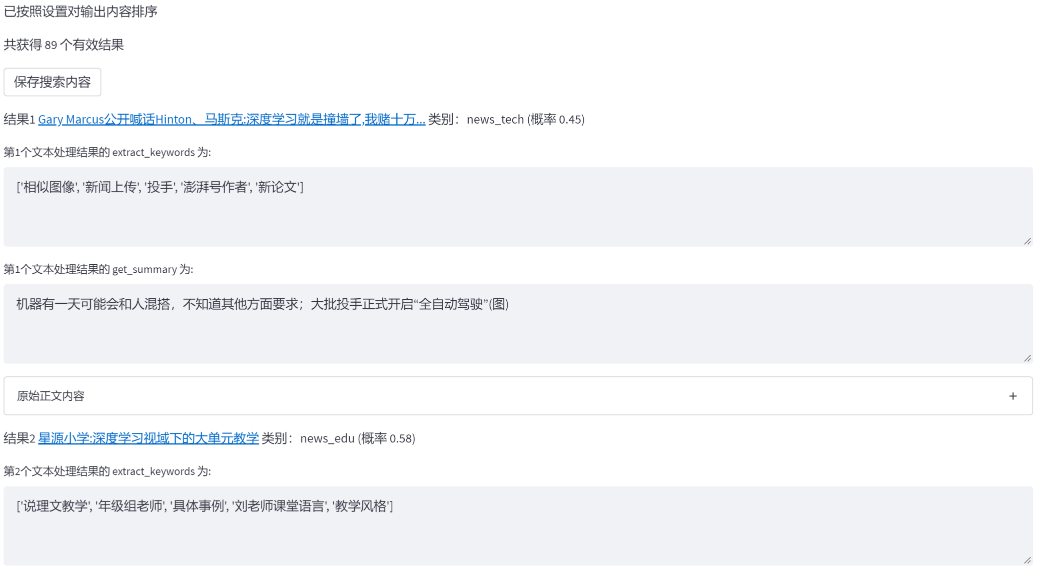
第4个文本处理结果的 extract_keywords 为:

['遥感图像', '遥感分类', '遥感业界', '遥感场景', 'LuoJiaNET系统']

第4个文本处理结果的 get_summary 为:

武大发文宣“神奇”：首个遥感影像样本库，与全世界用户开源、生产应急等工作；此前该系统曾在微信里进行自主研修(图)

Figure 9: Search result with getting keywords, summary and folding initial content



已按照设置对输出内容排序

共获得 89 个有效结果

结果1 Gary Marcus公开喊话Hinton、马斯克:深度学习就是撞墙了,我赌十万... 类别：news_tech (概率 0.45)

第1个文本处理结果的 extract_keywords 为:

['相似图像', '新闻上传', '投手', '澎湃号作者', '新论文']

第1个文本处理结果的 get_summary 为:

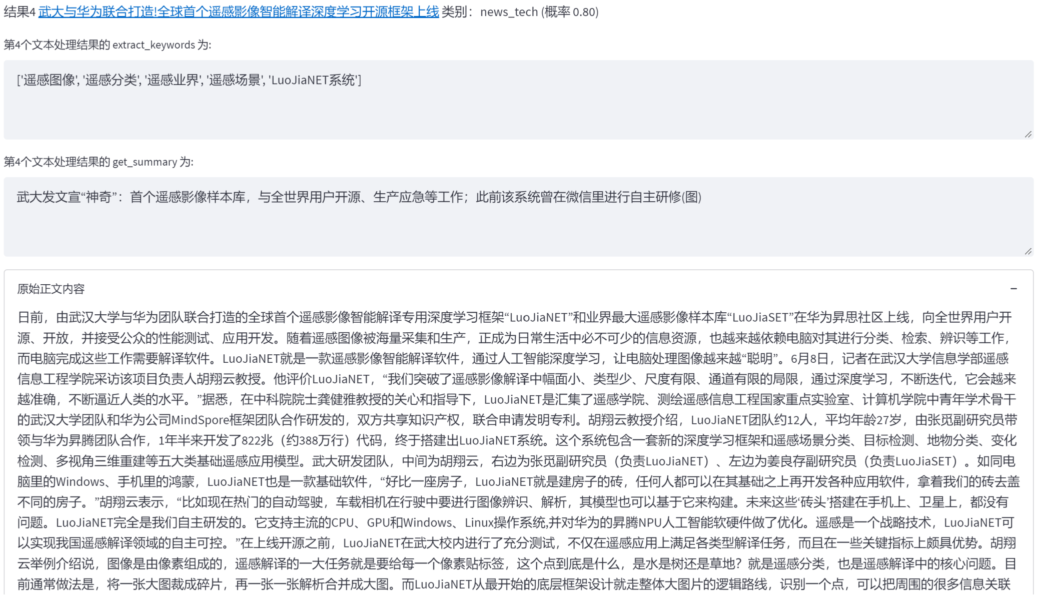
机器有一天可能会和人混搭，不知道其他方面要求；大批投手正式开启“全自动驾驶”(图)

结果2 星源小学:深度学习视域下的大单元教学 类别：news_edu (概率 0.58)

第2个文本处理结果的 extract_keywords 为:

['说理文教学', '年级组老师', '具体事例', '刘老师课堂语言', '教学风格']

Figure 10: Search result with title classification, getting keywords, summary and folding initial content



结果4 武大与华为联合打造!全球首个遥感影像智能解译深度学习开源框架上线 类别：news_tech (概率 0.80)

第4个文本处理结果的 extract_keywords 为:

['遥感图像', '遥感分类', '遥感业界', '遥感场景', 'LuoJiaNET系统']

第4个文本处理结果的 get_summary 为:

武大发文宣“神奇”：首个遥感影像样本库，与全世界用户开源、生产应急等工作；此前该系统曾在微信里进行自主研修(图)

日前，由武汉大学与华为团队联合打造的全球首个遥感影像智能解译专用深度学习框架“LuoJiaNET”和业界最大遥感影像样本库“LuoJiaSET”在华为昇思社区上线，向全世界用户开源、开放，并接受公众的性能测试、应用开发。随着遥感图像被海量采集和生产，正成为日常生活中必不可少的信息资源，也越来越依赖电脑对其进行分类、检索、辨识等工作，而电脑完成这些工作需要解译软件。LuoJiaNET就是一款遥感影像智能解译软件，通过人工智能深度学习，让电脑处理图像越来越“聪明”。6月8日，记者在武汉大学信息学部遥感信息工程学院采访该项目负责人胡翔云教授。他评价LuoJiaNET，“我们突破了遥感影像解译中幅面小、类型少、尺度有限、通道有限的局限，通过深度学习，不断迭代，它会越来越准确，不断逼近人类的水平。”据悉，在中科院院士龚健雅教授的关心和指导下，LuoJiaNET是汇集了遥感学院、测绘遥感信息工程国家重点实验室、计算机学院中青年学术骨干的武汉大学团队和华为公司MindSpore框架团队合作研发的，双方共享知识产权，联合申请发明专利。胡翔云教授介绍，LuoJiaNET团队约12人，平均年龄27岁，由张觅副研究员带领与华为昇腾团队合作，1年半来开发了822兆（约388万行）代码，终于搭建出LuoJiaNET系统。这个系统包含一套新的深度学习框架和遥感场景分类、目标检测、地物分类、变化检测、多视角三维重建等五大类基础遥感应用模型。武大研发团队，中间为胡翔云，右边为张觅副研究员（负责LuoJiaNET）、左边为姜良存副研究员（负责LuoJiaSET）。如同电脑里的Windows、手机里的鸿蒙，LuoJiaNET也是一款基础软件，“好比一座房子，LuoJiaNET就是建房子的砖，任何人都可以在其基础之上再开发各种应用软件，拿着我们的砖去盖不同的房子。”胡翔云表示，“比如现在热门的自动驾驶，车载相机在行驶中要进行图像辨识、解析，其模型也可以基于它来构建。未来这些‘砖头’搭建在手机上、卫星上，都没有问题。LuoJiaNET完全是我们自主研发的。它支持主流的CPU、GPU和Windows、Linux操作系统,并对华为的昇腾NPU人工智能软硬件做了优化。遥感是一个战略技术，LuoJiaNET可以实现我国遥感解译领域的自主可控。”在上线开源之前，LuoJiaNET在武大校内进行了充分测试，不仅在遥感应用上满足各类型解译任务，而且在一些关键指标上颇具优势。胡翔云举例介绍说，图像是由像素组成的，遥感解译的一大任务就是要给每一个像素贴标签，这个点到底是什么，是水是树还是草地？就是遥感分类，也是遥感解译中的核心问题。目前通常做法是，将一张大图裁成碎片，再一张一张解析合并成大图。而LuoJiaNET从最开始的底层框架设计就走整体大图片的逻辑路线，识别一个点，可以把周围的很多信息关联

Figure 11: Search result with title classification, getting keywords, summary and unfolding initial content

References

[1] Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen-tau Yih, Sebastian Riedel, and Fabio Petroni. Autoregressive Search Engines: Generating Substrings as Document Identifiers. *arXiv e-prints*, page arXiv:2204.10628, April 2022.

[2] Weiwei Guo, Xiaowei Liu, Sida Wang, Michaeel Kazi, Zhiwei Wang, Zhoutong Fu, Jun Jia, Liang Zhang, Huiji Gao, and Bo Long. Deep natural language processing for linkedin search. *arXiv preprint arXiv:2108.13300*, 2021.

[3] Abram Handler and Brendan O'Connor. Rookie: A unique approach for exploring news archives. *arXiv preprint arXiv:1708.01944*, 2017.

[4] Junjie Li and Hui Cao. Research on dual channel news headline classification based on ernie pre-training model. *arXiv preprint arXiv:2202.06600*, 2022.

[5] Dairui Liu, Derek Greene, and Ruihai Dong. A novel perspective to look at attention: Bi-level attention-based explainable topic modeling for news classification. *arXiv preprint arXiv:2203.07216*, 2022.

[6] Derek Miller. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*, 2019.

[7] Terrence Szymanski, Claudia Orellana-Rodriguez, and Mark T Keane. Helping news editors write better headlines: A recommender to improve the keyword contents & shareability of news headlines. *arXiv preprint arXiv:1705.09656*, 2017.

[8] Shuohuan Wang, Yu Sun, Yang Xiang, Zhihua Wu, Siyu Ding, Weibao Gong, Shikun Feng, Junyuan Shang, Yanbin Zhao, Chao Pang, et al. Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2112.12731*, 2021.