Project lecture

Amazon Food Review

Key Point

1.项目介绍

2.面试表述

3.项目答疑

1

项目介绍

- 项目基本介绍
- 项目目标要求
- 项目分析过程

项目基本介绍

项目背景

此项目（Project: Amazon Food Review ）涉及到自然语言分析相关内容，这是当下各个大公司的面试热点。这个知识点相对其他内容较为重要。因此，建议在平时强化编程方面的训练，不断增强对代码的理解，才能在该项目中增强对不同情境下的语言分析的技巧的深刻感悟。本项目的数据是亚马逊食品评价的语料。根据丰富的语料，可以对语义进行正面，负面的分析。然后根据已经学习过的机器学习进行分类。其次，结合在课上学过的对模型的评价指标来对不同的模型的精准度，自行进行鉴别。通过对模型的多次训练，来加深对数据建模的切割。

项目基本介绍

项目实用场景

情感分析并不只是对一段文本进行情感正负面的分类。情感分析在业界一般也称为观点挖掘（opinion mining），它是一系列技术概念的总称，旨在分析人类对一个目标对象所产生的情感极性（sentiment/valence）、情绪（emotion）、评价（evaluation）、态度（attitude）等等，目标对象包括但不限于商品、服务、组织、个人、事件等等。情感分析在不同场景可能表示不同的技术应用，例如情绪识别、情感分类、观点挖掘、观点分析、观点抽取、主观分析、情感计算、评价分析等等。综合来说，情感分析是分析人类对某个目标对象所蕴含的观点。

项目目标要求

基本要求

该项目重在掌握机器学习建模和数据可视化分析技能。主要是以亚马逊食品评价语料作为数据来源，基于词向量化、词云可视化、分类模型构建，通过ROC、AUC、 Precision、Recall、F1 score模型指标评价合理选择模型等分析方法理解自然语言处理过程，深入掌握数据建模与可视化的相关技能。

项目目标要求

具体任务目标

1.增强对自然语言模型代码的使用。理解对语料的各类必要处理。例如去除语气词，去除停用词，去除不必要的标点符号。推荐使用较为先进的处理包，通过数据预处理工具Texthero的使用增强数据分析能力。

项目目标要求

具体任务目标

2.推荐使用词云可视化word cloud, 学会使用词语在图表中的展示方式来分析词语的频率，词语的重要性，词语的情感，以及学会通过对词语的分析商业决策。

项目目标要求

具体任务目标

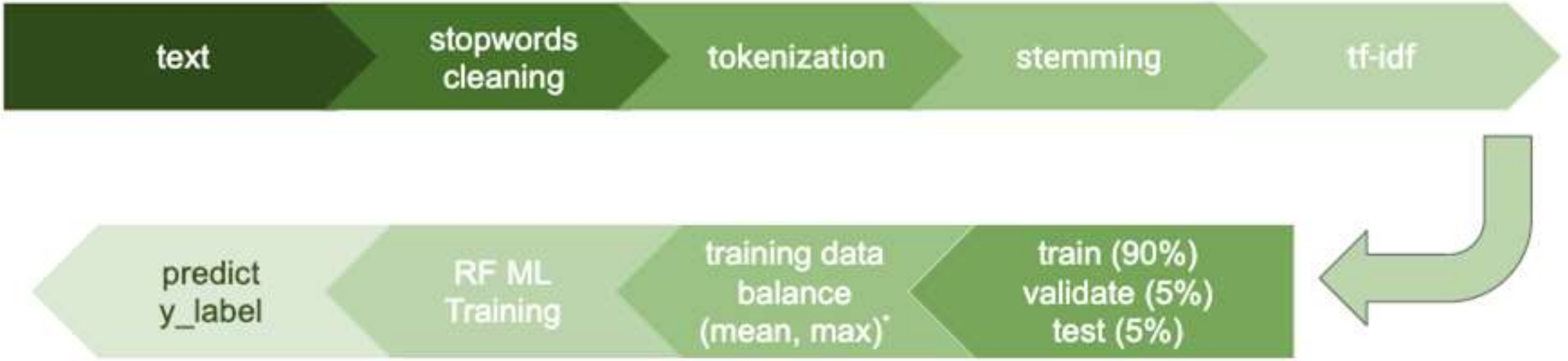
3.进一步强化并使用随机森林和逻辑回归模型。理解如何使用向量化来进一步分析自然语言语料。

项目目标要求

具体任务目标

4.强化对模型选择的实战经验。加强对ROC，AUC, Precision, Recall, F1 score在实际应用中的理解。

项目分析过程


text
stopwords cleaning
tokenization
stemming
tf-idf
train (90%) validate (5%) test (5%)
training data balance (mean, max)
RF ML Training
predict y_label

2

面试表述

项目考察重点

项目应用实例

项目面试表达

项目考察要点

重难点的应用实例

- NLP 的数据处理一般比较标准化，是种常见的面试类型
- NLP 模型的难点在于知道在何种情况下进行相对应的数据处理，并不能完全照章办事
- 在需要挖掘关键词的时候可能需要比较多种模型的输出，选择最有意义的结果

项目应用实例

重难点分析

- 数据处理
  - 大小写转换：bag of word 模型需要，Deep learning 模型不需要（例如BERT）
  - stopwords removal：切记去掉跟模型应用相关的词
  - 如果条件允许，至少试一下bi-gram，或者tf-idf
- 模型训练
  - 试验多种模型。
- 模型选择
  - 模型选择时不光要考虑accuracy，还要考虑可解释性

项目面试表达

I built a sentiment model based on XX to understand the public opinion on products sold on Amazon. The project includes crawling custom review data from various sources, including Kaggle, Amazon API, etc. Initial statistics analysis brought insights on customers' preference, product quality, and how to improve products, etc. Furthermore, a sentiment model was built to predict users' rating based on their feedback for better vendor selection and product recommendation.

3

项目答疑

项目疑问答疑

项目总结部分

项目疑问答疑

常见疑问

- Summary vs textbody as input?
  - 如果内存允且文学不是特别长，尽量使用textbody
- Accuracy过高
  - 可能性1: 过拟合。
  - 可能性2: 检查不同情况下的precision和recall
- 模型选择
  - tree base的model或者SVM比较适合在小数据情况下的NLP classification

项目征集疑问

其他疑问

- 如何决定x-gram?
  - data size ~ 10 x feature size

- Production model application
  - 3 bins method: low -> negative, high -> positive, median -> neutral

项目总结部分

整体总结

总体而言，此项目（Project3: Amazon Food Review ）重点考察了机器学习和数据可视化等内容。这些内容是数据科学的核心内容，也是极其重要的体现实战能力的面试考点。本项目将会重点学习机器学习建模和数据的可视化分析，并将相关技能应用于开源的亚马逊评论数据，在此过程中深入并且灵活的掌握数据建模与可视化的相关技能。

谢谢！