信息网络专题研究课（应用层）笔记

一、文献信息

1.论文题目：Using News Articles to Model Hepatitis A Outbreaks: A Case Study in California and Kentucky

2.作者：Marie Charpignon , Maria Mironova , Saeyoung Rho , Maimuna S.Majumder , Leo A.Celi

3.发表途径：AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.

4.发表时间：2019

二、问题意义

1.论文研究问题：利用新闻文章来模拟最近甲型肝炎爆发的可能性。

2.研究背景

美国最近爆发了一次大规模的甲型肝炎疫情，接近实时的发病率估计对政府预测有效的疫苗接种率，从而控制疫情的传播至关重要。这些估计值以往来自美国疾病控制中心提供的常规监测数据，但这些数据的传输延迟通常非常大，对控制疫情传播造成了阻力。人们正在寻找一些替代方法，例如新的数据源，以改善目前的困境。

3.研究意义

使用新闻报道数据作为数据源建模的疫情动态与使用传统数据源 CDC 的检测数据不同，它的实时性更强，但新闻文章的质量可能会影响建模结果。本文的研究将自然语言处理技术应用到与健康相关的新闻内容中，提取相关见解以反馈给决策者，帮助他们做出决策。这一方法是大数据时代利用新型数据解决问题的一个很好的例子。

三、思路方法

论文首先提出政府以往对疫情传播做预测使用的疾病控制与预防中心的实时监测数据往往出现延迟的问题既而引出使用新闻报道数据作为替代数据源的建议。随后文章对所使用的数据集、流行病学预测使用的模型、对数据集的文本分析、美国疾病控制与预防中心和使用新闻文章的数据采用的概念模型、对处理结果的文本分析、对采用两种数据源得到的结果的讨论、采用新闻文章的数据带来的限制和风险进行详细的分析。最后总结全文，并对未来的工作进行了展望。

1.数据和方法

论文研究重点关注加利福尼亚州和肯塔基州这两个在很大程度上代表了最高风险人群的疫情动态的州，利用两个不同的数据集，采用 IDEA 模型对他们进行性能比较并使用 NLP 技术分析文章主题和文本内容。

1.1 采用的数据

采用的数据集共有两个：一个是美国疾病控制与预防中心的流行病学研究在线数据 (WONDER)，另一个是 HealthMap 这个包含疾病相关新闻文章的公共数据库。对 CDC WONDER

数据集，提取其 2017 年 3 月 4 日至 2019 年 3 月 31 日关于甲型肝炎发病率的每周报告。对 HealthMap，提取同一时间从加利福尼亚和肯塔基州各大媒体获取的 563 篇新闻报道，仅删除在县级的报道并收集剩余新闻中报道的病例数。

1.2 发病率衰减和指数调整(IDEA)模型

IDEA 模型是一种取决于两个参数:基本繁殖数 R0 和折扣因子 d 的，用于短期流行病学预测的单方程模型。(I)随时间(T)的累积发生率满足公式 $I(\mathrm{t})=\sum_{t=0}^{T}\left(\frac{R_0}{(1+t)^t}\right)^t$。当 R0 小于 5 时，使用 IDEA 模型。当 R0 大于 5 时，应用了一个非线性优化过程，并使用 Python scipy.optimization curve_fit 方法将理论表示与经验数据拟合。

1.3 新闻的文本分析

所有的新闻文章都聚合在一个词包模型中，便于进行文本分析。预处理的重要步骤包括句子标记化、停止单词移除和单词词形化。使用 Spearman 进一步测量 CA 和 KY 两个词袋模型的交集，根据文字的相对频率来估计它们的相似度。

3.结果分析

3.1 概念模型

美国疾病控制与预防中心的数据使用可靠的串行间隔选择进行处理，HealthMap 的新闻文章主要注意点在处理丢失的数据上。CDC 数据的 IDEA 模型对 CA 和 KY 的 MAPE 拟合优度检验表明，模型的性能并不强烈依赖于序列区间的选择。对新闻文章采用的进位和线性平滑两种技术得到了相似的结果。两种数据源得到的分析结果近似一致。数据分析结果显示 CA 的病毒比 KY 传播速度要快很多，但 CA 能够迅速控制疫情。因此，对模型中参数的校准和疫苗接种的评估则需要考虑现实情况，以及估计的阈值和实际值之间的差异。

3.2 单词模型的文本分析

两个州数据的单词包模型术语交集量较低，同一个术语在两个州之间的相对频率也有差距。由此反映出两个州的疫情发展状况和疫情控制措施的侧重点不同。

4.结果讨论

以上结果证明了采用两种数据源得到的结果有较高的相似度，因此在 CDC 的数据暂不可用的情况下，新闻媒体数据有可能被用作政府评估疫情的另一种数据来源。新闻文章的质量也有可能影响结果，各地新闻媒体的态度不同，报道的侧重点也不尽相同，这就可能导致不同地区的媒体对不同类别的案例报道的数量不同。新闻报道的可获得性和质量将影响新闻媒体作为替代数据源的价值。

四、实验结论

新闻文章可以有效地用于甲型肝炎爆发的建模，且具有实时性更强的优点。各地新闻媒体对疫情的报道侧重点不同，从中可以看出各个州防疫措施和疫情情况的差别。未来的

工作则是将此方法应用于美国的其他州，但这需要采取有效措施来客服某些地理区域缺乏新闻文章和过度/不足报告趋势并进行谨慎调整。

五、启发思考

1. 大数据已经成为当下最热的关键词,它被广泛地应用于各个行业,比如交通、销售、医疗以及媒体行业。大数据的分析和挖掘在医疗领域的应用包含很多的方向，比如临床操作的比较效果研究、临床决策支持系统、医疗数据透明度、远程病人监控、对病人档案的先进分析；临床试验数据分析、个性化治疗、疾病模式的分析等；还有患者临床记录和医疗保险数据集等。

新闻数据作为大数据中的一种，目前也正在被用于医疗状况分析。对新闻中的数据进行抓取、挖掘、统计、分析，使得数据能够可视化呈现（比如通过复杂度交互式、动态化的图片和视频来呈现）。这使得大数据新闻能够对事件进行描述、判断、预测、信息定制，从而达到纯文字报道难以达到的更加清晰、说服力更强的效果。

2. 随着 21 世纪的来临，我们迎来了数据信息大爆炸的时代。移动互联、社交网络、电子商务等极大拓展了互联网的边界和应用范围，各种数据正在迅速膨胀并变大。互联网(社交、搜索、电商)、移动互联网(微博)、物联网(传感器，智慧地球)、车联网、GPS、医学影像、安全监控、金融(银行、股市、保险)、电信(通话、短信)都在疯狂产生着数据。

如何把这些海量数据进行收集与整合，然后进行数据处理使之产生额外的价值为人类服务，是一个重要的课题。在数据处理中涉及到的数据筛选、过滤、整合汇编等都需要扎实的数学功底和熟练的计算机编程知识，而这也正好顺应了当下的对数学和计算机越来越重视的教育趋向。

3. 云计算的出现带动了大数据应用的发展，对数据整合、分析与挖掘所产生的效果前所未有，社会和个人均因大数据的使用而获益，然而不容忽视的问题是隐私风险的存在。大数据背景下由于各种挖掘和整合技术的使用，导致个人的兴趣爱好、行为模式、社会习惯等隐私信息暴露。而这些信息往往是个人不想被他人所知的。抛开数据非法盗用不谈，一些无害的数据被大量收集后，也会暴露个人隐私。大数据如同一把双刃剑，在带来便利的同时也隐藏着风险。