阅读材料：

Using News Articles to Model Hepatitis A Outbreaks: A Case Study in California and Kentucky

姓名：王源

班级：通信 1710 班

学号：17211251

一、文献信息

作者：

Marie Charpignon、Maria Mironova、Saeyoung Rho、Maimuna S. Majumder、Leo A. Celi

论文题目：

Using News Articles to Model Hepatitis A Outbreaks: A Case Study in California and Kentucky

发表途径：

NeurIPS 2019 Workshop on AI for Social Good

二、问题意义

对于新出现的大规模流行疫情，及时估计有效的疫苗接种率、获得遏制其发展的增量疫苗接种对于政府控制疫情的蔓延至关重要。应用于近实时发病率估计的估计数的传统来源是疾病控制中心(CDC)提供的常规监测数据，但是，这些数据往往延迟到达。所以，本文提出新闻文章数据建模的爆发动力学在目标疫苗接种率方面产生的结果与 CDC 数据相当，并指出，自然语言处理技术应用于与健康相关的新闻内容，可以提取更快、更准确的信息，进一步提高对于社会的潜在影响。从而为疫情出现时政府确定增量疫苗接种所需的数据集采取过程提供新的选择，更好的遏制疫情蔓延。

三、思路方法

本文提出使用新闻文章作为替代数据来源来确定目标疫苗接种率，以达到有效遏制疫情的爆发的目标，并通过模拟甲型肝炎爆发，来验证理论的正确性。

文章选取了甲型肝炎爆发期间，加利福尼亚(CA)和肯塔基(KY)的爆发动力学作为研究对象，采用来自 CDC 的检测数据集作为传统数据，从健康地图中检索到的甲型肝炎相关新闻文章被用作非传统数据源；将发病率下降和指数调整模型(IDEA)模型应用于两个数据集，并对其性能进行比较；此外 ，还利用 NLP 技术对新闻文章的正文进行分析，测量语言相似性的程度，为决策提供基础。

1、IDEA 模型分析

使用的发病率下降和指数调整(IDEA)模型是一个用于短期流行病学预测的单方程模型，如下：

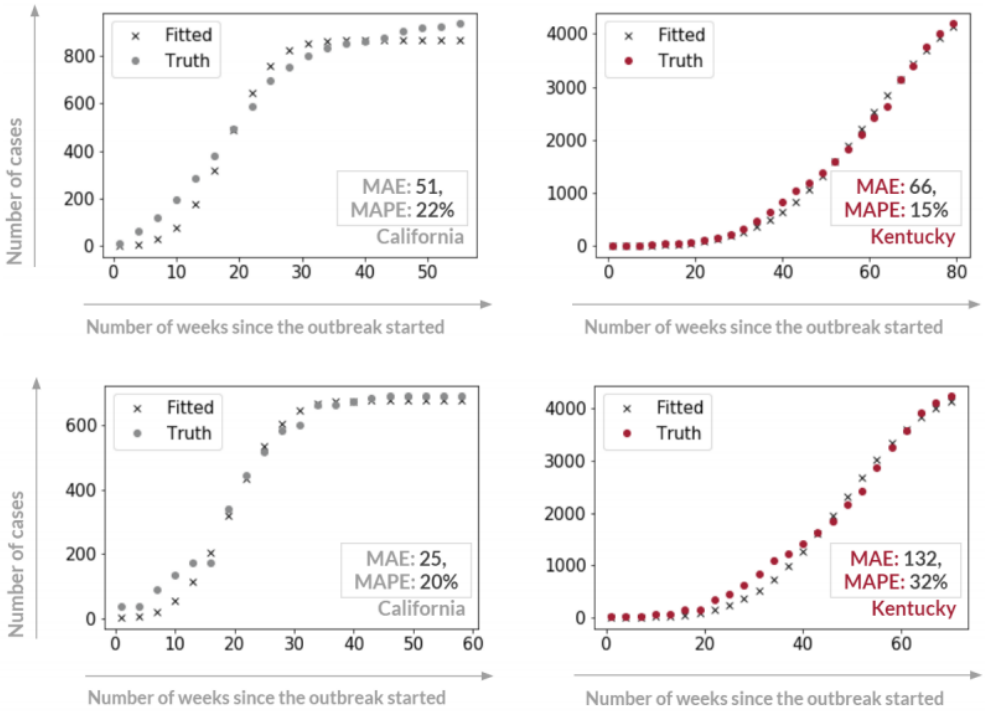
$$I(t)=\sum_{t=0}^{T}\left(\frac{R_0}{(1+d)^t}\right)^t$$

R_0 为基本繁殖数——表示每个感染者第一次进入完全易感人群时成功传播的数量，并描述爆发的初始指数增长

$I(t)$ 为随着时间变化的累积发病率

d 为折扣因子——$d.T$ 表示数据收集过程开始和结束之间的时间点数

利用 IDEA 模型分别利用 CDC 数据与新闻文章数据对 CA(左)与 KY(右)进行疫情曲线校准，得到图形如下



利用曲线判断两类数据集对于模型建立与处理的准确性，并分析估计阈值与实际疫苗接种百分比之间的差异，分析使用新闻文章数据集是否能够替代传统数据集。

2、新闻文章文本分析

将所有的新闻文章聚合在一个词袋模型中，进行句子标记化、停止词去除、词列化等预处理步骤，通过测量 CA 与 KY 两个词袋模型的交集，利用 Spearman 秩相关，根据单词的相对频率来估计它们的相似性。

四、实验结论

新闻文章数据集与传统数据集在评估目标疫苗接种百分比与当前疫苗接种百分比之间的差异方面具有一致性，此相对统计数字是政府规划疫苗接种相关政策的一个关键，所以在流行病持续发生和检测数据尚未获得的情况下，新闻媒体数据可能被用作替代数据。

五、启发思考

尽管新闻文章数据可以估计相对统计数据、目标和当前疫苗接种百分比之间的差异，但对于每一种疫苗的绝对估计使用新闻文章与使用 CDC 数据的估计不同。因此，如果决策者的

主要目标是获取个人统计数据，则不推荐采用此方法，此外，新闻对于流行病的报道的可用性和质量是进一步采用本方法的一个障碍。

本文的验证方法值得学习，为了验证理论对于流行病的应用情况，采用了较为典型的甲型肝炎作为研究样本，能够很好的得出相应结论，更有说服力。通过传统数据集与替代数据集分别探究并对结果进行比较的实验方法，让实验结论更明显。对于我今后的实验探究具有实验思想上的指导意义。

文章提出此方法的目标在于解决传统数据集提供数据不及时的问题，这也提醒了我，今后科技的发展必将向着高效性与低延时性前进。