项目一：人物言论提取

2018.12.10

问题描述： 在进行社会舆情分析、危机预测、观点态度分析、知识提取、观点提取、知识图谱等众多任务中，有一个基础性的任务我们需要完成。这个就是从每日的新闻中提取出来任务的言论。

我们以以下的一段社会新闻为例：

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

陕西金镝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。

我们再看一段实事新闻：

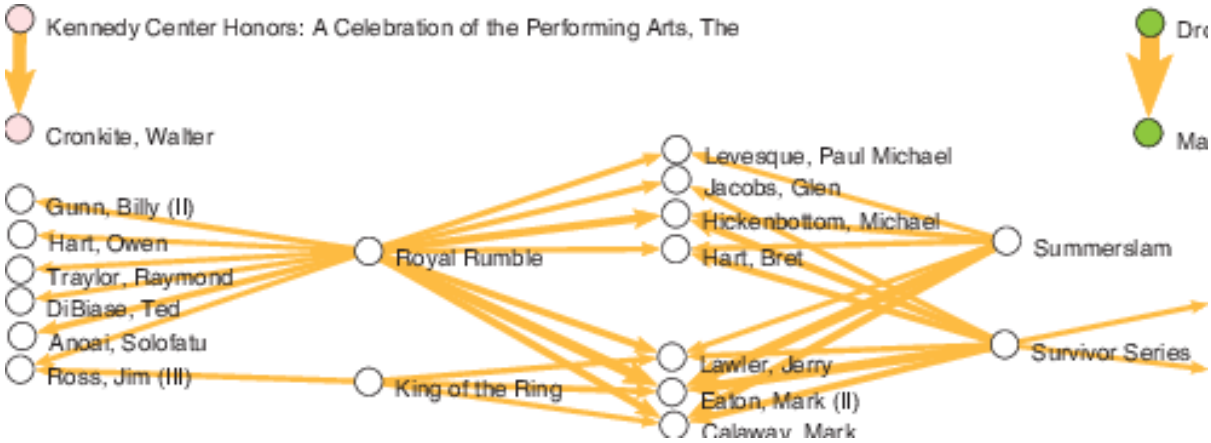
《中央日报》称，当前韩国海军陆战队拥有 2 个师和 2 个旅，还打算在 2021 年增设航空团，并从今年开始引进 30 余架运输直升机和 20 架攻击直升机。此外，韩军正在研发新型登陆装甲车，比现有 AAV-7 的速度更快、火力更猛。未来韩国海军陆战队还会配备无人机，“将在东北亚三国中占据优势”。

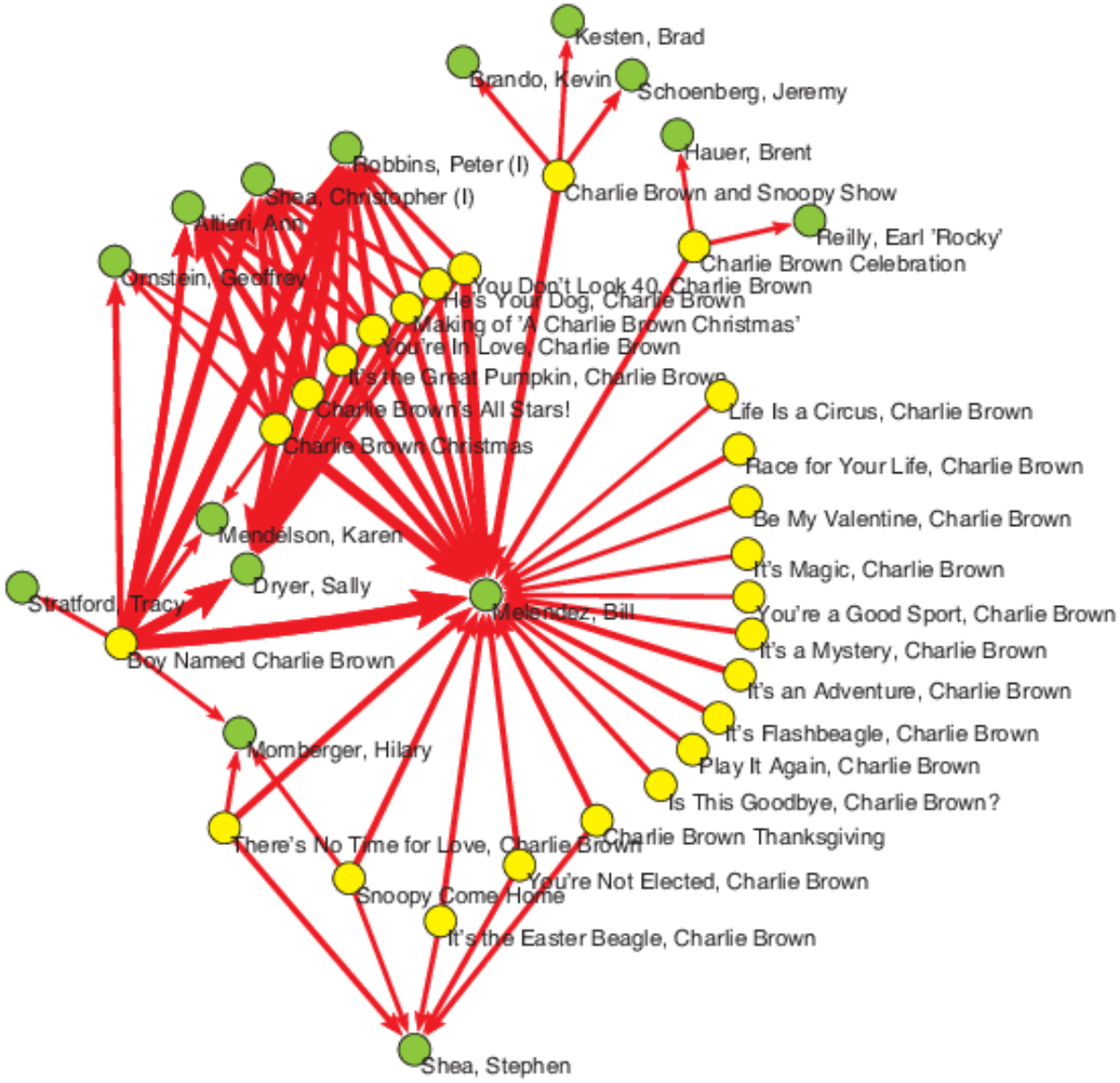
但韩国网友对“韩国海军陆战队世界第二”的说法不以为然。不少网友留言嘲讽称：“这似乎是韩国海军陆战队争取国防预算的软文”“现在很多韩国海军陆战队员都是戴眼镜、瘦豆芽体型，不知道怎么选拔的”“记者大概是海军陆战队退役的吧”。

所以，我们期望输入的基本结构是:

人物	言论
<XXX>	XXX
<XXX>	XXX
.....	.....

除此之外，如果我们进一步进行可视化，我们可以将其绘制为一个观点图，例如：





如果同学们是在学校读书还没毕业，推荐大家讲这个项目变成一个观点图谱(Opinion Graph) 的项目。 如果同学们已经毕业了，推荐大家先把这个信息能够提取出来，然后在此基础之上加上我们之后机器学习分类，做成一个危机预测(Risk Prediction) 的任务。

关键步骤：

我们要完成这个任务，有以下几个比较关键的点，大家可以去完成：

0. 下载数据源： https://github.com/Artificial-Intelligence-for-NLP/datasource/blob/master/export_sql_1558435.zip

1. 确定表示观点的 “说” 这个词，通过这个词确定可能是在进行观点表述的句子。

2. 为了获得表达 “说” 的这个词，我们可以使用词向量和图搜索的方法。 其中，词向量的训练推荐大家使用维基百科 + 新闻语料库 二者结合的方法[1]，在词向量的训练之后，大家要多多测试其性能，选择不同的参数，以确保得到 “较好” 的词向量[2]。

3. 词向量结合图搜索的时候，图搜索时候每个找到的词的权重如何定义？ 这个和广度优先搜索，A*搜索有何异同？

4. 在使用这几步获得了表示 “说” 的词汇之后，我们使用 NER，Dependency Parsing 的方式，获得是谁说了这个话，说了什么话。 其中，dependency parsing 我们有 Stanford CoreNLP， 哈工大的 LTP（这个 LTP 我们可以使用 python 的版本）这两个工具的安装过程可能会比较麻烦，大家要做好心理准备。

5. 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，但是有的话可能是夸了多个的。 那么这个如何确定多个呢？ 这个时候就是比较 tricky 了。首先，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。[3] 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[4]，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[5]，这个方法是 普林斯顿大家去年提出来的一个方法，很简单，但是效果也不错。

6. 完成了以上的几步，大家基本上就能够获得我们想获得的信息了。 这个时候可能还会有一些小问题，肯定会有一些错误信息。 这个时候就需要花时间打磨。做算法类的问题，我们可以花费 30%的时间达到 80%的效果。但是剩下的 20%，其实需要花费我们绝大多数的时间。

在 Slack 上我开了一个 Project-01 的 channel， 大家只会对于第一个项目有问题的就在这个里边进行讨论。

最后，我再给大家回答几个问题：

---

[1] 问题：为什么要结合？ 如果不结合会有什么不足？

[2] 问题：词向量的训练中，相关的参数有哪些？ 重要的参数有哪些？ 你是如何确定的？

[3] 这里的句子向量化，又是 sentence embedding 和单词的 word embedding 类似，是把一个句子变成一个固定长度的向量。

[4] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[5] https://openreview.net/pdf?id=SyK00v5xx

问题：1. 为什么不用微信进行讨论？

slack 的专门用来做即时通讯的，希望大家能够接受它。 如果在全世界的范围而言，slack 的使用范围肯定是高于微信，钉钉的。 基于微信和钉钉办公，这两个我也体验过，微信的问题是，微信上边的无关信息很多，在这个上边办公，很容易让人的思维分散(distract)； 钉钉的问题是，我们很多同学平时工作就是用的钉钉，而钉钉的这个“钉”的功能，只要一启用，内容就会悬浮在屏幕右上角。 公司的同事、领导很可能就看到了，这会很尴尬。

问题：2. 为什么不能把视频的录播放在国内的视频网站上？

国内的视频网站如果我们要上传，要经历这么几步： Zoom 录播转码: 大约 2 小时； Zoom 视频下载到本地 大约 1-2 小时； 视频本地转格式大约 1- 2 小时； 然后上传到国内视频网站 2 – 3 小时，然后等待国内视频网站的内容审核，大约 2 小时。 这样一次下来，就需要很久的时间。 如果大家 Zoom 录播看起来很慢，那很有可能是大家的网络被屏蔽了。

大家有问题就及时讨论，要“八仙过海各显神通”，也要“集思广益”。相信大家能做出来很好的结果！