编程作业：分词

wordSegment

任务描述

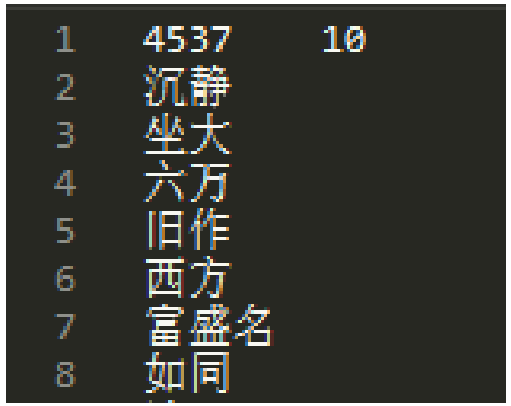
- 要求：编程平台不限（windows、linux），编程语言不限（C、C++，java，python）
- 任务：使用正向最大匹配算法、字典文件（corpus.dict），对语料（corpus.sentence）进行分词，将分词的结果输出到文件corpus.out中；对比corpus.answer和corpus.out，给出算法的P/R/F指标
- 输出：一个corpus.out文件（格式参照corpus.answer）

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

语料数据格式

语料文本都为utf8编码

（1）词典文件　corpus.dict格式



语料数据格式

（2）待分词文件corpus.sentence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）
```

语料数据格式

（3）给定的人工分词文件（即分词标准答案） corpus.answer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 。
```

加分项

（1）探索正向最大匹配和反向最大匹配哪个方法更适合中文分词？为什么（要实现反向最大匹配算法，通过实验数据来说明）？

（2）试学习使用Python第三方分词工具jieba进行分词，对比你实现的算法跟jieba分词的性能差异。

作业时间

- 6个课时