Homework 4

- 学术搜索引擎

  目标：

  – 1. 写一个Web爬虫，爬取文献网站的网页（及PDF文件）；
  – 2. 解析网页内容，对内容进行结构化，并存储到文件中；
  – 3. 为内容建立索引；
  – 4. 通过命令行进行内容检索，并展示内容列表
  – 可通过作者、标题、摘要、会议来检索

Homework 4

- ACL网站：https://www.aclweb.org/anthology/，其中某个会议的某一年 （具体给定学生-会议对应文件）

ACL Events

Venue	Present – 2010	2009 – 2000	1999 – 1990	1989 and older
ACL	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79
ANLP		00	97 94 92	88 83
CL	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 78 77 7
CoNLL	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97	
EACL	17 14 12	09 06 03	99 97 95 93 91	89 87 85 83
EMNLP	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96	
NAACL	19 18 16 15 13 12 10	09 07 06 04 03 01 00		
*SEMEVAL	19 18 17 16 15 14 13 12 10	07 04 01	98	
TACL	19 18 17 16 15 14 13			
WS	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 86 84 81 79 77
SIGs	ANN | BIOMED | DAT | DIAL | EDU | FSM | GEN | HAN | HUM | LEX | MEDIA | MOL | MORPHON | MT | NLL | PARSE | REP | SEM | SEMITIC | SLAV | SLPAT | UR | WAC			

Non-ACL Events

Venue	Present – 2010	2009 – 2000	1999 – 1990	1989 and older
ALTA	18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03		
COLING	18 16 14 12 10	08 06 04 02 00	98 96 94 92 90	88 86 82 80 73 69 67 65
HLT		06 05 04 03 01	94 93 92 91 90	89 86
IJCNLP	17 15 13 11	09 08 05		
JEP/TALN/RECITAL	14 13 12			
LREC	18 16 14 12 10	08 06 04 02 00		
MUC			98 95 93 92 91	

Annual Meeting of the Association for Computational Linguistics (2019)

Contents

↑up

pdf (full)

bib (full)

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics

One Time of Interaction May Not Be Enough: Go Deep with an Interaction-over-Interaction Network for Response Selection in Dialogues

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, Rui Yan

Abstract

Currently, researchers have paid great attention to retrieval-based dialogues in open-domain. In particular, people study the problem by investigating context-response matching for multi-turn response selection based on publicly recognized benchmark data sets. State-of-the-art methods require a response to interact with each utterance in a context from the beginning, but the interaction is performed in a shallow way. In this work, we let utterance-response interaction go deep by proposing an interaction-over-interaction network (IoI). The model performs matching by stacking multiple interaction blocks in which residual information from one time of interaction initiates the interaction process again. Thus, matching information within an utterance-response pair is extracted from the interaction of the pair in an iterative fashion, and the information flows along the chain of the blocks via representations. Evaluation results on three benchmark data sets indicate that IoI can significantly outperform state-of-the-art methods in terms of various matching metrics. Through further analysis, we also unveil how the depth of interaction affects the performance of IoI.

Anthology ID: P19-1001
Volume: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics
Month: July
Year: 2019
Address: Florence, Italy
Venue: ACL
SIG: –
Publisher: Association for Computational Linguistics
Note: –
Pages: 1–11
URL: https://www.aclweb.org/anthology/P19-1001.pdf
DOI: 10.18653/v1/P19-1001

Homework 4

- 关键技术：
  - 爬虫
  - 信息抽取
  - 索引建立
  - 查询

Homework 4

- Tips：
- 1. 如何在Eclipse中引入jar包

Homework 4

- Tips
- 2. JAVA爬虫
  - crawler4j
    - https://github.com/yasserg/crawler4j

crawler4j

build passing | maven-central v4.4.0 | chat online

crawler4j is an open source web crawler for Java which provides a simple interface for crawling the Web. Using it, you can setup a multi-threaded web crawler in few minutes.

  - JSOUP
    - https://blog.csdn.net/zbx931197485/article/details/78582407

    - jsoup 是一款 Java 的HTML 解析器，可直接解析某个URL地址、HTML文本内容。它提供了一套非常省力的API，可通过DOM，CSS以及类似于jQuery的操作方法来取出和操作数据，可以看作是java版的jQuery。
      jsoup的主要功能如下：
      从一个URL，文件或字符串中解析HTML；
      使用DOM或CSS选择器来查找、取出数据；
      可操作HTML元素、属性、文本；
      jsoup是基于MIT协议发布的，可放心使用于商业项目。官方网站：http://jsoup.org/

Homework 4

- 基于jsoup: Java HTML Parser来抽取信息 (如标题等，相同的网站同一个模板)，利用正则表达式来建立模板
  - https://jsoup.org/

```
File input = new File("/tmp/input.html");
Document doc = Jsoup.parse(input, "UTF-8", "http://example.com/");

Elements links = doc.select("a[href]"); // a with href
Elements pngs = doc.select("img[src$=.png]");
  // img with src ending .png

Element masthead = doc.select("div.masthead").first();
  // div with class=masthead

Elements resultLinks = doc.select("h3.r > a"); // direct a after h3
```

Homework 4

- Tips
- 3. 利用Lucene对文本进行索引，并进行检索

Apache Lucene™ is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

http://lucene.apache.org/core/

Homework 4

- 3. 利用Lucene对文本进行索引，并进行检索（输入检索词，查询得到相关的问题（或课程）列表，并显示详细信息。
  - 建索引和检索的简例

Homework 4

- 作业包括：java文件 + 文档 + 数据
- 作业打包上传到ftp homework/homework4下
- 文件：学号_姓名_homework4.rar

Homework 4

- 代码要求：
  - 遵守编程规范，如命名、注释等规范
  - 遵守面向对象的设计原则
  - 考虑异常处理等应用

Homework 4

- 文档要求：
  - 按附件格式样例，至少包括：引用、总体设计、详细设计、测试与运行、总结
  - 包括：数据格式说明
  - 附加：程序中包含的其他特色或改进，可加分
  - 附加：数据的丰富程度，可加分