hw3 邮件检索系统实现

0 WHAT

基于ElasticSearch + Python开发环境，对安然公司150位用户的50W封电子邮件进行检索系统实现。

1 WHERE

1.1 ElasticSearch工具

中文官网：https://www.elastic.co/cn/elasticsearch

英文官网：https://www.elastic.co/en/elasticsearch

1.2 安然邮件数据集

50W邮件数据集：http://www.cs.cmu.edu/~enron

```
Message-ID: <5468446.1075855378133.JavaMail.evans@thyme>
Date: Mon, 14 May 2001 13:39:00 -0700 (PDT)
From: phillip.allen@enron.com
To: outlook.team@enron.com
Subject: Re: 2- SURVEY/INFORMATION EMAIL 5-14- 01
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: quoted-printable
X-From: Phillip K Allen
X-To: Outlook Migration Team <Outlook Migration Team/Corp/Enron@ENRON>
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Sent Mail
X-Origin: Allen-P
X-FileName: pallen (Non-Privileged).pst
```

2 HOW

2.1 Python

原生Python：https://www.python.org

Anaconda：https://www.anaconda.com

Pycharm：https://www.jetbrains.com/pycharm

2.2 ElasticSearch（后简称ES）

1. 下载ES本体

你可以将ES安装到任何地方，比较直接的做法是扔到项目目录下`%HOME_PATH%/ElasticSearch`，但是在提交作业的时候请屏蔽它们。

2. 下载ES + Python API

> `pip install elasticsearch`

3. 部署ES

> `cd %ES_HOME%/bin`
>
> Windows: `./elasticsearch.bat`
>
> Linux: `./elasticsearch`

建议以单独窗口打开命令提示符，作为本地部署好的ES服务器。通过以下命令检查部署并访问ES：

> `curl http://localhost:9200`

4. 使用ES + Python API

仅提一点，学会使用漂亮的`json`格式文件传输ES需要的数据和指令。

```
{
    "Level 1" : {
        "Level 2 Attr" : "Val",
        // ...
    },
    "Level 1 Attr" : "Val",
    // ...
}
```

你可以使用下面的例子来检测前面步骤的正确性，

```
from datetime import datetime
from elasticsearch.client import Elasticsearch
es = Elasticsearch()
doc = {
    'Author': 'Information Retrieval',
    'Text': 'Test for ElasticSearch',
    'Time': datetime.now(),
}
if __name__ == "__main__":
    res = es.index(index="test-index", id=1, document=doc)
    print(res['result'])
    res = es.get(index="test-index", id=1)
    print(res['_source'])
    es.indices.refresh(index="test-index")
    res = es.search(index="test-index", query={"match_all": {}})
    print("Got %d Hits:" % res['hits']['total']['value'])
    for hit in res['hits']['hits']:
        print("%(timestamp)s %(author)s: %(text)s" % hit["_source"])
    print("Test OK")
```

如果输出内容符合预期，通过浏览器输入地址`http://localhost:9200/%INDEX_NAME%`访问到了正确的json结构，那么恭喜你，你已经成功完成了第一步。

更多内容参考ES + Python API Document：https://elasticsearch-py.readthedocs.io

3 HINTS

1. 可以按照收件人、发件人、标题、内容等进行邮件检索
2. 探索ES实现索引构建、向量空间模型等核心环节
3. 可以提取附件内容，进行附件检索
4. 垃圾邮件分类

5. 文本情感分析
6. GUI、Web呈现检索系统
7. 基于已学设计更多内容

4 SCORE

- 40%：索引设计与构建
- 30%：检索功能设计
- 10%：ES功能探索
- 10%：作业报告
- 10%：视频录制（功能展示和核心代码讲解，`≤10min`）

5 SUBMIT

截止时间为`11月21日23:59:59`，提交压缩包（包含项目文件、讲解视频、作业报告等）至公邮`nkuir2021fall@163.com`，命名为学号*姓名*hw3，如`1811412_戚晓睿_hw3.zip`.

请注意，邮件命名同作业命名。