大作业组织形式

- 以小组为单位， 10 月 30 日前将分组情况发至助教邮箱 mql_smile@126.com ，并抄送教师 benhe@ucas.ac.cn
  - 小组成员姓名、学号
  - 组长 Email 、手机号
- 每组不超过 5 人
- 若分组有困难找不到组员，可给教师发邮件
  - 若自己一个人一组，也请发邮件告知

大作业内容

- 分为两类
  - 学术类：面向学术研究的实验，包括基本索引器代码编写和检索评价实验
  - 工程类：面向工程应用，包括数据爬取和界面编写

学术类大作业

- 第一部分：编写索引器，构建 Shakespear-Merchant 语料索引 (15 分)
- 第二部分：在 WT10G 数据上进行检索竞赛（ 20 分）
- 第三部分：编写界面程序，
- 实验报告（ 15 分）

第一部分：编写索引器

- 内容：编写索引器，构建给定语料的词典与倒排索引
- Shakespear-Merchant 语料
  - 《威尼斯商人》剧本，预料规模极小，用于测试索引器
  - 下载地址：http://gucasir.org/ModernIR/shakespeare-merchant.trec.tgz
  - 解压命令： tar zxvf shakespear-merchant.trec.tgz

语料格式

- 按剧本场景分为 22 个文档，每个文档有如下格式：

  <DOC>

  <DOCNO>StringID</DOCID>

  <title>The Title</title>

  Content goes here

  <speaker> Name </speaker> Speech

  </DOC>

- 其中， DOC 标识文档起止位置， DOCNO 为文档字符串 ID ， title 为标题。

功能要求

- 基本要求：构建词典和倒排索引
  - 实现 Single-pass In-memory Indexing
  - 实现倒排索引的 Gamma 编码压缩 / 解压
  - 实现词典的单一字符串形式压缩 / 解压，任意数据结构（如哈希表、 B 树等）
  - 实现关键字的查找，命令行中 Print 给定关键字的倒排记录表
  - 给出以下语料统计量：词项数量，文档数量，词条数量，文档平均长度（词条数量）
  - 编程语言不限，但必须提交代码和说明文档
- 对停用词去除、词干还原等无要求，但应实现最基本的词条化功能
  - 例如：将所有非字母或数字字符转换为空格，不考虑纯数字词项

第二部分：检索竞赛

- 采用类似 TREC 竞赛的形式
  - 以小组形式在给定数据上进行实验
  - 鼓励创新思维
    - 评分：综合考虑实验结果和使用的新方法、提出的新思路
- Collection: WT10G
  - A small crawl of the Web, used in the TREC-9 & 10 Web track
  - 可从以下网址下载，包括查询 (topics) 和相关性标记 (qrels)

    http://gucasir.org/Data/WT10G.tar

使用的系统

- 不排斥使用等开源工具
- 也可以基于第一部分的索引器进一步开发实现检索功能，并用于第二部分的大作业
  - 鉴于此项工作的难度，评分时有加分

查询：只使用 title 域

<top>

<num> Number: 517
<title> titanic what went wrong

<desc> Description:
Find documents that discuss the reasons for or problems leading to the sinking of the Titanic.

<narr> Narrative:
A relevant document will discuss what caused the Titanic to sink.

</top>

竞赛规则参考资料

D. Hawking, N. Craswell. Overview of TREC-9 / 10 Web track. Proceedings of TREC-9/10. 2000/2011.

http://trec.nist.gov/pubs/trec9/papers/web9.pdf

http://trec.nist.gov/pubs/trec10/papers/web2001.ps.gz

检索竞赛评分规则

- 检索效果（20分）：
  - 一共100个查询，编号451-550
    - 前50个(451-500)用于训练
    - 后50个(501-550)用于最终评测
    - 不得在后50个查询上进行训练！违者视为作弊
  - 评价指标： Mean Average Precision (MAP)
  - 小组得分=20*(小组MAP-最低MAP)/(最高MAP-最低MAP)
  - 使用自己实现的系统有加分
- 实验报告（15分）：
  - 对索引器代码和运行方法进行说明
  - 详细描述实验中采用的技术
  - 对于提出的新方法、新技术有得分奖励
    - 新检索模型、新相关反馈方法等，或对现有模型、方法的提高和修正

可能需要采用的技术

- 检索模型
- 相关反馈 / 查询扩展
- 其它？
  - 链接分析？

实验报告

- 索引器系统结构、实现方案、主要代码类以及运行方法的说明
- 使用了什么技术？基于什么原理？分别给出公式
- 描述详细实验步骤
  - 训练
  - 测试
  - 要求能看出没有在测试查询集上进行训练
- 汇报最终在 50 个测试查询上获得的 MAP

结果提交

- 将所有材料做成一个压缩包， Email 至 benhe@ucas.ac.cn
  - 第一、二部分的源代码
  - 第一、二部分的可执行程序
  - 符合 TREC 格式的结果文件
    - 格式：参考 Hawking & Craswell TREC-9/10
  - 实验报告
  - 但不提交中间文件，避免附件过大
  - 提交时限： 12 月 31 日之前

工程类大作业内容

(一) 搜索型：

新闻搜索：定向采集3-4个新闻网站，实现这些网站信息的抽取、索引和检索。网页数目不少于10万条。能按相关度、时间、热度(需要自己定义)等属性进行排序，能实现相似新闻的自动聚类。

要求：有相关搜索推荐、snippet生成、结果预览(鼠标移到相关结果，能预览)功能

工程类大作业内容

(二)、分类型：

(1) 分类体系为：财经(类别号：1)、科技(类别号：2)、汽车(类别号：3)、房产(类别号：4)、体育(类别号：5)、娱乐(类别号：6)、其它类(类别号：7)，利用网站的新闻主页(可以下载Sogou语料)，训练一个分类器(训练集合不能少于5000篇文档)。能够实现新的网页的分类。支持交互式URL输入，或者输入一个文本，文本每行都是一个URL，系统输出结果文本，每行对应输入文本的类别号。

(2) 文本倾向性分析：下载餐馆的不少于2000篇评论信息进行训练，最后对餐馆的评价文本进行文本倾向性分析，首先分析该段文本是否涉及评价，如果是评价，则分析对餐馆的评价是褒还是贬。测试时，输入一篇文本，格式如下：

<docno>1</docno>

<text>这家餐馆的味道不错！</text>

<docno>2</docno>

<text>菜太贵了！</text>

<docno>3</docno>

<text>北京哪儿好玩？</text>

……

希望输出文本格式如下(每行中间空格,yes表示褒义、no表示贬义、na表示非餐馆评价)：

1 yes

2 no

3 na

……

要求：对于分类型任务，至少实现两种分类器并进行对比，至少实现IG这种特征选择方法并进行对比。画出在训练集合上10交叉测试的结果图。

其它任务

学生可以自行设计和选择本课程相关的其他任何有趣的题目，但是需要经过任课老师同意。

工程类作业提交

(一)、(二)任选一项任务

提交内容：

- 设计文档1份(内容包含并不仅仅限于设计方案、运行方法、自我测试情况、创新点、经验总结等等)
- 可运行程序
- 程序源码
- 能够支持运行的语料，请使用tar -c X | bzip2 -z9 >X.tbz 压缩后通过网盘发送，X为语料目录

Email至 benhe@ucas.ac.cn

说明：

不反对使用开源软件。但在设计文档中需要说明。

提交期限：12月31日之前