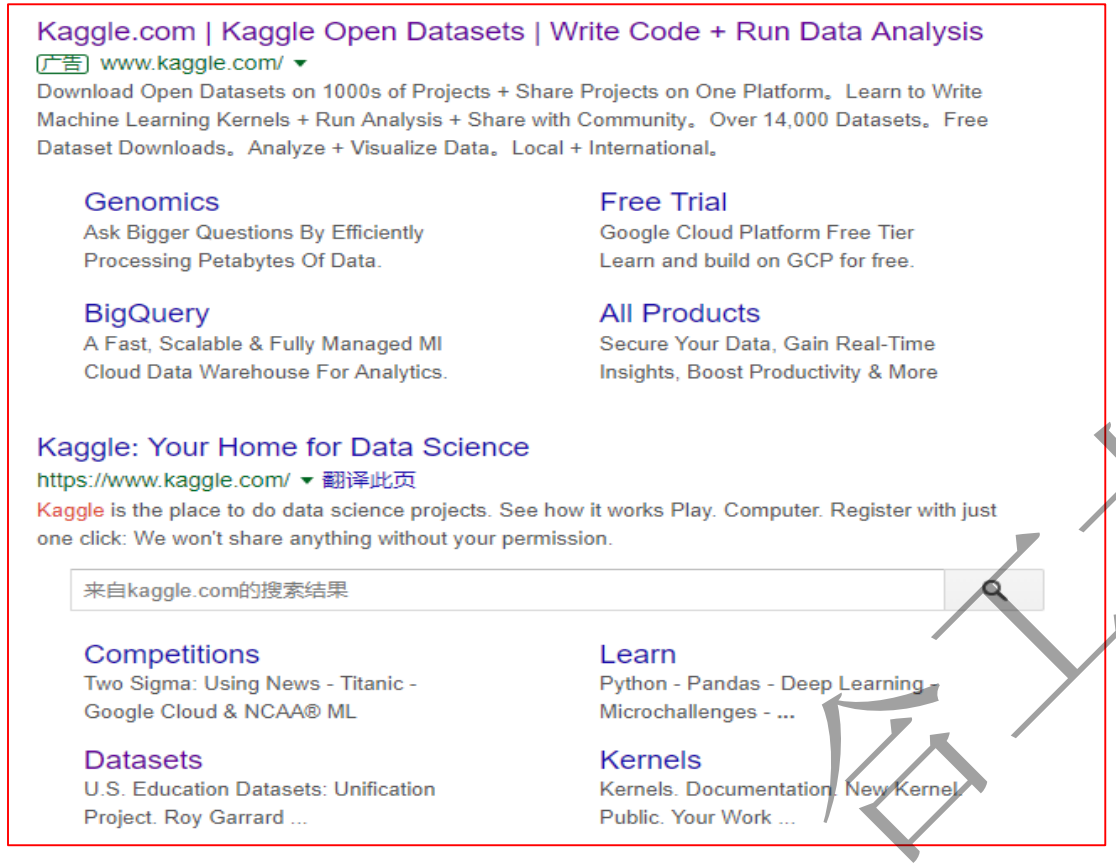
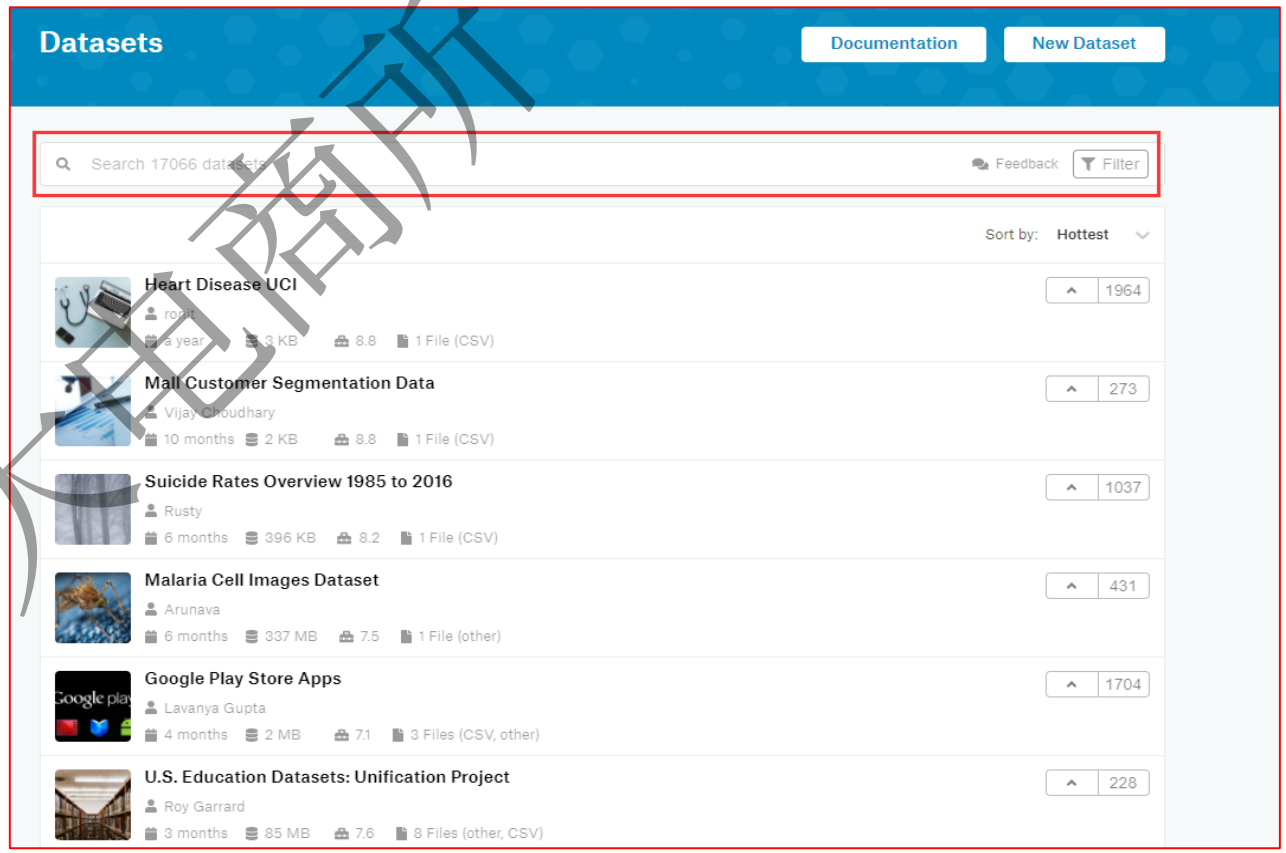
文本挖掘的两种简单方法

管理学院电子商务研究所
钱洋

数据来源

- 开源数据集(Open Source Data)
- 网络爬虫(Crawler Data)

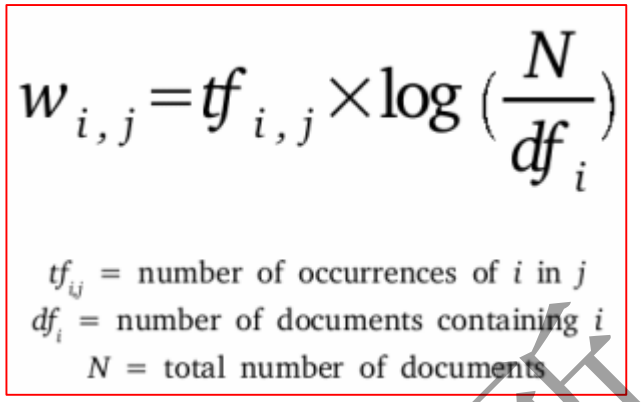


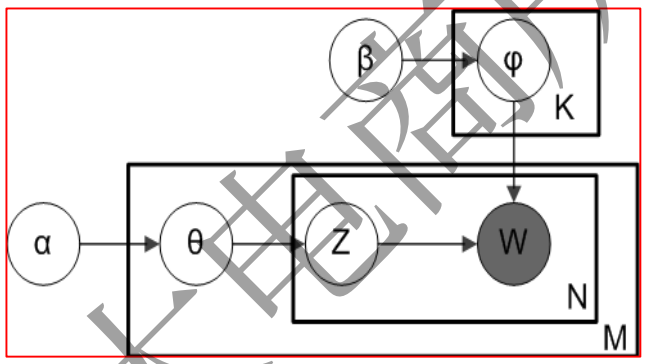


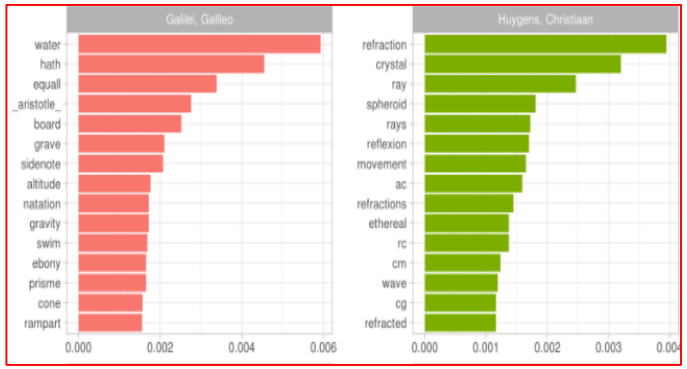
数据来源

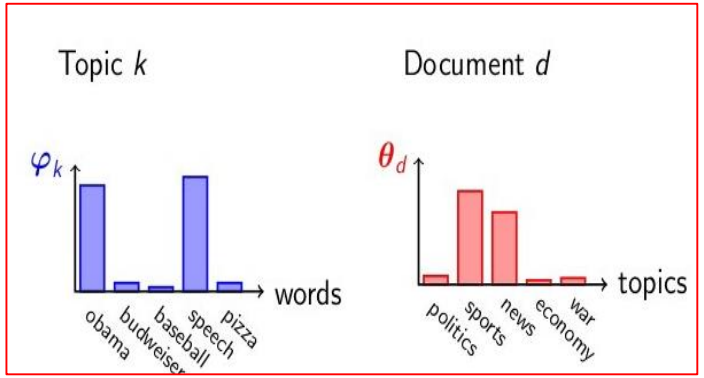
- 开源数据集(Open Source Data)
- 网络爬虫(Crawler Data)

案例









Input Data

Model

(TF-IDF 和 LDA)

Output

中文文本特征分析

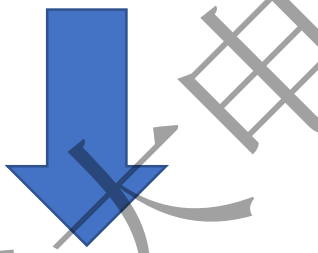
现身说法:移动互联网让买房更简单 http://t.cn/RLNbje0 就喜欢简单高效！ 哈哈哈哈👍我在用@Weico微博客户端，智能夜间模式、个性定制阅读环境、还有离线模式节省流量！想体验更好?★★★★戳链接下载Weico Pro 4:http://t.cn/RAPgR4n

- 以句子的形式呈现
- 存在标点符号
- 存在URL
- 存在非法字符
- 存在停用词

中文文本预处理

◆ 数据处理，尤其文本处理，大部分时间都花在预处理上。如：

- 将单个用户的数据进行合并。
- 分词、去停用词。
- 去标点符号、去URL链接、去非法字符等。



程序讲解

思考

◆ 英文数据处理与中文数据处理，存在哪些差异?

It's Not Just Coffee: Technology Now Fueling Massive Growth in Seattle

John Boitnott / AI, Entrepreneurs, Tech

Seattle has long been known for inciting the country's coffee culture, but now that we're all sufficiently caffeinated, these days it's the technology scene that's causing a buzz. Situated well north of the traditional tech enclave known as Silicon Valley, Seattle has come into its own as a magnet for innovative companies and a supportive [...]

How IoT Affects The Future Of Web Developments

Harshal Shah / AI, Code, IOT

ATMs, Apple Watch, Fitbit, Self-driving cars and more are part of a list of devices using IoT is endless. The concept of IoT began years ago; we just didn't use the term IoT. Kevin Ashton is credited with being the first to coin the phrase Internet of Things (IoT) in 1999. The Internet Of Things [...]

5 ways Artificial Intelligence is Revolutionizing eCommerce Marketing

Rameeza Yasin / AI, Marketing, Productivity

Marketing strategies are continuously changing due to increasing technological advancement. Technology is responsible for raising the exceptions of people all around the globe. With an increase in innovations, it's pretty obvious that customers are expecting more and companies are doing everything to can to remain in their good list. Consequently, competition between companies seems to [...]

https://readwrite.com/category/ai/

谢谢