2017 年第二届全国网络舆情分析技术邀请赛

评测大纲

关键词抽取

一、 任务说明

给定一组新闻文档，从每一篇文档中抽取出与该文档主题最相关的一些词或者短语。参赛者需要设计一个关键词抽取系统，用于抽取每篇文档的关键词，提供的关键词数量不允许超过 10 个。

二、 数据集

1. 新闻训练集

新闻训练集包含 30,000 篇新闻报道，文本未经过分词处理。

2. 新闻训练标注集

新闻训练标注集包含新闻训练集中每篇新闻的对应关键词。

3. 新闻测试集

新闻测试集共包含 20,000 篇新闻报道供评测阶段使用，文本未经过分词处理。

三、 参评系统输入输出文件格式

1. 输入文件格式

训练数据及测试数据示例

```
<Text>
    <ID>1</ID>
    <Title>习近平:扎实把"十三五"发展蓝图变为现实</Title>
    <Content>新华社北京1月30日电 中共中央政治局1月29日下午就
             "十三五"时期我国经济社会发展的战略重点进行第三十次集体学习。
             中共中央总书记习近平在主持学习时强调，发展战略重点，
             是"十三五"时期我国发展的"衣领子"、"牛鼻子"。
             抓准、抓住、抓好战略重点，是保证"十三五"发展开好头、起好步的关键，
             是保证全面建成小康社会决胜阶段获得全胜的关键。
             要准确把握"十三五"时期我国发展的战略重点，做到胸中有数、落实有策、行动有策，
             以奋发有为的精神状态、攻坚克难的拼搏意志、只争朝夕的紧迫劲头，
             通过抓好发展战略重点带动发展全局，
             把"十三五"发展宏伟蓝图一步一步变为现实。
    </Content>
</Text>
```

训练标注示例

```
<Class>
    <ID>1</ID>
    <kewords>习近平 十三五 发展蓝图 经济发展</kewords>
</Class>
```

2. 输出文件格式

每组输出结果为一个扩展名为 txt 的文本文件，结果文件名称由参赛者自定，只需在 config 配置文件（**config 配置文件请参见 stokis for text analysis 使用说明文档**）中注明。结果文件中的每一行对应一篇新闻关键词的抽取结果，每行中不同的关键词使用空格隔开。

中微子 基本粒子 日本 获得 成为 宇宙 获奖 领域 质量
美国 数据 市场 显示 黄金 位置 企业 指数 指标
公平 社会 学额 中国 阶层 科举 高考制度 考试 导致
病毒 评估 免疫 美国 反应 能否 研制 试验 英国
大学 哲人 创业 运营 工作 公司 学校 开始 时间
菲律宾 峰会 领导人 南海 会议 国际 成为 工商 经济
时间 高考 理工农 征集 志愿 职业 招生 教育 对口
招标 竞购 产品 企业 合作 竞标 认购 现场 签约
加沙 冲突 国际 委员会 以色列 人选 导致 英国 基本
做好 滑坡 安徽 地区 湖南 防御 电源 广西 地带
车手 车队 复赛 辽宁 比赛 衡驰 赛道 代表 障碍
加入 指数 市场 决定 额度 接受 董事 港股 解决
抵御 伊斯兰 监测 长城 隔离墙 范围 还有 能够 配有
变化 化学 中考 题型 分析 试卷 分数 北京 计算
效率 家长 大学 排行榜 填报 高考 指导 教师 宁夏
访谈 复合 婚变 女方 出轨 事情 看到 女明星 婚生子
决定 理事会 董事会 设立 协定 权力 成员 行长 银行业务

四、 评价指标

关键词抽取评价采用类似于 MRR(mean reciprocal rank)的评价方法。增加了关键词之间相似度的评分因素。对于文档 i，假设人工标注的关键词词典大小是 n，评测公式定义如下：

$$\mathrm{MRR}_i = \frac{1}{n}\sum_{\mathrm{j}=1}^{n}\frac{m_j}{p_{\mathrm{j}}}$$

其中m_j是关键词词典中第 j 个关键词k_j与抽取结果中最相似的词p_j的匹配程度，即两个关键词的相似度分值，$m_j = \frac{|\mathrm{lcs}(\mathrm{k_j},\mathrm{p_j})|}{\max\ (|\mathrm{k_j}|,|\mathrm{p_j}|)}$，其中 lcs 是两个字符串的最长公共子序列，|•|是字符串长度；p_i是关键词词典中每一个词在抽取结果中的排序位置，对于不在抽取结果中的关键词，$\frac{1}{p_i}$=0。

为了对关键词抽取系统进行综合评测，我们采用 MRR 的均值作为排名依据。假设一共有 K 篇文档，则：

$$\mathrm{MRR} = \frac{\sum_{i\in K} MRR_i}{K}$$

五、 测试步骤

1、在训练阶段，参赛方远程登录客户机，进行系统部署与环境搭建，并获取训练数据进行调试训练。参赛方可调用大赛提供的代理程序 stokis 对模型进行测试验证，测试数据由参赛方从训练数据中自行选取，代理程序可以反馈测试结果。调用代理程序需要参赛方提供可执行程序并填写代理程序的 config 文件。其中，使用 windows 虚拟机的队伍训练程序和测试程序分别提供 xxx.bat 的执行文件，使用 linux 虚拟机的队伍训练程序和测试程序分别提供 xxx.sh 的执行文件。具体调用步骤请参见**《stokis for text analysis 使用说明文档》**。

2、评测阶段开始前，参赛方需在 config 文件中配置最终版程序的路径，由大赛平台自动进行评测，并将评测结果反馈给参赛队伍。

说明：

1、代理程序的功能。代理程序会自动完成以下操作：1）切断客户机的访问连接，2）下载测试数据，3）运行测试任务产生结果文件，4）上传结果文件到服务器，5）提交评测服务获得评测结果，6）将整个运行结果存入数据库，7）恢复客户机访问连接。

2、技术指标评分。代理程序自动运行文本分类评测工具，给出相关性能指标，测试结果形式如下：

新闻评测结果，MRR___；

六、 注意事项

代理程序开始运行后，参赛方与客户机断开连接，无法通过调试程序来优化结果。代理程序运行完毕后，参赛方可重新登录客户机，通过调试程序来优化结果。如果代理程序出现问题，请与主办方工作人员及时沟通解决。