基于相似性的外卖-人群深层神经网络分类模型

邢城祎

网络餐饮,一个几年前还略显陌生的词汇,如今已经飞入寻常百姓家.近几年来,网络餐饮行业呈现爆发式增长,越来越多的人热衷于订外卖.的确,外卖为我们带来了不少便利,但是其所产生的垃圾污染也不可小觑.在订外卖的人群中,不同的人群种类所订外卖的数量有着很大差别.为定量确定外卖所带来的固体废弃物与人群特点的关系,本文通过小组讨论,准确制定本文中的外卖垃圾分类标准和人群特点及对应人群分布地区的种类.在模型建立方面,把外卖垃圾数量- 人群特点的关系和某地区月平均降水量- 该地气候类型的分类方式相类比.通过实地采集,得到外卖垃圾数量的第一手资料,并且使用历史降水量数据源查询所需位置的降水量数据,经过数据清洗后,得到较为充分准确的数据.同时,在分类方法上进行深入研究,采用深层神经网络模型训练对应的降水量- 气候类型数据.通过不断的模型设计与超参数调节,使模型在降水量- 气候类型和外卖数量- 人群种类两个数据集上均达到较高的分类正确率.保存训练后的模型作为之后的预测,达到使用时无需训练,仅需输入该地外卖垃圾数量就能够以较高的正确率推测该地人群特点的目的.此模型在实际应用上可以通过外卖数量推测人群特点,从而研究更有效,更有针对性的垃圾处理方案.

关键词 深层神经网络模型 相似性模型建构方法 分类模型

Contents

1 引言

1.1 问题背景

2016年,全国生活垃圾年清运量已经高达20362万吨[1]，而其中外卖垃圾所占的比例正逐年上升.随着”互联网+”的不断发展以及服务业领域的进步,外卖产业正逐渐成为现在热门的行业.每天大街小巷上都能见到骑手配送餐饮,这在给人们带来方便的同时,外卖包装产生的固体废弃物给我们的环境造成了巨大的破坏.根据外卖类平台”饿了吗”发布的数据,外卖服务业每天至少会产生2000万份废弃的一次性包装盒、塑料袋和一次性餐具[2].外卖所产生的垃圾已严重影响环境和和人民的生活.每年不正当的垃圾处理方式所带来的经济损失高达300亿元人民币[3].为此,提出确切实用的方案来缓解外卖垃圾所产生的问题.

1.2 问题分析

针对问题一,通过查询资料和小组讨论,确定外卖垃圾的特定分类标准.并将在所选区域内进行实地数据采集,最后使用采集的数据进行分类估算外卖所带来的固体废弃物的数量.

针对问题二,从网络上找寻有关垃圾回收处理的现行方案,找到处理方式的不足之处,进而相应地提出能够弥补不足的处理方案.

针对问题三,采用小组内单独思考后再进行讨论汇总的方式,明确人群种类与所处区域.再收集相应的数据,建立模型,找到外卖固体垃圾与人群特点的关系.

针对问题四,结合问题三的模型和问题二的处理方式,用通俗易懂的语言写出介绍短文.

2 分类标准

通过汇总组内每位组员的分类标准,得出以下带有一定普适性的分类标准

2.1 外卖固体废弃物

类别	举例
塑料	塑料餐具, 塑料盒, 塑料袋等
纸	纸袋, 纸巾等
纸皮	纸盒, 宣传单, 纸杯等
木头	木制筷子等
食品残余物	各种食物残渣

2.2 人群种类与所处区域

所处区域	人群种类
医院	病人,医护人员等
研究所	科研人员等
写字间	小型创业者等
住宅区	普通居民等
中学补课班	学生,老师等

3 外卖固体废弃物数据采集

3.1 采集方式

为了获取准确的第一手资料,收集数据采用实地采集的方法.因为在垃圾收集处,外卖垃圾与其他各种垃圾混合在一起,很难准确测量出仅属于外卖的垃圾质量.所以,我们采用单位骑手的外卖垃圾质量乘以外卖骑手数量作为外卖质量的估计值.对于单位骑手外卖垃圾质量,我们将通过模拟预定外卖,来测量每位骑手所带来的外卖垃圾质量.

3.2 居民区内外卖固体废弃物采集方法

鉴于外卖在居民区内主要分布于12点与17点,故将24小时分为如下三个时间段:8:00-12:30,12:30-17:00,17:00-次日8:00. 假设此三个时间段的外卖骑手数量相等.在此三个时间段内,选取10:00-10:30,12:00-12:30,14:30-15:00,16:30-17:00作为采样时间段,准确记录采样时间段内的外卖骑手数量,并加以计算分析.

3.3 各人群种类所在地区外卖固体废弃物采集方法

通过网络及实地调查,选取有代表性的人群种类所在地区,确定所在地区能使外卖骑手进入的大门数量,分散组员至各大门来记录7:30-19:30 每一个小时内各门外卖骑手的进入数量,准确测得数据后进行统一的数据汇总.

4 分类假设

1. 假设外卖固体废弃物仅分为如上类别,忽略其余垃圾种类.
2. 假设人群种类仅分为如上类别,忽略其他人群种类.
3. 假设所采集数据的地方具有强代表性,可以代表其他类似场所.
4. 假设所调查的地方每日所产生的外卖固体废弃物质量相同.
5. 假设模拟预定的外卖固体废弃物质量经过平均计算,可以代表每位骑手所带来的外卖固体废弃物质量.
6. 假设所有测量数据均没有测量误差

5 问题一的解决

5.1 区域概述

我们选定了自己的小区[4]为测量区域,该小区处在市中心附近,居民以已成家的家庭为主,容积率适中,如下是具体参数:

总户数	542
建筑面积	66500m^2
容积率	4.6
占地面积	14456.52m^2
绿化率	30%

5.2 估算每位骑手所带来的外卖固体废弃物质量

为估计每位骑手所带来的外卖垃圾质量,我们订了不同种类的外卖,分别称量种类不同的固体废弃物的质量,经多次称量,得出如下数据(单位:g):

外卖垃圾种类	塑料	纸	纸皮	食品残余物	木头
快餐:汉堡,圣代	28	6	22	8	0
盒饭:普通饭菜	46	0	0	71	5
面类:米粉	45	2	0	408	5
平均质量	40	3.6	7.3	162	3.3

5.3 估算区域内每天总外卖骑手数量

通过既定的外卖骑手数量统计方法,具体采集第一手数据如下(单位:辆):

10:00-10:30	1
12:00-12:30	4
14:30-15:00	3
16:30-17:00	0

5.4 估算区域内每天总外卖固体废弃物质量

每一时间段内采样1小时,所以总外卖垃圾质量$M = m * \frac{9}{2} * \frac{3}{2}$.

使用如上公式估算得(单位:g):

塑料	纸	纸皮	食品残余物	木头
2160	194.4	394.2	8748	178.2

6 问题二的解决

6.1 寻找问题

首先在网上寻找大量有关垃圾处理方式的资料,并对小区垃圾桶内的垃圾进行了跟踪调查.通过调查,我们发现外卖垃圾和其他垃圾一样，均需要经过产生,投放,收集,运输和处置5 个重要环节.所以外卖垃圾的主要流程为:投放到小区垃圾桶,由小区保洁员负责进行收集,送到生活垃圾压缩站点,运送到填埋场进行无害化处理.在这些过程中,我们发现了一些问题:

6.1.1 分类不彻底

外卖所产生的固体废弃物有很多种类,根据之前所述,可以将外卖固体废弃物具体分为5类.但是我们发现大部分居民没有对它们进行细致的分类,这导致垃圾混合收集,增大了处理难度,增加了处理成本,使外卖垃圾无法实现资源化,减量化.

6.1.2 处理方法单一,且方法不妥当

我市垃圾处理主要是靠卫生填埋 [5],这些固体废弃物中的大部分没有被相关人员分类,换言之,也就是没有回收利用,仅仅是被填埋.这加大了对部分外卖垃圾资源的浪费.同时,错误的垃圾处理方式也会产生反面效果,混合在一起的垃圾并不能保证垃圾的针对化处理.举例而言,混在一起的塑料遭到焚烧或者填埋,这会对空气和土壤造成不同程度的危害[6].

6.2 优化垃圾处理方案

基于上述问题,经过我们不断的讨论和对相关人员的询问,得到对外卖垃圾处理并回收利用的更优化可行的方式:

1. 从生产过程中阻止垃圾的产生.在设计和生产过程中要尽量使用更少的材料,减少不必要的包装.

2. 对外卖垃圾进行严格的分类处理,这样将减小对它们的处理难度,并且增大回收利用的效率。

3. 对物料的不经过再生产而直接重复利用,或是经过再生产的循环利用,将垃圾转化为新的物料或产品.

4. 在最终处置上,采取单纯的不发电焚烧.

7 问题三的解决

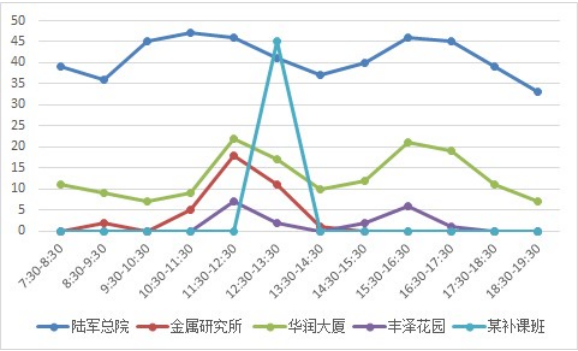
7.1 数据采集

通过既定的方法,选取中国人民解放军沈阳军区总医院 [7],中国科学院金属研究所 [8],沈阳市华润大厦 [9],丰泽花园与某补课班分别作为医院,研究所,写字楼,住宅区和中学补课班的具体场所,实地采集数据如下(单位:辆):

	陆军总院	金属研究所	华润大厦	丰泽花园	某补课班
7:30-8:30	39	0	11	0	0
8:30-9:30	36	2	9	0	0
9:30-10:30	45	0	7	0	0
10:30-11:30	47	5	9	0	0
11:30-12:30	46	18	22	7	0
12:30-13:30	41	11	17	2	45
13:30-14:30	37	1	10	0	0
14:30-15:30	40	0	12	2	0
15:30-16:30	46	0	21	6	0
16:30-17:30	45	0	19	1	0
17:30-18:30	39	0	11	0	0
18:30-19:30	33	0	7	0	0

7.2 模型建立思路

通过实地采集各地外卖骑手数量,将所得数据汇成折线统计图:



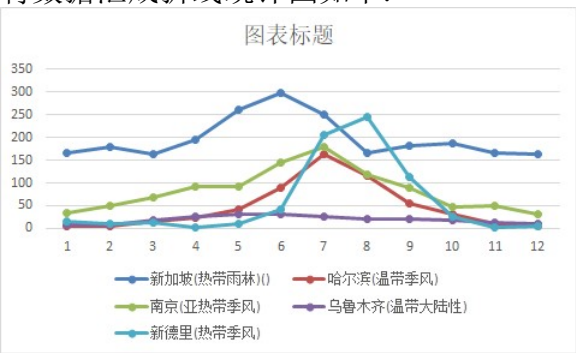
观察图像,发现各区域的每小时外卖骑手数量分布特点与一些气候类型所对应的每月平均降水量分布特点极为相似,故以相似为基础,将各个人群特点所处区域一一映射至如下的气候类型:

区域	数量特点	气候类型	降水量特点
陆军总院	数量较平均且均较高	热带雨林气候	全年多雨
金属研究所	在中午附近人群有所增加	温带季风气候	夏季湿润冬季干燥
华润大厦	数量较平均且中等	亚热带季风气候	全年湿润
丰泽花园	数量较平均且很少	温带大陆性气候	夏季湿润冬季干燥
某补课班	在中午附近有骤增	热带季风气候	旱雨季节分明

选取典型气候类型地区:新加坡,哈尔滨,南京,乌鲁木齐,新德里.收集历史平均月降水量数据 [10]如下(单位:mm):

	新加坡	哈尔滨	南京	乌鲁木齐	新德里
1	166	4	33	7	15
2	180	6	51	8	10
3	163	15	69	17	14
4	195	23	93	26	3
5	262	42	93	31	11
6	298	89	145	32	42
7	250	164	180	25	205
8	166	117	118	20	246
9	181	56	89	21	112
10	187	31	48	19	26
11	166	10	50	14	3

将数据汇成折线统计图如下:



由此可见,区域外卖骑手数量分布与与其相对应的地区降水量分布十分相似,故以此为基础建立模型.

7.3 模型目的

通过训练,模型能够接受一组从7:30-19:30每隔一小时的外卖骑手数量,输出五种人群特点的可信度.

7.4 数据收集及前期处理

7.4.1 数据收集

从GHCN数据库 [11],环境云 [12]和Global Weather Data for SWAT [13]上收集上述地区历年的降水量数据作为训练数据.选取一小部分数据作为降水量-气候类型验证数据,将区域外卖

骑手数量作为外卖骑手数量-人群特点验证数据.

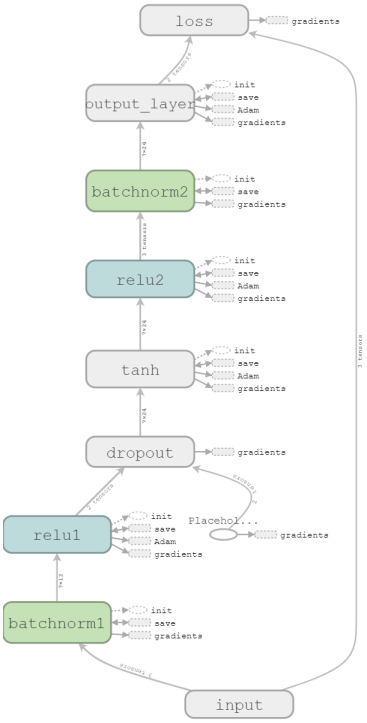
7.4.2 数据前期处理

由于降水量数据与外卖骑手数量量级并不相同.对于人眼,在观测折线统计图时,主要观测的是数据之间的相对大小关系以及数据(y) 随时间或者月份(x)的变化趋势.因此,为统一数据量级,防止深层神经网络模型出现学习方向错误.将每个数据除以12个单位时间或者月份(x),五种气候类型或者人群特点的总共60个数据的和,使降水量数据与骑手数据量级相似(使大部分数据在0-1之间).

7.5 模型建立

7.5.1 模型概述

此模型总共有六层,分别是:批规范层(Batch Normalization),全连接层,丢弃层(Dropout),全连接层,全连接层,批规范层.



7.5.2 输入层

输入层含有12个节点.与一年有12月相同,统计外卖骑手数量的时间也同样是12小时,这样就可以保证输入的维度相同.输入数据时,将每12个数据组成一个列表(List)并将其对应的答案放入验证列表中.

7.5.3 批规范层

深层神经网络模型强烈依赖于大量数据,相较于上百万的数据量,仅仅上万的数据量很容易导致模型过拟合.这样即使正确率很高,但模型实际正确率反而会降低.批规范层[14] 实现了对每批(Batch) 输入数据进行标准化处理,使其以较小的方差集中在均值附近,从而尽量防止出现梯度爆炸或是梯度消失的问题.

7.5.4 全连接层

本层全连接层含有24个节点,使用Relu作为激活函数,使用符合平均分布的随机数作为初始网络权重(最小值为0, 最大值为1).

7.5.5 丢弃层

同样是为了防止过拟合问题,丢弃层虽方法不同,但目的相似.丢弃层采用在网络中将某些节点随机赋值为0来防止过拟合,同时还可以节约训练时间.在这个模型中设置丢弃的比例为0.9.

7.5.6 全连接层

本层全连接层含有24个节点,使用Tanh作为激活函数,使用符合平均分布的随机数作为初始网络权重(最小值为0, 最大值为1).

7.5.7 全连接层

本层全连接层含有24个节点,使用Relu作为激活函数,使用符合平均分布的随机数作为初始网络权重(最小值为0, 最大值为1).

7.5.8 批规范层

为了防止过拟合,再次加入批规范层.

7.5.9 输出层

基于此模型的目的,输出是人群种类的置信度.所以,为保证输出是一个概率小数,即输出在0和1之间.把原本的输出经过Softmax回归处理后,再作为输出给损失函数.

7.5.10 损失函数

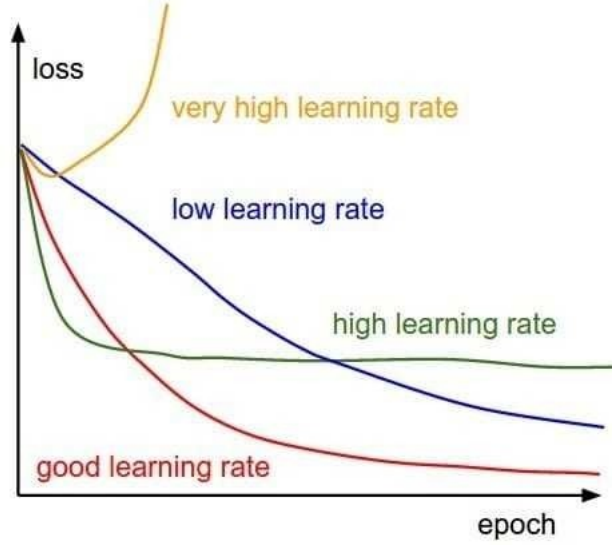
对于深层神经网络,优化器(Optimizer)需要得到一个数值进行反向传播来调整参数.交叉熵能够描述模型输出与正确答案之间的距离,故此模型使用交叉熵作为损失函数.

7.5.11 分类误差

与损失函数不同,分类误差(Classification Error)是对分类答案的确定误差.由每一个输入数据得出的输出,选择其中数值最大的答案(即可能性最大),用正确答案评判对错.与交叉熵不同,分类误差对于每一个输入仅能得出0 或1的结果.在实际模型训练中,把在一组验证中所有结果的平均值作为分类误差.

7.5.12 学习率

在深层神经网络模型训练中,学习率对于模型的训练速度和结果有很大的影响.



如图可见,优秀的学习率可以使模型的损失,以适当的速度降到最低值.相对的,过高的学习率可能会导致损失爆炸或者无法收敛到最小值,过低的学习率则会导致训练速度过慢,浪费训练时间.此模型经过不断的尝试,经过多次实验比对,确定初始学习率为0.035. 随着训练次数的增多,损失逐渐逼近极值,为防止误差错过极值点,采用指数衰减的方式,在达到规定的训练次数时,通过公式

$$R_1 = R_0 * decay\_rate^{global\_step/decay\_step}$$

,以特定的指数次幂降低学习率.经过多组控制变量的实验,此模型中decay_step设置为100,decay_rate设置为0.96.

7.5.13 优化器

深层神经网络模型使用梯度下降算法进行模型训练.在普通的梯度下降算法基础上,为进一步优化梯度,模型一般使用优化器来优化梯度下降.经过多种优化器的尝试,此模型采用了自适应矩估计(Adaptive Moment Estimation[15])优化器,即Adam优化器.Adam优化器能

够在给定学习率的基础上,为每个参数计算自适应的学习率.相较于随机梯度下降(SGD)优化器,Adam优化器梯度下降速度更快,并且很少出现梯度消失的情况.

7.6 模型训练

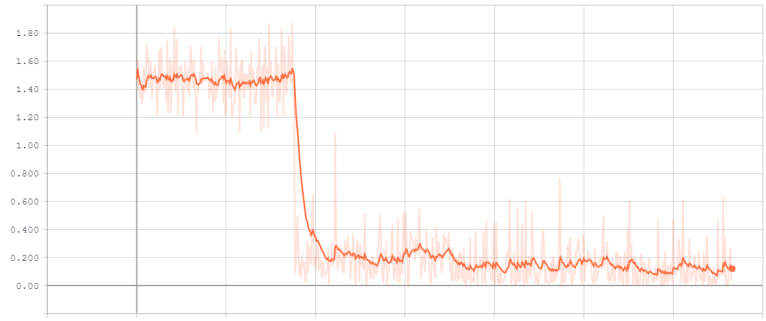
7.6.1 训练参数

经过多次实验比对,此模型采用全部数据以每10个为一批(Batch) 输入至网络,训练时将全部数据完整地训练100 次(epoch=100)的方法.将模型误差,降水量验证结果与外卖骑手数量验证结果输出.

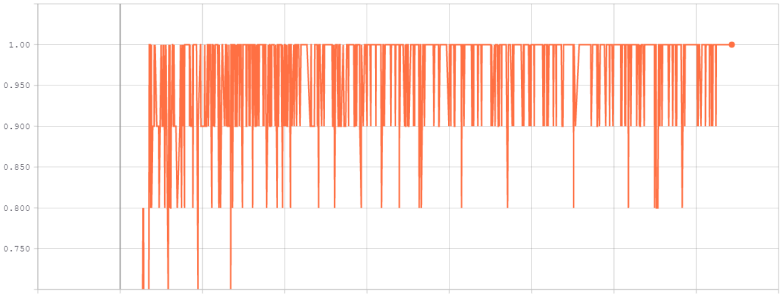
7.6.2 训练结果

训练后,保存训练日志.使用可视化工具TensorBoard[16]查看模型误差,并且使用TensorBoard画出降水量验证结果与外卖骑手数量验证结果折线图.

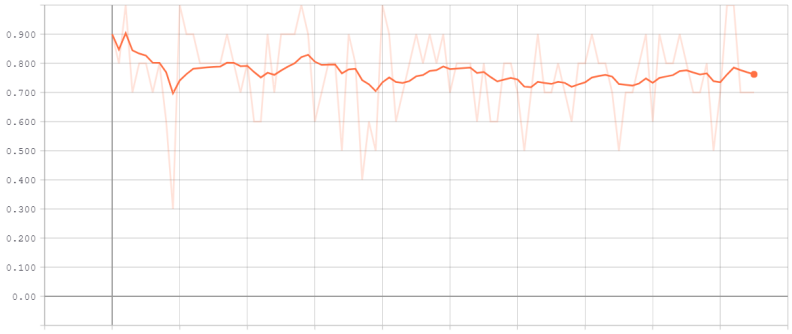
降水量模型训练误差如下(Smoothing=0.9):



降水量分类正确率如下(Smoothing=0)



外卖骑手数量分类正确率如下Smoothing=0.9



从图像中得出,经过100轮(epoch)的训练,降水量模型的训练误差在0和0.2之间,降水量的分类正确率几乎达到了100%,外卖骑手数量的分类正确率在75%到80%之间.

7.7 模型结论

基于所处地人群特点与气候类型,所处地12个小时外卖数量与特定气候类型月平均降水量的相似性,建立深层神经网络模型,通过不断的调整超参数,得出较为完美的模型,再进行适度的训练,使得此模型满足模型目的,即,输入12个小时的外卖骑手数量,可以得出一个正确率在75%到80%的人群特点.

7.8 附件文件结构

main.py:主程序

input.csv/ninput.csv:输入与清洗后的输入

val.csv/nval.csv:验证集与处理后的验证集

已经训练好的模型因为超过大小限制,以下部分可以从https://github.com/xingchengyi/modeling查看

prcp_log/prcp_save:降水量-气候类型模型训练日志/保存好的模型

takeaway_log/takeaway_log:外卖骑手数量-人群种类模型训练日志/保存好的模型

8 问题四的解决

外卖,降水量与人工智能的一碗汤

外卖,从专业人员才有所了解,到如今的遍布全中国的情景,可能才经历不过几十年,便利作为主要因素促使外卖不断发展进步.

但

殊不知,手中轻轻点下的确认,正在给这个世界涂上一笔白色.

那是一些比钻石还要永恒的东西.

不仅仅是塑料,外卖垃圾中的可以回收的木头,纸类,因为没有被回收而被压埋在无名的填埋场,抑或是飞扬在焚烧厂的浓烟中.

可能是没有分类或是并不看重这一些垃圾.但是,积少成多,外卖垃圾已经逐渐登上了固体废弃物火车的特等舱.

所以,在这里,我们的模型可以为这个世界做出一些贡献.

我们研究了外卖垃圾数量与人群种类的关系.

开始建立模型非常复杂,直到看到了那张折线统计图.像是有什么魔性的一样,从外卖垃圾中看出来了降水量的影子.依稀记得地理老师曾经讲过些什么"高温多雨","旱雨季节分明"这类的东西.没多过脑子便去网上四处查询降水量的数据.当把两张表对比之后,便明确了研究方向.

"我要从降水量中研究外卖数量"

先是寻找海量的数据.有了数据,倒是能研究什么呢?

idea真重要.

从图书馆回来,借了大大小小好几本人工智能的书籍,抱着开卷死不了人的心态,读了下去.

一本又一本

有趣的时间转眼就过去了,基本方向也明确了,艰苦奋斗的时候到了.

整夜整夜地调整模型结构.

看着梯度消失的loss愤恨地抛弃掉这任Optimizer.

或是同时开着几个页面同时确认节点个数

抑或是写着无奈的脚本疯狂地调着学习率

直到

”模型收敛了”

安静.

死一般的安静.

模型终于成功了,宛如伊丽莎白的国土续命论一样不可思议.

愉悦,激动,紧接着的是反思与记录

可能这只是一点微小的模型,但是如果真的有人在我们的模型上进行再研究,

外卖垃圾的现状可能真的会有一些改变吧.

9 参考文献

1. 中华人民共和国国家统计局. 生活垃圾清运量(万吨)[DB/OL].http://data.stats.gov.cn/easyquery.htm?cn=C01&zb=A0B09&sj=2016.

2. 解放日报. 外卖垃圾统计[DB/OL]. http://society.people.com.cn/n1/2017/0908/c1008-29523150.html.

3. 第一财经日报. 垃圾年产量增速与GDP匹敌 每年经济损失达300亿[EB/OL]. http://env.people.com.cn/n/2014/0521/c1010-25043415.html.

4. 百度文库. 丰泽花园百度百科[DB/OL]. https://baike.baidu.com/item/%E4%B8%B0%E6%B3%BD%E8%8A%B1%E5%9B%AD%E4%B8%80%E3%80%81%E4%BA%8C%E6%9C%9F/3777352?fr=aladdin.

5. MIT, sectors, solid-waste-landfills. What is a Sanitary Landfill?[EB/OL]. http://web.mit.edu/urbanupgrading/urbanenvironment/sectors/solid-waste-landfills.html.

6. 沈阳市规划设计研究院. 沈阳市垃圾分类及焚烧厂建设探析[EB/OL]. http://huanbao.bjx.com.cn/news/20171107/860012-2.shtml.

7. http://www.syjqzyy.com/

8. http://www.imr.cas.cn/

9. http://www.themixc.com/

10. https://en.climate-data.org/

11. https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GSOM

12. http://www.envicloud.cn/

13. https://globalweather.tamu.edu/

14. arXiv:1502.03167 [cs.LG]

15. arXiv:1412.6980 [cs.LG]

16. https://www.tensorflow.org/programmers_guide/graph_viz