文本检索作业要求

2021/4/20　　负责助教：郭效君、窦帅、陈翔、何语恬

在作业五中，只提供了 `李白`，`杜甫`，`白居易`，`王維` 四个作者的订阅选项。本次作业要求实现：

- 根据作者订阅
- 根据输入的关键词订阅，支持多关键词（1-3个）

作者订阅不限于李白、杜甫、白居易和王維四位诗人，数据库中包含的诗人都可以订阅。

关键词订阅类似“模糊匹配”，需要进行近义词挖掘。例如，输入关键词“明月”，除了返回包含“明月”的诗外，还要返回包含明月的近义词的诗。

功能要求

1. 需要按置信度将返回的诗歌**排序。**例如，若输入“明月”，且“山月”比“新月”更匹配，返回列表中包含“山月”的诗歌应排在包含“新月“的诗歌之前。
2. 订阅时，将返回结果的关键词/关键词的近义词**高亮。**只订阅诗人，不限关键词的情况下，不必高亮。
3. 可以**同时限制**订阅的诗人和关键词。例如，输入“李白”、“明月”，则返回李白写的“明月”相关的诗。
4. 订阅者在查看内容时可以**翻页**

模型要求

1. 每首诗的词的 TF-IDF 特征提取
2. 对词频超过10（所有诗中的总频度）的词进行近义词挖掘。推荐通过词在诗歌中的上下文分布向量（如TF-IDF，PMI等）的相似度来得到词汇间的相似关系。（注意：近义词挖掘只限制在词频大于10的词汇之间，但计算特征向量时要考虑所有词频大于1的词汇）
3. 在矩阵运算时若需要减少内存使用或加速计算，可以使用稀疏矩阵。

   参考 https://docs.scipy.org/doc/scipy/reference/sparse.html

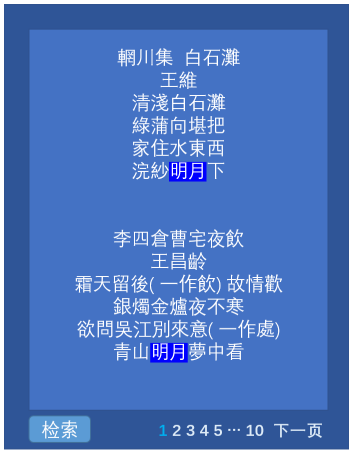
其他说明

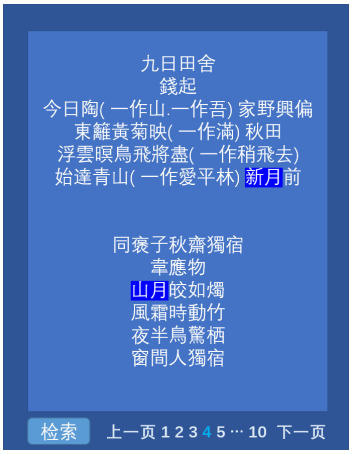
1. 基于第9次作业，按词表对诗歌语料进行分词（简繁版本任选其一）。可以假定检索词和订阅词都限定在分词得到的词表中，即不必考虑 Out-of-Vocabulary (OOV) 的情况
2. 发布者不必周期性推送，可一次性发布所有内容

UI示例

【新增订阅】

【查看订阅】





注：此实例只为直观地展示功能，返回结果非标准答案，UI样式也可自由发挥。

考察点

1. 近义词挖掘算法。加分项：设计合理的评价方案并据此设定**近义词的合理阈值方案**　10%
2. 通过合理加权得到订阅最终结果。加分项：设计**合理的评价方案**并据此调参　10%
3. 加分项：用Hits算法得到**检索主题向量**，据此进行检索优化　10%（总加分上限20%）
4. 本次作业不建议采用word2vec包，直接使用word2vec包做近义词最多可得95%

提交内容

1. 源代码 （打包为 `source-学号-姓名` 压缩文件）
    - 建议将TF-IDF特征提取、近义词挖掘算法两部分在notebook文件里单独实现
2. TF-IDF 特征提取的结果文件
3. 报告
    - 实现的功能
    - TF-IDF特征提取及近义词挖掘算法描述（不必大段地粘贴代码，文字描述清楚即可）
    - 关键词扩展前后结果的对比分析
    - 加分项的实现（可选）
4. demo 视频 （可选）

以上整体打包为 `TextHomework-学号-姓名` 压缩文件

截止时间

1. **2021/4/29 晚11：59** 截止，建议26号前提交，方便助教批改及推荐课堂交流。
2. 被邀请课堂交流的同学有额外5%加分（不受20%上限约束）。
3. 关于延期提交，5月6号前提交的最高90分，可加分；5月6号后提交的最高80分，不再有加分项。