

Morphological Analysis and lexical indexing of Chinese

汉语语素分析系统和词汇索引制作工具

HINO Yoshihiro (Mejiro University)

y.hino@mejiro.ac.jp

WORKSHOP的目的

- 本WORKSHOP的目的有以下三个。
- 学习汉语语素分析系统的使用方法
- 学习词汇索引制作工具的使用方法
- 使用“正则表达式” “Regular Expression”的文本检索。

今天一起操作电脑，练习相关内容。

今天的资料

<https://www.hinox.org/>



- 如果没安装TEXT EDITOR，请在网上检索一下ATOM。然后请下载并安装ATOM。

具体内容

1. 汉语的语素分析系统的概要和实践

介绍一下 中国、日本以及台湾地区能使用的语素分析网站。通过试用比较一下各个网页的特点。

2. 词汇索引制作工具

我们开发了一个词汇索引制作程序叫“**INDEX CONVERTER**”。试用这个程序做一个全文词汇索引。

3. 使用“正则表达式”的文本检索

网络上有很多汉语语料库，比如北大的**CCL**、北语的**BCC**等。但是其语料库没有自己的研究对象的文本，因此只好自己输入或收藏在自己的电脑里。今天体验一下这时候的检索方法。

学习汉语语素分析的系统使用方法

汉语的语素分析

语素分析是辨析词语很重要的基本功，使用语素分析程序，通过解析文本能把汉语的文章按照一个规则进行分词。语素分析在下面的领域被应用。

- 拼音转换
 - 词性分析
 - 文本分析
 - 提取关键字
 - 机械翻译 等等
-
- 下面看一下提供语素分析服务的网页

目前公开的汉语语素分析的网站

目前有以下提供语素分析服务的网站。

1. 语料库在线 (简体字)
2. NLP IR汉语分词系统 (简体字、繁体字)
3. THULAC (简体字)
4. CTA (简体字、日本)
5. HanLP (简体字)
6. CKIP Corporeal (繁体字)
7. 中文斷詞系統 (繁体字) (今天不讲)
8. 線上中文斷詞工具: Jieba-JS (繁体字) (今天不讲)

语料库在线

- 教育部语言文字应用研究所计算语言学研究室

<http://corpus.zhonghuayuwen.org/>

- 语料库在线的语料分析处理有以下特点
语料分词和词性标注
语料汉语拼音自动标注
语料字词频率统计

用以下文件进行分析

SAMPLE TEXT1

李明，你家在哪儿？

我家在北京。

你家有几口人？都有什么人？

我家有五口人，爸爸、妈妈、哥哥、妹妹和我。

你妹妹多大了？

我妹妹今年十三岁了。

汉语分词和词性自动标注


www.cncorpus.org

首页 语料库检索 语料分析处理 语料字词检索

汉语分词和词性自动标注 词类标记代码

文字内容 (最长100000字) :

李明, 你家在哪儿?
我家在北京。
你家有几口人? 都有什么人?
我家有五口人, 爸爸、妈妈、哥哥、妹妹和我。
你妹妹多大了?
我妹妹今年十三岁了。

自动分词&标注词性 重置

处理结果 :

李明/nh , /w 你/r 家/n 在/p 哪儿/r ? /w
我/r 家/n 在/p 北京/ns 。 /w
你/r 家/n 有/v 几/m 口/n 人/n ? /w 都/d 有/v 什么/r 人/n ? /w
我/r 家/n 有/v 五/m 口/n 人/n , /w 爸爸/n 、 /w 妈妈/n 、 /w 哥哥/n 、 /w 妹妹/n 和/c 我/r 。 /w
你/r 妹妹/n 多/a 大/a 了/u ? /w |
我/r 妹妹/n 今年/nt 十三/m 岁/nt 了/u 。 /w

语料库在线的 汉语分词和词性自动标注

分词后，后面加词性标注。标记代码的意思在下一页。

李明/nh ， /w 你/r 家/n 在/p 哪儿/r ？ /w

我/r 家/n 在/p 北京/ns 。 /w

你/r 家/n 有/v 几/m 口/n 人/n ？ /w 都/d

有/v 什么/r 人/n ？ /w

我/r 家/n 有/v 五/m 口/n 人/n ， /w 爸爸

/n 、 /w 妈妈/n 、 /w 哥哥/n 、 /w 妹妹/n 和

/c 我/r 。 /w

你/r 妹妹/n 多/a 大/a 了/u ？ /w

我/r 妹妹/n 今年/nt 十三/m 岁/nt 了/u 。 /w

词类标记代码

n 普通名词
nt 时间名词
nd 方位名词
Nl 处所名词
Nh 人名
Nhf 姓
Nhs 名
Ns 地名
Nn 族名
ni 机构名
Nz 其他专名
V 动词
Vd 趋向动词

VI 联系动词
Vu 能愿动词
A 形容词
F 区别词
M 数词
Q 量词
D 副词
R 代词
P 介词
C 连词
U 助词
E 叹词
O 拟声词

I 习用语
J 缩略语
H 前接成分
K 后接成分
G 语素字
X 非语素字
W 标点符号
Ws 非汉字字符串
Wu 其他未知的符号

字频统计

这是字频统计功能，能知道文章里出现多少同样的字。


www.chcorpus.org

[首页](#) [语料库检索](#) [语料分析处理](#) [语料字词检索](#)

● 字词频率统计

文字内容 (最长100000字) : 共 296 字符

李明/nh , /w 你/r 家/n 在/p 哪儿/r ?/w
我/r 家/n 在/p 北京/ns 。/w
你/r 家/n 有/v 几/m 口/n 人/n ?/w 都/d 有/v 什么/r 人/n ?/w
我/r 家/n 有/v 五/m 口/n 人/n , /w 爸爸/n 、 /w 妈妈/n 、 /w 哥哥/n 、 /w 妹
妹/n 和/c 我/r 。/w
你/r 妹妹/n 多/a 大/a 了/u ?/w
我/r 妹妹/n 今年/nt 十三/m 岁/nt 了/u 。/w

☐ 只要汉字

字频统计

词频统计

下载结果

序号	字词	频次	频率 %
1		108	37.7622
2	/	54	18.8811
3	n	19	6.6434
4	w	12	4.1958
5	r	9	3.1469
6	妹	6	2.0979
7	家	4	1.3986
8	我	4	1.3986
9	、	3	1.049
10	。	3	1.049
11	?	3	1.049
12	m	3	1.049
13	v	3	1.049
14	你	3	1.049
15	人	3	1.049
16	有	3	1.049
17	,	2	0.6993
18	a	2	0.6993
19	p	2	0.6993
20	t	2	0.6993
21	u	2	0.6993
22	爸	2	0.6993

词频统计

这是词频统计功能，能知道文章里出现多少同样的词。

语料库在线
www.chncorpus.org

首页

语料库检索

语料分析处理

语料字词检索

现代汉语语料库检索
古代汉语语料库检索

字词频率统计

文字内容 (最长100000字) : 共 296 字符

李明/nh , /w 你/r 家/n 在/p 哪儿/r ?/w
我/r 家/n 在/p 北京/ns 。/w
你/r 家/n 有/v 几/m 口/n 人/n ?/w 都/d 有/v 什么/r 人/n ?/w
我/r 家/n 有/v 五/m 口/n 人/n , /w 爸爸/n 、/w 妈妈/n 、/w 哥哥/n 、/w 妹
妹/n 和/c 我/r 。/w
你/r 妹妹/n 多/a 大/a 了/u ?/w
我/r 妹妹/n 今年/nt 十三/m 岁/nt 了/u 。/w

☐ 只要汉字
字频统计
词频统计
下载结果

序号	字词	频次	频率 %
1	家	4	10.8108
2	妹妹	3	8.1081
3	人	3	8.1081
4	有	3	8.1081
5	口	2	5.4054
6	了	2	5.4054
7	在	2	5.4054
8	爸爸	1	2.7027
9	北京	1	2.7027
10	大	1	2.7027
11	都	1	2.7027
12	多	1	2.7027
13	哥哥	1	2.7027
14	和	1	2.7027
15	几	1	2.7027
16	今年	1	2.7027
17	李明	1	2.7027
18	妈妈	1	2.7027
19	哪儿	1	2.7027
20	你	1	2.7027
21	什么	1	2.7027
22	十三	1	2.7027

汉语拼音自动标注

有拼音转换功能，但是没有声调符号。

语料库在线
www.cncorpus.org

首页语料库检索语料分析处理语料字词检索

汉语拼音自动标注

文字内容（最长100000字）：

李明，你家在哪儿？
我家在北京。
你家有几口人？都有什么人？
我家有五口人，爸爸、妈妈、哥哥、妹妹和我。
你妹妹多大了？
我妹妹今年十三岁了。

标注汉语拼音传统标调数字标调重置

处理结果：

李明/límíng，你/nǐ家/jiā在/zài哪儿/nǎr？
我/wǒ家/jiā在/zài北京/běijīng。
你/nǐ家/jiā有/yǒu几/jǐ口/kǒu人/rén？都/dōu有/yǒu什么/shénme人/rén？
我/wǒ家/jiā有/yǒu五/wǔ口/kǒu人/rén，爸爸/bàba、妈妈/māma、哥哥/gēge、妹妹/mèimei和/hé
我/wǒ。
你/nǐ妹妹/mèimei多/duō大/dà了/le？
我/wǒ妹妹/mèimei今年/jīnnián十三/shí sān岁/sui了/le。

NLPIR汉语分词系统

- 张华平，北京理工大学

<http://ictclas.nlpir.org>

- NLPIR汉语分词系统有自动对中文信息进行分词与词性标注的功能，涵盖了中文分词、词性标注、未登录词识别与用户词典等功能。
- 还有关键词提取功能、新词识别与自适应分词功能、用户专业词典功能等。

NLP IR汉语分词系统的分词标注

这是分词标注功能。不同的颜色代表不同的词类，一眼就能知道词的词性。

The screenshot displays the NLP IR Chinese word segmentation system interface. On the left, a sidebar contains a red circular icon with a sandglass symbol and the text "分词标注" (Word Segmentation and Tagging). Below this, a list of "词性类别" (Word Class Categories) is shown, each with a corresponding color-coded box: 名词 (Noun, orange), 动词 (Verb, blue), 介词 (Preposition, pink), 后缀 (Suffix, light blue), 代词 (Pronoun, light green), 数词 (Numeral, yellow), 连词 (Conjunction, light orange), 助词 (Particle, light purple), 叹词 (Interjection, light green), 前缀 (Prefix, light blue), 量词 (Measure word, light purple), 副词 (Adverb, light blue), 语气词 (Interjection, light blue), 拟声词 (Onomatopoeia, light blue), 字符串 (String, light green), 形容词 (Adjective, light green), 时间词 (Time word, light purple), 处所词 (Location word, light purple), 区别词 (Distinguishing word, light orange), 方位词 (Directional word, light blue), 状态词 (State word, light purple), 标点符号 (Punctuation mark, light purple), and 自定义词 (Custom word, red).

The main area displays a text input field with the following text: "李明/nr , /wd 你/rr 家/n 在/p 哪儿/rys ? /ww 我/rr 家/n 在/p 北京/ns 。 /wj 你/rr 家/n 有/vyou 几/m 口/q 人/n ? /ww 都/d 有/vyou 什/ng 么人/nr ? /ww 我/rr 家/n 有/vyou 五/m 口/q 人/n , /wd 爸爸/n 、 /wn 妈妈/n 、 /wn 哥哥/n 、 /wn 妹妹/n 和/cc 我/rr 。 /wj 你/rr 妹妹/n 多/m 大/a 了/y ? /ww 我/rr 妹妹/n 今年/t 十三/m 岁/qt 了/y 。 /wj". Each word is highlighted with a color-coded box indicating its part of speech.

On the right side, there is a section titled "新词发现" (New Word Discovery) and a section titled "用户自定义词语" (User-defined Words) with a text input field and a "导入" (Import) button.

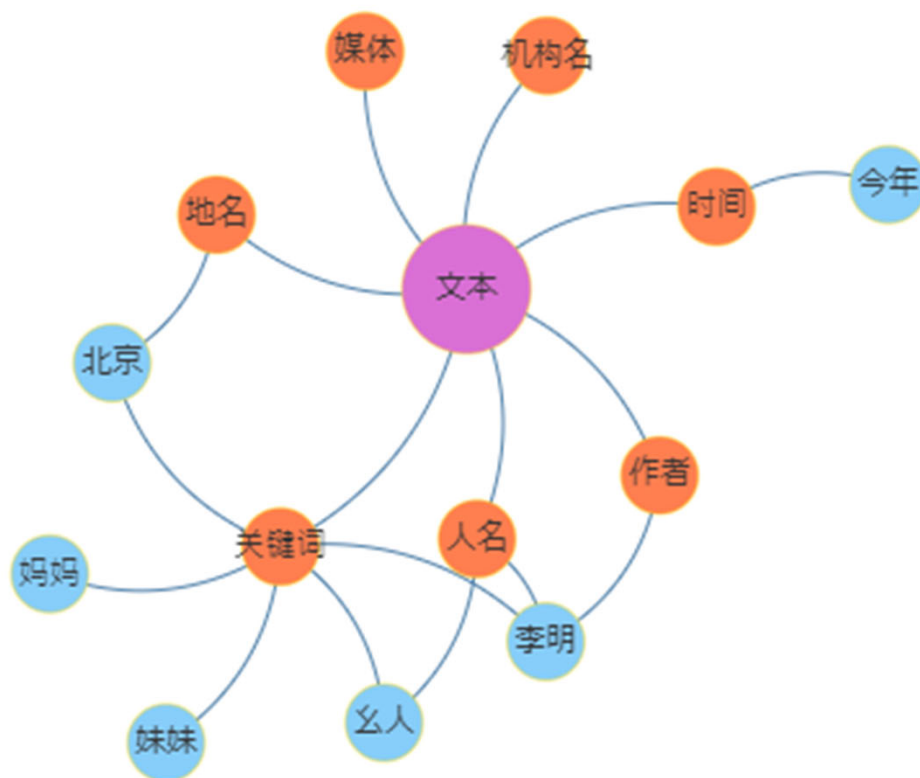
NLPIR汉语分词系统的分析

分词后，后面也加词性标注。标注后没有空格。

李明/nr, /wd你/rr家/n在/p哪儿/rys? /ww 我/rr家/n
在/p北京/ns。 /wj 你/rr家/n有/vyou几/m口/q人/n?
/ww都/d有/vyou什/ng么人/nr? /ww 我/rr家/n有
/vyou五/m口/q人/n, /wd爸爸/n、 /wn妈妈/n、 /wn
哥哥/n、 /wn妹妹/n和/cc我/rr。 /wj 你/rr妹妹/n多/m
大/a了/y? /ww 我/rr妹妹/n今年/t十三/m岁/qt了/y。
/wj

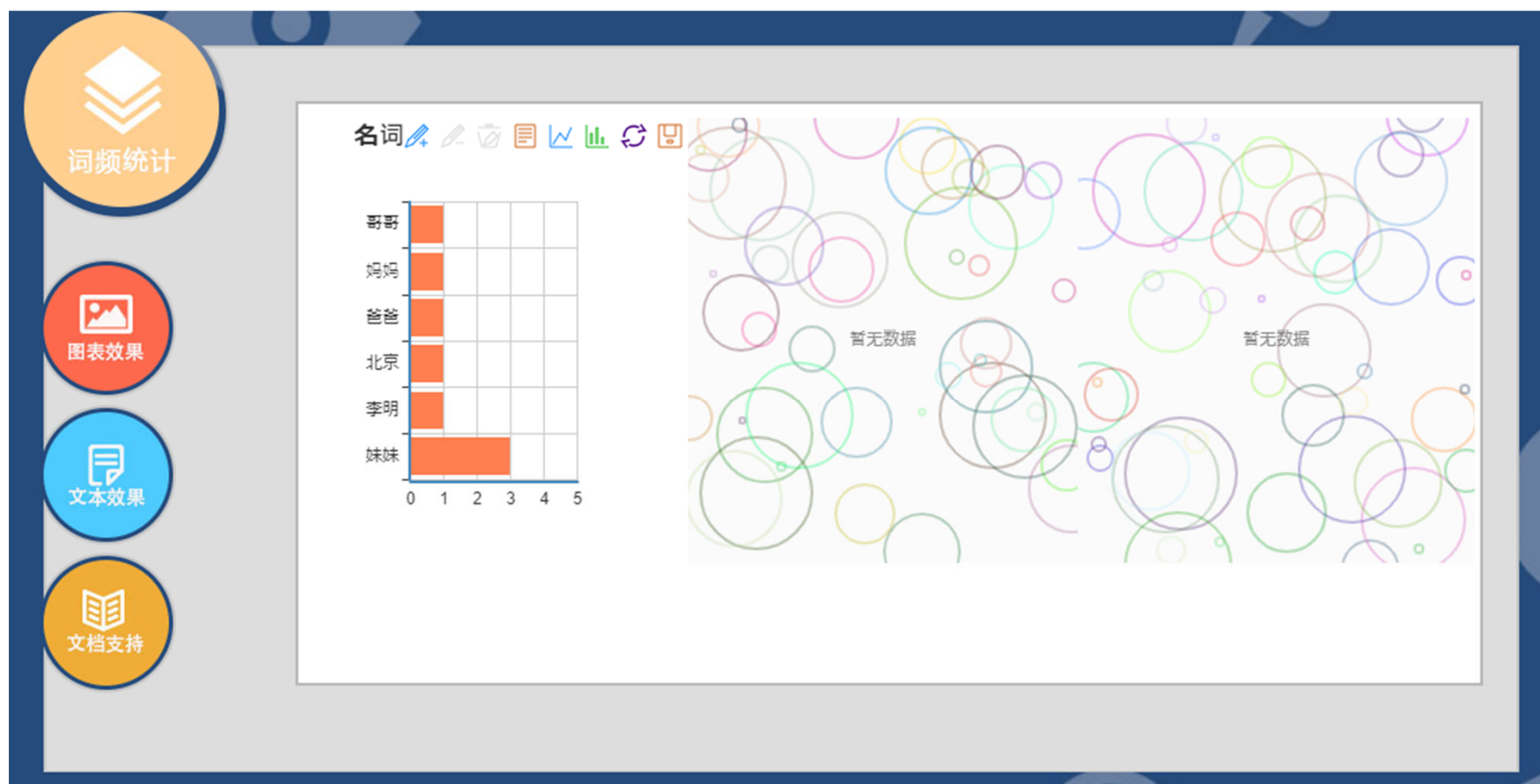
实体抽取

把词汇分类为人名、地名、机构名、作者、关键词、时间、媒体关键词、地名等相关联的词汇进行抽取。



词汇统计

把词汇进行词性分类后，统计词汇的出现频率。



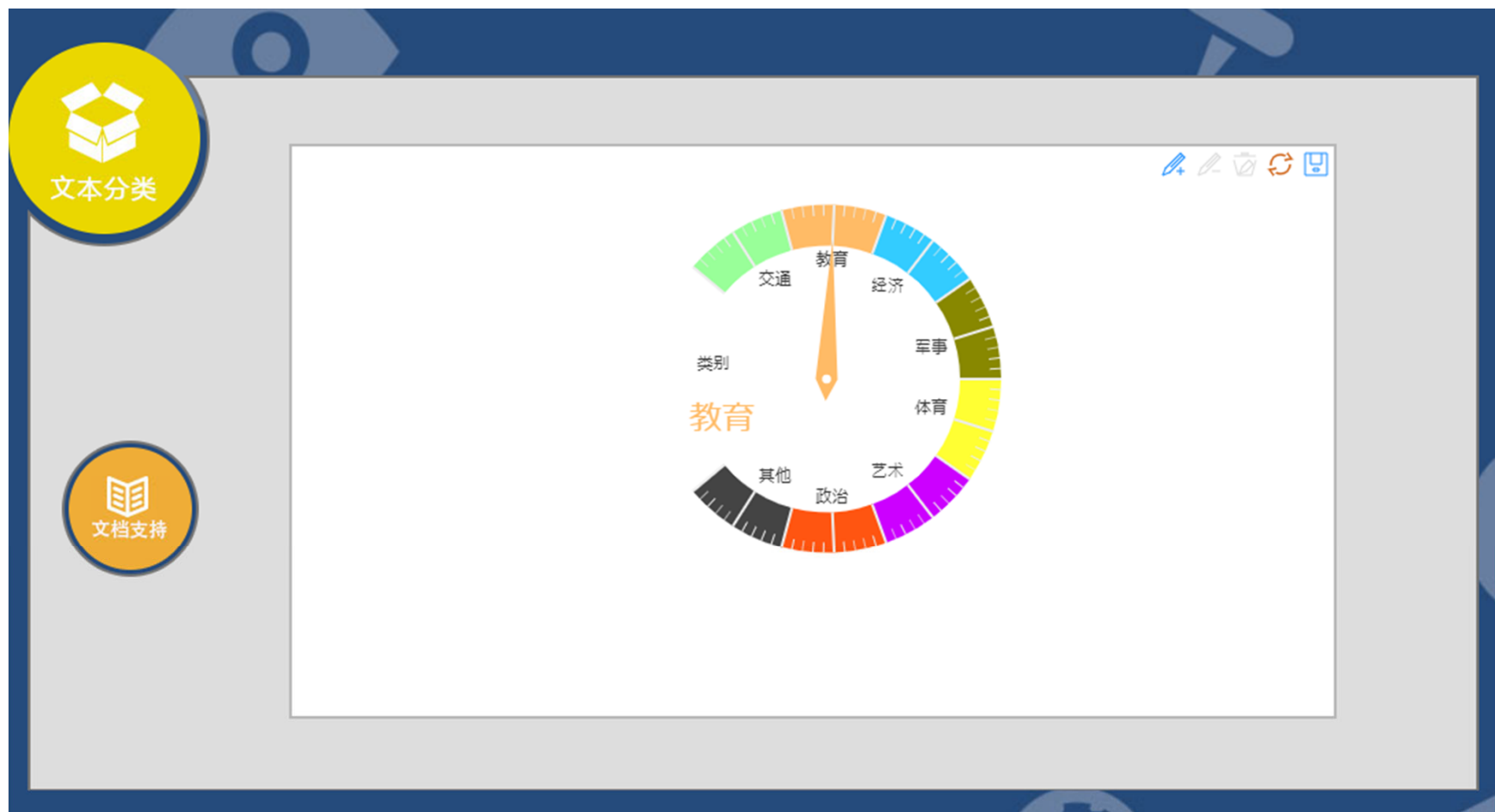
词汇统计

这是用其他比较长一点的文章做的词汇统计的结果。



文本分类

分析文本的内容，显示文本内容的种类。



关键词提取

显示关键词图。



简繁转换 拼音转换功能

简繁转换相当好，但是拼音转换后，各个词间还是没有空格。

繁
拼

李明，你家在哪儿？
我家在北京。
你家有几口人？都有什么人？
我家有五口人，爸爸、妈妈、哥哥、妹妹和我。
你妹妹多大了？
我妹妹今年十三岁了。

繁
拼

李明，你家在哪兒？
我家在北京。
你家有幾口人？都有什麼人？
我家有五口人，爸爸、媽媽、哥哥、妹妹和我。
你妹妹多大了？
我妹妹今年十三歲了。

繁
拼

LiMing, NiJiaZaiNaEr ?
WoJiaZaiBeiJing.
NiJiaYouJiKouRen ? DuYouShiMeRen ?
WoJiaYouWuKouRen, BaBa、MaMa、GeGe、MeiMei
NiMeiMeiDuoDaLiao?
WoMeiMeiJinNianShiSanSuiLiao.

3. THULAC

- 清华大学自然语言处理与社会人文计算实验室
<http://thulac.thunlp.org>
- THULAC (THU Lexical Analyzer for Chinese) 由清华大学自然语言处理与社会人文计算实验室研制推出的一套中文词法分析工具包，具有中文分词和词性标注功能。利用他们集成的目前世界上规模最大的人工分词和词性标注中文语料库训练而成，模型标注能力强大。

THULAC的分析

分词后，后面也加词性标注。形式与其他网站不一样。但是基本功能，显示的内容都一样。

李明_np ， _w 你_r 家_n 在_p 哪儿_r ？ _w

我_r 家_n 在_p 北京_ns 。 _w

你_r 家_n 有_v 几_m 口_q 人_n ？ _w 都_d 有_v 什么_r
人_n ？ _w

我_r 家_n 有_v 五_m 口_q 人_n ， _w 爸爸_n 、 _w 妈妈
_n 、 _w 哥哥_n 、 _w 妹妹_n 和_c 我_r 。 _w

你_r 妹妹_n 多_d 大_a 了_u ？ _w

我_r 妹妹_n 今年_t 十三_m 岁_q 了_u 。 _w

THULAC的词性解释

N	名词	V	动词	P	介词
Np	人名	Vm	能愿动词	U	助词
Ns	地名	Vd	趋向动词	Y	语气助词
Ni	机构名	A	形容词	E	叹词
Nz	其它专名	D	副词	O	拟声词
M	数词	H	前接成分	G	语素
Q	量词	K	后接成分	W	标点
Mq	数量词	I	习语	X	其它
T	时间词	J	简称		
F	方位词	R	代词		
S	处所词	C	连词		

4. CTA

- HINO Yoshihiro

<http://ch-station.org/cta/>

- 在日本编辑初级汉语课本或编辑教材的时候，分词和加拼音特别麻烦。而且中国语素分析网站的拼音转换功能教学上都不好用。所以自己开发了语素分析系统。
- CTA是以日语分词系统MeCab为基础的汉语分词系统。
- 有分词、拼音转换、词汇对照功能等。
- 拼音转换按照《汉语拼音正词法规则》。

CTA的分析

分词后，下面显出拼音。颜色表示词汇级别（参照HSK的级别）。

汉字+ピンイン

汉字

ピンイン

分割修正

私案処理

エクスポート

印刷

クリア

「漢字+ピンインモード」では複数の読み方がある漢字は青色で表示されます。クリックして正しいピンインを選択してください。分割が間違っている場合は「分割修正」をクリックして修正してください。

李明	,	你	家	在	哪儿	?	我	家	在
Lǐmíng	,	nǐ	jiā	zài	nǎr	?	Wǒ	jiā	zài
北京	。	你	家	有	几	口	人	?	都
Běijīng	,	Nǐ	jiā	yǒu	jǐ	kǒu	rén	?	Dōu
有	什么	人	?	我	家	有	五	口	人
yǒu	shénme	rén	?	Wǒ	jiā	yǒu	wǔ	kǒu	rén
,	爸爸	、	妈妈	、	哥哥	、	妹妹	和	我
,	bàba	,	māma	,	gēge	,	mèimei	hé	wǒ
。	你	妹妹	多	大	了	?	我	妹妹	今年
,	Nǐ	mèimei	duō	dà	le	?	Wǒ	mèimei	jīnnián
十	三	岁	了	。					
shí	sān	sui	le	,					

照ropen

New HSK

New HSK

☒Lv1 (19)

☒Lv2 (2)

☒Lv3 (1)

☒Lv4 (0)

☒Lv5 (0)

☒Lv6 (0)

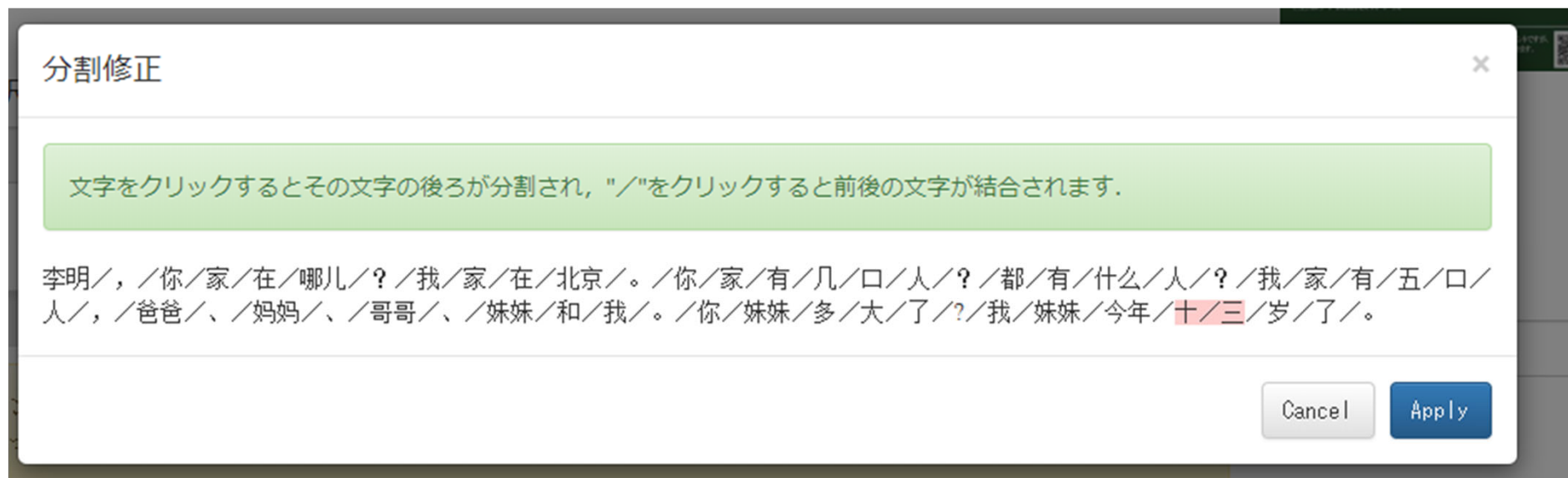
☒無 (4)

☐未

破線

CTA的错误修改功能

分词万一有错误，很容易修改。



/你/家/在/哪儿/? /我/家/在/北京/. /你/家/有/几/口/人/? /都/有/什么/人/? /我/家/有/五/口/爸爸/, /妈妈/, /哥哥/, /妹妹/和/我/. /你/妹妹/多/大/了/? /我/妹妹/今年/十三/岁/了/。

CTA的错误修改功能

若发音有错误，很容易修改。

有异读的汉字，只按其字就换另一个拼音。

北京	。	你	家	有	几	口	人	？	都
Běijīng	.	Nǐ	jiā	yǒu	jǐ	kǒu	rén	?	Dū

北京	。	你	家	有	几	口	人	？	都
Běijīng	.	Nǐ	jiā	yǒu	jǐ	kǒu	rén	?	Dōu

CTA的分析

CTA目前还未提供词性分类功能，预计今年底开始提供该功能。

李明，你 家 在 哪儿 ？

我 家 在 北京 。

你 家 有 几 口 人 ？ 都 有 什么 人 ？

我 家 有 五 口 人 ， 爸爸 、 妈妈 、
哥哥 、 妹妹 和 我 。

你 妹妹 多 大 了 ？

我 妹妹 今年 十三 岁 了 。

5. HanLP

- HanLP

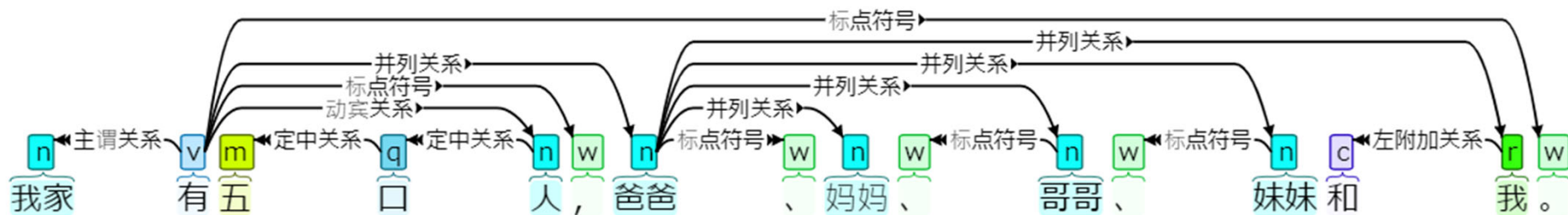
<https://github.com/hankcs/HanLP>

HanLP有词法分析（中文分词、词性标注、命名实体识别）、句法分析、文本分类和情感分析等功能。

HanLP公开的工具包<http://hanlp.com/>可以进行句法分析。

HanLP的分析

n v m q n w n w n w n w n c r w
我家 有 五 口 人 , 爸爸 、 妈妈 、 哥哥 、 妹妹 和 我 。



6. CKIP CoreNLP

- 中央研究院 中研院资讯所
<http://ckip.iis.sinica.edu.tw/service/corenlp/>
- CKIP CoreNLP有分词（断词）分析系统、实体分析等功能。
- 目前简体字的词汇无法分析？

CKIP CoreNLP的分析

分词后，后面加词性标注。

斷詞系統

☒ Show POS tagging

李明(Nb) , (COMMACATEGORY) 你(Nh) 家(Nc) 在(P) 哪兒(Ncd) ? (QUESTIONCATEGORY)

我(Nh) 家(Nc) 在(P) 北京(Nc) 。 (PERIODCATEGORY)

你(Nh) 家(Nc) 有(V_2) 幾(Neqa) 口(Nf) 人(Na) ? (QUESTIONCATEGORY) 都(D) 有(V_2) 什麼(Nep) 人(Na)

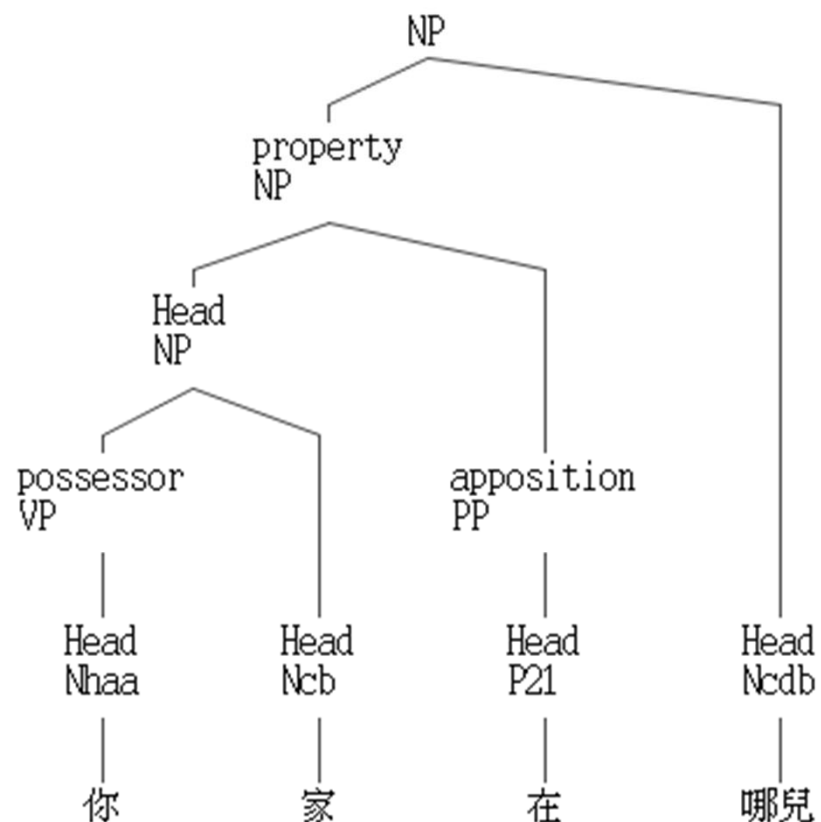
我(Nh) 家(Nc) 有(V_2) 五(Neqa) 口(Nf) 人(Na) , (COMMACATEGORY) 爸爸(Na) 、 (PAUSECATEGORY) 媽媽
(PAUSECATEGORY) 妹妹(Na) 和(Caa) 我(Nh) 。 (PERIODCATEGORY)

你(Nh) 妹妹(Na) 多(Dfa) 大(VH) 了(T) ? (QUESTIONCATEGORY)

CKIP CoreNLP的分析

还有句法分析功能。

#2:1.[0] NP(property:NP(Head:NP(possessor:VP(Head:Nhaa:你)|Head:Ncb:家)|apposition:PP(Head:P21:在))|He:



CKIP CoreNLP的分析

指代消解功能。这功能其他网站没见过。主要分析代词指的是谁。

指代消解

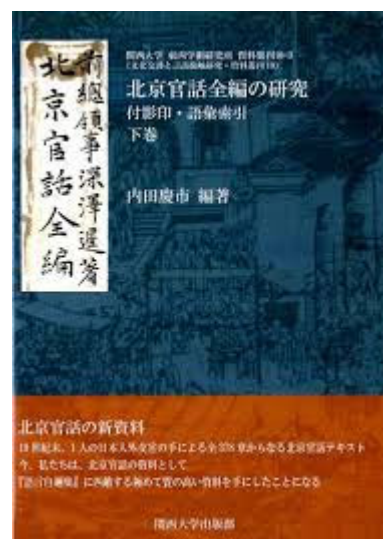
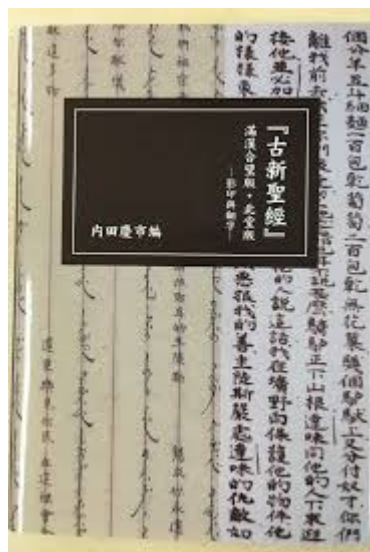
李明，你李明家在哪兒？我李明家在北京。你李明家有幾口人？NULL李明都有什麼人？我李明家有五口人，妹多大了？我李明妹妹今年十三歲了。

- 提供语素分析的网站很多。接下来用语料库在线和INDEXCONVER做全文词汇索引。

实践 INDEX CONVERTER

全文词汇索引是什么？

检索词汇是有自己想检索的词汇才能检索。
很难掌握一篇文章里有什么样的词汇。所以
过去很多日本的学者为了词汇或语言研究，
做自己的研究对象的书的词汇索引集。



全文词汇索引是什么？

左边的图示全文词汇索引。比如 官名 在这本书上只出现一次，1-3-a1b-1表示其页数和行数

官話：4 1-17-A4b-9, 1-17-A4b-10,

1-17-A4b-10, 4-18-B53b-2

官帽兒：1 3-6-B7a-3

官名：1 1-3-A1b-1

官兒：10 2-12-A26a-7, 2-12-A26a-8,

2-12-A26a-8, 2-16-A34

2-19-A39b-1, 2-32-A6

2-35-A68a-3, 2-35-A6

2-35-A69b-1, 2-38-A7

4-19-B55b-3, 4-19-B55b-4

官員：1 4-18-B54a-6

官座兒：3 3-11-B13a-10, 3-11-B13b-2,

3-11-B13b-4

關：7 2-2-A14b-7, 2-9-A21b-3, 2-9-A21b-4,

2-17-A36a-6, 2-23-A47a-10, 2-29-A57b-2,

4-13-B48a-3

賤姓張，官名呼守先。
我還小，哪今年二十
買賣，我和你令叔相
號，小號信昌，
渴想得很，今兒早起
當勞您駕，我本要到
的行李各件還沒拾

幾，我居長，
喜在那兒，
來請安，不
到了，特過來
去，就因為昨
子也還沒打

1 1 3 A1 B 1
1回／ 第1卷第3章 第1葉 裏 1行目

准备工作

1. 输入全文。（平时用OCR或手工输入，今天由于时间关系，请用hinox.org上的sample text）
 2. 使用语素分析系统，分词、加词性标注。
（如果不需要词性信息的话只分词就行） 以TEXT形式保存文件，源代码是UTF-8。
 3. 使用INDEX-CONVER网站制作全文词汇索引
（只上传上面做的文件按一个按钮就可以。）
- 下面逐一介绍。

STEP1 准备文件

输入全文。（平时用OCR或手工输入，今天由于时间关系，请用hinox.org上的sample text）

- 简繁都可以。
- 利用Text Editor做UTF-8形式的文件。我平时用ATOM Editor。

STEP2 分词/加标注

使用语素分析系统，分词、加词性标注。以TEXT形式保存文件，源代码是UTF-8。

- 用语料库进行在线分词和标注。
- 按照书的信息加页数、章数、改行等的标注。这目前只能手工，还没有自动化功能。

STEP1 准备文件

SAMPLE 利用语料库在线分词

李明，你家在哪儿？

我家在北京。

你家有几口人？都有什么人？

我家有五口人，爸爸、妈妈、哥哥、妹妹和我。

你妹妹多大了？

我妹妹今年十三岁了。

在hinox.org下载以下压缩数据。

[Simple data for practice](#)

STEP2 分词/加标注

SAMPLE利用语料库在线分词

李明/nh ， /w 你/r 家/n 在/p 哪儿/r ？ /w

我/r 家/n 在/p 北京/ns 。 /w

你/r 家/n 有/v 几/m 口/n 人/n ？ /w 都/d 有/v 什么/r 人/n ？ /w

我/r 家/n 有/v 五/m 口/n 人/n ， /w 爸爸/n 、 /w 妈妈/n 、 /w 哥哥/n 、
/w 妹妹/n 和/c 我/r 。 /w

你/r 妹妹/n 多/a 大/a 了/u ？ /w

我/r 妹妹/n 今年/nt 十三/m 岁/nt 了/u 。 /w

分词后词汇之间一定需要半角的空格，推荐
使用语料库在线的分词功能。

STEP2 分词/加标注

SAMPLE按照书的版面改行

李明/nh , /w 你/r 家/n 在/p 哪儿/r ? /w
我/r 家/n 在/p 北京/ns 。 /w
你/r 家/n 有/v 几/m 口/n 人/n ? /w 都/d 有/v
什么/r 人/n ? /w
我/r 家/n 有/v 五/m 口/n 人/n , /w
爸爸/n 、 /w 妈妈/n 、 /w 哥哥/n 、 /w
妹妹/n 和/c 我/r 。 /w
你/r 妹妹/n 多/a 大/a 了/u ? /w
我/r 妹妹/n 今年/nt
十三/m 岁/nt 了/u 。 /w

按照书版面的该行位置加该行。

第一课 P. 12

李明你家在哪儿？
我家在北京。
你家有几口人？都有
什么人？
我家有五口人，
爸爸、妈妈、哥哥、
妹妹和我。
你妹妹多大了？

第一课 P. 13

我妹妹今年
十三岁了。

书的版面

加标注

INDEX Converter的标注有以下四个。

P 页数

V 章数

L是行数

= 改行地点断的词汇的标注

比如，P. 12 表示 <P 12>。第1课、第1章等用 <V 1> <V 第1课> <V 第1章>来表示。这段第3行开始的话用 <L 3>来表示。PVL后面一定要半角的空格。

加标注

如果改行的地方有断行的词汇的加 = 。

第一课 P. 12

李明你家在哪儿？我家在**北**
京。你家有几口人？都有
什么人？我家有五口人，**爸**
爸、妈妈、哥哥、妹妹和我。
你妹妹多大了？我妹妹今年
十三岁了。

第一课 P. 12

李明你家在哪儿？我家在**北=**
京。你家有几口人？都有
什么人？我家有五口人，**爸=**
爸、妈妈、哥哥、妹妹和我。
你妹妹多大了？我妹妹今年
十三岁了。

加标注

<P 12>

<V 1>

<L 3>

李明/nh , /w 你/r 家/n 在/p 哪儿/r ? /w

我/r 家/n 在/p 北京/ns 。 /w

你/r 家/n 有/v 几/m 口/n 人/n ? /w 都/d 有/v

什么/r 人/n ? /w

我/r 家/n 有/v 五/m 口/n 人/n , /w

爸爸/n 、 /w 妈妈/n 、 /w 哥哥/n 、 /w

妹妹/n 和/c 我/r 。 /w

你/r 妹妹/n 多/a 大/a 了/u ? /w

<P 13>

<L 1>

我/r 妹妹/n 今年/nt

十三/m 岁/nt 了/u 。 /w

第一课 P. 12

李明你家在哪儿？

我家在北京。

你家有几口人？都有什么人？

我家有五口人，
爸爸、妈妈、哥哥、
妹妹和我。

你妹妹多大了？

第一课 P. 13

我妹妹今年
十三岁了。

加标注

<P 12> 页数 P和V的顺序反过来也没问题。

<V 1> 课数

<L 3> 行数

李明/nh , /w 你/r 家/n 在/p 哪儿/r ? /w

我/r 家/n 在/p 北京/ns 。 /w

你/r 家/n 有/v 几/m 口/n 人/n ? /w 都/d 有/v

什么/r 人/n ? /w

我/r 家/n 有/v 五/m 口/n 人/n , /w

爸爸/n 、 /w 妈妈/n 、 /w 哥哥/n 、 /w

妹妹/n 和/c 我/r 。 /w

你/r 妹妹/n 多/a 大/a 了/u ? /w

<P 13>

<L 1> L1的话, 不加也可以

我/r 妹妹/n 今年/nt

十三/m 岁/nt 了/u 。 /w

第一课 P. 12

李明你家在哪儿?

我家在北京。

你家有几口人? 都有什么人?

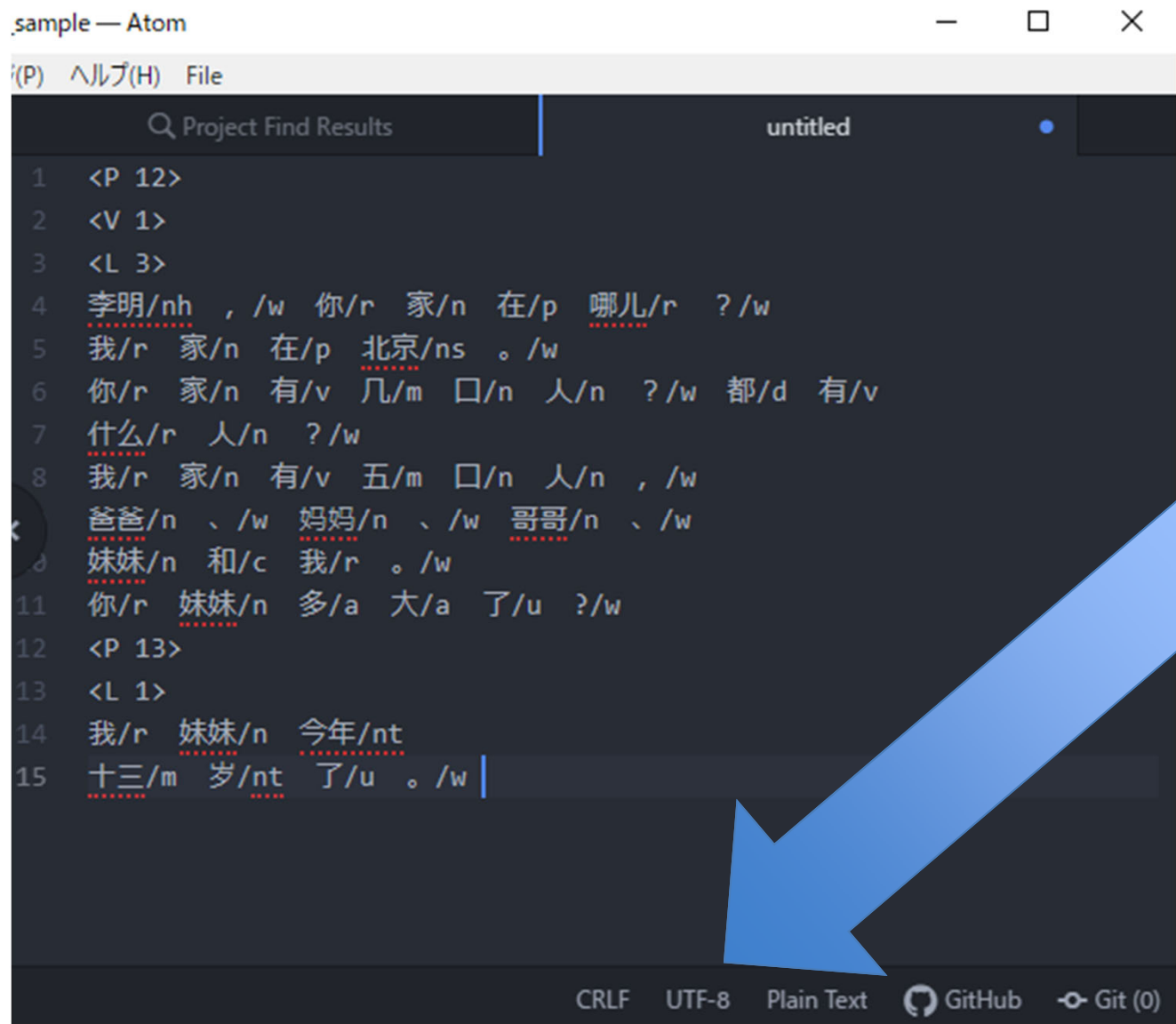
我家有五口人,
爸爸、妈妈、哥哥、
妹妹和我。

你妹妹多大了?

第一课 P. 13

我妹妹今年
十三岁了。

确认源代码后保存



sample — Atom

(P) ヘルプ(H) File

Project Find Results untyped

```
1 <P 12>
2 <V 1>
3 <L 3>
4 李明/nh , /w 你/r 家/n 在/p 哪儿/r ?/w
5 我/r 家/n 在/p 北京/ns 。/w
6 你/r 家/n 有/v 几/m 口/n 人/n ?/w 都/d 有/v
7 什么/r 人/n ?/w
8 我/r 家/n 有/v 五/m 口/n 人/n , /w
9 爸爸/n 、/w 妈妈/n 、/w 哥哥/n 、/w
10 妹妹/n 和/c 我/r 。/w
11 你/r 妹妹/n 多/a 大/a 了/u ?/w
12 <P 13>
13 <L 1>
14 我/r 妹妹/n 今年/nt
15 十三/m 岁/nt 了/u 。/w
```

CRLF UTF-8 Plain Text GitHub Git (0)

Step3 用网站的系统来分析

使用INDEX-CONVER网站制作全文词汇索引
打开保存的文件，按一下“ソート実行”



してソートを実行



ダウンロード



STEP1：テキストデータの作成

サイズ各自のパソコンでテキストデータを作成してUTF-8で保存する。

STEP2：ファイルをアップロードしてソートを実行

作成したテキストファイルを読み込む。



C:\fakepath\sample.txt

```
<P 1>
<V 1>
<L 1>
李明/nh , /w 你/r 家/n 在/p 哪儿/r ? /w
我/r 家/n 在/p 北京/ns 。 /w
你/r 家/n 有/v 几/m 口/n 人/n ? /w 都/d 有/v
什么/r 人/n ? /w
我/r 家/n 有/v 五/m 口/n 人/n , /w
```

用网页的系统来分析

下面显示分析结果。目前句号、问号没有拼音
所以出现提示警告。如果不要的话自己删掉。

STEP3：ソート形式を選択

ローマ字ソート

漢字ソート

辞書ソート

```
/w:*ピンインが格納されていません*:8 1-1-1, 1-1-2, 1-1-5, 1-1-6, 1-1-6, 1-1-6, 1-1-7, 1-2-4
?/w:*ピンインが格納されていません*:1 1-2-2
? /w:*ピンインが格納されていません*:3 1-1-1, 1-1-3, 1-1-4
爸爸/n:1 1-1-6
北京/ns:1 1-1-2
大/a:1 1-2-2
都/d:1 1-1-3
多/a:1 1-2-2
哥哥/n:1 1-1-6
和/c:1 1-1-7
```

怎么看分析结果

这个文章的词汇总数是25个。比如“我”出现4次。分别出现在第1课 第12页 第4、7、9行和第1课第13页第1行。

我/r:4 1-12-4, 1-12-7, 1-12-9, 1-13-1

五/m:1 1-12-7

有/v:3 1-12-5, 1-12-5, 1-12-7

在/p:2 1-12-3, 1-12-4

Total words:25

我（r代词）出现数4

第一课 P. 12

李明你家在哪儿？

我家在北京。

你家有几口人？都有什么人？

我家有五口人，爸爸、妈妈、哥哥、妹妹和我。

你妹妹多大了？

第一课 P. 13

我妹妹今年十三岁了。

排序功能

INDEXCONVERTER有三个排序功能。

STEP3：ソート形式を選択

ローマ字ソート

漢字ソート

辞書ソート

```
/w:*ピンインが格納されていません*:8 1-1-1, 1-1-2, 1-1-5, 1-1-6, 1-1-6, 1-1-6, 1-1-7, 1-2-4
?/w:*ピンインが格納されていません*:1 1-2-2
? /w:*ピンインが格納されていません*:3 1-1-1, 1-1-3, 1-1-4
爸爸/n:1 1-1-6
北京/ns:1 1-1-2
大/a:1 1-2-2
都/d:1 1-1-3
多/a:1 1-2-2
哥哥/n:1 1-1-6
和/c:1 1-1-7
```

3つのソート機能

- 按字母顺序排序
- 按汉字顺序排序
- 按词典顺序排序

ローマ字ソート 漢字ソート 辞書ソート

一個/r:2 0-2, 0-9
一切/r:1 0-30
一向/d:1 0-4
一帶/n:1 0-16
一点/mq:1 0-33
一百/m:1 0-30
上/nd:8 0-2, 0-2, 0-6, 0-7, 0-8, 0-13, 0-31, 0-33
上月/nt:1 0-16
不/d:14 0-1, 0-2, 0-2, 0-2, 0-3, 0-13, 0-13, 0-13, 0-14, 0-18,
不好/a:1 0-2

ローマ字ソート 漢字ソート 辞書ソート

阿公/n:1 0-33
阿鹤亭/ns:3 0-22, 0-32, 0-33
阿妈/n:2 0-33, 0-33
啊/e:4 0-3, 0-4, 0-5, 0-14
哀/a:2 0-2, 0-16
埃/n:1 0-2
挨/v:2 0-2, 0-16
挨次/d:1 0-2
挨打/v:3 0-1, 0-2, 0-13
挨饿/v:2 0-2, 0-3

STEP3 : ソート形式を選択

ローマ字ソート 漢字ソート 辞書ソート

阿哥/n:1 0-1
阿公/n:1 0-33
阿鹤亭/ns:3 0-22, 0-32, 0-33
阿妈/n:2 0-33, 0-33
啊/e:4 0-3, 0-4, 0-5, 0-14
哀/a:2 0-2, 0-16
哀告/v:1 0-2
哀号/v:1 0-30
哀哭/v:2 0-13, 0-33
哀怜/v:2 0-2, 0-3

练习题

- 在线上下载比较长的文章试一下。
- Simple data for practice
- 1900前后 《北京官话全编》 1-10章
- 2018 《START DASH》 1-14课的课文

- INDEXCONVERTER是利用JavaScript开发的系统，所以完全依靠PC的能力。字多了需要相当长的分析时间。
- 我们推荐使用Firefox进行处理。

“正则表达式” “Regular Expression” 的文本检索

建设自己的语料库

网络上有很多汉语语料库，比如北大的CCL、北语的BCC等。但是有时语料库没有自己研究对象的文本，可能只好自己输入或收藏在自己的电脑里。今天体验一下TextEditor的检索方法。



准备工作

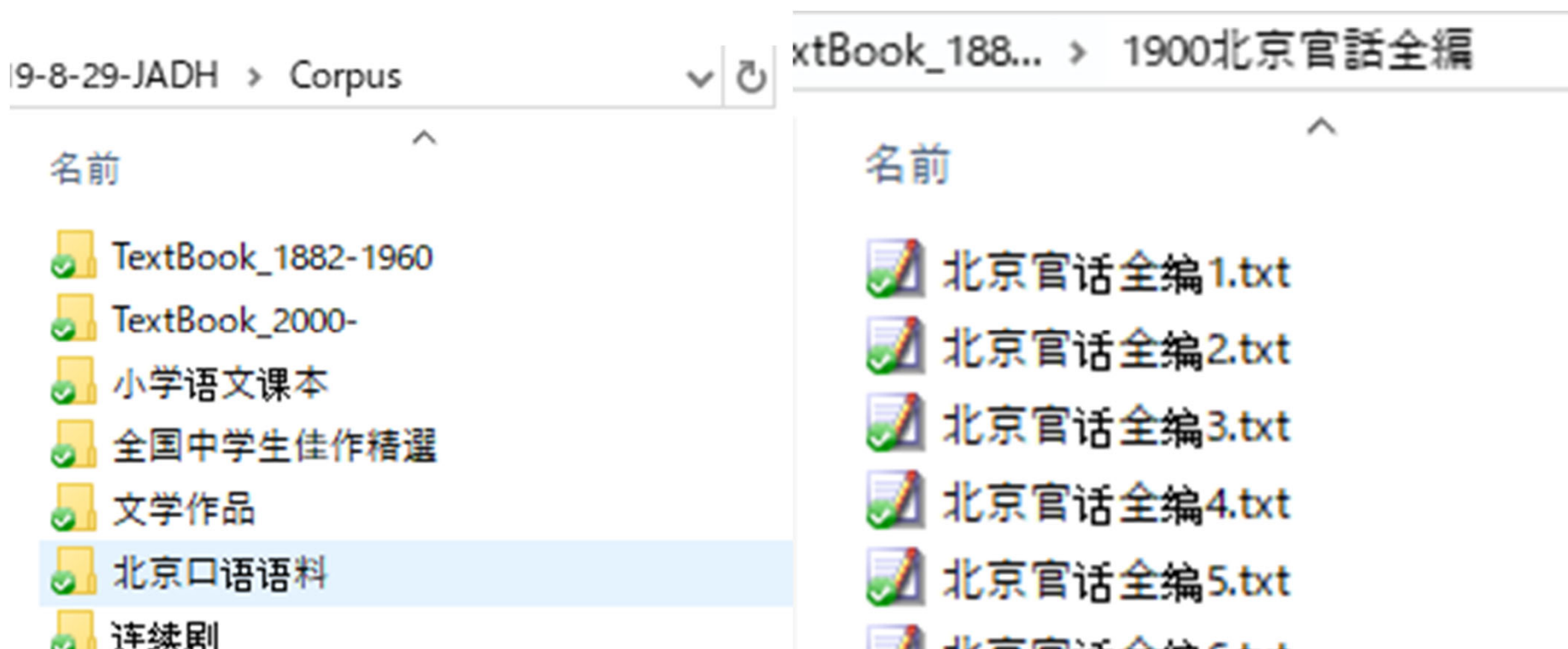
- TextEditor Atom <https://atom.io/>
- 收集与自己研究对象有关的文本

比如 小说/报纸/课文/连续剧的剧本等等。

- 今天我提供的sample数据是几本课本、老舍的小说以及19世纪的会话集（繁体字）的数据。请在hinox.org上下载。
- [Sample data for practice \(compressed file\)](#)

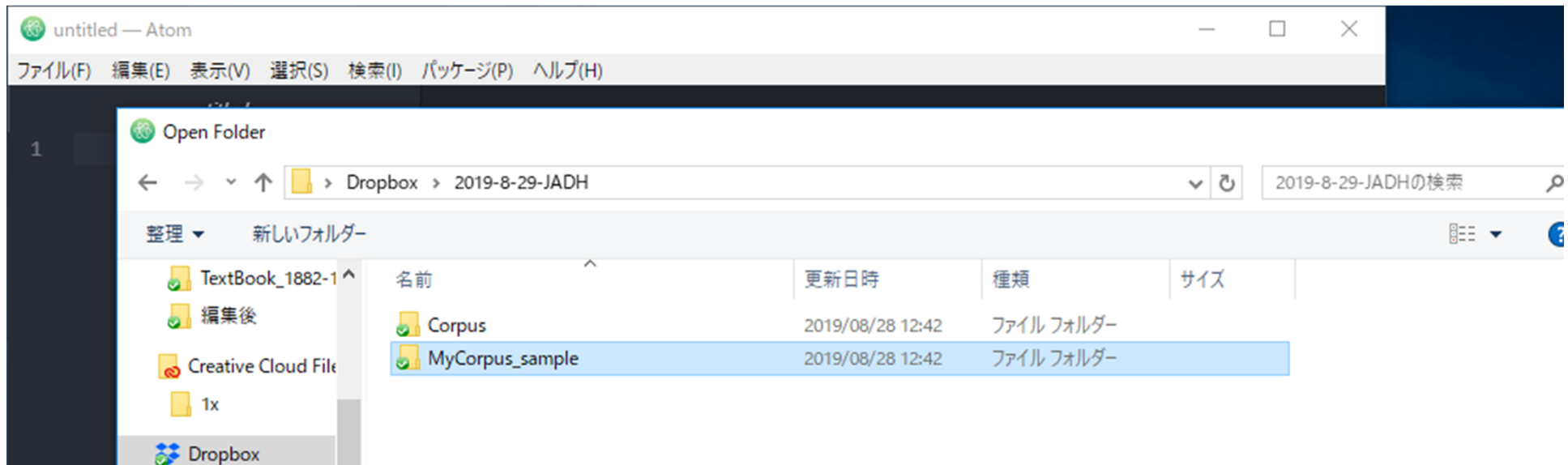
准备

下载、解压后放在电脑的任意地方。如果自己在网络上收集数据的话，在电脑上随便建个文件夹保存数据。



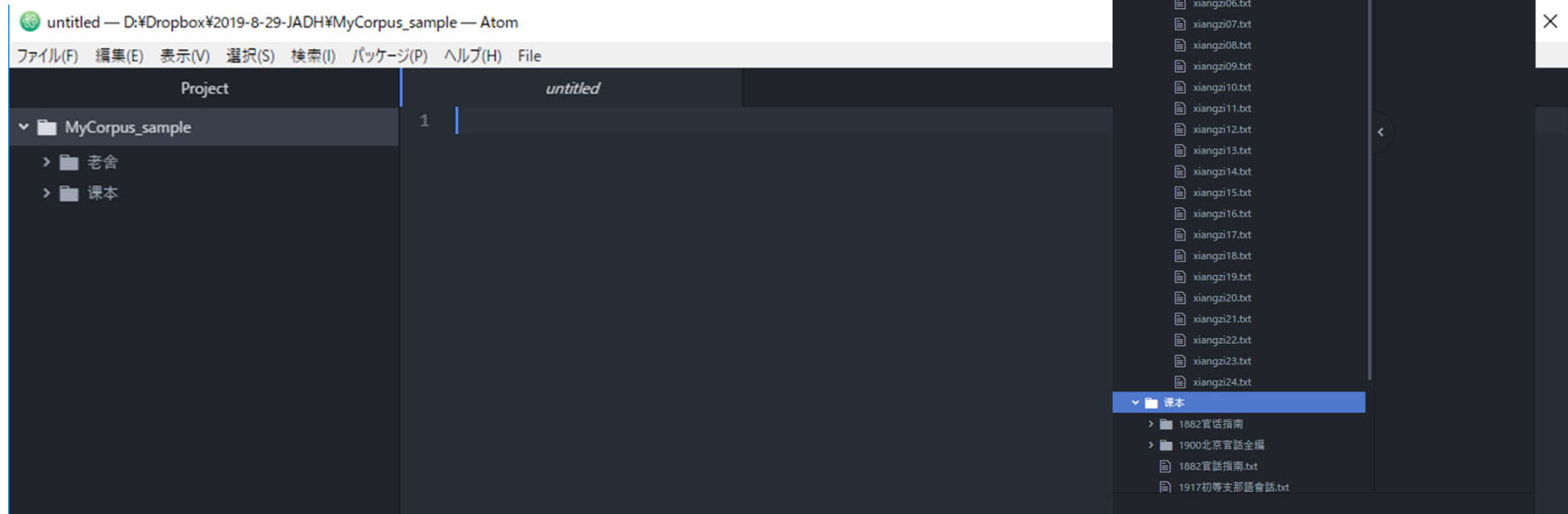
Project（文件夹）检索

Atom等Text Editor基本上都能用文件夹检索。
就是能检索文件夹里面的所有文件的功能。
按 CTL+SHIFT+A 加add project folder。
选择包含txt文件的文件夹。



Project（文件夹）检索

选择文件夹以后左边表现文件夹里面的文件和文件夹。



Project（文件夹）检索

按CTL+SHIFT+F开始检索。



检索结果

检索结果如下：



使用“正则表达式”检索

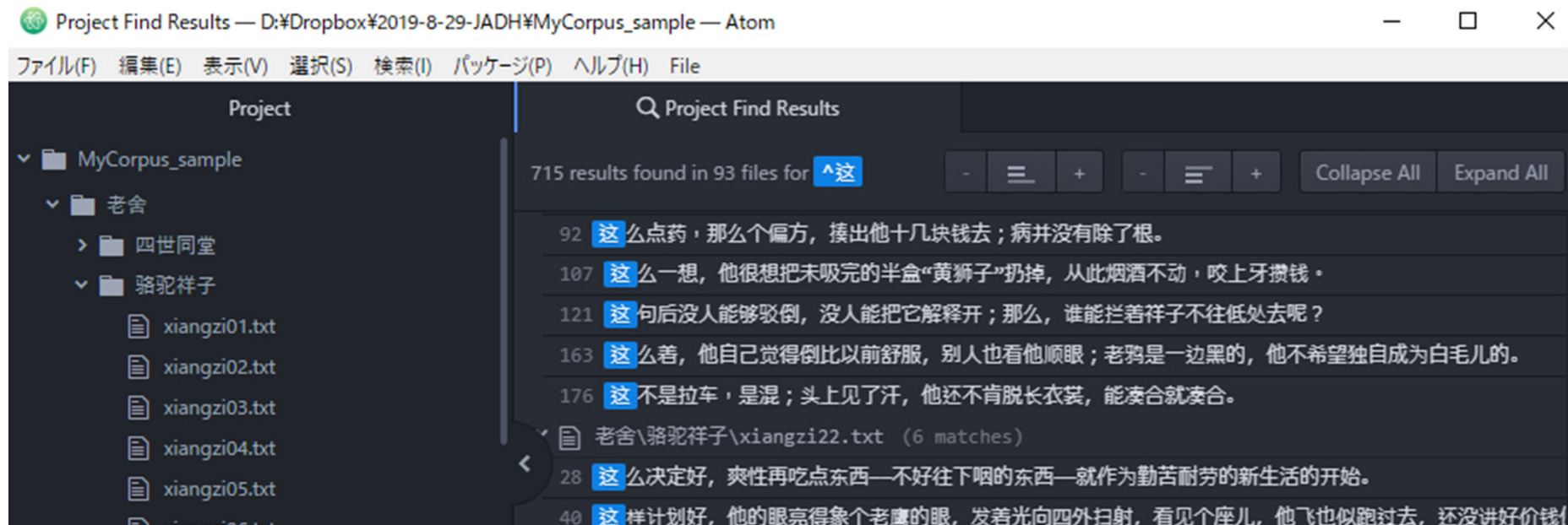
- 正则表达式检索是使用句号、星号等特殊文字来检索的方法。用这些文字能检索比如结婚 结过婚 结一次婚 结了婚等一般的检索方法不能检索的词组。
- 检索汉语词汇时可以发挥很大的功用。
- 下面介绍几个简单的“正则表达式”的文字。

^（克拉）

^对应文字列的前头和行的前头。

^ABC是指检索符合行头为ABC的文字列。

^对应行头。



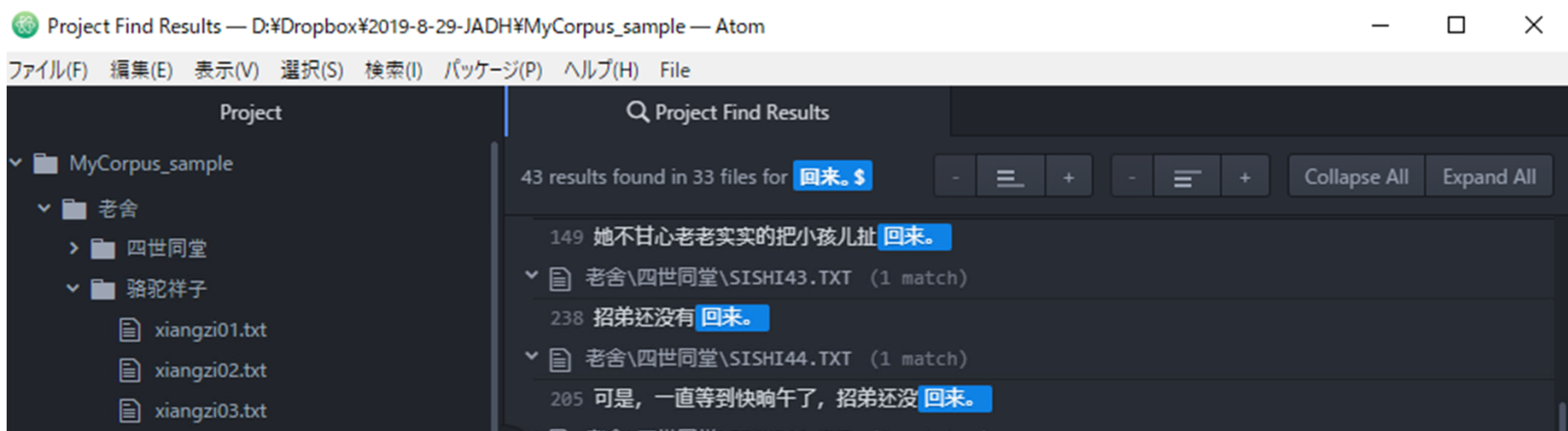
\$（美元记号）

\$对应文字列末或行尾。

ABC\$对应行尾是ABC的文字列。

“回来。\$”对应行尾是”回来。”的文字列

“。\$”对应段落的句点“。”（不对应段落以外的“。”）

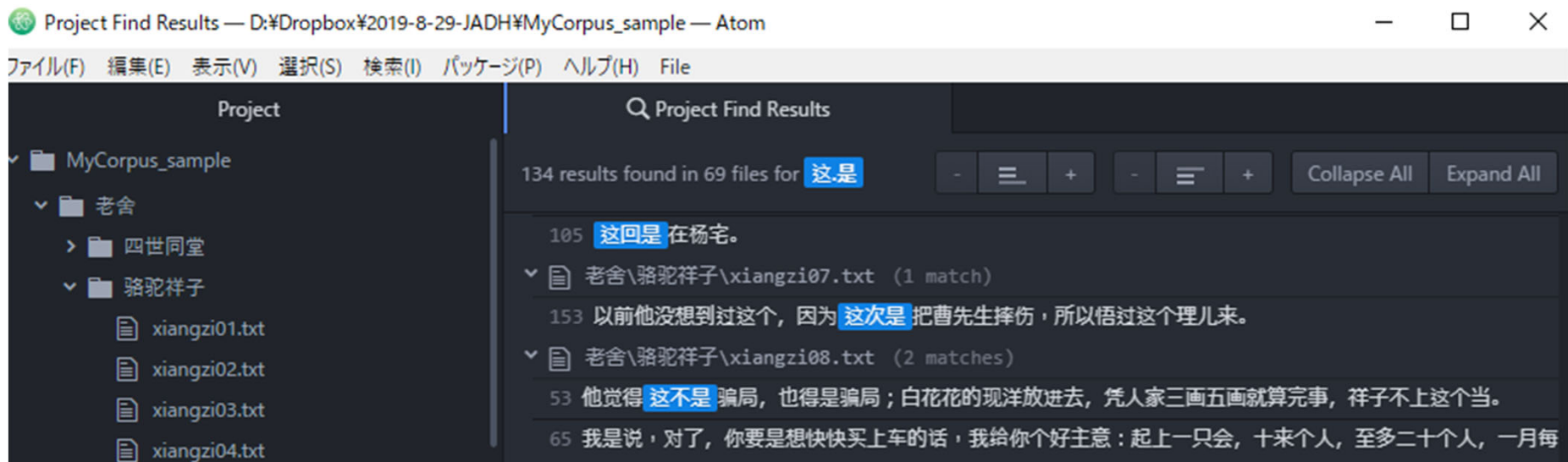


. (句号)

对应.和换行代码（n）以外的任意的1个文字。
三个句号比如...的话对应任意的三个文字。

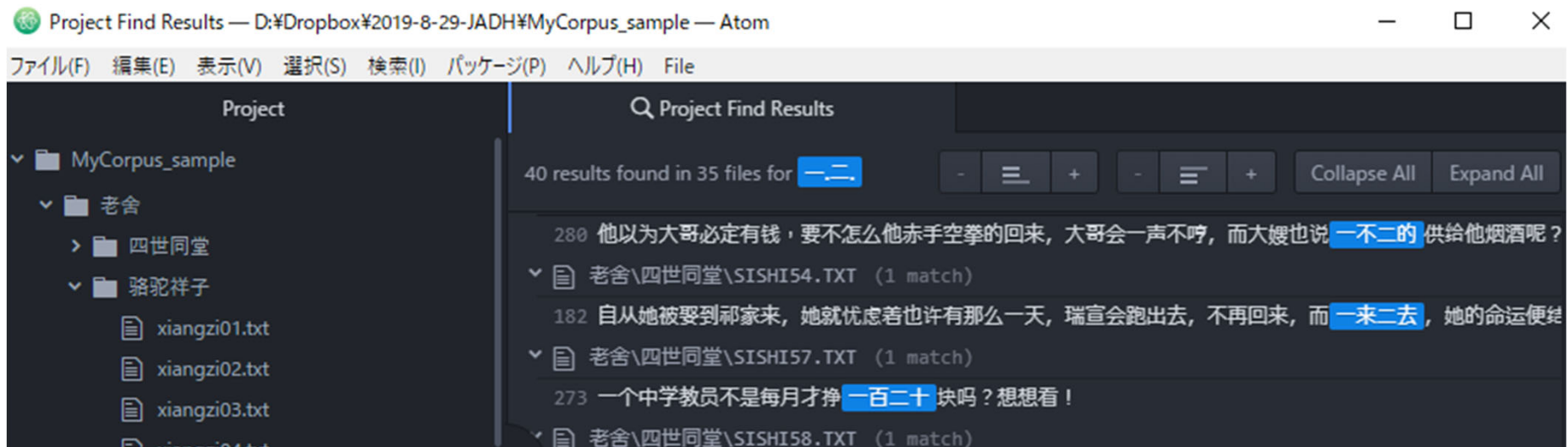
a.C对应abc、acc、adc...等

比如“这.是”对应“这不是，这也是，这都是，这还是，这个是，这本是”...等文字列。



. (句号)

查“离合词”有用。比如，“结.婚”、“生..气”对应“结了婚、结过婚”；“生他的气”、“生儿子气”等文字列。另外“一.二.”则对应“一干二净、一清二白、一穷二白”等文字列。一个句号能表示换行以外的任意的一个文字。还有像后述的，句号并用“*+?”等的记号能做到更有效率的检索。



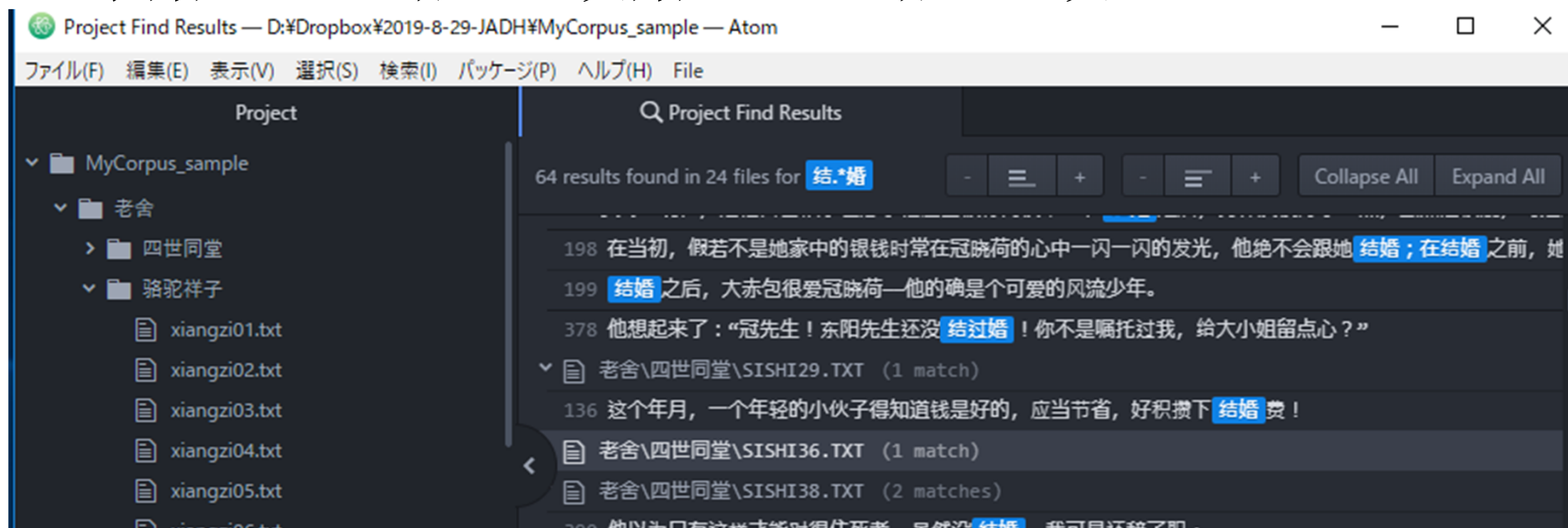
*（星号）

* 对应前面的一个文字（正则表达式）的0次以上重复。（包含0）

ab*c对应ac、abc、abbc、abbbc、...

. *对应 包含空文字列的任意文字列。

比如”结.*婚”对应”结婚”、”结了婚”、”结完了婚”、”结过一次婚”、”结过一次”



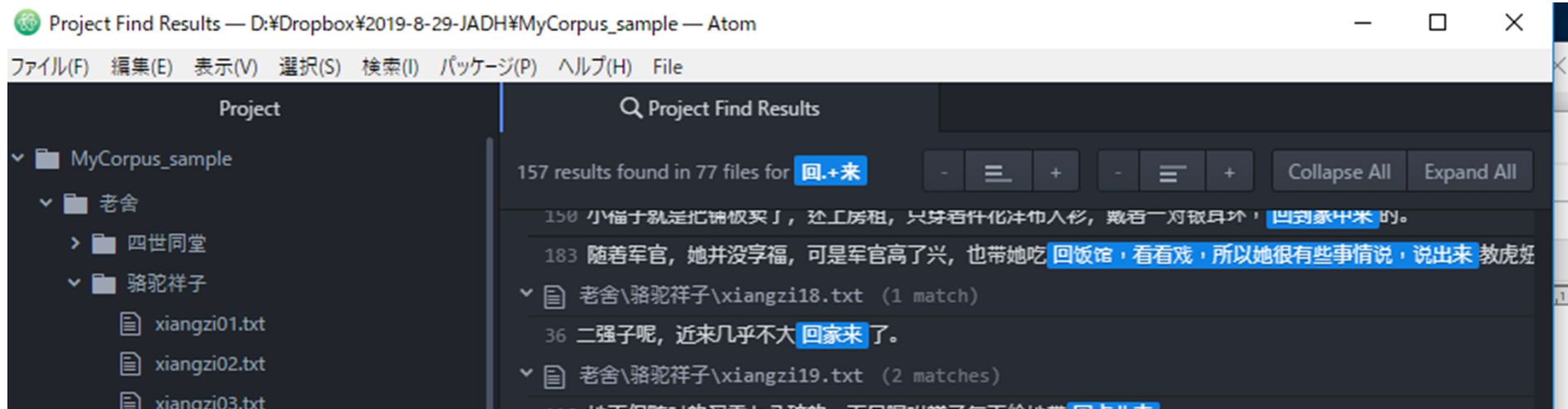
+（加号）

+对应前面的一个文字（或正则表达式）一次以上的重复（不含0次）

ab+c对应abc、abbc、abbbc、…（不对应ac）

.+对应任意的一个文字。

比如”回.+来”对应”回家来”、”回北京来”、”回到了阔别已久的故乡来”、…等文字列,而不对应“回来”。

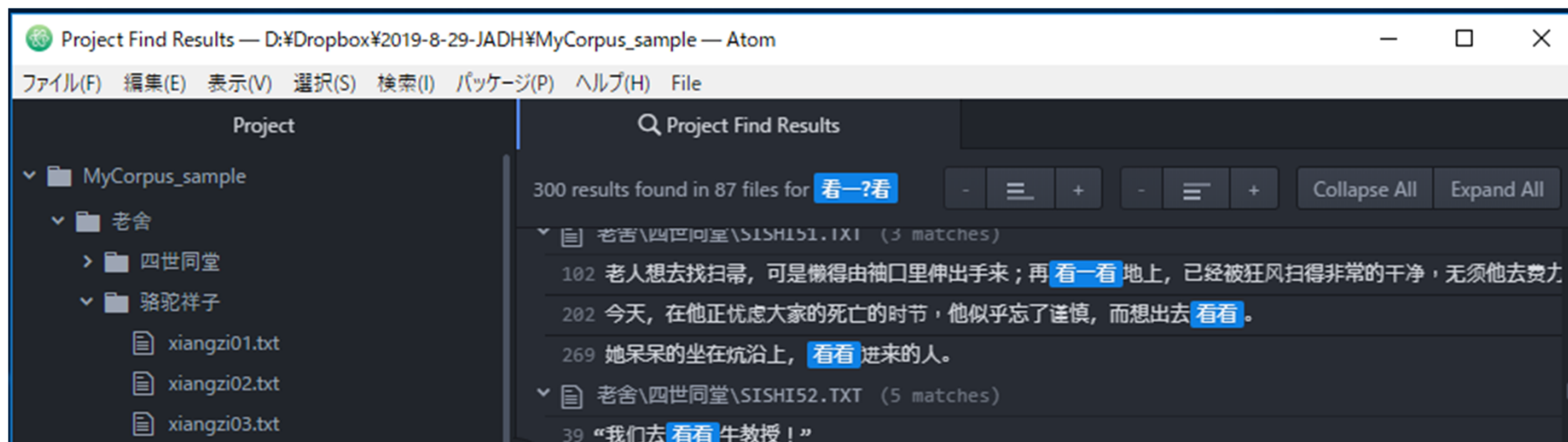


? (问号)

?表示正前面的一个文字（或正则表达式）的0次以上的出现。?被叫做重复的隐喻文字。其实不重复两次以上的。

ab?c对应ac、abc,

比如“看一?看”对应“看看”、“看一看”

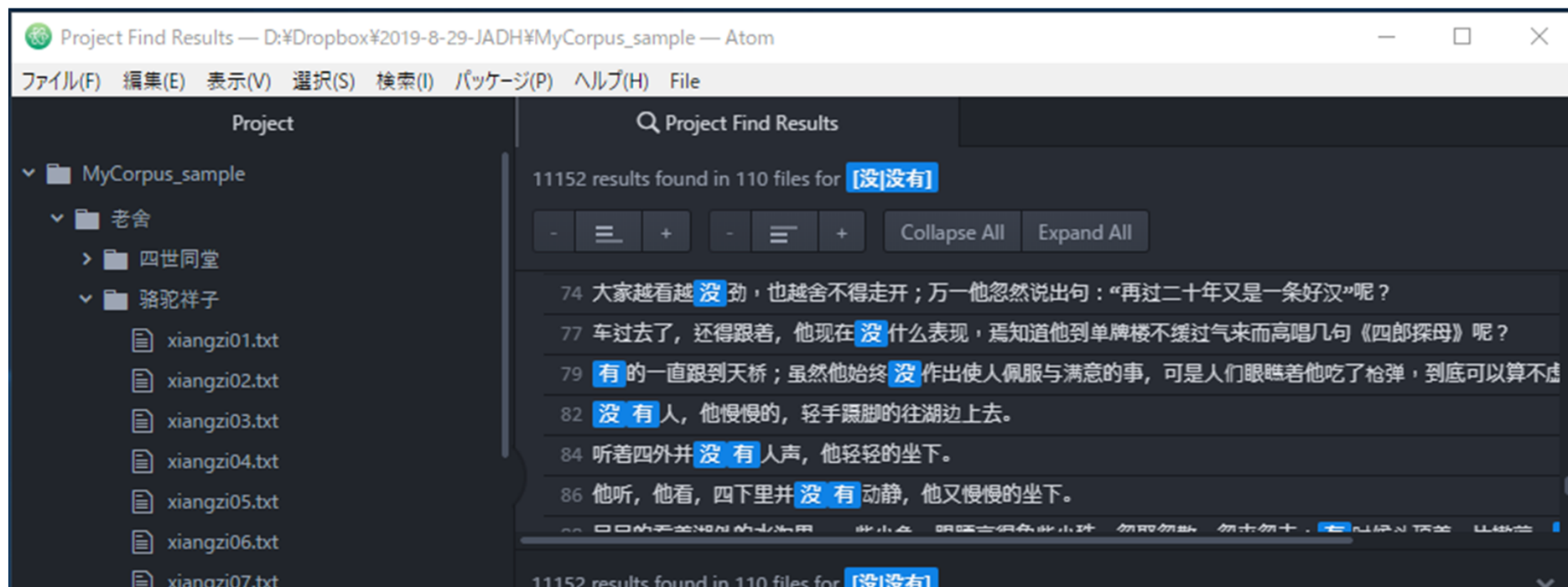


|（选择）

|表示选择文字列。

“没 | 没有” 对应 “‘没’ 或 ‘没有’ ”。

用[没 | 没有]的方式来检索。



[]（括号）

表示括号内的任意的一个文字。也可以指定范围。

[abcdef]对应“a～f”的任意1个文字

[あいうえお]对应「あ～お」的任意1个文字还有一种是用“-”指定范围。“-”拥有特殊意思。

[0-9] 表达任意一个数字

[A-Za-z] 表达任意一个英文字母

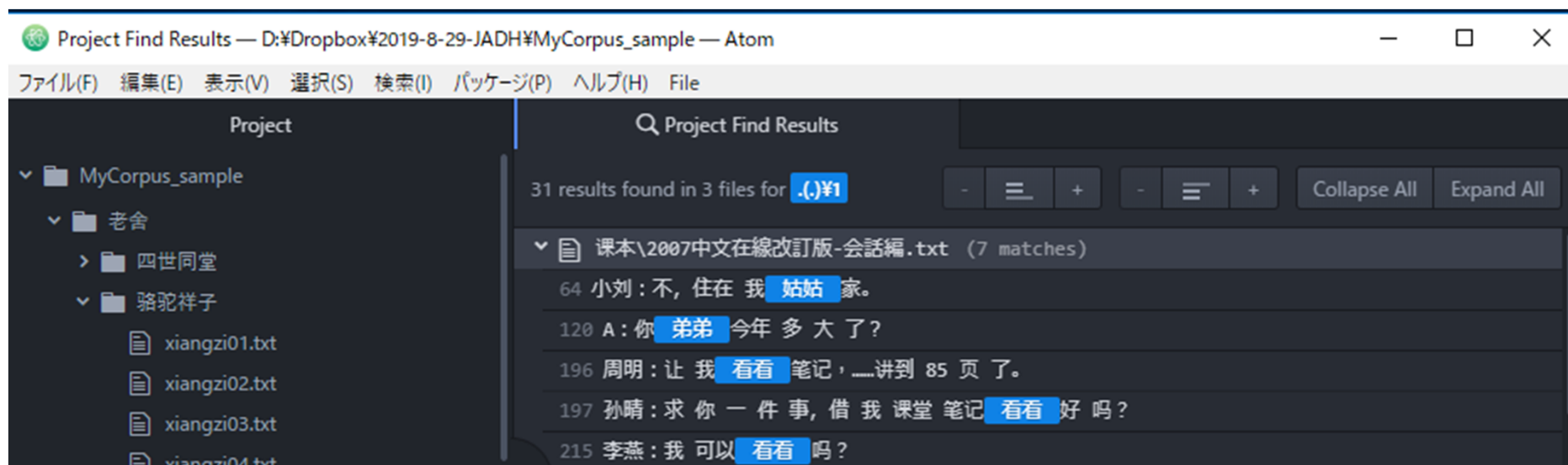
() (圆括弧)

() 有两种意思，一个是集合化的意思，比如李
(先生 | 同学 | 师傅) 对应李先生，李同学，李师傅
等。

(高兴) + 对应高兴高兴，高兴高兴高兴…等

还有一种用法称作“参照后方” (back reference)，
指定引用¥1～¥9的部分。数字表示对应第几个 ()。

a. (.)¥1 对应AA、BB、看看、说说、多多…等文字列



{ } (重复)

{ } 是指定某规定次数以上的重复隐喻文字。{n} 对应正前面的一个文字（或正则表达式）的n次以上的重复。{1,3} 对应正前面的一个文字（或正则表达式）一次到三次的重复。

[0-9]{5} 对应5位数的数字

a{1, 3} 对应a和aa、aaa

{min, max} 对应正前面的一个文字（或正则表达式）的min次~ma次的重复。如果不输入Min的数量则被解释为0次，同样不输入max的数量则被解释为∞次（无穷）

*和+、?、{min, max} 作为重复的模式表示最大次数的重复、正后面添加? 能停止在最小次数重复。

*?表示正前面的正则表达零次以上的重复（最小次数，就是说优先零次）

+?表示正前面的正则表达一次以上的重复（最小次数，就是说优先一次）

??表示正前面的正则表达的零次或一次以上的重复。（最小次数，就是说优先零次）

{min, max} ?对应正前面的正则表达的min次~max次的重复（最小次数）

检索汉语的技巧

- 单音节词的检索

- 量词的检索

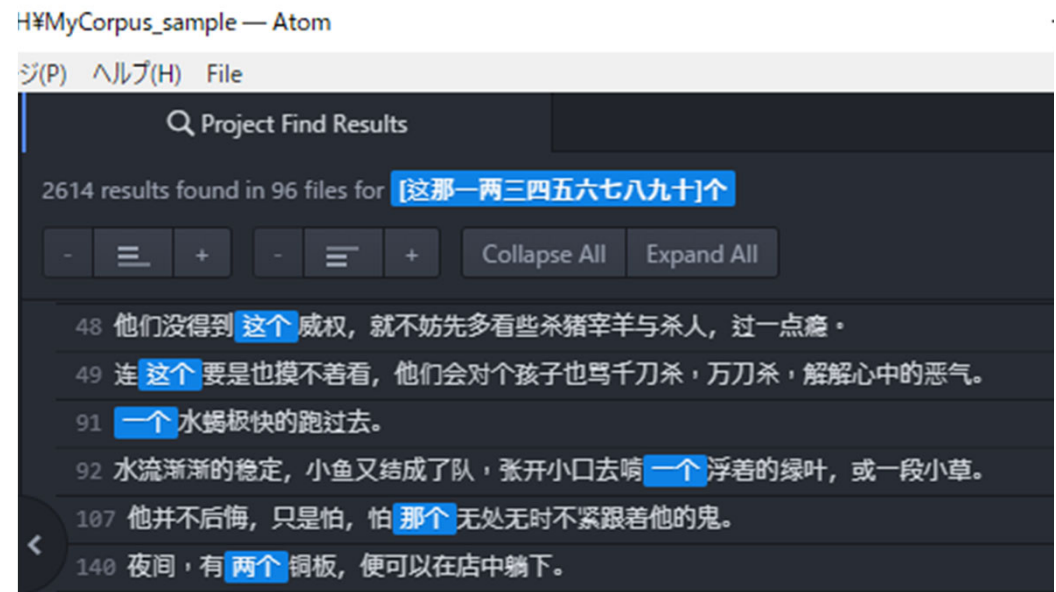
如果要检索使用“个”的名词怎么查好？

[这那一两三四五六七八九十]个

下面检索什么词汇？

“[很真太]大”

“这”[个次人回]是”



检索汉语的技巧

- 检索单音节动词的技巧

跟“着、了、过”等的助词一起检索
“看一看、尝了尝”

正则表达式	检索词
(.)一¥1	看一看, 尝了尝...
(.)了¥1	想了想, 说了说...
	是不是, 大不大...

检索汉语的技巧

离合词的检索

“结婚、生气、请客、洗澡”那样的形式，“结了婚、结过一次婚、生闷气、生孩子的气、请他客、请老王的客、洗完澡、洗一个热水澡”

检索汉语的技巧

使用例	正则表达式的意思	对应词
结.婚	出现任意的一个文字	结了婚，结过婚，结完婚...
结.?婚	此限零次或一次的任意的一个文字。	结婚，结了婚，结完婚...
结..婚	出现任意的两个文字	结不起婚，结不了婚...
结...婚	出现任意的三个文字	结了两次婚...
结.*婚	出现零到任意的文字列	结婚，结过婚，结了一次很草率的婚...
结.+婚	出现一次到任意的文字列	结了婚，结过三，四次婚...

检索汉语的技巧

不连续成分的检索

“因为…所以”，“虽然…但是”，“虽然…可是”，
“虽然…不过”

“因为.+所以” “是…的…” 句法

“是.{5,8}的(。|,)”

“是.{4,8}的.+ [,。?] ”

- 是.{4,8}的.+ [,。?]
- 他是昨天去的北京。
- 我说，“前天不是我们一起打的电报？”

检索汉语的技巧

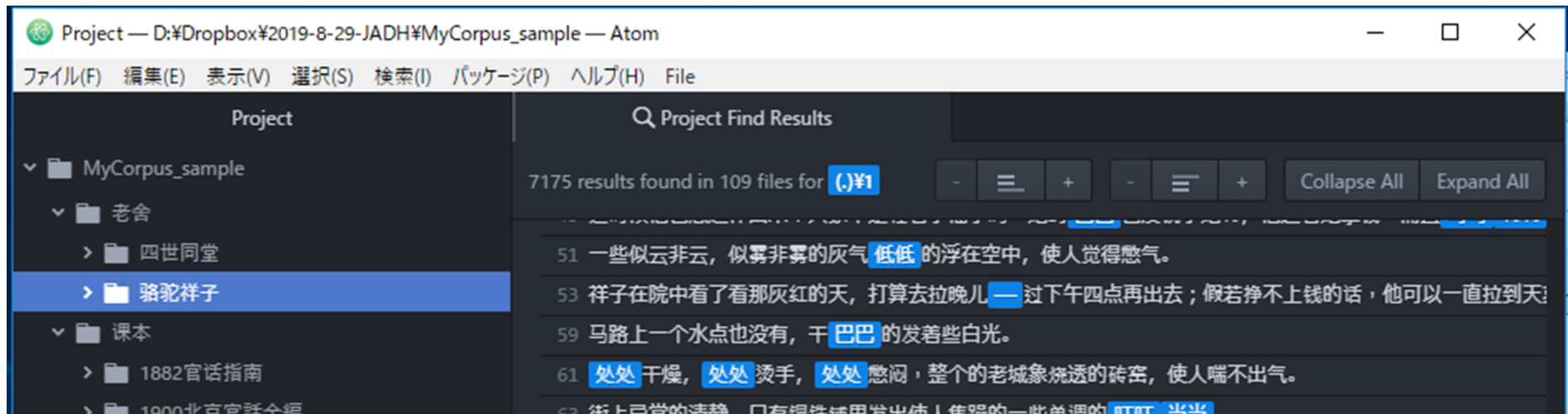
正则表达式	对应形式	例子
(.) ¥1	AA	茫茫, 尝尝, 说说
.(.) ¥1	ABB	白胖胖, 恶狠狠
(..) ¥1	ABAB	介绍介绍, 热闹热闹
(.) ¥1(.) ¥2	AABB	热热闹闹, 高高兴兴
(.)+ ¥1	A(B)A(B)	说说, 说明说明
(.)— ¥1	A—A	看一看, 写一写
(.)—? ¥1	A(—)A	看看, 看一看
(.)了 ¥1	A了A	尝了尝, 看了看
(.)了? ¥1	A(了)A	尝尝, 尝了尝
(..)了 ¥1	AB了AB	解释了解释
(.)不 ¥1	A不A	是不是, 来不来
(..)不 ¥1	AB不AB	可能不可能, 热情不热情
(.) ¥1.	AAB	开开心, 点点头
(.) ¥1[了的地]	AA[了的地]	谢谢了, 红红的, 狠狠地

检索汉语的技巧

重叠词的检索

“看看、说明说明；大大的、高高兴兴”

“(.)¥1”。这正则表达是符合再一次引用括号里面的，能找出AA式的重叠所检索结果，动词和形容词以外的不是重叠的也找能出来。



检索汉语的技巧

正则表达式	对应形式	例子
(.) ¥1	AA	茫茫, 尝尝, 说说
. (.) ¥1	ABB	白胖胖, 恶狠狠
(..) ¥1	ABAB	介绍介绍, 热闹热闹
(.) ¥1(.) ¥2	AABB	热热闹闹, 高高兴兴
(.) ¥1(.) ¥1	A(B)A(B)	说说, 说明说明
(.) ¥1—¥1	A—A	看一看, 写一写
(.) ¥1—? ¥1	A(—)A	看看, 看一看
(.) ¥1了¥1	A了A	尝了尝, 看了看
(.) ¥1了? ¥1	A(了)A	尝尝, 尝了尝
(..) ¥1了¥1	AB了AB	解释了解释
(.) ¥1不¥1	A不A	是不是, 来不来
(..) ¥1不¥1	AB不AB	可能不可能, 热情不热情
(.) ¥1.	AAB	开开心, 点点头
(.) ¥1[了的地]	AA[了的地]	谢谢了, 红红的, 狠狠地

谢谢！

- 今天的内容就到这里。以上内容都在网络上公开，有兴趣的话看一下。