

项目——新闻人物言 论自动提取

2019.April.27

项目背景

- 随着每天涌入的新闻信息越来越多，获得不同人物对于不同事件的观点，获得重要人物每日对于不同事项的观点描述，这个功能对于新闻阅读、观点总结能够起到很大的辅助作用。
- 例如，我们现在如果有一款新闻阅读的 app，我们能够把文中的每个人的核心观点整理出来，总结成表格，那么对于读者来说，就容易看清楚多了。
例如：



人物	观点	来源
人社部信息中心有关负责人	今年要在所有地市实现签发应用全国统一标准的电子社保卡，至少1亿人领取电子社保卡，所有地市均开通移动支付服务	

- 上文中是比较短的一个新闻片段，事实上，为了进行舆情分析，危机预测，知识图谱等等各种任务，我们往往需要采集很多任务的观点，尤其是事实描述的时政、社会新闻，其主要信息往往其实在不同人物的言论中，例如：

昨日，雷先生说，交警部门罚了他 16 次，他只认了一次，交了一次罚款，拿到法院的判决书后，会前往交警队，要求撤销此前的处罚。

律师：不依法粘贴告知单

有谋取罚款之嫌

陕西金镒律师事务所律师骆裕德说，这起案件中，交警部门在处理交通违法的程序上存在问题。司机违停了，交警应将处罚单张贴在车上，并告知不服可以行使申请复议和提起诉讼的权利。这既是交警的告知义务，也是司机的知情权利。交警如果这么做了，本案司机何以被短时间内处罚 16 次后才知晓被罚？程序违法，为罚而罚，没有起到教育的目的。

我们再看一段实时新闻：

《中央日报》称，当前韩国海军陆战队拥有 2 个师和 2 个旅，还打算在 2021 年增设航空团，并从今年开始引进 30 余架运输直升机和 20 架攻击直升机。此外，韩军正在研发新型登陆装甲车，比现有 AAV-7 的速度更快、火力更猛。未来韩国海军陆战队还会配备无人机，“将在东北亚三国中占据优势”。

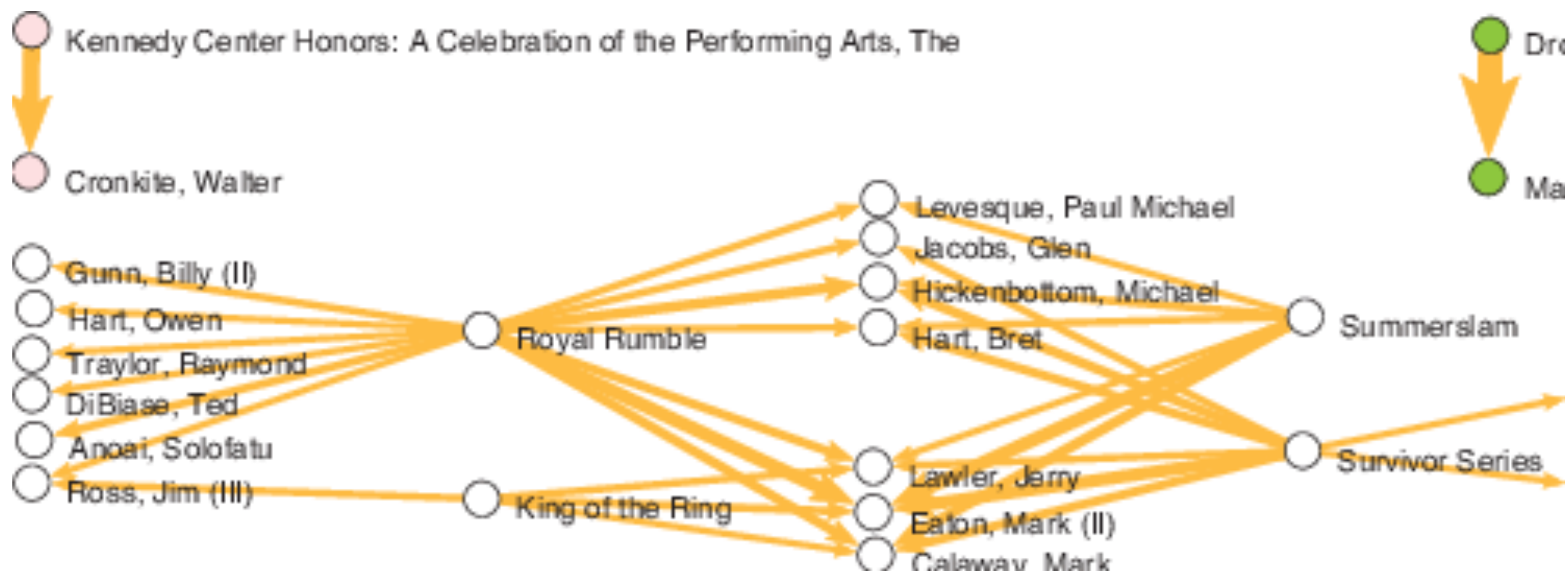
但韩国网友对“韩国海军陆战队世界第二”的说法不以为然。不少网友留言嘲讽称：“这似乎是韩国海军陆战队争取国防预算的软文”“现在很多韩国海军陆战队员都是戴眼镜、瘦豆芽体型，不知道怎么选拔的”“记者大概是海军陆战队退役的吧”。

任务描述

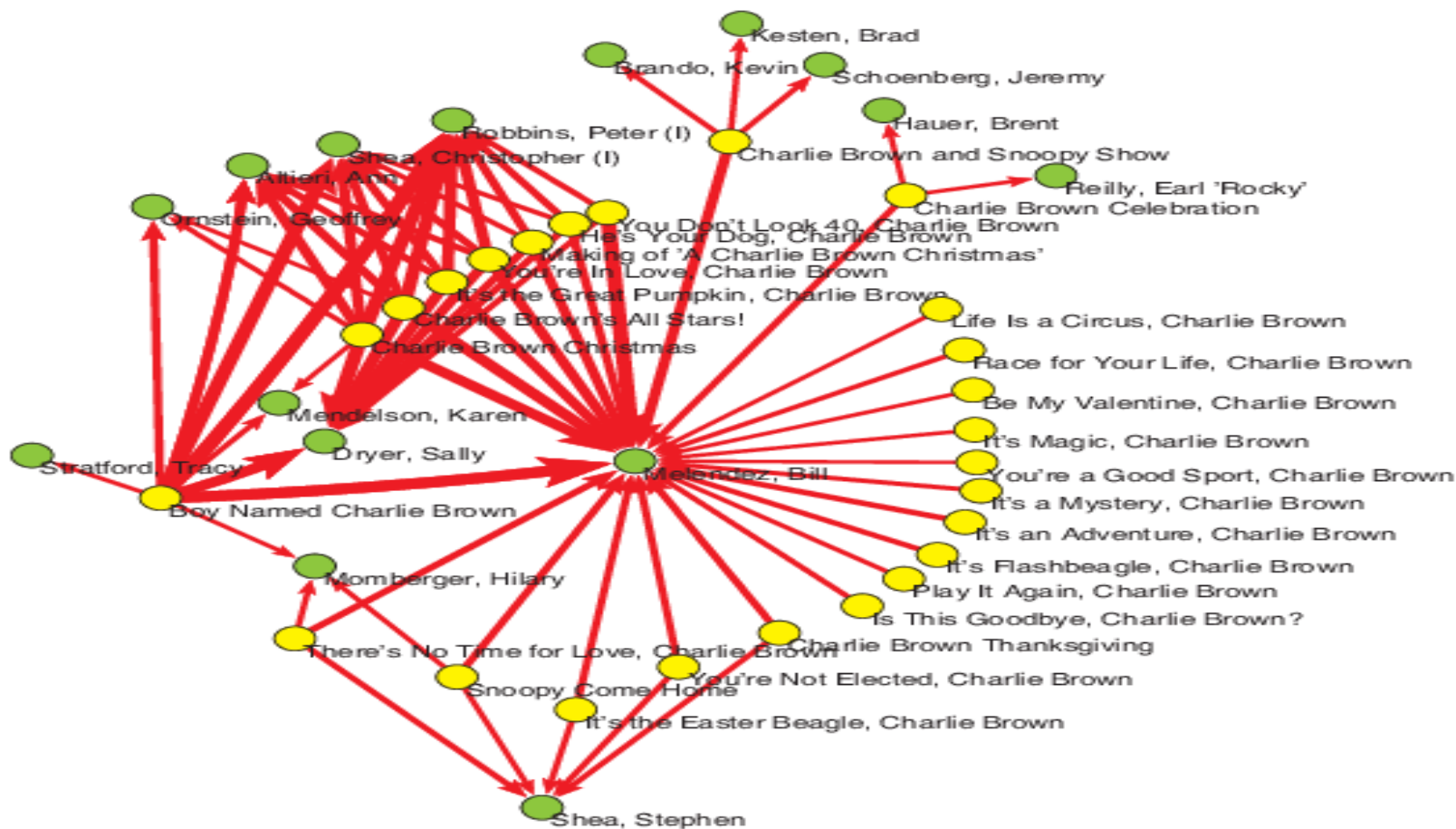
- 所以，我们面对的是这样一个任务：
 - 输入：TEXT，一段新闻文字
 - 输出：List，文中每个人物/实体的观点

人物	言论
韩国网友	不以为然
雷先生	交警部门罚了他 16 次，他只认了一次，交了一次罚款，拿到法院的判决书后，会前往交警队，要求撤销此前的处罚。
..	...

除此之外，我们还可以进一步把表格进行可视化，将其画为一个观点图：



又或者是可视化成这样的图



- 1. 如果同学还没有毕业，那么可以在把人物的言论提取出来之后，进一步做成“知识图谱”，“人物观点图谱”等偏向学术类的应用；
- 2. 如果同学已经毕业，面临找工作的要求，那么可以把任务的言论提取出来之后，加上我们项目 2 的情感分类，对言论进行极性分析，变成一个能够依据网络信息，获得群众对该事件的危机预警应用。

这个项目实际被用在哪里？

- 1. 被用在新闻 app 中；
- 2. 被用在公司内部的事件跟踪中；
- 3. 被用在商用的舆情监督系统中；
- 4. 被用在学术研究中；

我该如何完成？

- 这个项目大家要能完成，需要综合这么3 大块：

视图层(使用 HTML, Python Web 服务进行网页展示)

模型层(构建自然语言处理模型，能够提取出文章中
客户的言论)

数据层(能够使用数据库操作，对数据库中的信息进行访问)

关键步骤：1，获取数据

- 数据库为云数据库，配置为 Mysql；请大家下载 DataGrip 或者 MySQL WorkBench 进行数据库的访问；
- 访问地址和密码为：
- 数据库地址：cdb-q1mnsxjb.gz.tencentcdb.com: 10102
- 用户名：root
- 用户密码：A1@2019@me
- 数据库名：news_chinese
- 如果在 DataGrip 下配置，则配置的界面如图：

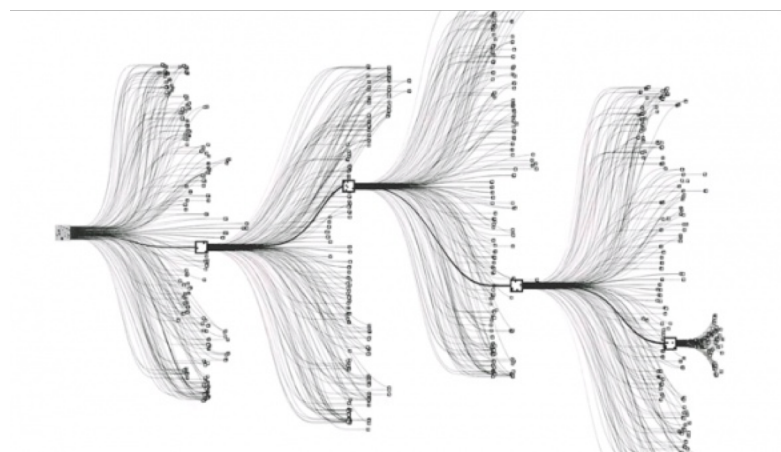
The screenshot shows the 'Data Sources and Drivers' dialog box in DataGrip. The 'General' tab is selected. The configuration is as follows:

- Name:** news_chinese@cdb-q1mnsxjb.gz.tencentcdb.com (with a 'Reset' link)
- Comment:** (empty text box)
- Host:** cdb-q1mnsxjb.gz.tencentcdb.com
- Port:** 10102
- Database:** news_chinese
- User:** root
- Password:** (masked with dots, with a 'Remember password' checkbox checked)
- URL:** jdbc:mysql://cdb-q1mnsxjb.gz.tencentcdb.com:10102/news_chinese (with a 'default' dropdown menu)
- Test Connection:** Successful (with a 'Details' link)
- Driver:** MySQL
- no objects** (text indicating no database objects are visible)
- Tx:** Auto (dropdown menu)
- Read-only:** (unchecked checkbox)
- Auto sync:** (checked checkbox)
- Update driver files:** (warning icon and text)

At the bottom right, there are three buttons: 'Cancel', 'Apply', and 'OK'.

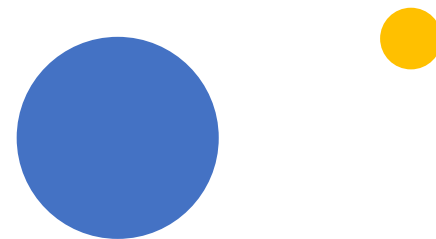
2. 获得所有表示“说”的意思的单词

- 使用维基百科+新闻语料库制作的词向量，在基于第一课，第二课讲过的搜索树 + 第四课的动态规划，结合第五课所讲的内容，获得出所有与“说”意思接近的单词。
- 思考：词向量结合图搜索的时候，每个找到的单词如何赋其权重，这个和广度优先，A*搜索有何异同？



- 使用 NER, Dependency Parsing 等方式, 获得是谁说了话, 说了什么话。其中 Dependency Parsing 我们有 Stanford 的 CoreNLP 和哈工大的LTP, 这两个工具的安装过程会比较麻烦, 大家要做好心理准备。
 - Stanford
CoreNLP: <https://stanfordnlp.github.io/CoreNLP/>
 - 哈工大 LTP: <https://github.com/HIT-SCIR/pyltp>
-

3. 使用 NER, Dependency Parsing等对句子形式进行解析



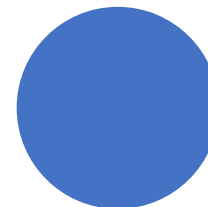
4. 确定言论 的结束

- 在确定了谁说的，说了什么之后，我们要做的就是确定这个话语的结束。要确定这个话语如何结束，最简单的方式解释碰见句号的时候就停止，但是有的话可能是跨了多个的。那么这个如何确定多个呢？这个时候就比较 tricky 了。在有的时候，我们可以使用 tfidf 等关键字，或者使用 tfidf 关键字获得句子的向量然后使用向量进行对比的。获得句子向量之后，那么我们就可以把判断两句话是不是类似的、说得同一个主题这个问题变成这两个句子的距离是不是小于某个阈值。Tfidf 的句子向量化是一种比较基础的向量化方式，长久以来也是大家用的。但是 tfidf 不能变成不相同的单词的语义相似性，在词向量提出来之后，有一个比较好的方式解释基于词向量进行句子的向量化。基于词向量获得句子的向量化也是现在的一个研究课题，这里给大家推荐一个简单性和高效性两者比较平衡的方法，其原理就是使用单词的词向量加权 + PCA 降维 这个方法是普林斯顿大家2017年提出来的一个方法，很简单，但是效果也不错。

- 普林斯顿句子向量原始论文
Paper: <https://openreview.net/pdf?id=SyK00v5xx>
- Scikit-learning TFIDF句子向量化: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

- 基于以上几步，相信大家已经能够输入一段新闻，获得新闻中每个人说了什么话了，最后一步就是我们要能有一个展示自己作品的环境。
- 大家使用 Flask 或者 Bottle，使用 Bootstrap + HTML 构建一个简单的网页，在这个网页中，我们能够提交文本内容，然后会生成表格，表格里边能够显示这个文章中每个人的观点。
- 如果你有兴趣，还可以使用 D3 工具，做成网络状的示意图。
- Bootstrap: <https://getbootstrap.com/>
- Bottle: <https://bottlepy.org/>
- D3: <https://d3js.org/>

5. 展示自己的作品

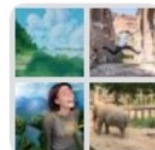


5.1 发布到服务器上

- 为了能够让大家进行协同工作，我们给大家提供了Linux服务器，大家代码写完之后，可以把自己的项目发布在服务器上。
- 服务器已经安装好了 anaconda，大家在上边操作的时候，每个组先 create 一个虚拟环境，然后在 home 目录下的 project-01 下边，建立一个自己 team 的名称的文件夹，把代码置于该文件夹下。
- 服务器为了访问安全，目前只能向外暴露 8800 - 8899 这 99 100 个端口。
- 服务器地址：stu@39.97.187.115
- 服务器密码：A1@2019@me



6. 分组



AI-For-NLP-项目01

- 看完了以上内容，你应该也会知道我们这个项目不是一个简单的项目，所以我推荐大家以小组的方式进行，因为我们的同学背景各异，我们的建议是小组以 2-4 人为佳。组员尽可能包含曾经有过工程项目背景的(对 MySQL, Python Web 较为熟悉的人)，有是数学、物理等相关背景的同学。
- 如果你已经有了比较熟悉的小伙伴，那你们可以直接组队，如果还没有，请扫描屏幕下方二维码，我们在群里为大家召集小伙伴。
- 如果你已经有小伙伴了，那也请扫描二维码，我们可以在里边讨论这个项目相关的问题。
- 确定好分组的同学，请到这个链接中登录信息：
https://docs.google.com/spreadsheets/d/1XVp8Lfh_1A92YF27DQqRJncBX1txZh1Pz3JNJfFL6Tc/edit?usp=sharing



Valid until 5/3 and will update upon joining group