考题1：

场景：<https://aclanthology.coli.uni-saarland.de/>主页上有1965年来ACL协会组织的各项会议的论文集，目前收集了4万多篇学术论文。

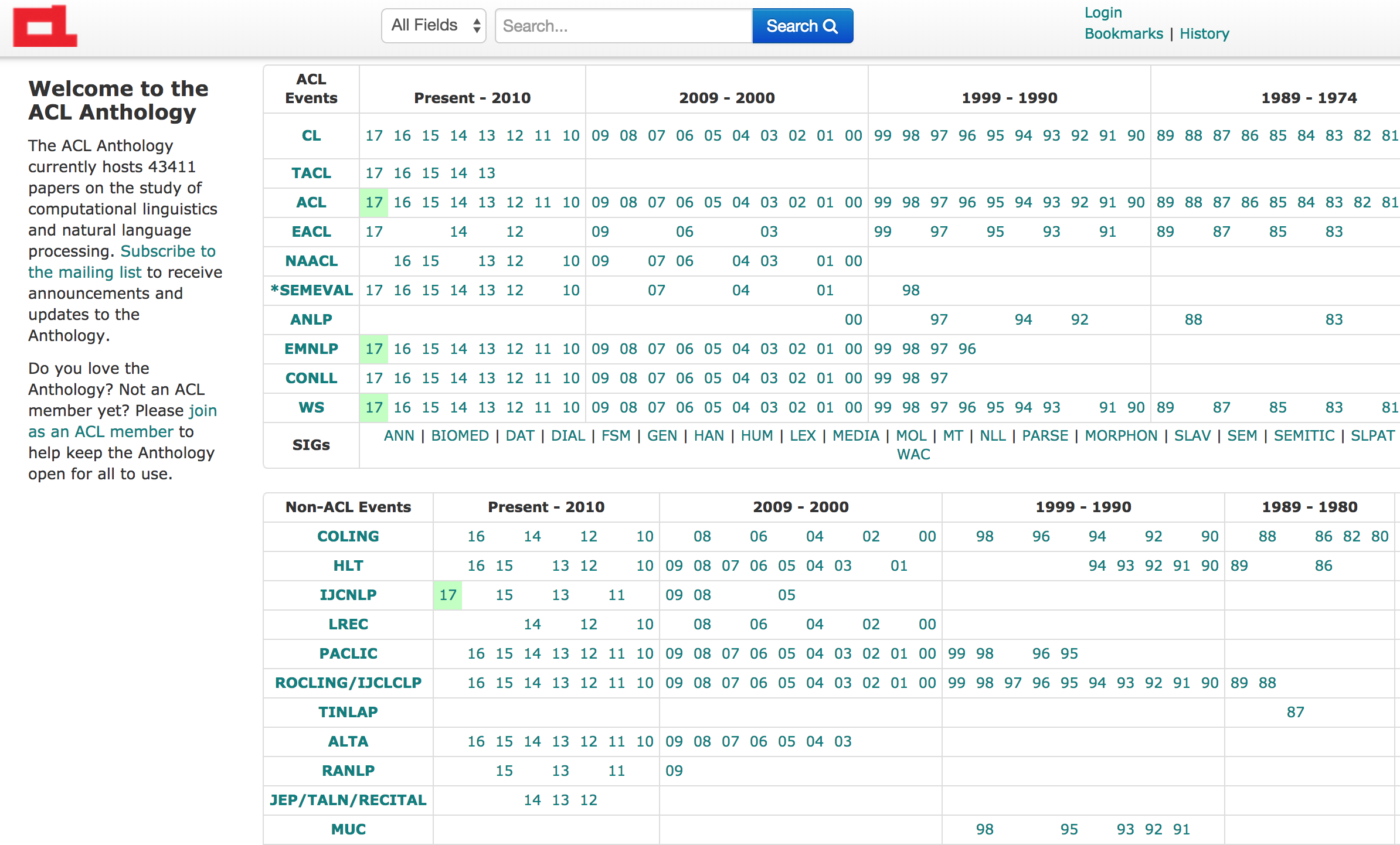


图1：主页

用户可点击任一会议，进入查看该会议的论文数据。

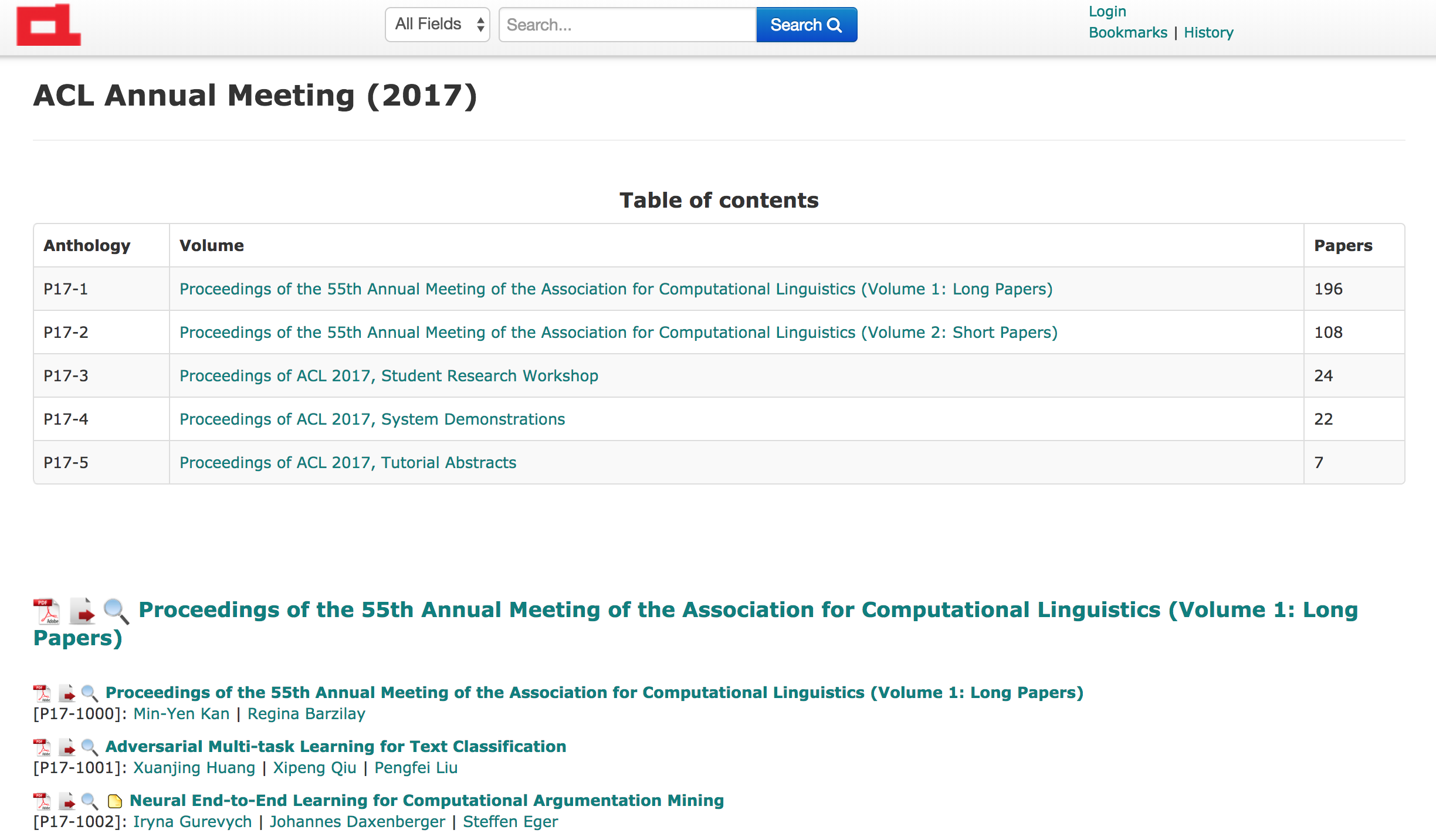


图2：ACL2017论文

每一篇论文都可以导出bib格式的索引文件以及提供了论文的pdf预览并提供下载功能。Bib格式以及论文题头如下面两图所示。

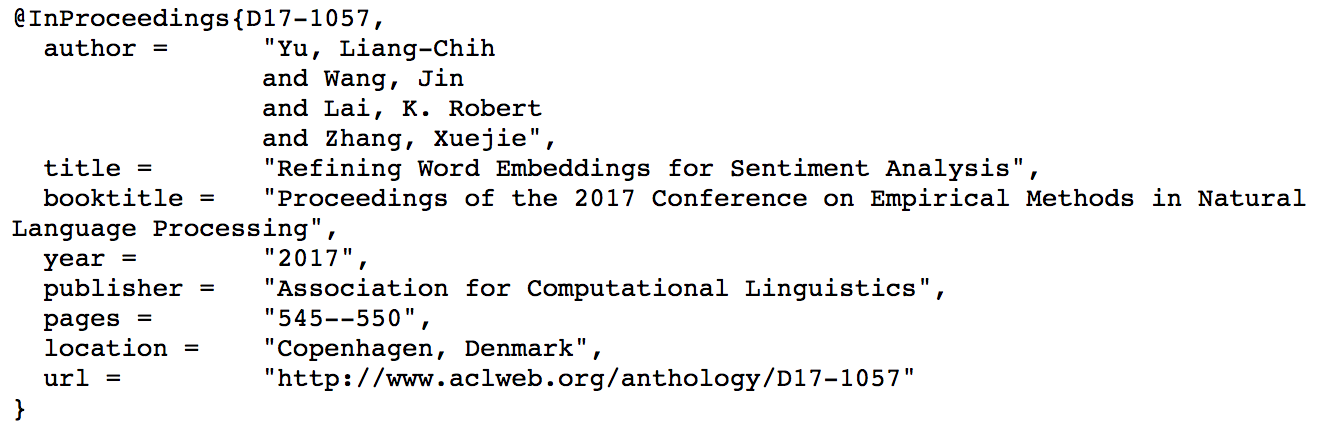


图3：论文bib格式样例

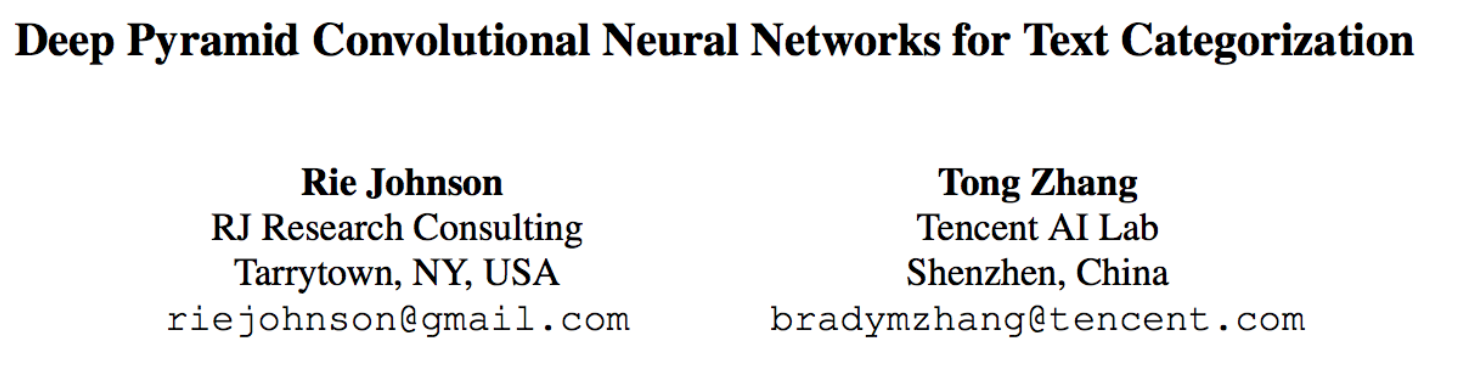


图4：论文PDF首页上部格式样例

结合自己对题干理解、实际项目经验以及已掌握的公开数据资源，简要回答以下问题。要求：有相关工程经验的，尽可能深入得阐述实现方式，包括开发语言、使用模块以及过程中出现的问题和解决方法。无相关经验的可阐述实施思路。

1、回答以下三个问题。

（1）您是否有web页面和pdf文件数据抽取的工作经验。如果有，请简要叙述，突出自己工作重点；如果没有，请填无。

（2）结合上述已知的bib文件和PDF样例，给出元数据（bib中所有字段，pdf中作者单位）抽取方法和思路。

（3）给出遍历采集4万余篇论文并抽取相关字段的技术方案。

2、简述构建ACL领域内论文集知识库的思路和实施路线。应细化到实体与关系的设计（实体类别、实体属性字段设计、关系种类、关系描述），数据库建设方案（使用的数据存储方案，以满足实体和关系的表达、描述）等。

3、根据自身理解，给出ACL领域作者重要度排序算法思路。注意：考虑重要性特征时可不局限于题干给出的论文和作者相关字段属性等特征，可结合自身经验和掌握的数据资源，发散性作答。

4、重要人物属性扩充方面的见解和可行方案。提示：ACL论文集网站范围内，重要人物的属性无非姓名、单位、相关论文、合著作者等。但扩展至整个互联网，重要人物的媒体曝光率相对较高、个人站点博客建设较为完备。因此，其人物属性还可包括其个人站点、twitter、微博、媒体报道、重要言论等。上述提示仅为抛砖，请结合自身经验和理解，给出答案。

考题2：

阅读下面2篇材料，并回答问题。

材料1：IFlyTek and other Chinese companies say they follow China’s laws and protect user data. But they agree that the sheer number of users in China, plus the government’s single-minded drive to dominate the new technology, puts them at an advantage.

“China has entered the artificial intelligence age together with the U.S.,” said Liu Qingfeng, iFlyTek’s chairman, at the Beijing conference. “But due to the advantage of a huge amount of users and China’s social governance, A.I. will develop faster and spread from China to the world.”

……

“We work with the Ministry of Public Security to pin down the criminals,” said Liu Junfeng, the general manager of the company’s automotive business, at a conference in September.

Where iFlyTek gets its data is not clear. But one of its owners is China Mobile, the state-controlled cellular network giant, which has more than 800 million subscribers. IFlyTek preloads its products on millions of China Mobile phones and runs the hotline service for China Mobile, which did not respond to a request for comment.

“Data is gold,” said Anil Jain, a professor who studies biometrics at Michigan State University. “These days you cannot design an accurate and robust recognition system for anything” without data.

节选自：《China’s A.I. Advances Help Its Tech Industry, and State Security》，《纽约时报》英文版，2017年12月3日

材料2：The world’s automakers are just starting to bet on an electric car future — and already, one of the most powerful people in the industry says that future belongs to China.

The Ford Motor Company said on Tuesday that it planned to introduce 15 battery electric or plug-in gasoline-electric hybrid car models in China by 2025. Speaking in Shanghai, William C. Ford Jr., Ford Motor’s longtime executive chairman, outlined why in an unusually blunt comment.

“When I think of where E.V.s are going,” he said, using an abbreviation for electric vehicles, “it’s clearly the case that China will lead the world in E.V. development.”

……

Daimler AG, the maker of Mercedes-Benz and Smart cars, said at the Guangzhou auto show last month that it also planned a sharp increase in its electric car production investments, particularly for making the batteries.

“China will, in my view, be the biggest market for passenger cars for a very long time,” said Hubertus Troska, the company’s management board member responsible for China. Many of them, he added, will be electric.

节选自：《China Will Lead an Electric Car Future, Ford’s Chairman Says》，《纽约时报》英文版，2017年12月5日

问题：设计相关算法，满足上述两篇材料作为系统输入，输出“人物-言论-人物职位”字符串的需求。请尽可能详细的阐述技术解决方案细节。

简述构建ACL领域内论文集知识库的思路和实施路线。

要求：

1、数据获取、解析、抽取等技术环节，到

需要补齐作者单位信息，由于bib文件中没有作者单位信息，因此需要设计自动化工具从pdf中抽取。如果有pdf抽取经验，可简要介绍作者单位抽取思路和需要注意的关键技术问题。

2、构建的知识库，需要包括实体和关系两类数据。简要说明实体种类

答题要点：

最终构建的知识库体系应实体（包括属性）和关系。

实体为：作者、论文、作者单位、会议名称。作者单位除作为单独实体外，其具体值还应该作为作者的属性项。论文的属性项应至少包含bib文件中的所有字段。

实体间的关系有：论文与作者之间的一对多关系；作者与作者单位之间的一对一或一对多关系；会议与论文之间的一对多关系。

需要重点解决的技术点：

1、元数据解析与格式化问题。涉及bib文件获取（包括html文件解析、文件路径匹配、url构造）、bib文件字段解析、人名格式对齐等。

2、作者单位需要从pdf中进行抽取。

3、数据组织结构，借助图数据库或者常规结构化数据库。给出简单库表或文件字段和属性设计样例。

4、