dblp预处理

XML解析

格式问题

现有的同学是用已有的python的sax库对xml进行解析。这样解析的结果清晰，但是任然存在着一个问题就是，没有对格式进行检查。这里我使用了官方的dtd文件对8688326条数据的格式进行了检查，并且发现了一个问题。其实这也说明了dblp的可靠，800w条数据只有一条有着明显的格式错误。

如下，可以看见有两个年份

```
<dblp>
<phdthesis key="books/daglib/0079308" mdate="2019-07-19">
<author>Sai-Lai Lo</author>
<title>A modular and extensible network storage architecture.</title>
<school>University of Cambridge, UK</school>
<pages>I-XIX, 1-132</pages>
<year>1994</year>
<publisher>Cambridge University Press</publisher>
<year>1995</year>
<series>Distinguished dissertations in computer science</series>
<volume>11</volume>
<isbn>978-0-521-55115-1</isbn>
</phdthesis>
</dblp>
```

数据问题

- 人名的末尾有数字，这是因为有重名的存在

- 有拉丁字母的话，文字保存有问题

在前面的dtd文件中有将特殊的字符给描述出来，如下

```
        <!ENTITY Agrave  "&#192;" ><!-- capital A, grave accent -->
    <!ENTITY Aacute  "&#193;" ><!-- capital A, acute accent -->
    <!ENTITY Acirc   "&#194;" ><!-- capital A, circumflex accent -->
    <!ENTITY Atilde  "&#195;" ><!-- capital A, tilde -->
    <!ENTITY Auml    "&#196;" ><!-- capital A, dieresis or umlaut mark -->
    <!ENTITY Aring   "&#197;" ><!-- capital A, ring -->
    <!ENTITY AElig   "&#198;" ><!-- capital AE diphthong (ligature) -->
    <!ENTITY Ccedil  "&#199;" ><!-- capital C, cedilla -->
    <!ENTITY Egrave  "&#200;" ><!-- capital E, grave accent -->
    <!ENTITY Eacute  "&#201;" ><!-- capital E, acute accent -->
    <!ENTITY Ecirc   "&#202;" ><!-- capital E, circumflex accent -->
    <!ENTITY Euml    "&#203;" ><!-- capital E, dieresis or umlaut mark -->
    <!ENTITY Igrave  "&#204;" ><!-- capital I, grave accent -->
    <!ENTITY Iacute  "&#205;" ><!-- capital I, acute accent -->
    <!ENTITY Icirc   "&#206;" ><!-- capital I, circumflex accent -->
```

```
    <!ENTITY Iuml    "&#207;" ><!-- capital I, dieresis or umlaut mark -->
    <!ENTITY ETH     "&#208;" ><!-- capital Eth, Icelandic -->
    <!ENTITY Ntilde  "&#209;" ><!-- capital N, tilde -->
    <!ENTITY Ograve  "&#210;" ><!-- capital O, grave accent -->
    <!ENTITY Oacute  "&#211;" ><!-- capital O, acute accent -->
    <!ENTITY Ocirc   "&#212;" ><!-- capital O, circumflex accent -->
    <!ENTITY Otilde  "&#213;" ><!-- capital O, tilde -->
    <!ENTITY Ouml    "&#214;" ><!-- capital O, dieresis or umlaut mark -->
    <!ENTITY Oslash  "&#216;" ><!-- capital O, slash -->
    <!ENTITY Ugrave  "&#217;" ><!-- capital U, grave accent -->
    <!ENTITY Uacute  "&#218;" ><!-- capital U, acute accent -->
    <!ENTITY Ucirc   "&#219;" ><!-- capital U, circumflex accent -->
    <!ENTITY Uuml    "&#220;" ><!-- capital U, dieresis or umlaut mark -->
    <!ENTITY Yacute  "&#221;" ><!-- capital Y, acute accent -->
    <!ENTITY THORN   "&#222;" ><!-- capital THORN, Icelandic -->
    <!ENTITY szlig   "&#223;" ><!-- small sharp s, German (sz ligature) -->
    <!ENTITY agrave  "&#224;" ><!-- small a, grave accent -->
    <!ENTITY aacute  "&#225;" ><!-- small a, acute accent -->
    <!ENTITY acirc   "&#226;" ><!-- small a, circumflex accent -->
    <!ENTITY atilde  "&#227;" ><!-- small a, tilde -->
    <!ENTITY auml    "&#228;" ><!-- small a, dieresis or umlaut mark -->
    <!ENTITY aring   "&#229;" ><!-- small a, ring -->
    <!ENTITY aelig   "&#230;" ><!-- small ae diphthong (ligature) -->
    <!ENTITY ccedil  "&#231;" ><!-- small c, cedilla -->
    <!ENTITY egrave  "&#232;" ><!-- small e, grave accent -->
    <!ENTITY eacute  "&#233;" ><!-- small e, acute accent -->
    <!ENTITY ecirc   "&#234;" ><!-- small e, circumflex accent -->
    <!ENTITY euml    "&#235;" ><!-- small e, dieresis or umlaut mark -->
    <!ENTITY igrave  "&#236;" ><!-- small i, grave accent -->
    <!ENTITY iacute  "&#237;" ><!-- small i, acute accent -->
    <!ENTITY icirc   "&#238;" ><!-- small i, circumflex accent -->
    <!ENTITY iuml    "&#239;" ><!-- small i, dieresis or umlaut mark -->
    <!ENTITY eth     "&#240;" ><!-- small eth, Icelandic -->
    <!ENTITY ntilde  "&#241;" ><!-- small n, tilde -->
    <!ENTITY ograve  "&#242;" ><!-- small o, grave accent -->
    <!ENTITY oacute  "&#243;" ><!-- small o, acute accent -->
    <!ENTITY ocirc   "&#244;" ><!-- small o, circumflex accent -->
    <!ENTITY otilde  "&#245;" ><!-- small o, tilde -->
    <!ENTITY ouml    "&#246;" ><!-- small o, dieresis or umlaut mark -->

    <!ENTITY oslash  "&#248;" ><!-- small o, slash -->
    <!ENTITY ugrave  "&#249;" ><!-- small u, grave accent -->
    <!ENTITY uacute  "&#250;" ><!-- small u, acute accent -->
    <!ENTITY ucirc   "&#251;" ><!-- small u, circumflex accent -->
    <!ENTITY uuml    "&#252;" ><!-- small u, dieresis or umlaut mark -->
    <!ENTITY yacute  "&#253;" ><!-- small y, acute accent -->
    <!ENTITY thorn   "&#254;" ><!-- small thorn, Icelandic -->
    <!ENTITY yuml    "&#255;" ><!-- small y, dieresis or umlaut mark -->
```

于是我们就可以建立翻译的规则，如下：

```
{
    "&Agrave;":"À",
    "&Aacute;":"Á",
    "&Acirc;":"Â",
    "&Atilde;":"Ã",
    "&Auml;":"Ä",
```

```
    "&Aring;":"Å",
    "&AElig;":"Æ",
    "&Ccedil;":"Ç",
    "&Egrave;":"È",
    "&Eacute;":"É",
    "&Ecirc;":"Ê",
    "&Euml;":"Ë",
    "&Igrave;":"Ì",
    "&Iacute;":"Í",
    "&Icirc;":"Î",
    "&Iuml;":"Ï",
    "&ETH;":"Ð",
    "&Ntilde;":"Ñ",
    "&Ograve;":"Ò",
    "&Oacute;":"Ó",
    "&Ocirc;":"Ô",
    "&Otilde;":"Õ",
    "&Ouml;":"Ö",
    "&Oslash;":"Ø",
    "&Ugrave;":"Ù",
    "&Uacute;":"Ú",
    "&Ucirc;":"Û",
    "&Uuml;":"Ü",
    "&Yacute;":"Ý",
    "&THORN;":"Þ",
    "&szlig;":"ß",
    "&agrave;":"à",
    "&aacute;":"á",
    "&acirc;":"â",
    "&atilde;":"ã",
    "&auml;":"ä",
    "&aring;":"å",
    "&aelig;":"æ",
    "&ccedil;":"ç",
    "&egrave;":"è",
    "&eacute;":"é",
    "&ecirc;":"ê",
    "&euml;":"ë",
    "&igrave;":"ì",
    "&iacute;":"í",
    "&icirc;":"î",
    "&iuml;":"ï",
    "&eth;":"ð",
    "&ntilde;":"ñ",
    "&ograve;":"ò",
    "&oacute;":"ó",
    "&ocirc;":"ô",
    "&otilde;":"õ",
    "&ouml;":"ö",
    "&oslash;":"ø",
    "&ugrave;":"ù",
    "&uacute;":"ú",
    "&ucirc;":"û",
    "&uuml;":"ü",
    "&yacute;":"ý",
    "&thorn;":"þ",
    "&yuml;":"ÿ"
}
```

最后保存的使用utf-16。

数据压缩

我们首先注意到一个现象就是，如果我们去寻求一个人的和其他人的合作次数超过50次以上的话，就需要这个人首先在我们的数据中出现超过50次。所以根据这个现象我们就可以对我们的数据进行压缩，首先统计一遍超过50次的人名，把小于50次的人名进行删除就可以了。最后我们将100w次数据压缩为了32573条。