汉语分词：最大匹配方法
(4学时)

陈文亮

2017年9

UTF-8编码

- UTF-8是不定长的，根据左侧位1的个数来决定占用了几个字节，中文一般占2-4个字节

utf-8可以根据字的第一个字节移位推出长度的
0xxxxxxx占1个字节
110xxxxx 10xxxxxx占2个字节
1110xxxx 10xxxxxx 10xxxxxx占3个字节
11110xxx 10xxxxxx 10xxxxxx 10xxxxxx占4个字节

1.分词任务

中文分词的目的是将汉字序列切分为词序列

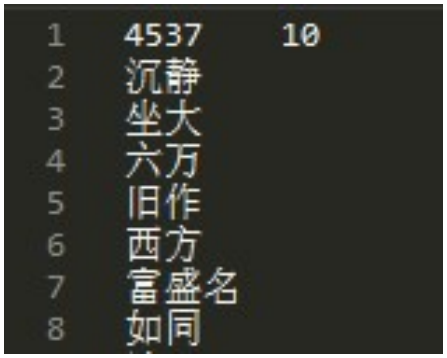
举例说明：
输入句子：他是研究生物化学的。

可能的分词：他 是 研究生 物化 学 的 。
他 是 研究生 物 化学 的 。
他 是 研究 生物 化学 的 。
合理答案：他 是 研究 生物 化学 的 。

2.最大匹配算法

从左到右寻找词的最大匹配（每次都从字典中贪心的找一个最长的词）

我们有一个**词典**，用于存放所有可能的词语，即除了单字，分词结果中的每个词均要在词典中出现。



文件的格式：

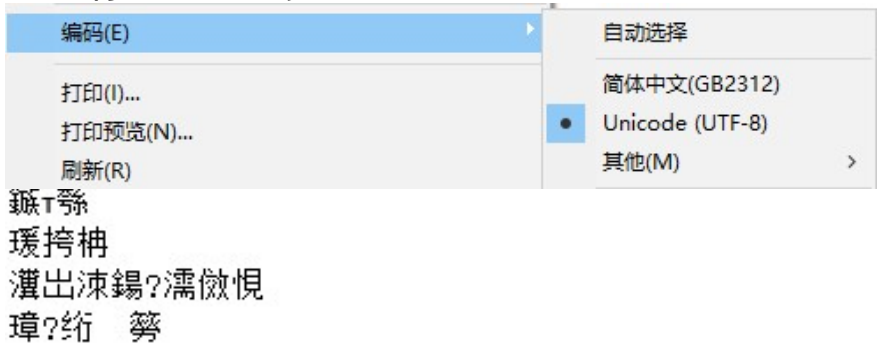
词总个数\t词的最大长度

word1

word2

…

如果你打开乱码……



UTF8文件记事本下显示乱码，可以在浏览器（或其他高级编辑器）中打开

2.最大匹配算法

从当前位置开始，向右截取最大长度，组成当前词；

和字典中的词逐一进行匹配；

若匹配成功，则进行下次匹配，下次匹配的当前位置则为这次词后面的那个字。

如果未能匹配，就缩短长度（长度减一）重新截取，直到当前词与词典中的词匹配或者当前词是单字；

2.最大匹配算法

- 举例：

  给定句子：我是中国人

  字典：中国、中国人

  指定：词的最大长度m=3

2.最大匹配算法

- 句子：我是中国人

第一轮：

第一次："我是中"是选取的词，在词典中未找到匹配项

第二次："我是"是选取的词，在词典中未找到匹配项

第三次："我"是选取的词，是单字，匹配成功

2.最大匹配算法

- 句子：我 是中国人

第二轮：

第一次："是中国"是选取的词，在词典中未找到匹配项

第二次："是中"是选取的词，在词典中未找到匹配项

第三次："是"是选取的词，是单字，匹配成功

2.最大匹配算法

- 句子：我 是 中国人

第三轮：

第一次："中国人"是选取的词，在词典中找到匹配项，匹配成功

至此，短句中所有字匹配结束，该短句分词结束。

3.分词算法评价

给定人工标注的分词答案，评价某一算法给出的结果。

戴相龙 说 中国 经济 发展 为 亚洲 作出 积极 贡献
新华社 福冈 5月 11日 电 （ 记者 乐绍延 ）

正确率(Precision) =正确识别的个体总数 / 识别出的个体总数

召回率(Recall) = 正确识别的个体总数 / 测试集中存在的个体总数

F值 = 正确率 * 召回率 * 2 / (正确率 + 召回率)

思考：评价程序应该怎么写？

3.分词算法评价

- 例子：
  - 句子：我是中国人
  - 分词：我是 中国人
  - 答案：我 是 中国人

  识别出的个体总数：2　　正确识别的个体总数：1

  测试集中存在的个体总数：3　正确识别的个体总数：1

  正确率(Precision) = 1/2 = 50.00%

  召回率(Recall) = 1/3 = 33.33%

  F值 = (1/2) * (1/3) * 2 / (1/2 + 1/3) = 40.00%

4.语料数据格式

corpus.sentence.txt格式

```
1  戴相龙说中国经济发展为亚洲作出积极贡献
2  新华社福冈５月１１日电（记者乐绍延）
3  中国人民银行行长戴相龙今天在亚洲开发银行第３０届年会的“亚洲未来３０年”研讨会上说，中国的经济发展为亚洲的繁荣与发展作出了积极贡献。
4  戴相龙在发言时说，中国的发展得益于亚洲国家和地区的经济发展与合作，与亚洲的繁荣息息相关。
5  他指出，随着经济的持续增长和改革开放政策的深入，中国将在亚洲经济区域合作中发挥更积极的作用。
6  中国经济的快速增长将为亚洲地区创造更多的贸易机会，在今后四年中，中国将为世界提供将近７０００亿美元的市场。
7  关于香港回归中国后的国际金融地位问题，戴相龙强调，香港的国际金融地位不但能够维持，而且还会得到加强。
8  在谈到亚洲经济的发展前景时，戴相龙认为，亚洲经济将继续保持稳定的发展势头，仍将成为推动世界经济发展的主导力量。
9  戴相龙同时指出，亚洲经济发展中还存在工资上涨过快削弱竞争力；高级研究、管理人才严重匮乏；能源、交通等基础设施相对落后等制约经济发展的因素，解决这些问题是亚洲经济发展的当务之急。
10 戴相龙认为，要保持亚洲地区经济增长，既需要亚洲各国继续开发利用自身的经济潜力，也需要进一步加强区域经济合作。
11 亚洲国家和地区今后除了在商品、投资领域加强合作外，还应在科技和环保以及货币政策和金融监管方面加强合作。
12 亚洲开发银行总裁佐藤光夫主持了这次研讨会。
13 日本前首相宫泽喜一、印度财政部长奇丹巴拉姆和芬兰环境部长佩卡·哈维斯托也在研讨会上发了言。
14 （完）
```

4.语料数据格式

corpus.answer.txt格式

```
1 戴相龙 说 中国 经济 发展 为 亚洲 作出 积极 贡献
2 新华社 福冈 5月 11日 电 （ 记者 乐绍延 ）
3 中国 人民 银行 行长 戴相龙 今天 在 亚洲 开发 银行 第30 届 年会 的“ 亚洲 未来 30 年 ”研讨会 上 说 ，
  中国 的 经济 发展 为 亚洲 的 繁荣 与 发展 作出 了 积极 贡献 。
4 戴相龙 在 发言 时 说 ， 中国 的 发展 得益于 亚洲 国家 和 地区 的 经济 发展 与 合作 ， 与 亚洲 的 繁荣
  息息相关 。
5 他 指出 ， 随着 经济 的 持续 增长 和 改革 开放 政策 的 深入 ， 中国 将 在 亚洲 经济 区域 合作 中 发挥 更
  积极 的 作用 。
6 中国 经济 的 快速 增长 将 为 亚洲 地区 创造 更 多 的 贸易 机会 ， 在 今后 四 年 中 ， 中国 将 为 世界 提供
  将近 7000亿 美元 的 市场 。
7 关于 香港 回归 中国 后 的 国际 金融 地位 问题 ， 戴相龙 强调 ， 香港 的 国际 金融 地位 不但 能够 维持 ，
  而且 还 会 得到 加强 。
8 在 谈到 亚洲 经济 的 发展 前景 时 ， 戴相龙 认为 ， 亚洲 经济 将 继续 保持 稳定 的 发展 势头 ， 仍 将 成为
  推动 世界 经济 发展 的 主导 力量 。
```

5.编程作业

- 要求：编程平台不限（windows、linux），编程语言不限（C、C++）
- 任务：使用最大匹配算法、字典文件（corpus.dict.txt），对语料（corpus.sentence.txt）进行分词

--将分词的结果输出到文件corpus.out.txt中；

--对比corpus.answer.txt和corpus.out.txt，给出算法的P/R/F指标

- 输出：一个corpus.out.txt文件（格式参照corpus.answer.txt）

  P/R/F指标(格式类似于：Precision = 36 / 100 = 36.00%)