### 程序设计：

#### 1， 爬虫。

##### A）Python库选择：

request是常用python抓取网页的库程序。如果用selenim库则更为完美，selenim可以应对使用JavaScript解析才能查看的网站，比如我们常见的只有翻到web页面最后，更多新闻或者评论才会被javascript加载。再比如以下文章生成网站

<https://suulnnka.github.io/BullshitGenerator/index.html>  ，如果用request抓取该类javascript网站将无法抓到真实的显示内容，而是javascript内容。

回避以上问题带来的编程复杂度，选择较为简便的方案，通过request抓取以下网站，新闻内容，<https://zh.wikinews.org/> 抓取过程进行一定简化处理。

在本地架设apache置入新闻内容。1.6万篇新闻，235MB字节容量。Apache 服务IP地址为192.168.5.5。

##### B）抓取内容保存。

新闻页面内容为繁体，抓取后及时转换成简体，并用lzo python库压缩，后存入hdfs中。

##### C) 爬虫程序代码

#!/usr/bin/env python

# -\*- coding: utf-8 -\*-

import requests #引入抓取页面库

from bs4 import BeautifulSoup #网页tag分析查找库

from opencc import OpenCC #简繁体转换库

from hdfs import \* #hdfs读写库

import lzo #文本压缩后保存用

import subprocess #hdfs内容手动输出到stdin时用

import sys

from simhash import Simhash #文章相似度判定

import jieba #“结巴”分词库，配合上面的simhash使用

class wikinews(object):

def \_\_init\_\_(self):

self.headers = {

'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/87.0.4280.66 Safari/537.36'

}

self.url = 'http://192.168.5.5/news/X/'

self.cc = OpenCC('t2s')

self.root\_path = "/"

self.client = InsecureClient('http://192.168.5.20:9870',user='hdfs',root=self.root\_path)

#转换繁体为简体字

def conv\_lang(self,text):

converted = self.cc.convert(text)

return(converted)

#获取新闻列表页

def get\_html(self):

response = requests.get(self.url, headers=self.headers)

html = response.content.decode('utf-8')

return html

#获取新闻内容页

def get\_news(self,url):

response = requests.get(url, headers=self.headers)

html = response.content.decode('utf-8')

return html

#写新闻文本并压缩lzo后到hadoop

def write\_news(self,filename,url):

html=self.get\_news(url)

soup = BeautifulSoup(html, 'lxml')

text= lzo.compress(self.conv\_lang(soup.text),9)

sfilename = self.conv\_lang(filename)

self.client.write('/user/input/'+sfilename, data=text, overwrite=True)

# file = open("./news/"+sfilename, 'w', encoding='utf-8')

# file.write(text)

# file.close()

def get\_list(self):

html = self.get\_html()

soup = BeautifulSoup(html, 'lxml')

file = open('list.txt', 'w', encoding='utf-8')

file2 = open('filenamelist.txt', 'w', encoding='utf-8')

x=soup.find\_all(name='a')

del(x[0])

for link in x:

title = link.string

url = self.url+link.get('href')

file.write(url)

file.write('\n')

file2.write(self.conv\_lang(title))

file2.write('\n')

try:

self.write\_news(title,url)

except ValueError:

print(title)

if \_\_name\_\_ == '\_\_main\_\_':

web = wikinews()

web.get\_list()

#### 2，文章相似度判断

取出hdfs里的压缩文件内容，使用了命令行输出到stdin，见下面程序subprocess一行。

解压缩后比较两篇文章距离，完全一样的文章距离为0，测试另一篇不一样的文章距离报告为28。

需要说明的是，文章经过了处理，取出了html tag，已经是纯文本文件，这部分应该在这一节完成或者在mapreduce中完成。为了简化这个步骤，页面初级处理安排在了抓取页面时，压缩保存到hdfs中的已经是纯文本。这部分代码只为了测试之用，本应该放在下一节程序中。

from simhash import Simhash #文章相似度判定

import jieba #“结巴”分词库，配合上面的simhash使用

def test(self):

# data1 = lzo.compress(str1, 5)

# client.write('/tmp/list.txt', data=data1, overwrite=True)

# client.download('/tmp/list.txt', '/root/list.txt1',overwrite=True)

# obj=client.read("/tmp/list.txt")

output1 = subprocess.check\_output("hdfs dfs -cat "+"\""+"/user/input/ .Asia全球抢滩潮\_申请量单日逼近三十万"+"\"", shell=True)

data1 = lzo.decompress(output1)

data2 = lzo.decompress(output2)

print(self.test\_simhash(data1.decode(),data2.decode()))

# client.upload('/tmp','./list.txt',True)

def test\_simhash(self,text1,text2):

words1 = jieba.lcut(text1, cut\_all=True)

words2 = jieba.lcut(text2, cut\_all=True)

# print(Simhash(words1).distance(Simhash(words2)))

print(words1)

#### 3，MapReduce程序

##### A） 调试错误信息

这部分程序调试最终未能通过。

错误提示如下，代码见后面。从错误提示可知，任务被分成了6个，因为测试用例就是6篇文章。下一步可以回退到较简单的任务重新测试。比如不依赖hadoop-streaming提供stdin上的压缩hex输出，而是主动读取压缩后的文件。(base) [root@cm1 pycharm\_project\_842]# su - hdfs -c "/home/hdfs/anaconda3/bin/python /tmp/pycharm\_project\_842/yn69.py -r hadoop --hadoop-streaming-jar /opt/cloudera/parcels/CDH-6.2.0-1.cdh6.2.0.p0.967373/lib/hadoop-mapreduce/hadoop-streaming.jar hdfs:///tmp/input/\* hdfs:///tmp/output/"

No configs found; falling back on auto-configuration

No configs specified for hadoop runner

Looking for hadoop binary in $PATH...

Found hadoop binary: /bin/hadoop

Using Hadoop version 3.0.0

Creating temp directory /tmp/yn69.hdfs.20211228.103211.416644

uploading working dir files to hdfs:///user/hdfs/tmp/mrjob/yn69.hdfs.20211228.103211.416644/files/wd...

Copying other local files to hdfs:///user/hdfs/tmp/mrjob/yn69.hdfs.20211228.103211.416644/files/

Running step 1 of 1...

WARNING: Use "yarn jar" to launch YARN applications.

packageJobJar: [] [/opt/cloudera/parcels/CDH-6.2.0-1.cdh6.2.0.p0.967373/jars/hadoop-streaming-3.0.0-cdh6.2.0.jar] /tmp/streamjob3958123564344371990.jar tmpDir=null

Connecting to ResourceManager at cm1.example.com/192.168.5.20:8032

Connecting to ResourceManager at cm1.example.com/192.168.5.20:8032

Disabling Erasure Coding for path: /user/hdfs/.staging/job\_1640645331135\_0003

Total input files to process : 6

number of splits:6

yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled

Submitting tokens for job: job\_1640645331135\_0003

Executing with tokens: []

resource-types.xml not found

Unable to find 'resource-types.xml'.

Submitted application application\_1640645331135\_0003

The url to track the job: http://cm1.example.com:8088/proxy/application\_1640645331135\_0003/

Running job: job\_1640645331135\_0003

Job job\_1640645331135\_0003 running in uber mode : false

map 0% reduce 0%

Task Id : attempt\_1640645331135\_0003\_m\_000000\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:465)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:349)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:174)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1875)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:168)

Task Id : attempt\_1640645331135\_0003\_m\_000001\_0, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:465)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:349)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:174)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1875)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:168)

Task Id : attempt\_1640645331135\_0003\_m\_000000\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:465)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:349)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:174)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1875)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:168)

Task Id : attempt\_1640645331135\_0003\_m\_000001\_1, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:465)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:349)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:174)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1875)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:168)

Task Id : attempt\_1640645331135\_0003\_m\_000000\_2, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:465)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:349)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:174)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1875)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:168)

Task Id : attempt\_1640645331135\_0003\_m\_000001\_2, Status : FAILED

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:465)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:349)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:174)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1875)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:168)

map 100% reduce 100%

Job job\_1640645331135\_0003 failed with state FAILED due to: Task failed task\_1640645331135\_0003\_m\_000000

Job failed as tasks failed. failedMaps:1 failedReduces:0 killedMaps:0 killedReduces: 0

Job not successful!

Streaming Command Failed!

Counters: 11

Job Counters

Data-local map tasks=2

Failed map tasks=7

Killed map tasks=5

Killed reduce tasks=1

Launched map tasks=7

Other local map tasks=5

Total megabyte-milliseconds taken by all map tasks=40361984

Total time spent by all map tasks (ms)=39416

Total time spent by all maps in occupied slots (ms)=39416

Total time spent by all reduces in occupied slots (ms)=0

Total vcore-milliseconds taken by all map tasks=39416

Scanning logs for probable cause of failure...

Looking for history log in /var/log/hadoop-yarn...

Probable cause of failure:

Error: java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 127

at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:325)

at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:538)

at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:130)

at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)

at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:34)

at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:465)

at org.apache.hadoop.mapred.MapTask.run(MapTask.java:349)

at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:174)

at java.security.AccessController.doPrivileged(Native Method)

at javax.security.auth.Subject.doAs(Subject.java:422)

at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1875)

at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:168)

Step 1 of 1 failed: Command '['/bin/hadoop', 'jar', '/opt/cloudera/parcels/CDH-6.2.0-1.cdh6.2.0.p0.967373/lib/hadoop-mapreduce/hadoop-streaming.jar', '-files', 'hdfs:///user/hdfs/tmp/mrjob/yn69.hdfs.20211228.103211.416644/files/wd/mrjob.zip#mrjob.zip,hdfs:///user/hdfs/tmp/mrjob/yn69.hdfs.20211228.103211.416644/files/wd/setup-wrapper.sh#setup-wrapper.sh,hdfs:///user/hdfs/tmp/mrjob/yn69.hdfs.20211228.103211.416644/files/wd/yn69.py#yn69.py', '-input', 'hdfs:///tmp/input/\*', '-input', 'hdfs:///tmp/output/', '-output', 'hdfs:///user/hdfs/tmp/mrjob/yn69.hdfs.20211228.103211.416644/output', '-mapper', '/bin/sh -ex setup-wrapper.sh python3 yn69.py --step-num=0 --mapper', '-reducer', '/bin/sh -ex setup-wrapper.sh python3 yn69.py --step-num=0 --reducer']' returned non-zero exit status 256.

(base) [root@cm1 pycharm\_project\_842]#

##### B) 代码部分：

#!/usr/bin/env python

# -\*- coding: utf-8 -\*-

from mrjob.job import MRJob #mrjob是python 运行Hadoop MapReduce的库

from mrjob.step import MRStep

from simhash import Simhash #文章相似度判断

import jieba #分词库，配合上面库使用，提高准确度

class xiangsi(MRJob):

def steps(self):

return [

MRStep(mapper=self.mapper\_get\_ratings,

reducer=self.reducer\_count\_ratings)

]

def mapper\_get\_ratings(self, \_, line):

data1 = "Asia全球抢滩潮\_申请量单日逼近三十万"

data2 = lzo.decompress(line)

words1 = jieba.lcut(data1, cut\_all=True)

words2 = jieba.lcut(data2.decode(), cut\_all=True)

yield Simhash(words1).distance(Simhash(words2))

def reducer\_count\_ratings(self, word):

yield word

if \_\_name\_\_ == '\_\_main\_\_':

xiangsi.run()