

P3 - 情感细粒度分类 指导文件


开课吧人工智能学院

2019-10-13

目录

1. 问题背景描述
2. 所需要的数据与输入输出定义
3. 所需要的相关环境
4. 相关的技术栈
5. 数据可视化建议
6. 总结与建议
7. 如何提交项目
8. 优秀学员奖励

01. 问题背景描述

- 
- ▶ 在自然语言处理中，有一个常见的问题就是对客户的评价进行分析。这些用户评论中，包含了大量的有用信息，例如情感分析，或者相关事实描述。例如：

“味道不错的面馆，性价比也相当之高，分量很足～女生吃小份，胃口小的，可能吃不完呢。环境在面馆来说算是好的，至少看上去堂子很亮，也比较干净，一般苍蝇馆子还是比不上这个卫生状况的。中午饭点的时候，人很多，人行道上也是要坐满的，隔壁的冒菜馆子，据说是一家，有时候也会开放出来坐吃面的人。”

- ▶ 首先情感是正向的，除此之外我们还能够进行知道这个的几个事实描述：1. 性价比比较高；2. 装修比较好；3. 分量足。

用户评价

**梦**r**

05-15 口味:香辣250g+五香250g

东西质量非常好,很好吃的。与卖家描述的完全一致,非常满意,真的很喜欢,完全超出期望值,发货速度非常快,包装非常仔细、严实,物流公司服务态度很好,运送速度很快,很满意的一次购物。



正宗网红麻辣牛肉干 1斤装四川500G内蒙古散装零食手撕风...

¥ 178.00 已好评

知乎 @子子

网友点评(112) 搜索评论

大家认为

菜品健康(24) 回头客(3) 干净卫生(5) 分量足(7) 味道赞(28)

环境优雅(14) 服务热情(51) 价格实惠(8) 性价比(5)

停车信息(1) 有图片(59) 4星

**影随心尚**   口味: 2 环境: 2 服务: 2 人均: 120

出去旅游了回来一般都是顿火锅伺候😂火了一段时间的“小龙坎”层出不穷的出了很多同名的,不晓得哪个才是正宗地道的,朋友住在这附近就选了这家,去晚了吃的第二轮,火锅的魅力真的是无以抵挡,尽管都知道这种美食并不怎么健康,但还是忍不住要吃。这家味道不错,虽然是重庆老火锅但味道还是偏香辣,还放了酒糟提味,菜品也比较新鲜,就是肥牛太肥了,太饿了好多菜品忘了拍照。饭后还赠送了自制的冰激淋,由于生意太火爆服务有些跟不上,总体还是可以的。



收起 ^

06-02 姚记小龙坎老... 赞(1) 回应(1) 收藏 举报

商家回应

06-04 20:59:26

情感识别分类的应用场景非常多

- 这些信息是非常重要的，不论是对于公司进行商业分析或者要建立一个搜索引擎排序，这些信息都是重要的参考因素。那么在这个时候，我们就需要进行文本的情感分类了



KUTA'S KITCHEN(淮海中...

订 惠

★★★★☆ ¥207/人

音乐学院 日式烧烤/...

380m

别名:披头士串烧日本料理



BONOBO

团 订 惠

★★★★☆ ¥154/人

音乐学院 西餐

8.9折起

440m

别名:吃喝玩乐 百乐吧



小巴黎花园咖啡馆

团 订 惠

★★★★☆ ¥69/人

音乐学院 西式简餐

4.9折起

250m

别名:Petit Paris



V Cafe (Sasha Garden)

团 惠

★★★★☆ ¥47/人

音乐学院 咖啡厅

6.6折起

130m



辣螺重庆美蛙火锅

团 促 订

★★★★★ ¥100/人

5.6折起

2. 所需要的数据与输入输出定义

2. 所需要的数据与输入输出定义

- ▶ 这个问题我们希望的是，输入一句话，输出是这句话对于以下6大类，20小类进行打标，对于每个小类而言，都会有<正面情感, 中性情感, 负面情感, 情感倾向未提及> 这4个类别。
- ▶ 总得来说，我们现在这6大类，20小类的类别如下：
- ▶ （见下页）

2. 所需要的数据与输入输出定义

- ▶ 位置(location)
 - ▶ 交通是否便利(traffic convenience)
 - ▶ 距离商圈远近(distance from business district)
 - ▶ 是否容易寻找(easy to find)
- ▶ 服务(service)
 - ▶ 排队等候时间(wait time)
 - ▶ 服务人员态度(waiter's attitude)
 - ▶ 是否容易停车(parking convenience)
 - ▶ 点菜/上菜速度(serving speed)
- ▶ 价格(price)
 - ▶ 价格水平(price level)
 - ▶ 性价比(cost-effective)
 - ▶ 折扣力度(discount)
- ▶ 环境(environment)
 - ▶ 装修情况(decoration)
 - ▶ 嘈杂情况(noise)

2. 所需要的数据与输入输出定义

- ▶ 而为了方便训练数据的标注，训练数据中，<** 正面情感, 中性情感, 负面情感, 情感倾向未提及 > ** 分别对应与 (1, 0, -1, -2).
- ▶ 例如说，“味道不错的面馆，性价比也相当之高，分量很足～女生吃小份，胃口小的，可能吃不完呢。环境在面馆来说算是好的，至少看上去堂子很亮，也比较干净，一般苍蝇馆子还是比不上这个卫生状况的。中午饭点的时候，人很多，人行道上也是要坐满的，隔壁的冒菜馆子，据说是一家，有时候也会开放出来坐吃面的人。”

2. 所需要的数据与输入输出定义

这句话在训练数据中的标签就是：

- 交通是否便利(traffic convenience) -2
- 距离商圈远近(distance from business district) -2
- 是否容易寻找(easy to find) -2
- 排队等候时间(wait time) -2
- 服务人员态度(waiter's attitude) -2
- 是否容易停车(parking convenience) -2
- 点菜/上菜速度(serving speed) -2
- 价格水平(price level) -2
- 性价比(cost-effective) 1
- 折扣力度(discount) -2
- 装修情况(decoration) 1

2. 所需要的数据与输入输出定义

训练数据在哪里：

- 该数据集合大众点评公开给创新工厂的2018 AI全球挑战赛的数据集，因为开课吧和创新工厂工程院的合作，在数据下线之后，创新工厂将数据分享给我们。
- 这个比赛最终的第一名f1 score只有0.76，大家可以以此为标准，观察自己模型的表现
- 数据存放在服务器: student@39.100.3.165/dataset 中，请大家下载或者直接在远程服务器进行使用

3. 所需要的相关环境

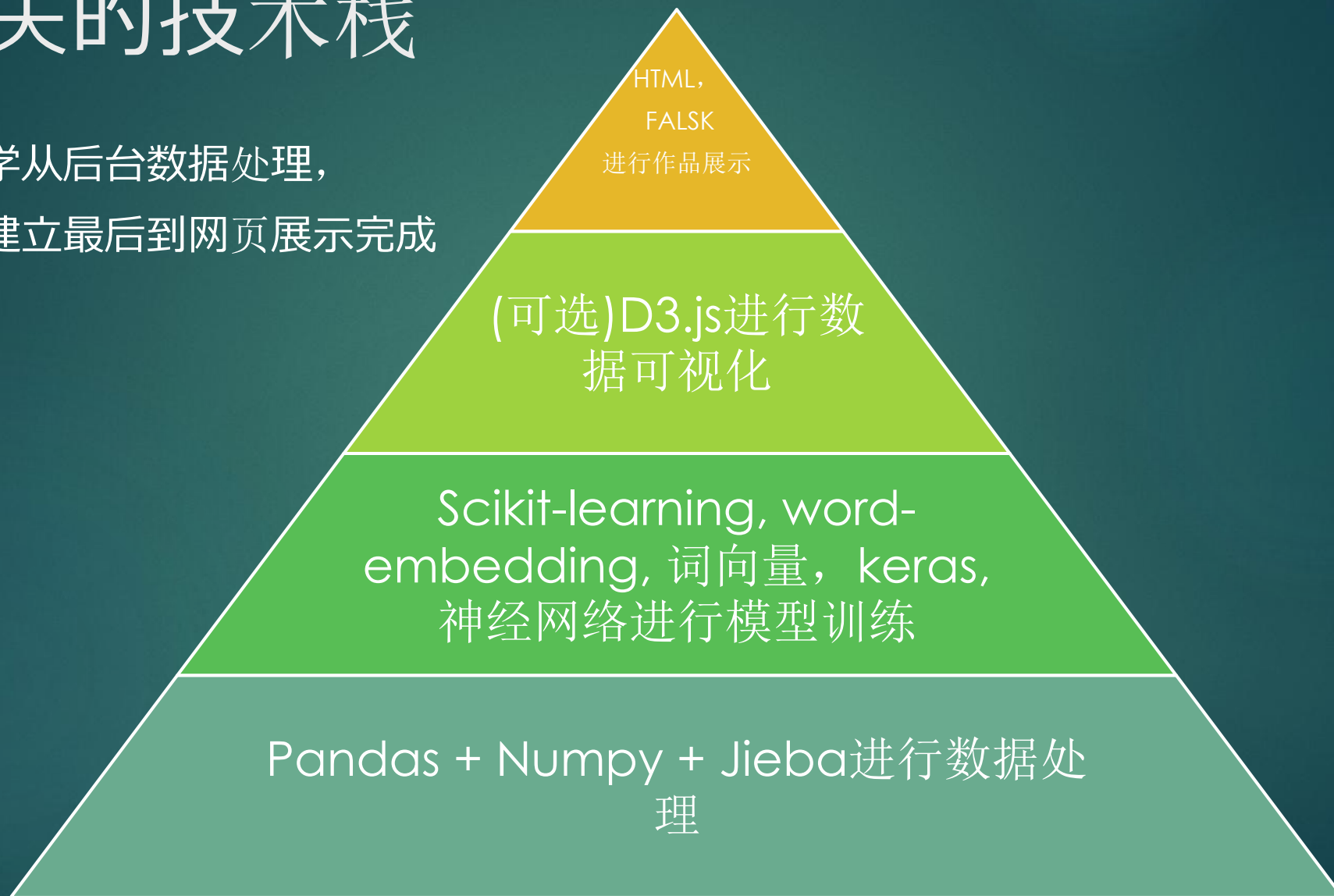
- ▶ 本次项目用到的环境为:
 - ▶ Python3.6, Pycharm, Jupyter Notebook
 - ▶ Keras, Tensorflow, Pandas
 - ▶ Linux Ubuntu 服务器
- ▶ Ubuntu服务器的用户名和密码为:
 - ▶ student@[39.100.3.165](#)
 - ▶ AI@2019@ai

2. 所需要的数据与输入输出定义 - Evaluation

- 你现在需要考虑一个问题，如何定义该模型的表现：
 - Precision? Recall? AUC? 还记得这几个有什么区别吗？
 - Baseline是多少？例如，5分类的问题，那么准确率最低也应该是20%（为什么？）
 - 在Keras里边，选择合适的metric，然后观察validation和training的变化，就可以知道我们的结果了
 - 当你知道了输入输出 + 模型的Evaluation Metric 你就可以动手Coding了

4. 相关的技术栈

- ▶ 需要同学从后台数据处理，
- ▶ 到模型建立最后到网页展示完成
- ▶ 该项目

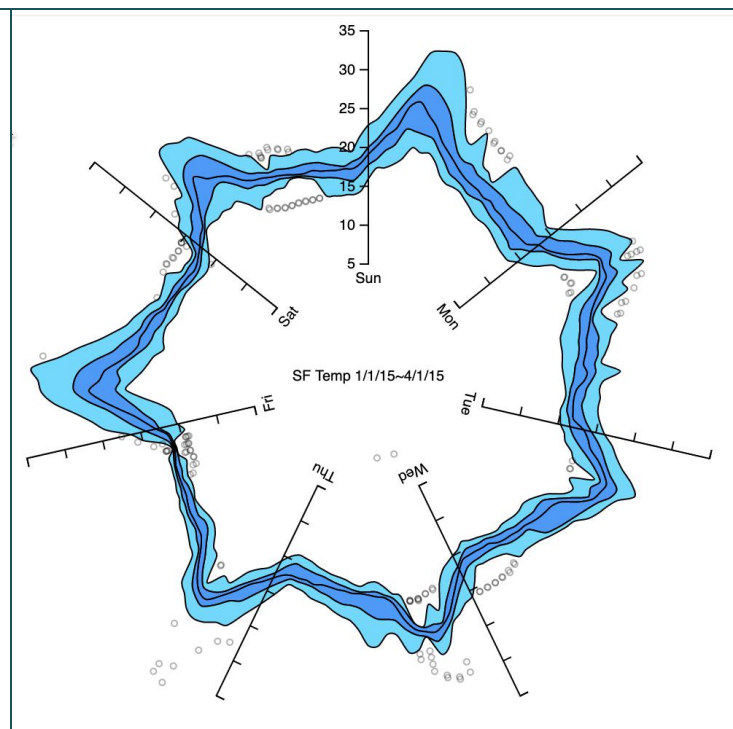
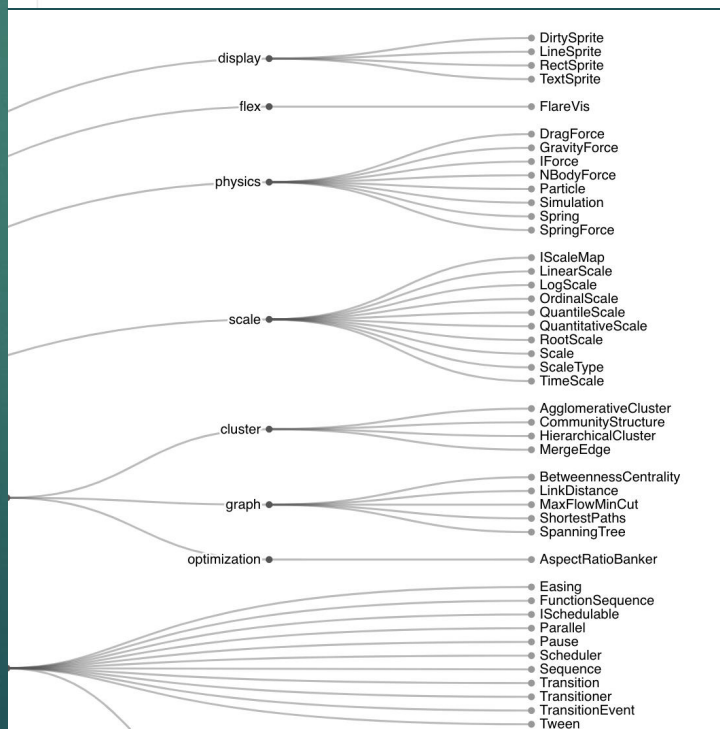


5. 数据可视化建议

- ▶ 良好的可视化是展示我们的能力的一个非常好的帮手，很多同学的代码往往停留在“黑框框”的Terminal里边，这对于展示自己实力是很不好的。
- ▶ 我们之前给大家介绍了bottle, Flask, Bootstrap等工具，这些是非常好的，但是，对于这个项目，维度这么高的分析，光光有一个页面还不够。
- ▶ 我们给大家推荐D3库，这个库是数据可视化最广泛的库了。
- ▶ <https://github.com/d3/d3/wiki/Gallery>
- ▶ 可以用在我们这个项目中的可视化工具有：
 - ▶ <https://bl.ocks.org/mbostock/4061961>
 - ▶ <https://observablehq.com/@d3/cluster-dendrogram>
 - ▶ <https://bl.ocks.org/davidwclin/ad5d13db260caeffe9b3>
- ▶ D3的使用需要对Javascript有所了解，大家如果自己不能完成，请在群里找到合适的队友一起完成。

5. 数据可视化建议

右边的图是我们给到大家的D3实例，大家如果熟悉JS，那么把里边的数据语言进行修改，就能得到这样的可视化图形了，这些图形都是可以运行在HTML页面的。



6. 总结与建议

- ▶ 1. 看到这个项目，我们首先想到的是，这个Case和之前Kaggle，豆瓣中遇到的问题比较相似，那么，我们首先要基于前人的经验进行复现；
- ▶ 2. 基于前人的经验和模型，修改自己的代码，能够end-to-end完成项目初版；
- ▶ 3. 修整、改变此模型，从词向量、模型结构等层面进行分析。Overfitting, Underfitting?
- ▶ 4. 依照自己所观察到的欠拟合、过拟合现象，进行模型调整；
- ▶ 5. 模型调整，不是眯着眼，随机瞎调参，要有一个 分析现象 -> 提出假设 -> 改变模型做实验 -> 观察是否符合期望这样一个过程

6. 总结与建议-2

例如，我们可以按照以下的类似思路进行调参

- ▶ 第1次优化:
- ▶ 存在的问题: loss下降太慢;
- ▶ 准备进行的优化: 减小模型的神经元数量;
- ▶ 期待的结果: loss下降加快;
- ▶ 实际结果: loss下降的确加快(或者并没有加快)
- ▶ 原因分析: 模型神经元数量减小, 收敛需要的次数减少, loss下降加快
- ▶ ——你的实验优化结构记录在此——
- ▶ 第1次优化:
- ▶ 存在的问题:
- ▶ 准备进行的优化:
- ▶ 期待的结果:
- ▶ 实际结果:
- ▶ 原因分析:

7. 如何提交项目

- ▶ 提交项目应该是一个压缩包，该压缩包包含以下内容：
 - ▶ 1. 项目源代码（不需要包含数据）
 - ▶ 2. 项目的PPT效果展示
 - ▶ 3. 你的参数调整记录表
 - ▶ 4. 该项目能够访问的网站链接
 - ▶ 5. 该项目的优缺点和模型分析报告，包含模型的 precision, AUC, recall等关键性指标
- ▶ 之后将该Zip压缩包以发送至 ai-college@kaikeba.com
- ▶ 项目接受截止日期：201911月2日

8. 优秀学员奖励

- ▶ 这是我们第 3 次项目，15天之后第 4 个项目“面向服务的对话机器人”即将开始。我们按照 1 – 5分给每位同学每个项目打分，4个项目的占比分别是：
 - ▶ 20%， 20%， 30%， 30%
 - ▶ 团队成员以团队整体成绩为计算
- ▶ 对综合排名前6名的同学，我们提供以下奖励：
 - ▶ 第 1 名： Kindle Oasis 阅读器（或同等价值奖品）
 - ▶ 第 2， 3 名： Kindle Paperwhite 阅读器（或同等价值奖品）
 - ▶ 第4， 5， 6 名： Lamy钢笔 + 精选图书一册
 - ▶ 此6名同学可直接获得阿里巴巴（蚂蚁金服），百度，字节跳动，微软，IBM内推机会，



好了，大家加油吧！