

### Report Date

Report Date	07/04/2020	Name	Yiming Zhang
-------------	------------	------	--------------

### Period covered by this report

Start Date	28/03/2020	End Date	07/04/2020
------------	------------	----------	------------

## 一、本周研究内容

研究内容	本周的研究内容主要为：实现 3315 DEMO 的可视化部分。 项目 GitHub 地址： <a href="https://github.com/yiming95/3315project">https://github.com/yiming95/3315project</a> 网页链接： <a href="https://yiming95.github.io/3315project/">https://yiming95.github.io/3315project/</a>
------	---

## 二、项目实施当前状态

项目进度实施情况	目前正在对 3315 DEMO 可视化的部分。
项目整体进度完成情况	预计下周将进一步对 3315 DEMO 可视化的前端部分进行完善。

## 三、本周成果

本周研究的主要成果为实现了网站前端的可视化，一共分为“项目简介”、“可视化疫情预警”、“可视化疫情预测”和“科研创新”四个部分。

1. “项目简介”：包含一个图片、项目的内容介绍以及团队成员介绍。



图 1：项目简介页面- 图片部分

 UNNC 3315传染病智能预警预测系统

项目简介 可视化疫情预警 可视化疫情预测 科研创新

### 项目简介

对于新型冠状病毒（COVID-19），除了病毒的传染性强等客观因素外，我国的传染病预警系统没有进行及时有效的预警，也是此次疫情如此严重的原因之一。传染病的监测预警是公共卫生领域的重要组成部分。高效的传染病自动预警信息系统能够早期发现传染病爆发的异常情况并对其进行预警，以帮助有关决策部门尽早采取相应的防控措施。由于此次疫情的预警能力，我国将加强在公共卫生以及疫情预警预控方面的支出，对传染病预测预警系统进行升级，以加强我国在疫情预测预警的能力。目前基于大数据与人工智能技术的场景融合对于传染病进行预测预警的系统还属于新兴技术领域，因此本课题所研发的产品具有深远的社会意义与重大的经济效益，以及广阔的市场前景。

随着国家信息化建设的进步，以及大数据和人工智能等技术的发展，建立基于大数据和人工智能技术的新型传染病监测预警预测系统是十分有必要的。新型的智能化传染病预警预测系统，将不仅仅针对此次新型冠状病毒，而是可以针对所有传染病，尤其是传染性较强的呼吸道类传染病进行精准的预警与预测。新型的预警预测系统将对当前的预警预测系统进行有效的补充与升级，帮助流行病学家更快和更准确地发现和追踪传染病，并提升政府对传染病的防控能力。

基于大数据与人工智能技术面向新型传染病的重大场景融合预测预警系统的研发可以作为第三方软件平台对现有的国家预测预警系统进行很好的补充，而且也很符合当前的市场发展趋势。关于基于大数据和人工智能技术面向新型传染病的重大场景融合预测预警系统研发，具体研究内容包括：（1）通过网络爬虫等技术对传染病（新冠肺炎，H1N1，SARS）等关键信息进行大数据搜集。（2）对获取到的多源数据进行集成并且将处理后的大数据存入分布式数据库。（3）在算法层面，通过神经网络构建预警预测模型。可利用自然语言处理技术中的文本分类模型来对传染病信息进行预警；在预测方面，我们将通过传染病动力学模型SEIR模型进行改进，并利用图神经网络对传染病疫情进行精准预测等。（4）在软件与服务层面，通过构建可视化应用软件，实现人工智能模型与疫情预警预测的场景融合。

### 团队成员

Supervisor: Dr Ying Weng  
 Research: Xindi Zhao, Hejia Qiu, Yiming Zhang  
 Engineering: Yiming Zhang

---

**图 2：项目简介页面- 项目介绍和团队成员**

**2. 可视化疫情预警部分：一个可交互的地图：用户将鼠标放在地图上会看到该省的预警等级，预警等级分为“低、中、高”三个档次，颜色不同代表预警等级不同。**



**图 3：可视化疫情预警部分**

当用户点击黄色的“数据详情”按钮，会显示出所有新闻与微博等预警的内容详情。

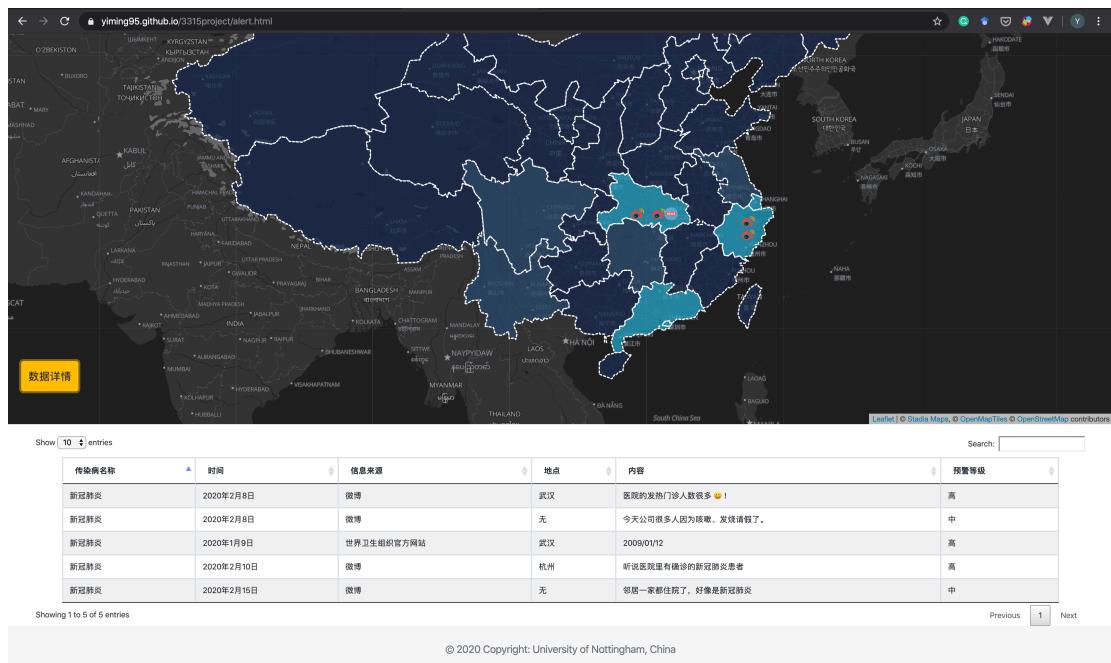


图 4：可视化疫情预警部分- 展开数据详情

此外，当用户点击左上方的“扩大”按钮，可以将地图最大化。地图上有图标，如“新浪微博”图标和“新闻”图标，用户点击图标可以看到作为预警来源的微博和新闻。



图 5：可视化疫情预警部分 – 最大化地图以及点击微博图标

**3. 可视化预测部分：一共有四个折线图，分别显示：确诊人数、痊愈人数以及死亡人数。左上图为全国，右上图为浙江省、左下图为宁波市、右下图为鄞州区。**

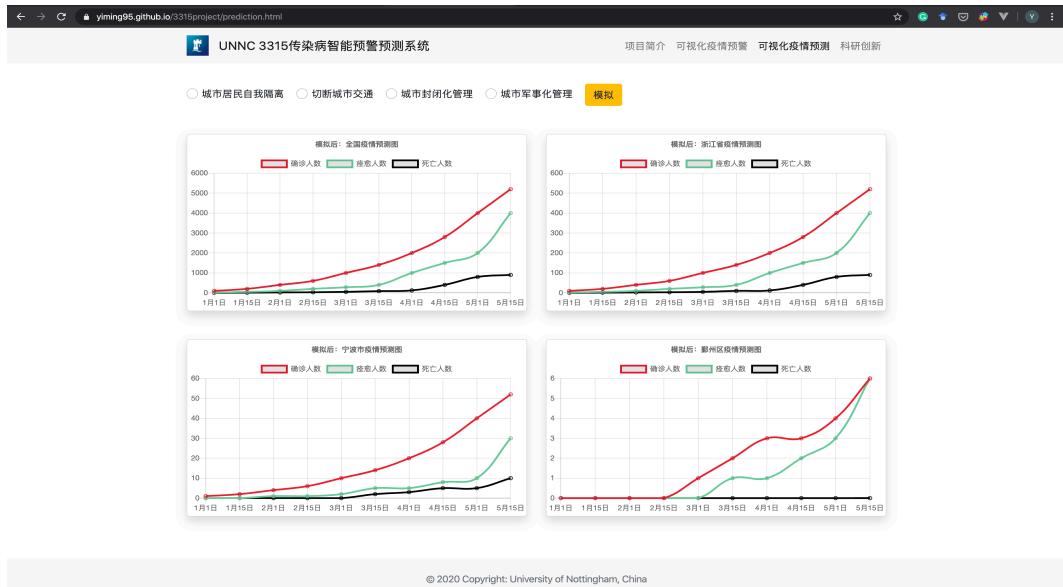


图 6：可视化疫情预测

此外，当用户选取“四个场景融合”的措施后，可以进行进一步模拟。



图 7：可视化疫情预测 – 选取措施后的模拟图

## 4. 科研创新部分：里面为“NLP 大数据方向创新”V3 版本里的内容。

在智能预警的科研方向上，创新点主要为数据源的创新，自然语言处理文本分类模型的创新以及传染病预警分析方法的创新，具体描述如下。

1. 传染病相关的多源数据集(语料库):  
目前在 text classification 这个方向上比较流行的数据集为: Reuters 数据集 和 arXiv Academic Paper 数据集等。就我们所知，在医疗健康方向，特别是传染病相关的文本分类方向，并没有公开的数据集。因此本项目通过多源数据收集模块、使用网络爬虫、Twitter 数据、CDC 数据等，将构建一个中文英文两种语言的传染病相关的多源数据集。如果项目将来开源一部分数据集，即可为后续相关研究人员研究医疗健康类的文本分类 (text classification) 提供来源可靠的数据集，在领域内作出贡献。

2. 自然语言处理文本分类模型:  
自然语言处理文本分类(text classification)是指为句子或文档分配适当类别的任务。模型的输入为一个文件 (通常表示为特征向量形式)，以及预定的输出类别  $C = \{c_1, c_2, \dots, c_k\}$  (在本项目中为“与传染病有关”和“与传染病无关”两类)。模型的输出为预测结果。一个文本分类模型通常包括特征提取 (Feature Extraction)、降维 (Dimensional Reduction)、分类 (Classification) 以及模型评估 (Evaluation)。本项目在自然语言处理文本分类模型上的创新点为在分类器上创建新的基于CNN+MLP的深度学习网络模型并用来进行传染病的文本分类任务。

比较常见的分类模型如朴素贝叶斯分类器 (Naïve Bayes Classifier), K 近邻 (K-Nearest Neighbor) 以及支持向量机 (Support Vector Machine) 等。在本项目中，我们将以这些机器学习的方法作为 baseline，并在我们构建的数据集上进行实验。深度学习中的卷积神经网络 (CNN) 常是用来应用在机器视觉中的图像处理任务，但是近年来，CNN 在自然语言处理中的文本分类任务上，也表现出了出色的成绩。比如，Alexis Conneau 等提出了一个在词层面的 CNN 模型在文本分类上取得了不错的成绩；Re Johnson 等则提出了一个使用 LSTM 的局部词嵌入的 CNN 网络用来进行文本分类。Alexis Conneau 的模型目前在文本分类任务上是现有技术的最高水平之一 (SOTA)，因此我们拟基于 Alexis Conneau 模型上提出四点创新：

2.1: Alexis Conneau 使用的 Encoded Characters 的特征，而我们将在特征抽取上使用 TF-IDF 与 Encoded Characters 相结合的方式。  
特征提取的基本思路是根据某个评价指标独立的对原始特征项 (词汇) 进行评分排序，从中选择得分最高的特征项，过滤掉其余的特征项。Encoded characters 采用的是字符嵌入与词嵌入相结合的方法，目的是将文本数据转化为数值数据，获取一个单词的数值表示。并使用这些数值向量得到句子/段落/文本等数值表示。然而训练字符嵌入在计算上十分昂贵，因此我们考虑结合 TF-IDF 方法。Kowarik 等人在他们提出的文本分类模型中，就用到了 TF-IDF 为特征提取方法 [7]。TF-IDF 方法中 TF 为 Term Frequency 用来表示一个词的重要性与在类别内的词频成正比，而 IDF 为 Inverse Document Frequency 用来表示一个词的重要性与所有类别出现的次数成反比。TF-IDF 可以很好的表达特征权重，因此我们认为 TF-IDF 与 Encoded Characters 相结合的特征提取方式会取得更好的效果。

2.2: Alexis Conneau 在池化方法中使用 K-Max pooling。我们将使用 Chunk-Max pooling 的方法进行改进实验，看能否取得更好的效果。  
池化 (Pooling) 是卷积神经网中的一个重要的概念。它实际上是一种形式的降采样。有许多不同形式的非线性池化函数。池化会不断地减小数据的空间大小，因此参数的数量和计算量也会下降。这在一定程度上也起到了过拟合。Max-pooling 是最大池化是对领域内特征只求最大，而 K-max-pooling 可以取每一个 filter 抽取的一个特征值中得分在前 k 的值，并保留他们的相对的先后顺序。Chunk-Max pooling 是把某个 filter 抽取到的特征向量进行分段，切割成若干段后，在每个分段里各自取得一个最大特征值。比如将某个 filter 的特征向量切成 3 个 chunk，那么就在每个 chunk 里面取一个最大值，于是获得 3 个特征值。相比于 K-Max pooling，Chunk-Max pooling 的方法能够保留了大部分局部最大特征值的相对顺序信息。因此我们期望使用 Chunk-Max pooling 的方法能够取

### 1. 图 4 : 科研创新

## 四、下周计划

下周计划为：

1. 继续优化前端可视化的部分。