Atitit attilax擅长项目解析与大数据采集提取 词法分析 电话号码提取

**package** vcfvcardprj;

**import** java.util.Collection;

**import** java.util.List;

**import** java.util.regex.Matcher;

**import** java.util.regex.Pattern;

**import** com.alibaba.fastjson.JSON;

**import** com.attilax.fsm.TokenEndEx;

**import** com.attilax.parser.Token;

**import** com.google.common.collect.Lists;

**public** **class** **mblFetch** {

**public** **static** **void** **main**(**String**[] args) {

// **TODO** Auto-generated method stub

**String** **s** = "周何琪\_\_学校郑州大学\_\_联系方式15538130516\_\_身高体重162cm,47k";

s="天津广播影视职业学院 韩震宇 15641656234 161cm，44kg";

List<Token> **process** = **new** mblFetch().getTokens(s);

**System**.***out***.println( **JSON**.*toJSONString*(process,**true**) );

**String** **cp**=*getMblCp*(process);

**System**.***out***.println(cp);

}

**private** **static** **String** **getMblCp**(List<Token> process) {

**for** (**Token** **token** : process) {

**if**(**new** mblFetch().isnum(token.value))

**return** token.value;

}

**return** "";

}

**private** **char**[] **process**(**String** s) {

// **TODO** Auto-generated method stub

**return** **null**;

}

**int** charIndex;

**char** cur\_char;

**char**[] code\_char\_arr;

**private** **String** curStat ="start";

**private** List<Token> tokens\_tmp;

**private** **String** curTokenTxt="";

***@SuppressWarnings***("unchecked")

**public** List<Token> **getTokens**(**String** codeStr) {

List<Token> **li** = **Lists**.*newArrayList*();

code\_char\_arr = codeStr.toCharArray();

**while** (**true**) {

**Object** **tk**;

**try** {

tk = nextTokens();

} **catch** (**TokenEndEx** **e**) {

**break**;

}

**if** (tk **instanceof** **Token**)

li.add((**Token**) tk);

**else** **if** (tk **instanceof** List)

li.addAll((Collection<? **extends** **Token**>) tk);

**else**

**throw** **new** RuntimeException("token type err,curchar:" + cur\_char + ",colidx:" + charIndex);

}

**return** li;

}

/\*\*

\*

\* **@return** token or list<token>

\* **@throws** TokenEndEx

\*/

**public** **Object** **nextTokens**() **throws** **TokenEndEx** {

// code\_char\_arr = code.toCharArray();

charIndex++;

**if** (charIndex > code\_char\_arr.length - 1)

**throw** **new** TokenEndEx(**new** String(code\_char\_arr));

cur\_char = code\_char\_arr[charIndex];

// cur\_char=cur\_char;

// if (this.curTokenTxt.equals("1598"))

// System.out.println("dbg");

// if(this.gColumn==30)

// System.out.println("dbg");

// get next char,,then change stat

// jude cur char and cur stat...then if or not chage stat

**if** (ishanzi(cur\_char))

**return** hanziEvt();

**else** **if** (isnum(cur\_char))

**return** numEvt();

**else**

**return** splitorCharEvt();

// break;

}

**private** **Object** **numEvt**() **throws** **TokenEndEx** {

**if** (**this**.curStat.equals("start")) {

**this**.curStat="numStat";

**return** gaziStat();

}

**if** (**this**.curStat.equals("numStat")) {

**return** gaziStat();

}

**if** (**this**.curStat.equals("hanziStat")) {

**this**.curStat="numStat";

**return** retNumtoken();

}

**if** (**this**.curStat.equals("splitorStat")) {

**this**.curStat="numStat";

**return** retSplitorToken();

}

**return** **null**;

}

**private** **Object** **hanziEvt**() **throws** **TokenEndEx** {

**if** (**this**.curStat.equals("start")) {

**this**.curStat="hanziStat";

**return** gaziStat();

}

**if** (**this**.curStat.equals("hanziStat")) {

**return** gaziStat();

}

// if is hanzi && cur is numstat

**if** (**this**.curStat.equals("numStat")) {

**this**.curStat="hanziStat";

**return** retNumtoken();

}

**if** (**this**.curStat.equals("splitorStat")) {

**this**.curStat="hanziStat";

**return** retSplitorToken();

}

**this**.curStat="hanziStat";

**return** **null**;

}

**private** **Object** **splitorCharEvt**() **throws** **TokenEndEx** {

**if** (**this**.curStat.equals("start")) {

**this**.curStat="splitorStat";

**return** gaziStat();

}

**if** (**this**.curStat.equals("hanziStat")) {

**this**.curStat="splitorStat";

**return** retHeziToken();

}

**if** (**this**.curStat.equals("numStat")) {

**this**.curStat="splitorStat";

**return** retNumtoken();

}

//gazi

**this**.curStat="splitorStat";

**return** gaziStat();

}

**private** **Object** **retHeziToken**() {

**Token** **tk** = **new** Token();

tk.Text = curTokenTxt.toString();

tk.Type = "hezi";

tk.value = curTokenTxt.toString();

curTokenTxt=**String**.*valueOf*(cur\_char);

**return** tk;

}

**private** **Object** **retNumtoken**() {

**Token** **tk** = **new** Token();

tk.Text = curTokenTxt.toString();

tk.Type = "num";

tk.value = curTokenTxt.toString();

curTokenTxt="";

curTokenTxt=**String**.*valueOf*(cur\_char);

**return** tk;

}

**private** **Object** **retSplitorToken**() {

**Token** **tk** = **new** Token();

tk.Text = curTokenTxt.toString();

tk.Type = "splitor";

tk.value = curTokenTxt.toString();

curTokenTxt=""; curTokenTxt=**String**.*valueOf*(cur\_char);

**return** tk;

}

**private** **Object** **gaziStat**() **throws** **TokenEndEx** {

curTokenTxt = curTokenTxt + **String**.*valueOf*(cur\_char);

**return** nextTokens();

}

**private** **boolean** **ishanzi**(**char** cur\_char2) {

**return** *isChinese*(**String**.*valueOf*(cur\_char2));

}

**private** **boolean** **isnum**(**char** cur\_char2) {

**String** **str** = **String**.*valueOf*(cur\_char2);

**return** isnum(str);

}

**private** **boolean** **isnum**(**String** str) {

**for** (**int** **i** = str.length(); --i >= 0;) {

**if** (!**Character**.*isDigit*(str.charAt(i))) {

**return** **false**;

}

}

**return** **true**;

}

**public** **static** **boolean** **isChinese**(**String** str) {

**String** **regEx** = "[\u4e00-\u9fa5]";

**Pattern** **pat** = **Pattern**.*compile*(regEx);

**Matcher** **matcher** = pat.matcher(str);

**boolean** **flg** = **false**;

**if** (matcher.find())

flg = **true**;

**return** flg;

}

}