



互联网搜索引擎

项目一实验报告

**题目： 搜索引擎文本预处理**

**学院： 信息学院**

**教师： 吴晓**

**姓名： 周雕**

**学号： 2016112396**

**班级： 软件1班**

1. 实验要求

* 通过下载引擎（Web Crawler/Spider）自动下载至少500个英文文档/网页，以及500个中文文档/网页，并保留原始的文档/ 网页备份（如：News\_1\_Org.txt）
* 编程对所下载文档进行自动预处理：
  + 将各个单词进行字符化，完成删除特殊字符等操作
  + 调研并选择合适的中文分词技术和工具实现中文分词，可参看《Lucene分词器比较》
  + 删除英文停用词（Stop Word）
  + 删除中文停用词
  + 调用或者编程实现Porter Stemming功能
  + 将中文文档进行字符化，即可被搜索引擎索引的字符单元
  + 对于英文文档，经过以上处理之后，将经过处理之后所形成简化文档保存（如：News\_1\_E.txt），以备以后的索引处理
  + 对于中文文档，经过以上处理之后，将经过处理之后所形成简化文档保存（如：News\_1\_C.txt），以备以后的索引处理

2实验内容

1. 通过网络爬虫下载网络文档
2. 使用ansj分词器对下载的文档进行分词
3. 读取停用词表（将特殊字符加入停用词表）
4. 去除停用词

本次实验使用Java语言编写

3.实验环境

操作系统：Windows10

IDE：IntelliJ IDEA

4.实验步骤（过程）

（1）学习网络爬虫技术

（2）学习使用Jsoup编写网络爬虫及HTML文档内容提取

（3）待爬取中文网站选取网易新闻—科技频道（IT、科学、互联网分类）

（4）用Chrome浏览器分析网页源码标签类和id找出需要的url规律

（5）编写爬虫爬取网易科技频道的url并存于数据库中

MyCrawler.java代码如下：

package crawler;

/\*\*

\* 爬取URL并存于数据库表record中

\*/

import java.io.BufferedReader;

import java.io.IOException;

import java.io.InputStreamReader;

import java.net.HttpURLConnection;

import java.net.URL;

import java.sql.PreparedStatement;

import java.sql.ResultSet;

import java.sql.SQLException;

import java.sql.Statement;

import java.util.ArrayList;

import java.util.regex.Matcher;

import java.util.regex.Pattern;

public class MyCrawler {

/\*\*

\* 要分析的网页

\*/

String htmlUrl;

/\*\*

\* 分析结果

\*/

ArrayList<String> hrefList = new ArrayList();

/\*\*

\* 网页编码方式

\*/

String charSet;

public static DB db=new DB();

public MyCrawler(String htmlUrl) {

// TODO 自动生成的构造函数存根

this.htmlUrl = htmlUrl;

}

/\*\*

\* 获取分析结果

\*

\* @throws IOException

\*/

public ArrayList<String> getHrefList() throws IOException {

parser();

return hrefList;

}

/\*\*

\* 解析网页链接

\*

\* @return

\* @throws IOException

\*/

private void parser() throws IOException {

URL url = new URL(htmlUrl);

HttpURLConnection connection = (HttpURLConnection) url.openConnection();

connection.setDoOutput(true);

String contenttype = connection.getContentType();

charSet = getCharset(contenttype);

InputStreamReader isr = new InputStreamReader(

connection.getInputStream(), charSet);

BufferedReader br = new BufferedReader(isr);

String str = null, rs = null;

while ((str = br.readLine()) != null) {

rs = getHref(str);

if (rs != null)

hrefList.add(rs);

}

}

/\*\*

\* 获取网页编码方式

\*

\* @param str

\*/

private String getCharset(String str) {

Pattern pattern = Pattern.compile("charset=.\*");

Matcher matcher = pattern.matcher(str);

if (matcher.find())

return matcher.group(0).split("charset=")[1];

return null;

}

/\*\*

\* 从一行字符串中读取链接

\*

\* @return

\*/

private String getHref(String str) {

Pattern pattern = Pattern.compile("https://tech.163.com/19/.\*html");

Matcher matcher = pattern.matcher(str);

if (matcher.find())

return matcher.group(0);

return null;

}

public static void processUrl(String url) throws SQLException,IOException {

//检查一下是否给定的URL已经存在数据库中

String sql="select \* from Record where URL='"+url+"'";

ResultSet rs=db.runSql(sql);

if(!rs.next()) {

//将URL存储到数据库中避免下次重复

String sql2 = "insert into record (URL) values " + "(?)";

PreparedStatement stmt = db.conn.prepareStatement(sql2, Statement.RETURN\_GENERATED\_KEYS);

stmt.setString(1, url);

stmt.execute();

System.out.println(url);

}

}

public static void main(String[] arg) throws IOException, SQLException {

for(int i=2;i<=20;i++){

String str="http://tech.163.com/special/it\_2016\_02";

if(i<=9){

str="http://tech.163.com/special/it\_2016\_0"+i+"/";

}

if(i>=10){

str="http://tech.163.com/special/it\_2016\_"+i+"/";

}

MyCrawler a = new MyCrawler(str);

ArrayList<String> hrefList = a.getHrefList();

for (int j = 0; j < hrefList.size(); j++){

processUrl(hrefList.get(j));

}

}

for(int i=2;i<=9;i++){

String str="http://tech.163.com/special/techscience\_0"+i+"/";

MyCrawler a = new MyCrawler(str);

ArrayList<String> hrefList = a.getHrefList();

for (int j = 0; j < hrefList.size(); j++){

processUrl(hrefList.get(j));

}

}

for(int i=2;i<=9;i++){

String str=" http://tech.163.com/special/tele\_2016\_0"+i+"/";

MyCrawler a = new MyCrawler(str);

ArrayList<String> hrefList = a.getHrefList();

for (int j = 0; j < hrefList.size(); j++){

processUrl(hrefList.get(j));

}

}

for(int i=2;i<=20;i++){

String str="http://tech.163.com/special/internet\_2016\_02/";

if(i<=9){

str="http://tech.163.com/special/internet\_2016\_0"+i+"/";

}

if(i>=10){

str="http://tech.163.com/special/internet\_2016\_"+i+"/";

}

MyCrawler a = new MyCrawler(str);

ArrayList<String> hrefList = a.getHrefList();

for (int j = 0; j < hrefList.size(); j++){

processUrl(hrefList.get(j));

}

}

}

}

DB.java

package crawler;

import java.sql.\*;

public class DB {

public Connection conn = null;

/\*\*

\* 构造对象时加载驱动

\*/

public DB() {

try {

Class.forName("com.mysql.jdbc.Driver");

String url = "jdbc:mysql://localhost:3306/Crawler";

conn = DriverManager.getConnection(url, "root", "123456");

System.out.println("conn built");

} catch (ClassNotFoundException e) {

e.printStackTrace();

} catch (SQLException e) {

e.printStackTrace();

}

}

/\*\*

\* 数据库查询语句

\*

\* @param sql

\* @return

\* @throws SQLException

\*/

public ResultSet runSql(String sql) throws SQLException {

Statement sta = conn.createStatement();

return sta.executeQuery(sql);

}

/\*\*

\* 数据库操作语句

\*

\* @param sql

\* @return

\* @throws SQLException

\*/

public boolean runSql2(String sql) throws SQLException {

Statement sta = conn.createStatement();

return sta.execute(sql);

}

/\*\*

\* 关闭数据库连接

\*/

protected void finalClose() {

try {

if (conn != null || !conn.isClosed()) conn.close();

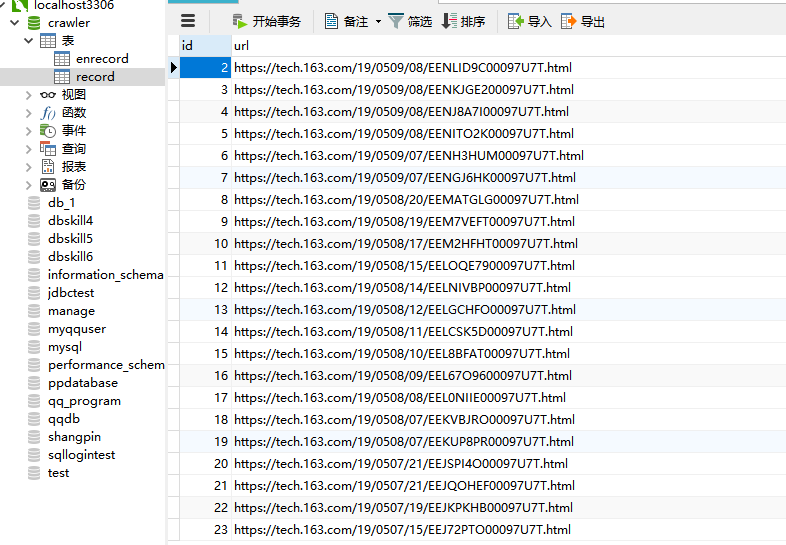
} catch (SQLException e) {

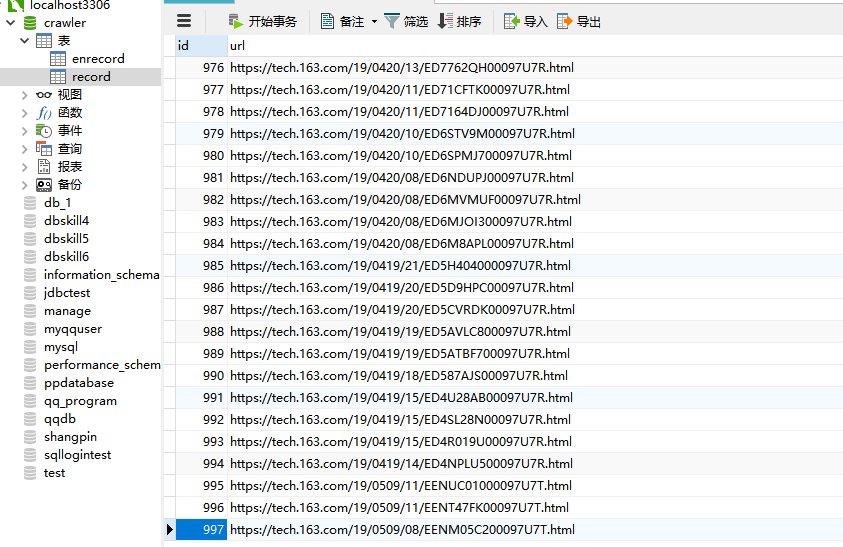
e.printStackTrace();

}

}

}





总共爬取了997个URL

（6）编写下载引擎，从数据库中读取所有链接，访问链接，Jsoup解析HTML，抽取出文本内容并下载到本地

DownloadTxt.java：

package crawler;

import org.jsoup.\*;

import org.jsoup.Connection;

import org.jsoup.nodes.Document;

import org.jsoup.nodes.Element;

import java.io.BufferedWriter;

import java.io.File;

import java.io.FileWriter;

import java.io.IOException;

import java.sql.ResultSet;

import java.sql.SQLException;

import java.util.ArrayList;

public class DownloadTxt {

private static DB db=new DB();

//从数据库获取所有URL并存入urlList中

public static ArrayList<String> getUrl() throws SQLException {

ArrayList<String> urlList=new ArrayList<>();

String sql="select \* from record";

ResultSet rs=db.runSql(sql);

while(rs.next()){

String url=rs.getString("URL");

urlList.add(url);

}

return urlList;

}

public static void getContent(String url,int i) throws InterruptedException, IOException {

//链接到目标地址

Connection connect = Jsoup.connect(url);

//设置useragent,设置超时时间，并以get请求方式请求服务器

Document document = connect.userAgent("Mozilla/4.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0)").timeout(6000).ignoreContentType(true).get();

Thread.sleep(1000);

//获取指定标签的数据

Element elementById = document.getElementById("endText");

//输出文本数据

System.out.println(elementById.text());

String fileName="F://CrawlerSourceTxtC/News\_1\_"+i+"rg.txt";

File file=new File(fileName);

if(!file.exists()){

file.createNewFile();

}

try {

writeCharBuffer(elementById.text(),fileName);

} catch (Exception e) {

e.printStackTrace();

}

}

/\*\*

\* 字符缓冲流写入文件

\* @param contentString

\* @param fileName

\* @throws Exception

\*/

private static void writeCharBuffer(String contentString, String fileName) throws Exception {

FileWriter fw=new FileWriter(fileName);

BufferedWriter bw=new BufferedWriter(fw,40);

bw.newLine();

bw.write(contentString);

bw.close();

fw.close();

System.out.println("写入成功");

}

public static void main(String[] args) throws Exception {

for(int i=0,j=getUrl().size();i<j;i++){

getContent(getUrl().get(i),i);

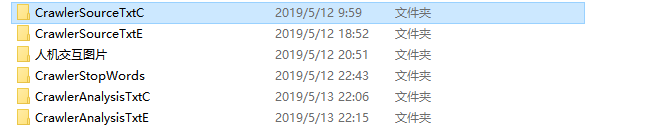
}

}

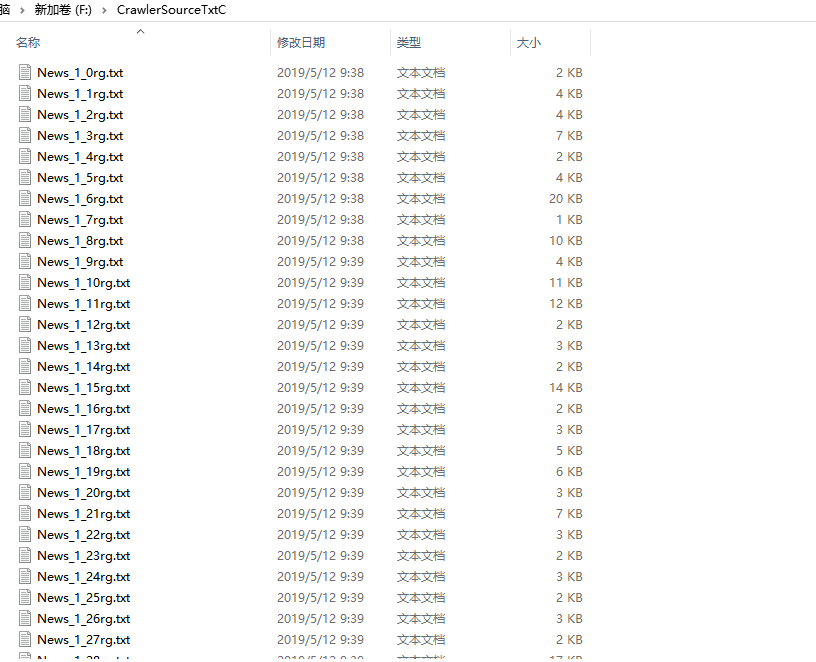
}

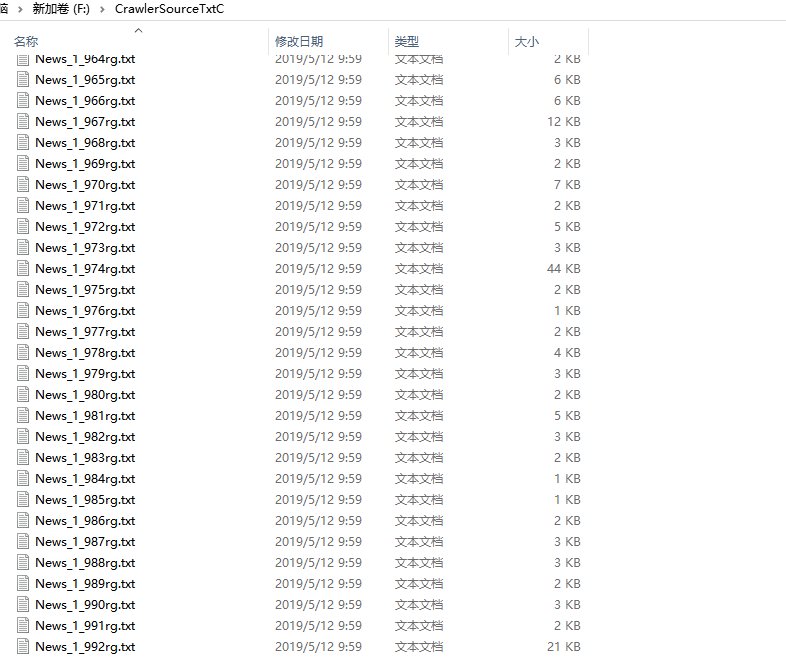
总共下载了993个中文文档，目录如下

选中为下载的中文原文档目录CrawlerSourceTxtC：



文档索引：



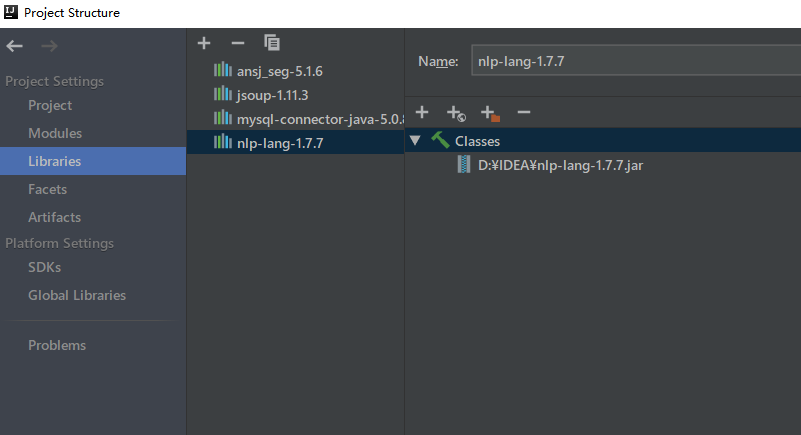


中文文档内容摘要：

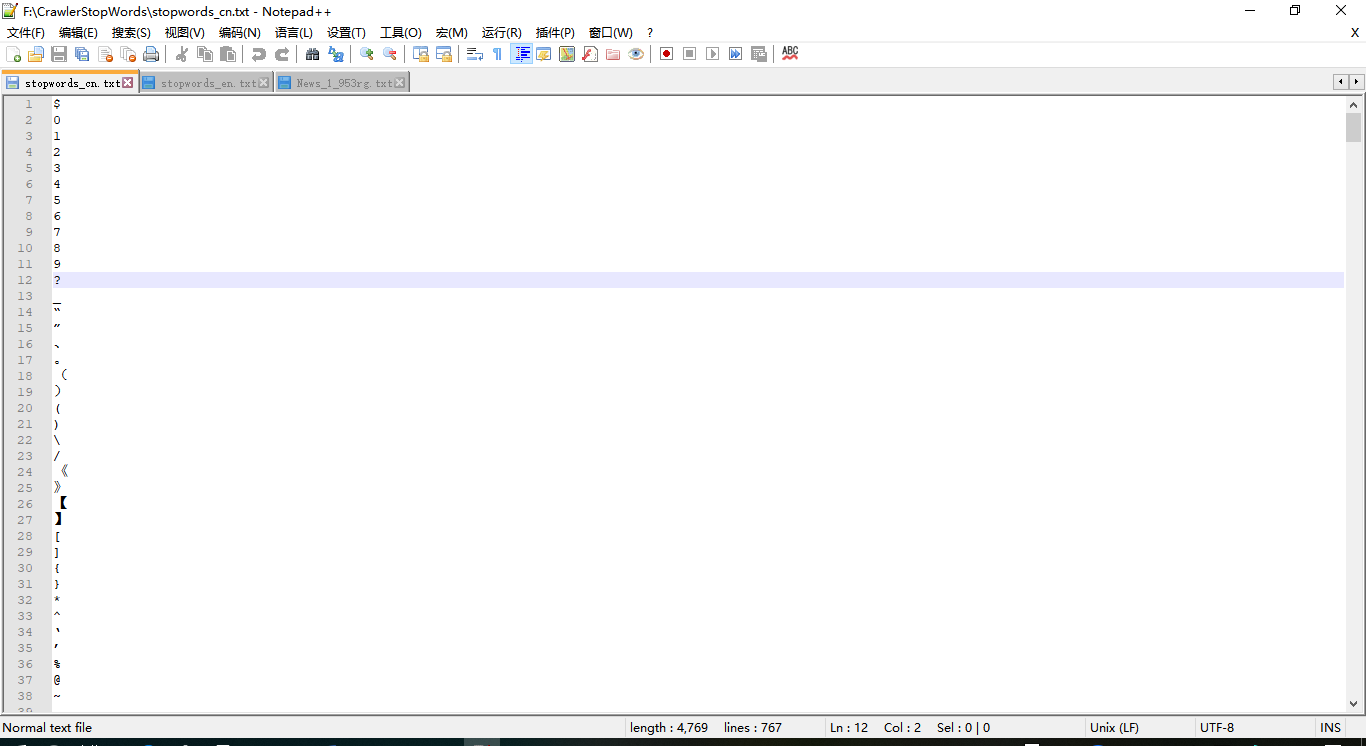


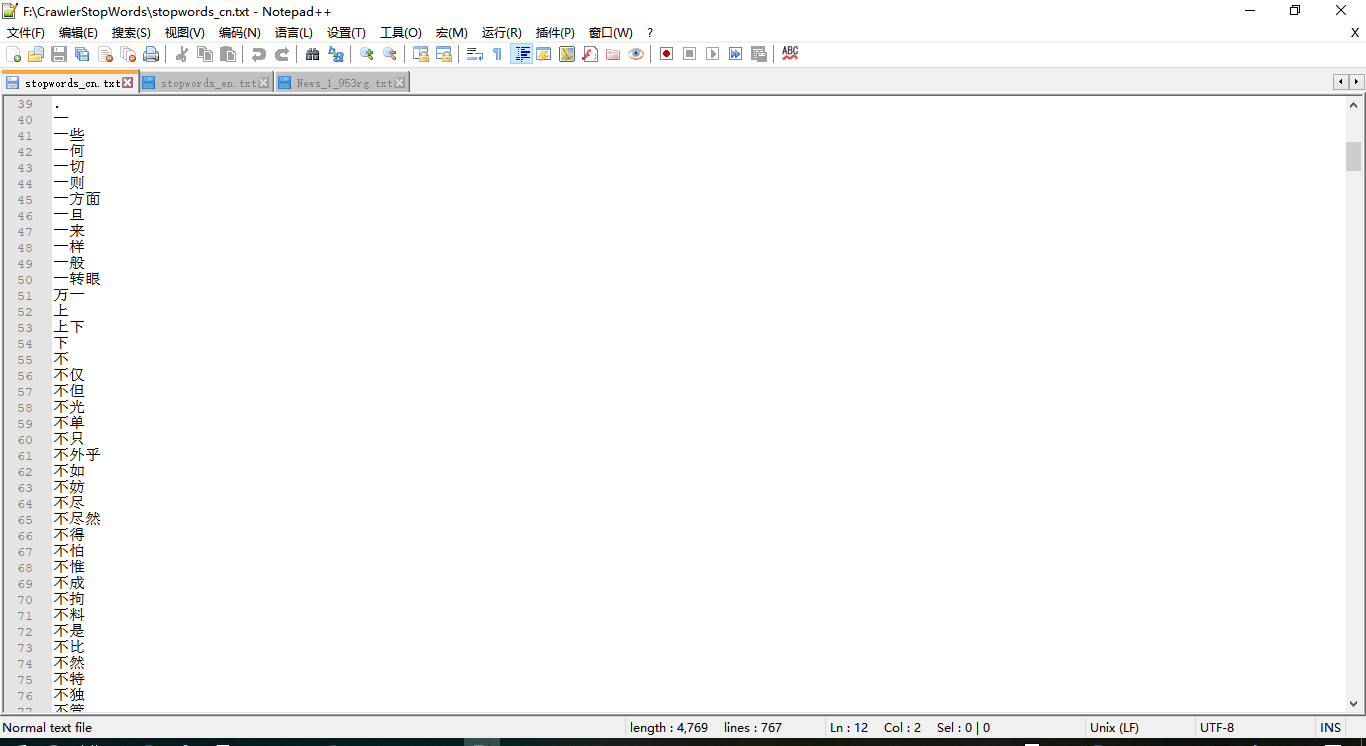
（7）对下载的中文文档分词、去停用词、去特殊字符

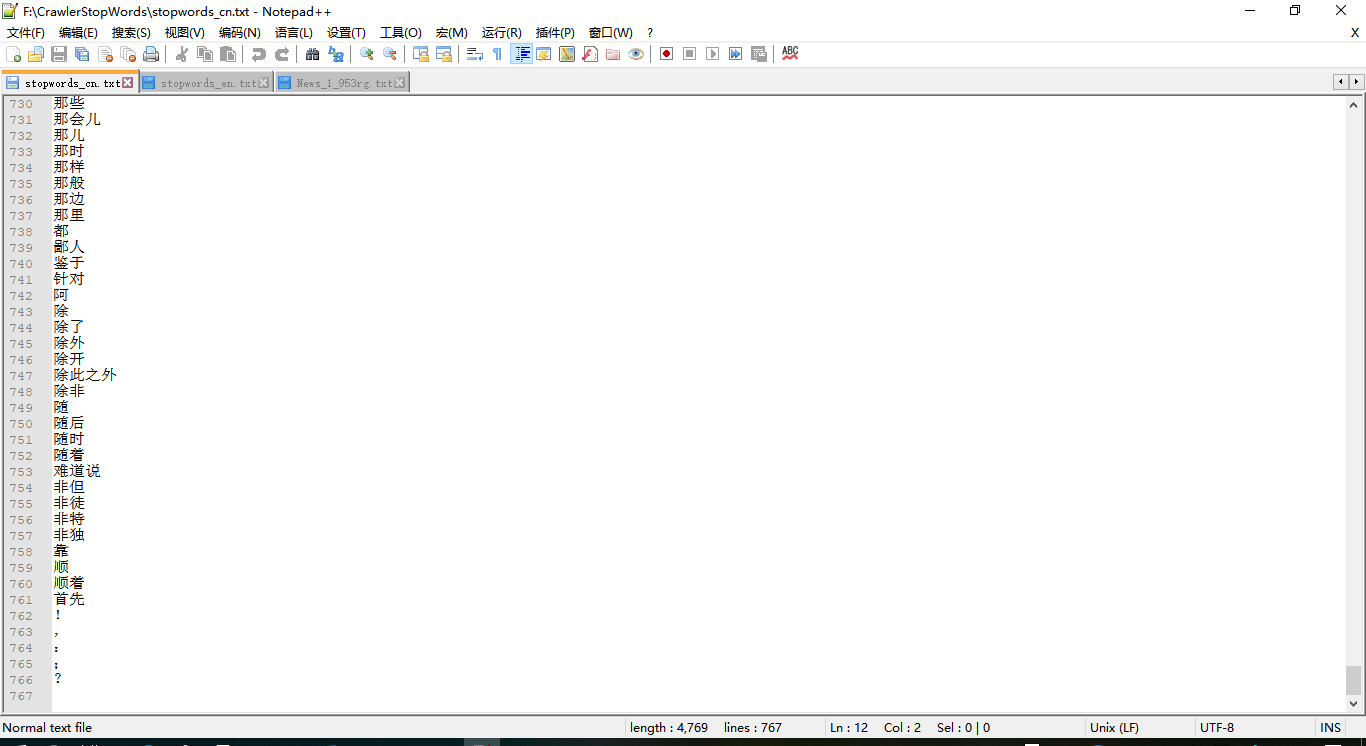
本次项目使用的分词器为ansj，下载ansj\_seg-5.1.6.jar包和nlp-lang-1.7.7.jar包，再IDEA导入包：



将特殊字符和标点符号加入停用词表：







用ansj分词并去除停用词：

Participle.java:

package participle;

import org.ansj.domain.Result;

import org.ansj.domain.Term;

import org.ansj.recognition.impl.StopRecognition;

import org.ansj.splitWord.analysis.ToAnalysis;

import java.io.\*;

import java.util.\*;

public class Participle {

public static void writeCharBuffer(String contentString, String fileName) throws Exception {

FileWriter fw=new FileWriter(fileName);

BufferedWriter bw=new BufferedWriter(fw,40);

bw.newLine();

bw.write(contentString);

bw.close();

fw.close();

System.out.println("写入成功");

}

public static void getAnalysis(){

String path="F://CrawlerSourceTxtC";

File file=new File(path);

File[] fs=file.listFiles();

List<String> stopword=new ArrayList<>();

String fpath="F://CrawlerStopWords/stopwords\_cn.txt";

try {

BufferedReader b=new BufferedReader(new FileReader(fpath));

String s=null;

while((s=b.readLine())!=null){

stopword.add(s);

}

b.close();

for(String stop:stopword){

System.out.println(stop);

}

} catch (FileNotFoundException e) {

e.printStackTrace();

} catch (IOException e) {

e.printStackTrace();

}

int num=0;

//遍历目录下的所有Txt

for(File f:fs){

if(!f.isDirectory()){

StringBuilder sb = new StringBuilder();

BufferedReader br=null;

try {

br=new BufferedReader(new FileReader(f));

String str=null;

while((str=br.readLine())!=null){

sb.append(str);

}

//ansj分词器精准分词

StopRecognition s=new StopRecognition();

s.insertStopWords(stopword);

Result result= ToAnalysis.parse(sb.toString()).recognition(s);

List<Term> terms=result.getTerms();

System.out.println(f);

// System.out.println(terms.size());

StringBuilder sWord=new StringBuilder();

for(int i=0;i<terms.size();i++){

String word=terms.get(i).getName()+" ";

String natureStr=terms.get(i).getNatureStr();

sWord.append(word);

System.out.print(word);

}

String fileName="F://CrawlerAnalysisTxtC/News\_"+num+"\_C.txt";

num+=1;

File fileout=new File(fileName);

if(!fileout.exists()){

fileout.createNewFile();

}

try{

writeCharBuffer(sWord.toString(),fileName);

}catch(Exception e){

e.printStackTrace();

}

// System.out.println(sb.toString());

} catch (IOException e) {

e.printStackTrace();

}finally{

try {

br.close();

} catch (IOException e) {

e.printStackTrace();

}

}

}

}

}

public static void getAnalysisE(){

String path="F://CrawlerSourceTxtE";

File file=new File(path);

File[] fs=file.listFiles();

List<String> stopword=new ArrayList<>();

String fpath="F://CrawlerStopWords/stopwords\_en.txt";

try {

BufferedReader b=new BufferedReader(new FileReader(fpath));

String s=null;

while((s=b.readLine())!=null){

stopword.add(s);

}

b.close();

for(String stop:stopword){

System.out.println(stop);

}

} catch (FileNotFoundException e) {

e.printStackTrace();

} catch (IOException e) {

e.printStackTrace();

}

int num=0;

//遍历目录下的所有Txt

for(File f:fs){

if(!f.isDirectory()){

StringBuilder sb = new StringBuilder();

BufferedReader br=null;

try {

br=new BufferedReader(new FileReader(f));

String str=null;

while((str=br.readLine())!=null){

sb.append(str);

}

//ansj分词器精准分词

StopRecognition s=new StopRecognition();

s.insertStopWords(stopword);

Result result= ToAnalysis.parse(sb.toString()).recognition(s);

List<Term> terms=result.getTerms();

System.out.println(f);

// System.out.println(terms.size());

StringBuilder sWord=new StringBuilder();

for(int i=0;i<terms.size();i++){

String word=terms.get(i).getName()+" ";

String natureStr=terms.get(i).getNatureStr();

sWord.append(word);

System.out.print(word);

}

String fileName="F://CrawlerAnalysisTxtE/News\_"+num+"\_E.txt";

num+=1;

File fileout=new File(fileName);

if(!fileout.exists()){

fileout.createNewFile();

}

try{

writeCharBuffer(sWord.toString(),fileName);

}catch(Exception e){

e.printStackTrace();

}

// System.out.println(sb.toString());

} catch (IOException e) {

e.printStackTrace();

}finally{

try {

br.close();

} catch (IOException e) {

e.printStackTrace();

}

}

}

}

}

public static void main(String[] args){

getAnalysis();

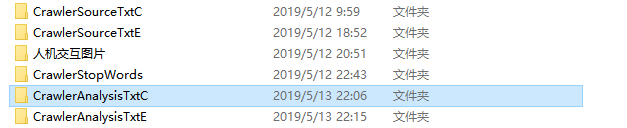
getAnalysisE();

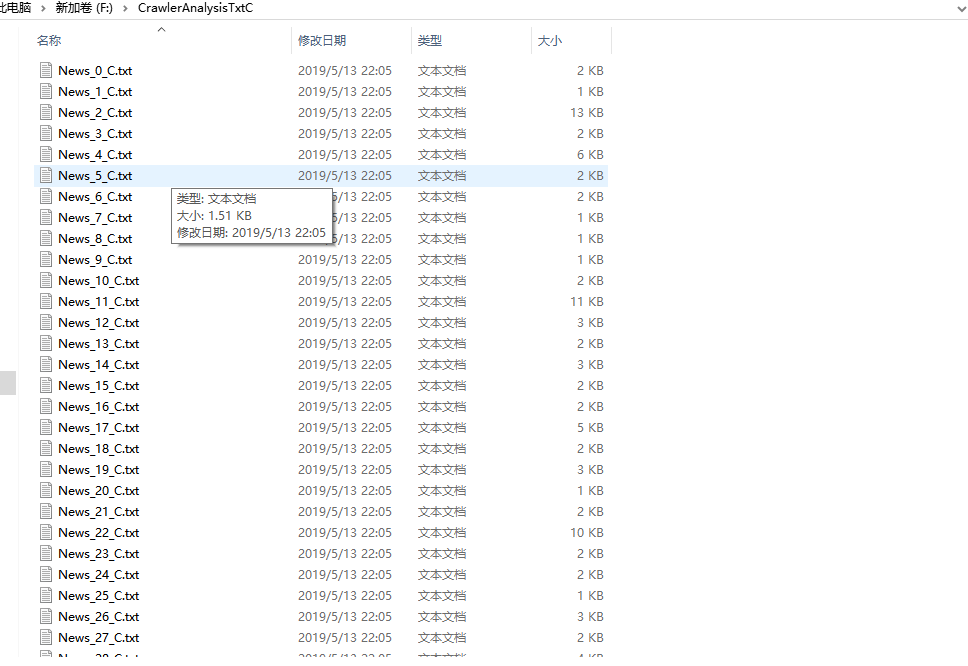
}

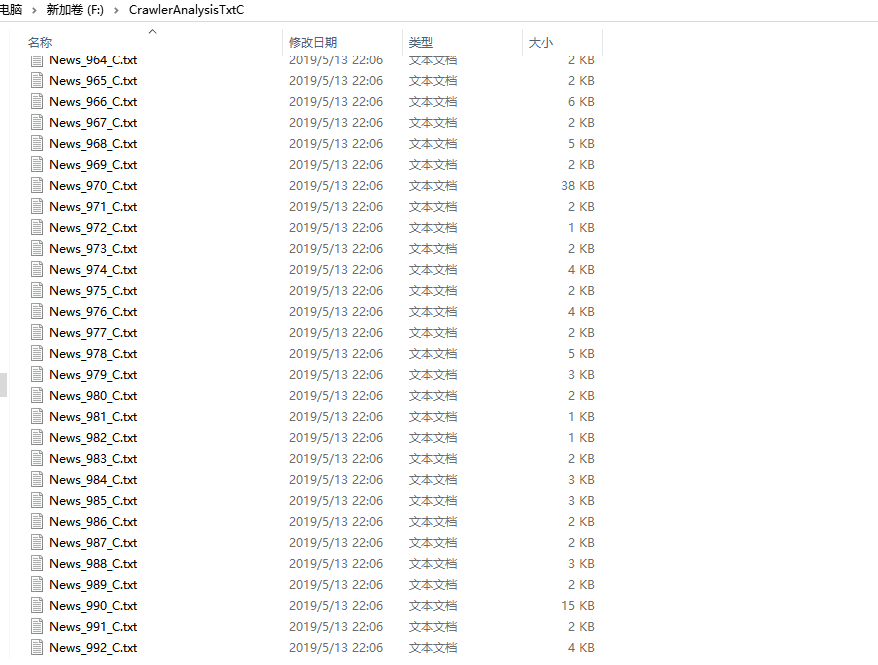
}

getAnalysis()方法用于对中文文档分词，分词结果如下：

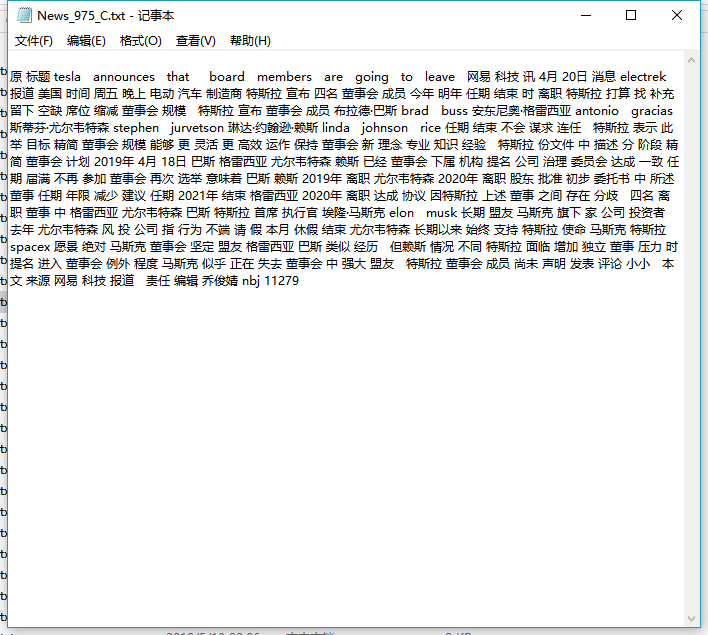
分词结果文档目录：

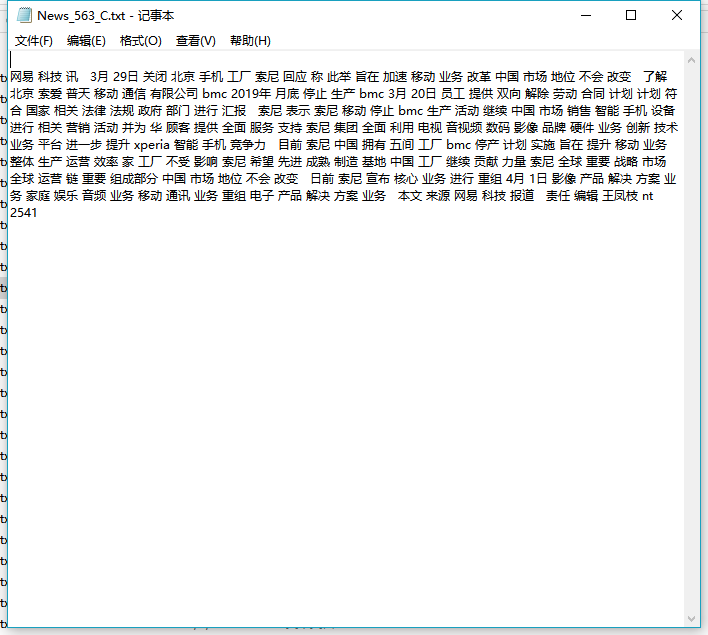






分词结果举例：





（8）选取英文网站WuxiaWorld，爬取一部小说的全部内容

找出该小说URL规律，编写爬虫：

DownloadTxtE.java

package crawler;

import org.jsoup.\*;

import org.jsoup.Connection;

import org.jsoup.nodes.Document;

import org.jsoup.nodes.Element;

import org.jsoup.select.Elements;

import java.io.BufferedWriter;

import java.io.File;

import java.io.FileWriter;

import java.io.IOException;

import java.sql.ResultSet;

import java.sql.SQLException;

import java.util.ArrayList;

public class DownloadTxtE {

private static DB db=new DB();

//从数据库获取所有URL并存入urlList中

public static ArrayList<String> getUrl() throws SQLException {

ArrayList<String> urlList=new ArrayList<>();

String sql="select \* from enrecord";

ResultSet rs=db.runSql(sql);

while(rs.next()){

String url=rs.getString("URL");

urlList.add(url);

}

return urlList;

}

public static void getContent(String url,int i) throws InterruptedException, IOException {

//链接到目标地址

Connection connect = Jsoup.connect(url);

//设置useragent,设置超时时间，并以get请求方式请求服务器

Document document = connect.userAgent("Mozilla/4.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0)").timeout(6000).ignoreContentType(true).get();

Thread.sleep(1000);

//获取指定标签的数据

Elements elementByClass = document.getElementsByClass("fr-view");

//输出文本数据

System.out.println(elementByClass.text());

String fileName="F://CrawlerSourceTxtE/Stories\_1\_"+i+"rg.txt";

File file=new File(fileName);

if(!file.exists()){

file.createNewFile();

}

try {

writeCharBuffer(elementByClass.text(),fileName);

} catch (Exception e) {

e.printStackTrace();

}

}

private static void writeCharBuffer(String contentString, String fileName) throws Exception {

FileWriter fw=new FileWriter(fileName);

BufferedWriter bw=new BufferedWriter(fw,40);

bw.newLine();

bw.write(contentString);

bw.close();

fw.close();

System.out.println("写入成功");

}

public static void main(String[] args) throws Exception {

for(int i=475;i<=1314;i++){

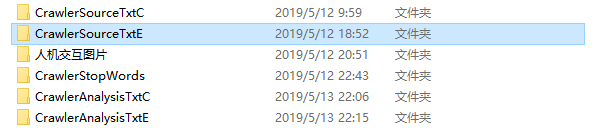
String url="https://www.wuxiaworld.com/novel/a-will-eternal/awe-chapter-"+i;

getContent(url,i);

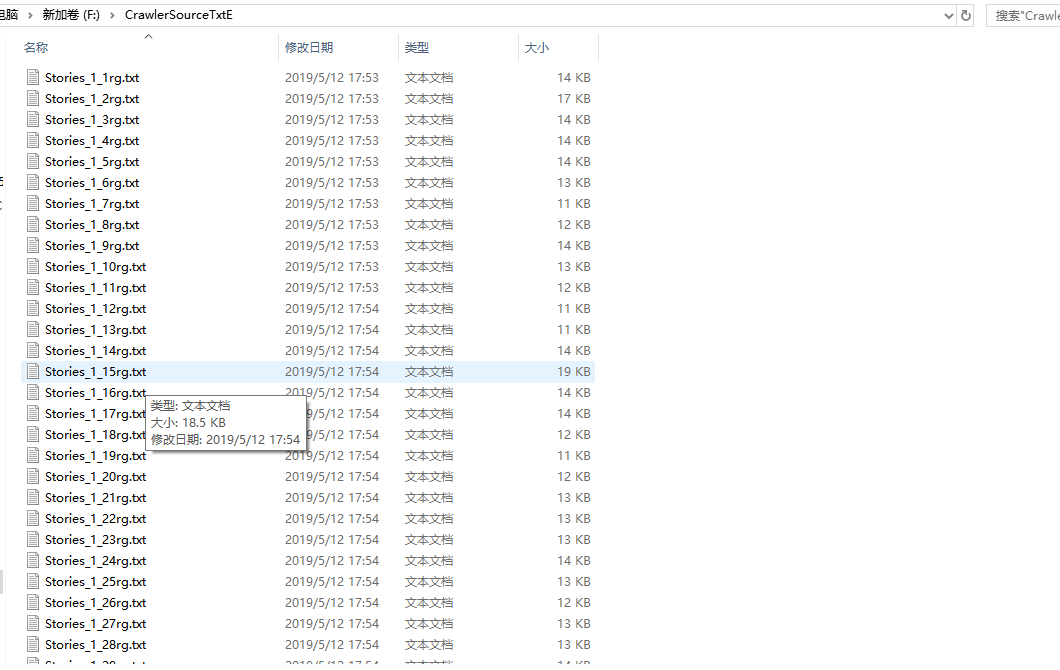
}

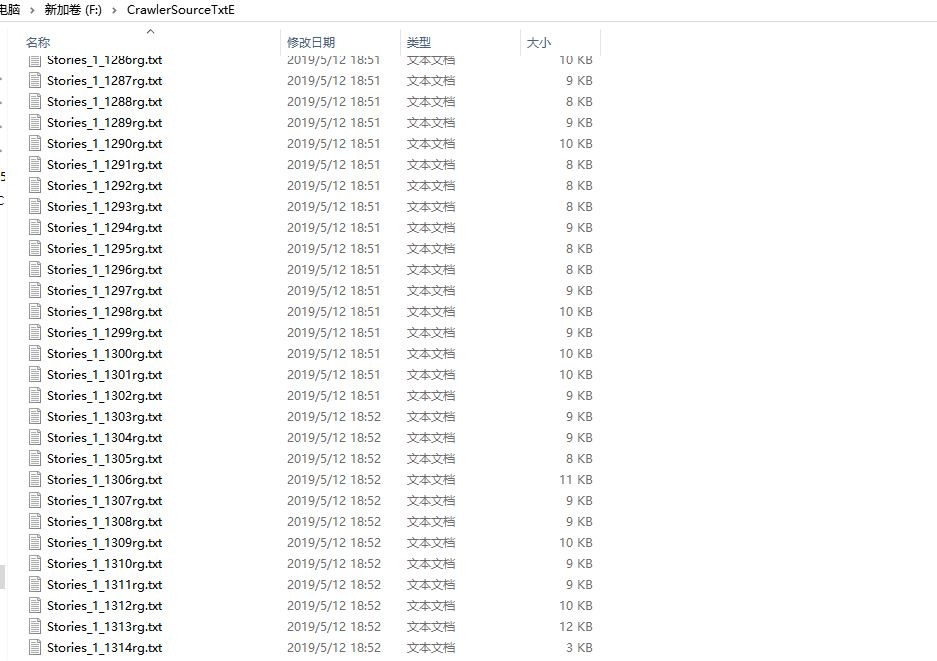
}

}

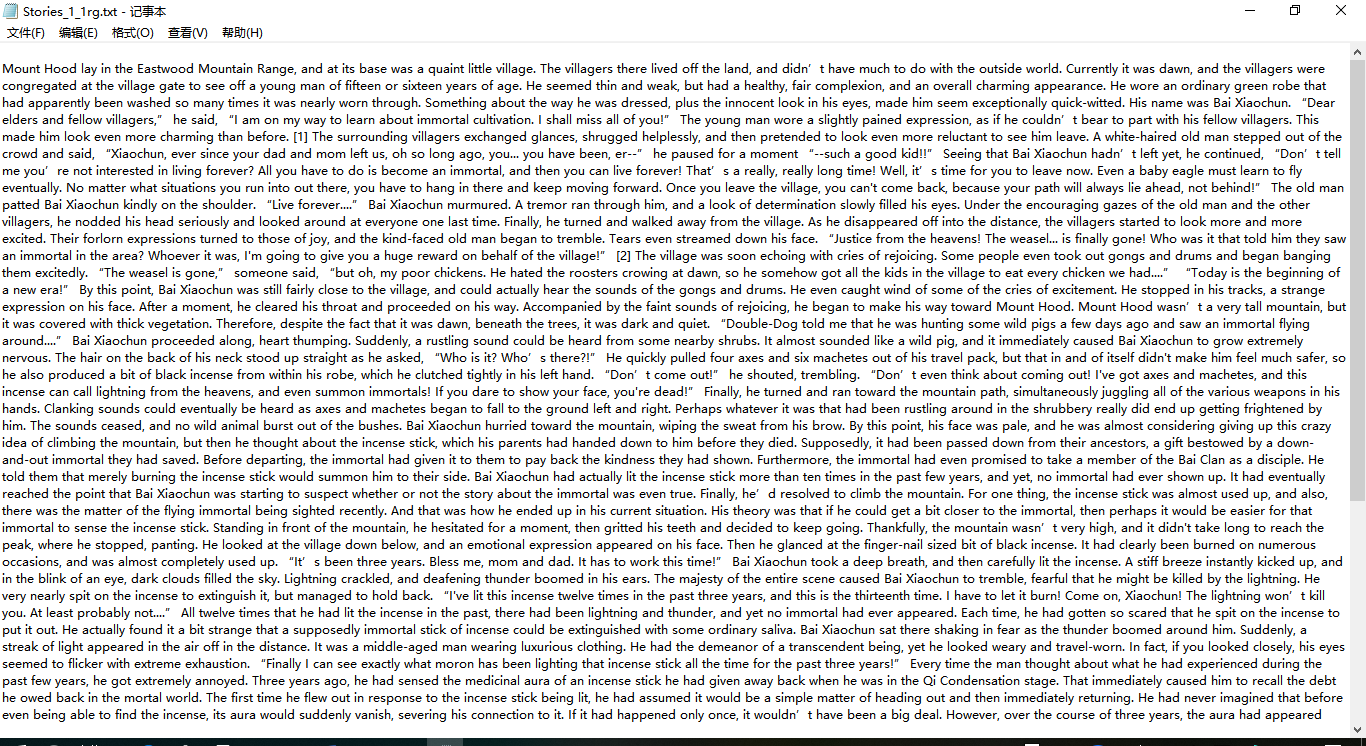


英文原文档目录：





英文原文档内容举例：



Participle类的getAnalysisE()方法为英文分词方法

public static void getAnalysisE(){

String path="F://CrawlerSourceTxtE";

File file=new File(path);

File[] fs=file.listFiles();

List<String> stopword=new ArrayList<>();

String fpath="F://CrawlerStopWords/stopwords\_en.txt";

try {

BufferedReader b=new BufferedReader(new FileReader(fpath));

String s=null;

while((s=b.readLine())!=null){

stopword.add(s);

}

b.close();

for(String stop:stopword){

System.out.println(stop);

}

} catch (FileNotFoundException e) {

e.printStackTrace();

} catch (IOException e) {

e.printStackTrace();

}

int num=0;

//遍历目录下的所有Txt

for(File f:fs){

if(!f.isDirectory()){

StringBuilder sb = new StringBuilder();

BufferedReader br=null;

try {

br=new BufferedReader(new FileReader(f));

String str=null;

while((str=br.readLine())!=null){

sb.append(str);

}

//ansj分词器精准分词

StopRecognition s=new StopRecognition();

s.insertStopWords(stopword);

Result result= ToAnalysis.parse(sb.toString()).recognition(s);

List<Term> terms=result.getTerms();

System.out.println(f);

// System.out.println(terms.size());

StringBuilder sWord=new StringBuilder();

for(int i=0;i<terms.size();i++){

String word=terms.get(i).getName()+" ";

String natureStr=terms.get(i).getNatureStr();

sWord.append(word);

System.out.print(word);

}

String fileName="F://CrawlerAnalysisTxtE/News\_"+num+"\_E.txt";

num+=1;

File fileout=new File(fileName);

if(!fileout.exists()){

fileout.createNewFile();

}

try{

writeCharBuffer(sWord.toString(),fileName);

}catch(Exception e){

e.printStackTrace();

}

// System.out.println(sb.toString());

} catch (IOException e) {

e.printStackTrace();

}finally{

try {

br.close();

} catch (IOException e) {

e.printStackTrace();

}

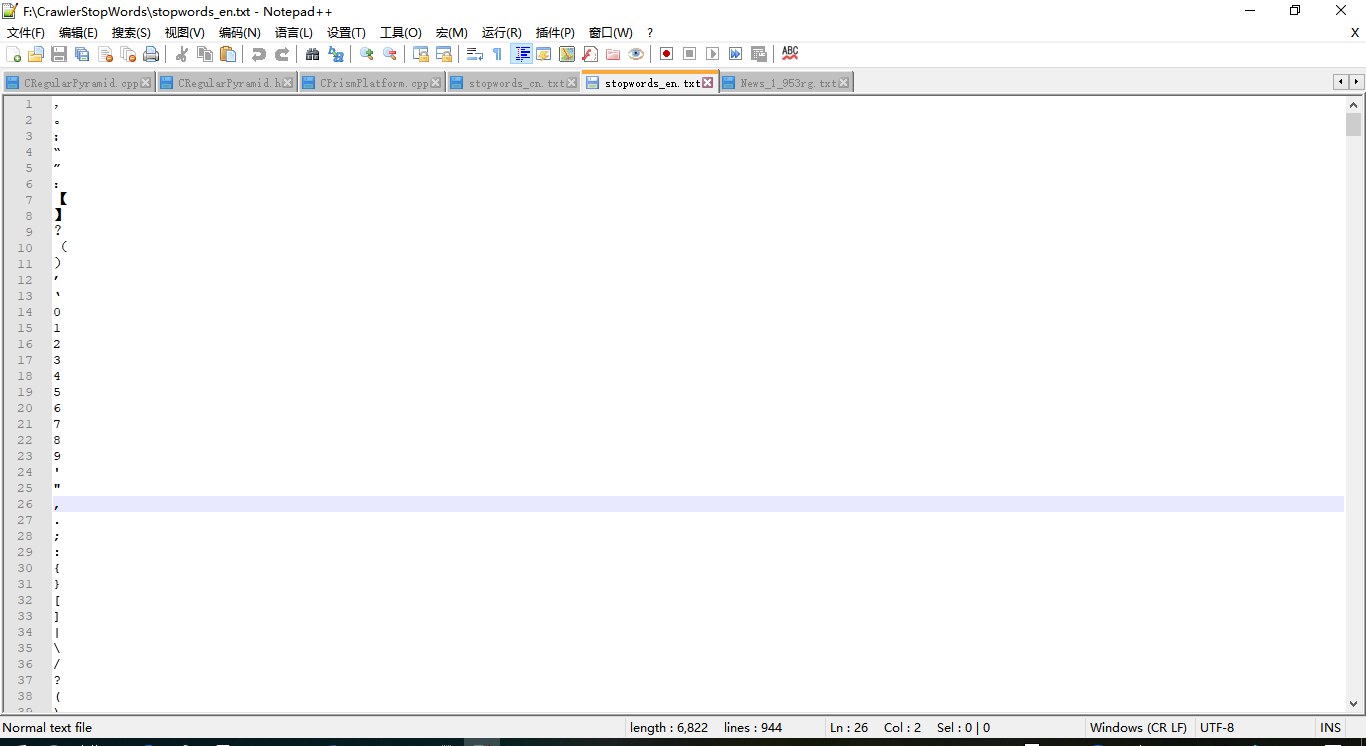
}

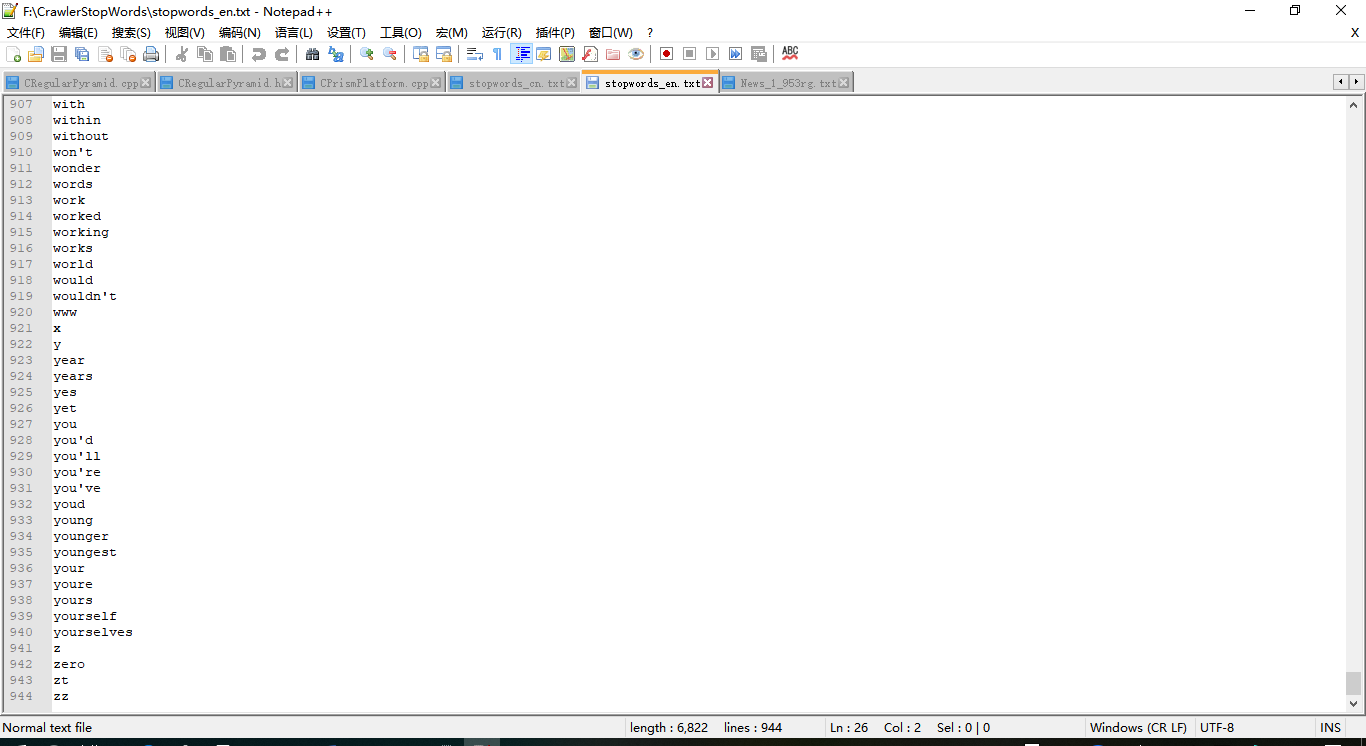
}

}

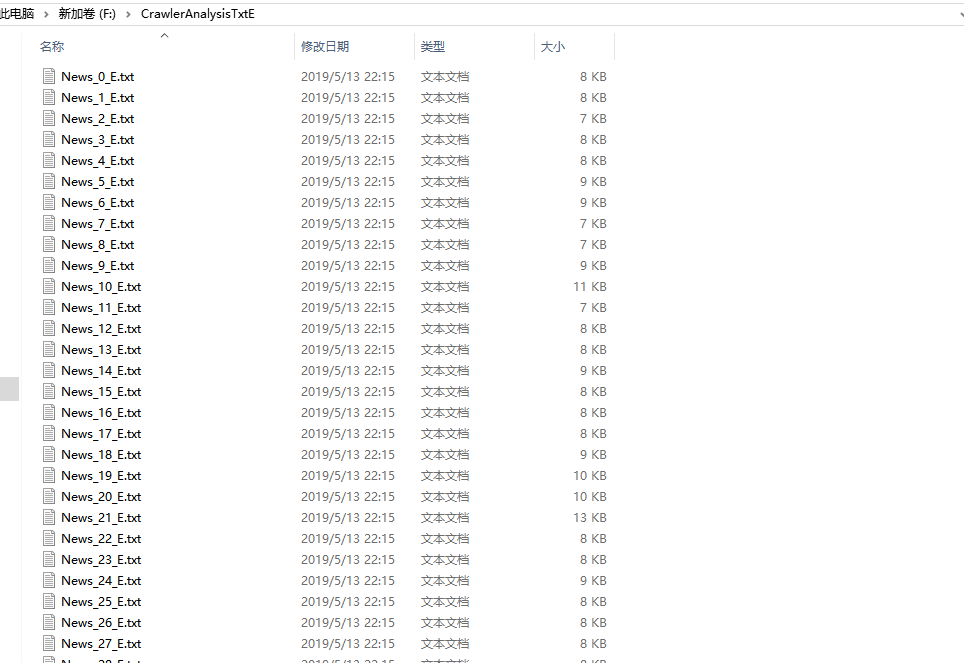
}

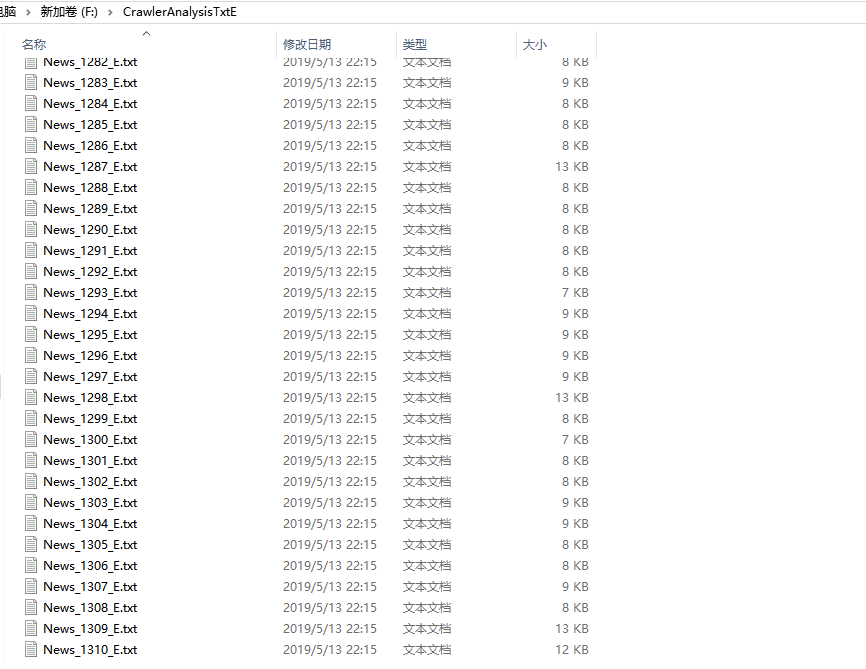
英文停用词表：





英文分词结果目录：





分词结果举例：



5.实验体会

感谢吴老师给了我这次项目的实践，通过本次项目实验，我的编程能力得到很大的提升，也对搜索引擎文本预处理有了更加深刻的理解，果然动手其乐无穷，在实验过程中，所有东西都自己从不同渠道学习，遇到过很多困难，不过都自己一个一个解决了，在过程中感觉很有成就感，看到URL写入数据库、文档从网上下载到本地、文档分词结果，种种这些都让人欣喜，特别是看到自己写的程序的控制台打印的一行行“写入成功”和各种URL，内心真的无比高兴，再次感谢吴老师给的机会