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、 评测数据集

主办方提供：新闻调试集、新闻调试标注集、新闻训练集、新闻训练标注集、新闻测试集。

1. 新闻调试集

新闻调试集包含 100 篇左右新闻报道，每篇新闻报道有不同的关键词，用于调试参赛系统。

2. 新闻调试标注集

新闻调试标注集包含新闻调试集中每篇新闻对应的关键词列表。

3. 新闻训练集

新闻训练集包含 20,000 篇新闻报道，文本未经过分词处理。

4. 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。

5. 新闻测试集

新闻测试集共包含 30,000 篇新闻报道，文本未经过分词处理。新闻测试集分为 3 份，每一份测试数据的评测分值占最终评测分值的 10%、30%、60%。

三、 参评系统输入输出文件格式

1. 输入文件格式

```
<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>


<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。

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，

$|\cdot|$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、在调试阶段，参赛方远程登录客户机，进行系统部署与环境搭建，并获取调试数据进行调试。

2、在系统测评阶段，参赛方提供可执行程序。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。参赛方需要填写代理程序的 config 文件，利用代理程序 stokis 提交评测任务。比赛数据和测试步骤请参见**《stokis for text analysis 使用说明文档》**。

3、参赛方允许在规定时间范围内完成关键词抽取任务，包括 3 组新闻测试数据集。每组测试集可以提交任意次测试结果，但只有第一次结果的评测成绩作为最终排名依据。如果参赛方没有提交一组测试集的运行结果，则该组测试集的得分为 0。

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）下载训练数据，3）运行训练任务产生模型文件，4）下载测试数据，5）运行测试任务产生结果文件，6）上传结果文件到服务器，7）提交评测服务获得评测结果，8）将整个运行结果存入数据库，9）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：

新闻第 1 组结果，MRR___；

新闻第 2 组结果，MRR___；

新闻第 3 组结果，MRR___；

最终评测结果按照 1:3:6 的比例对三次测试 MRR 值进行加权平均。

六、 注意事项

1、参赛方需要在指定时间内完成全部 3 组测试集的运行测试。每组测试集可以提交多次测试任务，但只有第 1 次提交的测试任务的评测成绩作为排名依据。如果某一组测试集没有对应的输出结果，则该组测试集的评测成绩为 0。该项目最终得分为 3 组评测成绩的加权平均。

2、代理程序开始运行后，如果参赛程序中断，则视为一次失败的提交。参评程序执行过程中，应记录程序中断提示信息。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。

3、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。