对从 Web 中抽取出来的列表进行实体链接

第一步：生成候选实体集合

构建 Dictionary（<key, key.value>集合）。4 种资源用于构建 Dictionary：

（1） Entity page：key 为 page 的 title，key.value 为该页面所描述的实体；

（2） Redirect page：key 为 page 的 title，key.value 为该页面中 key 所指向的具体页面的 entity；

（3） Disambiguation page：key 为 page 的 title，key.value 为该页面中所列出来的 entities；

（4） Hyperlink in Wikipedia article：key 为超链接的瞄文本，key.value 为该超链接所指向的 entity。

有了 Dictionary 就可以直接查询 Dictionary 以获得某一字符串对应的候选实体。

第二步：实体消歧

链接质量评估：两方面知识

（1） Prior probability：实体的流行度，通过指向该实体的 link 数来表示。计算公式如下：

$$P_{pr}(r_{i,j}) = \frac{count(r_{i,j})}{\sum_{u=1}^{|R_i|} count(r_{i,u})}$$

其中$|R_i|$表示列表中第 i 项的候选实体个数。对于不同的候选实体，其流行度是不一样的。

（2） Coherence: 候选实体间的语义一致性。表示同一列中候选实体中语义相似性（semanticlly similar）。计算公式如如下：

$$Coh(r_{i,j}) = \frac{1}{|L|-1} \sum_{u=1,u\neq i}^{|L|} Sim(r_{i,j}, m_u)$$

其中|L|表示列表中的 item 数目（行数）。

Sim（a,b）的值由两部分组成：

$$Sim_{hr}(r_{i,j}, m_u)$$

和

$$Sim_{ds}(r_{i,j}, m_u)$$

其中 Sim_{hr} 表示 type hierarchy based similarity，Sim_{ds} 表示 distributional context similarity。

1） Sim_{hr} 的计算

计算两个类别 t1 和 t2 的 hierarchy based similarity 的公式为：

$$Sim_{hr}(t_1, t_2) = \frac{2 \times log(P(t_0))}{log(P(t_1)) + log(P(t_2))}$$

其中 P（t）为随机选择一个实体 e，该实体 e 属于类别 t 的概率值。P（t）的计算在文献“An information-theoretic definition of similarity”中进行了介绍。

因此，两个候选实体的 type 集合的相似度计算如下所示：

$$Sim_{hr}(T(e_1) \rightarrow T(e_2)) = \frac{\sum_{t_1 \in T(e_1)} Sim_{hr}(t_1, \varepsilon(t_1))}{|T(e_1)|}$$

其中ε(t1)表示在实体 e2 的 type 集合中，使得值 Sim_{hr}（t1，t2）最大的 type t2。

最终，两个候选实体 e1 和 e2 的 hierarchy based similarity 的计算公式为：

$$Sim_{hr}(e_1, e_2) = \frac{Sim_{hr}(T(e_1) \rightarrow T(e_2)) + Sim_{hr}(T(e_2) \rightarrow T(e_1))}{2}$$

2） Sim_{ds} 的计算

Sim_{ds} 表示的是两个候选实体的上下文语义一致性。

候选实体的上下文的获得：收集维基百科页面中实体 e 的上下文。e 的上下文为 e 出现的位置的前后若干个词（停用词除外），综合实体 e 的所有上下文（实体 e 可能出现在不同的页面中，或者同一页面出现多次），并记录其上下文单词出现的次数。

Sim_{ds} 的计算公式为：

$$Sim_{ds}(e_1, e_2) = \frac{\sum_{i=1}^{g} a_i * b_i}{\sqrt{\sum_{i=1}^{g} a_i^2} * \sqrt{\sum_{i=1}^{g} b_i^2}}$$

其中，a_i，b_i 表示实体 e1 和 e2 的上下文中某一个词出现的次数。g 表示两个候选实体 e1 和 e2 的上下文的交集的词的个数。

链接质量可由如下公式表示：

$$LQ(r_{i,j}) = \alpha * P_{pr}(r_{i,j}) + \beta * \frac{1}{|L| - 1} \sum_{u=1, u \neq i}^{|L|} Sim_{hr}(r_{i,j}, m_u)$$

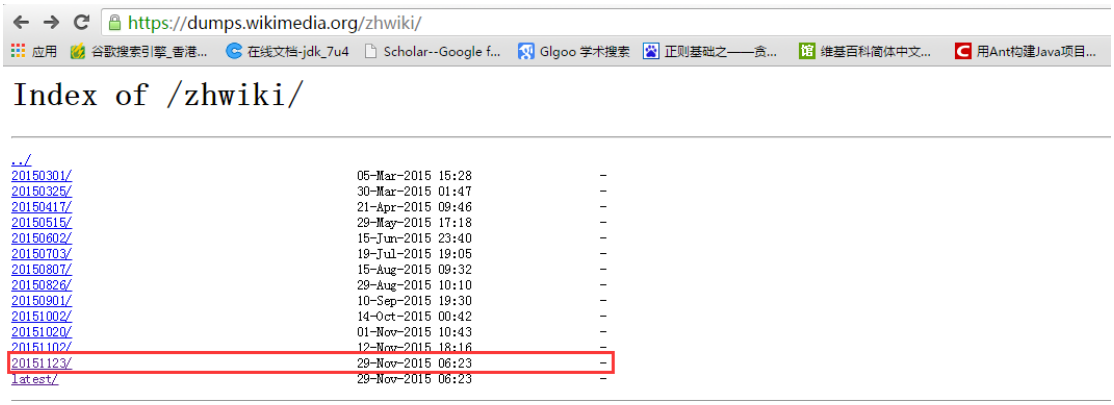
$$+\gamma * \frac{1}{|L| - 1} \sum_{u=1, u \neq i}^{|L|} Sim_{ds}(r_{i,j}, m_u)$$

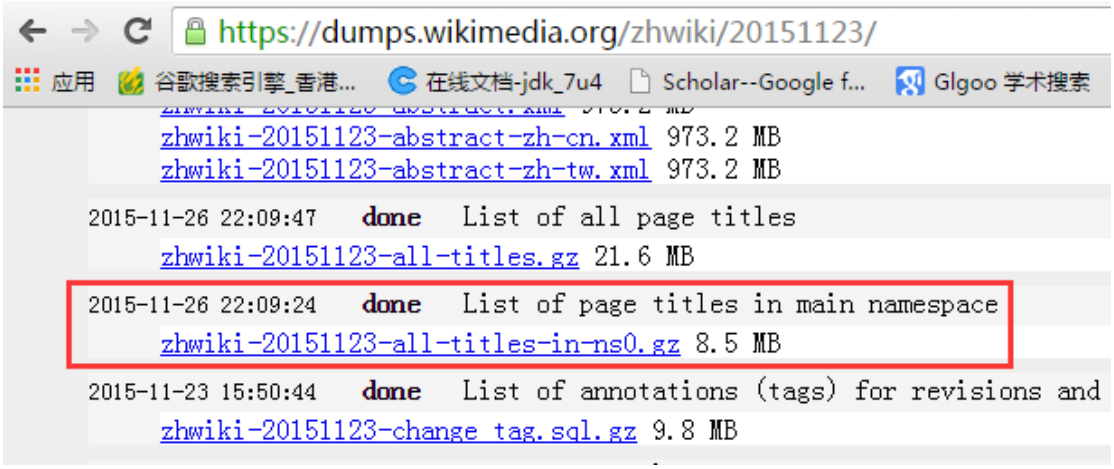
其中三个和项前面的参数为权重，三个权重之和为 1。实体消歧的最终结果就是使得链接质量取得最大值。

权重的学习：通过 max-margin technique 来得到。具体形式如下：

$$\forall l_i, \forall r_{i,j} \neq m_i \in R_i : LQ(m_i) - LQ(r_{i,j}) \geq \xi_{l_i}$$

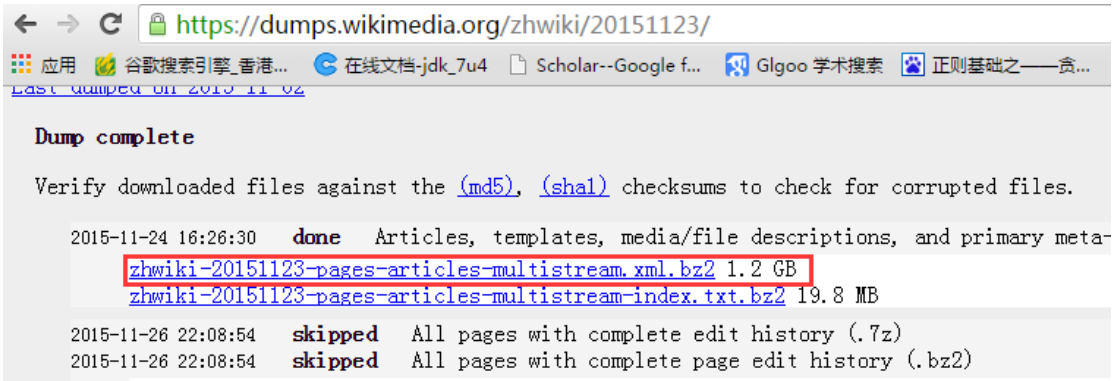
Dump 文件的下载地址：https://dumps.wikimedia.org/zhwiki/



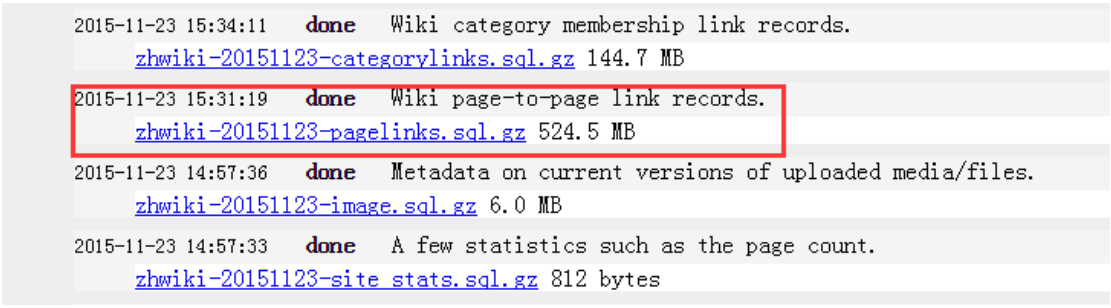


Entity page 信息

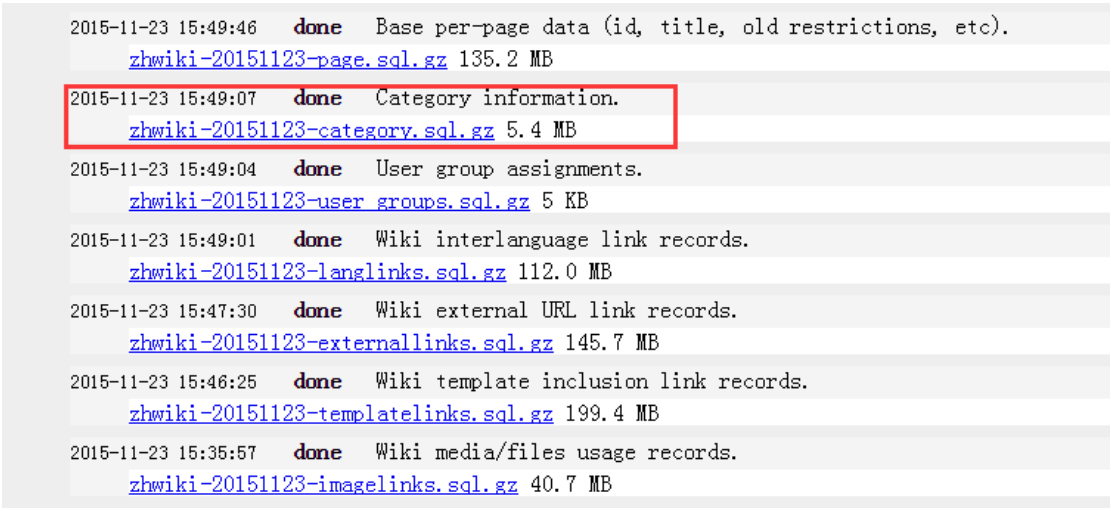
Redirect page 信息



Hyperlink 信息以及实体上下文信息



Hyperlink 信息？



实体的 category（type）信息

http://licstar.net/archives/262

entityandname.txt

```
569358 [[黑沙环|黑沙环]]
569359 [[工业街|工业街]]
569360 [[中国大陆|中国大陆]]
569361 [[北京|北京]]
569362 [[百货公司|百货公司]]
569363 [[维兹亚州立大学|维兹亚州立大学]]
569364 [[中华人民共和国教育部|中华人民共和国教育部]]
569365 [[中华人民共和国教育部|教育部]]
569366 [[北京|北京]]
569367 [[清华大学|清华大学]]
569368 [[上海|上海]]
569369 [[上海交通大学|上海交大]]
569370 [[湖北|湖北]]
569371 [[武汉大学|武汉大学]]
```

entityNameDistribution.txt

```
707606 2008青宁心曲:2008青宁心曲 (1)
707607 三岛站:三岛 (2)
707608 马格南:马格南 (2)
707609 李昭皇:李昭皇（李佛金） (1) 昭圣皇后 (1) 昭圣公主 (1) 李昭皇 (23) 李佛金 (2)
707610 马桥乡:马桥乡 (4)
707611 乐华邨:乐华北邨 (1) 乐华南邨 (1) 乐华 (2) 乐华邨 (36)
707612 蕴着:蕴着 (2)
707613 横滨新都市交通:横滨新都市交通 (1)
707614 平潭镇:平潭镇 (3)
707615 乌尔班六世:乌尔班六世 (6)
707616 福特冰原岛峰:福特冰原岛峰 (1)
707617 北海舰队:北海舰队 (77)
707618 李景源:李景源 (3)
707619 加姆萨尔:加姆萨尔 (1)
```

wikidump0after.txt

```
354802 江苏　海峡&&、&&之后&&在&&中国&&广东&&及&&福建&&沿岸&&登陆&&；&&亦&&可能&&经过&&
354803 台风菲特 2013年　江&&]&&]&&、&&江苏&&一带&&沿海&&登陆&&。&&这种&&路径&&的&&台风&
354804 关岛　热带&&气旋&&在&&菲律宾&&以东&&或&&附近&&形成&&后&&，&&先&&向&&西北&&偏&&西
354805 朝鲜半岛　带&&气旋&&在&&较&&北&&、&&较&&西&&的&&地方&&转向&&，&&途径&&可以&&影响
354806 日本　，&&途径&&可以&&影响&&台湾&&、&&东海&&或&&中国&&大陆&&东部&&沿岸&&浙江&&、
354807 台风范斯高 2013年　岸&&浙江&&、&&江苏&&等&&地&&，&&转向&&后&&可以&&影响&&朝鲜&&半
354808 藤原效应　部&&份&&热带&&气旋&&因&&外围&&引导&&气流&&不&&明&&，&&或&&受&&其他&&天
354809 季候风　部&&份&&热带&&气旋&&因&&外围&&引导&&气流&&不&&明&&，&&或&&受&&其他&&天气
354810 台风天秤 2012年　影响&&如&&藤原&&效应&&或&&季&&候&&风&&，&&路径&&出现&&打转&&、&
354811 台风罗莎 2013年　]&&，&&路径&&出现&&打转&&、&&停滞&&等&&，&&例如&&2012年&&的&&台
354812 热带气旋　台风&&预警&&是&&盛行&&地区&&，&&于&&风暴&&可能&&侵袭&&期间&&，&&由&&各
354813 热带　热带&&气旋&&是&&发生&&在&&、&&亚热带&&地区&&海面&&上&&的&&气旋&&性&&环流&&
354814 亚热带　热带&&气旋&&是&&发生&&在&&热带&&、&&地区&&海面&&上&&的&&气旋&&性&&环流&&
354815 气旋　热带&&气旋&&是&&发生&&在&&热带&&、&&亚热带&&地区&&海面&&上&&的&&性&&环流&&
```