SLDS 2025: Project2

Introduction to Project and Background Knowledge

2024.4.30

目录

1 Task1： 寿命预测建模

Task1. Life Expectancy (40 points)

In this task, we are working with a dataset that includes 12 different features of 211 countries with their corresponding life expectancy. The dataset is named `life_indicator_2008-2018.xlsx`. Our goal is to predict `Life Expectancy at Birth` using these features. The main task involves using data from 2008 to train a model and then predicting life expectancy for 2018 based on this trained model.

Task1: 寿命预测建模

- 农业、林业和渔业增加值（占GDP的百分比）
- 每年抽取淡水总量（占内部资源的百分比）
- 当前保健支出（占国内生产总值的百分比）
- …

预测 → **寿命**

	Country Name	Agriculture, forestry, and fishing, value added (% of GDP)	Annual freshwater withdrawals, total (% of internal resources)	Current health expenditure (% of GDP)	Forest area (% of land area)	GDP (current US$)	Immunization, measles (% of children ages 12-23 months)	Income share held by lowest 20%	Industry (including construction), value added (% of GDP)	Population, total	Prevalence of underweight, weight for age (% of children under 5)	Research and development expenditure (% of GDP)	School enrollment, secondary (% net)	Life expectancy at birth, total (years)
0	Afghanistan	29.297501	43.015907	9.818487	1.852782	1.241615e+10	60.0	NaN	21.897122	27385307	NaN	NaN	NaN	60.364000
1	Albania	16.794384	4.454944	5.727621	28.496095	1.204421e+10	97.0	NaN	24.413778	2927519	6.3	NaN	NaN	77.781000
2	Algeria	9.343365	64.729099	5.359398	0.791060	1.372110e+11	92.0	NaN	46.948177	35196037	NaN	NaN	NaN	73.620000
3	Angola	6.621197	0.476824	3.842608	58.324425	7.030720e+10	46.0	NaN	44.056374	22507674	NaN	NaN	9.63002	55.752000
4	Antigua and Barbuda	1.491637	9.038462	4.177328	20.106818	1.228330e+09	99.0	NaN	19.460936	84534	NaN	NaN	92.48712	76.669000
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
205	Virgin Islands (U.S.)	NaN	NaN	NaN	53.240000	4.201000e+09	NaN	NaN	NaN	108404	NaN	NaN	NaN	77.514634
206	West Bank and Gaza	10.021396	42.339491	NaN	1.638372	8.085700e+09	97.0	7.7	17.558158	3689099	NaN	0.35646	85.04096	72.608000
207	Yemen, Rep.	10.019070	169.761905	5.986200	1.039832	2.513028e+10	65.0	NaN	52.785666	24029589	NaN	NaN	NaN	67.196000
208	Zambia	11.552784	1.960100	4.426805	62.863100	1.532834e+10	90.0	NaN	30.224469	13318087	NaN	NaN	NaN	55.300000
209	Zimbabwe	10.742550	28.762235	NaN	46.403903	9.665793e+09	76.0	NaN	21.454292	12679810	11.7	NaN	NaN	48.063000

210 rows × 14 columns

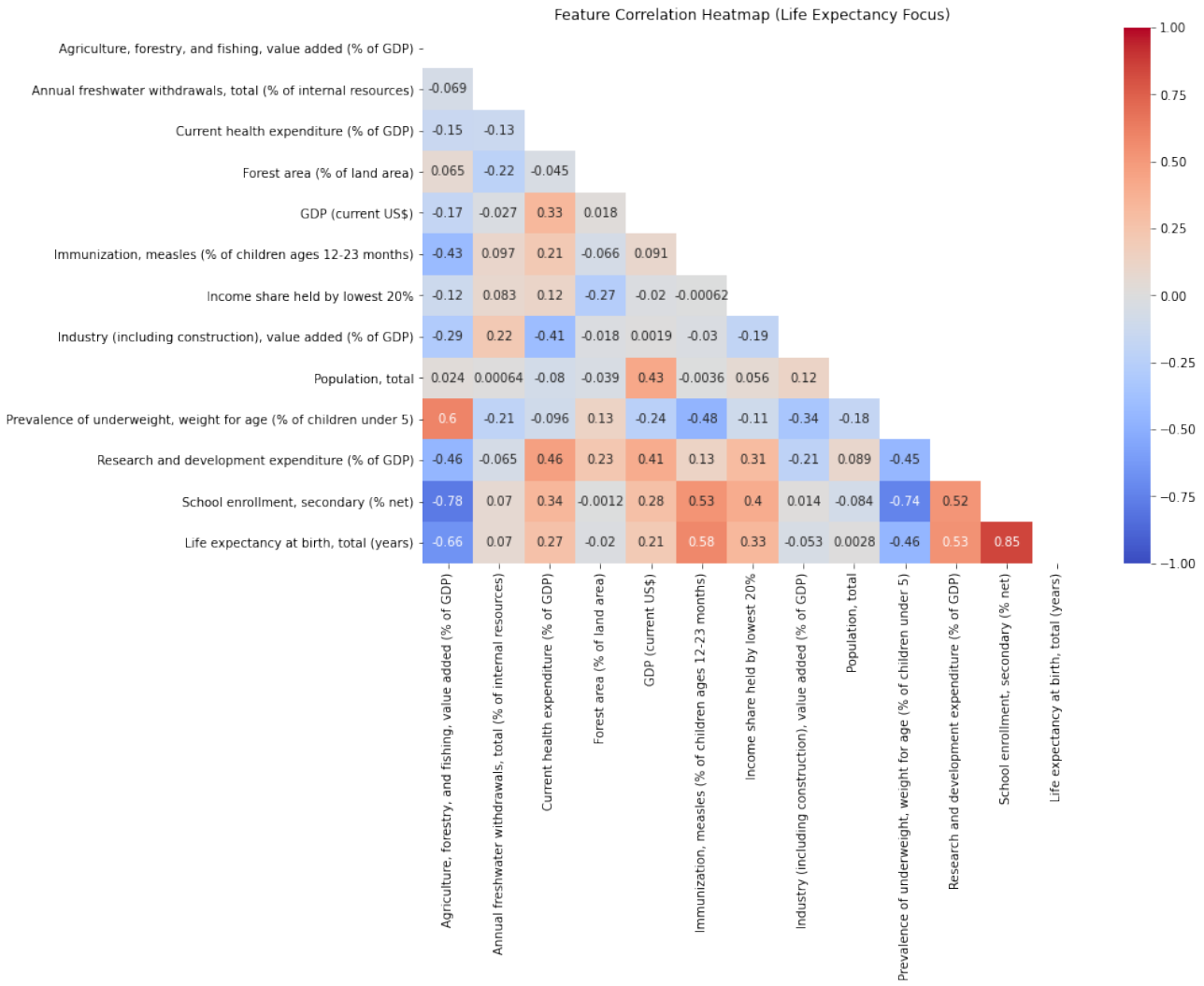
a. Data Understanding

- Understand each of the 12 features and make a guess which ones are likely to have a significant impact on life expectancy.

```
count    194.000000
mean      11.387296
std       11.445366
min        0.019907
25%        2.321262
50%        7.645451
75%       16.655000
max       58.035747
Name: Agriculture, forestry, and fishing, value added (% of GDP), dtype: float64
```

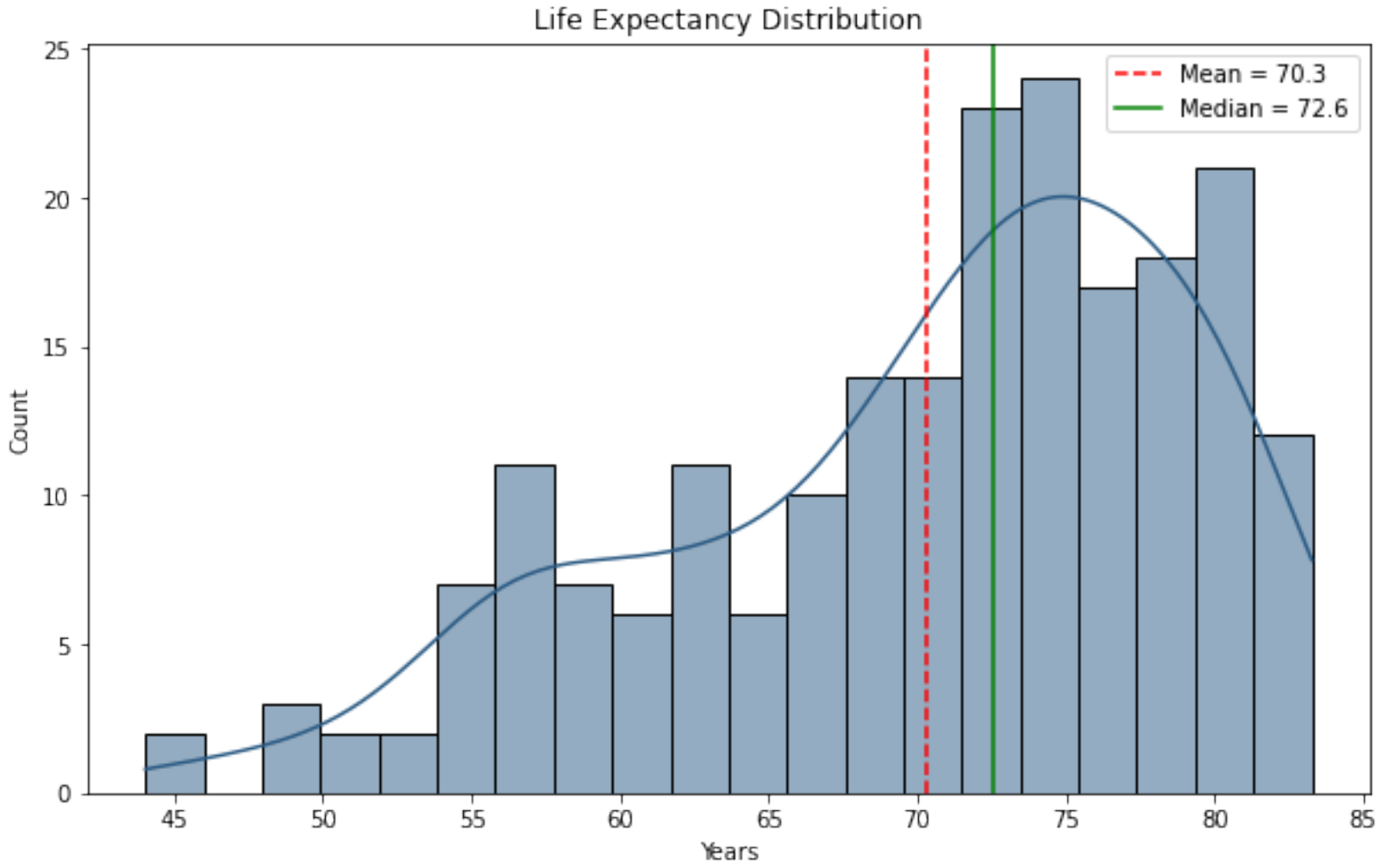
a. Data Understanding

- Visualize the relationships between the features using a heatmap to see how they correlate with each other.



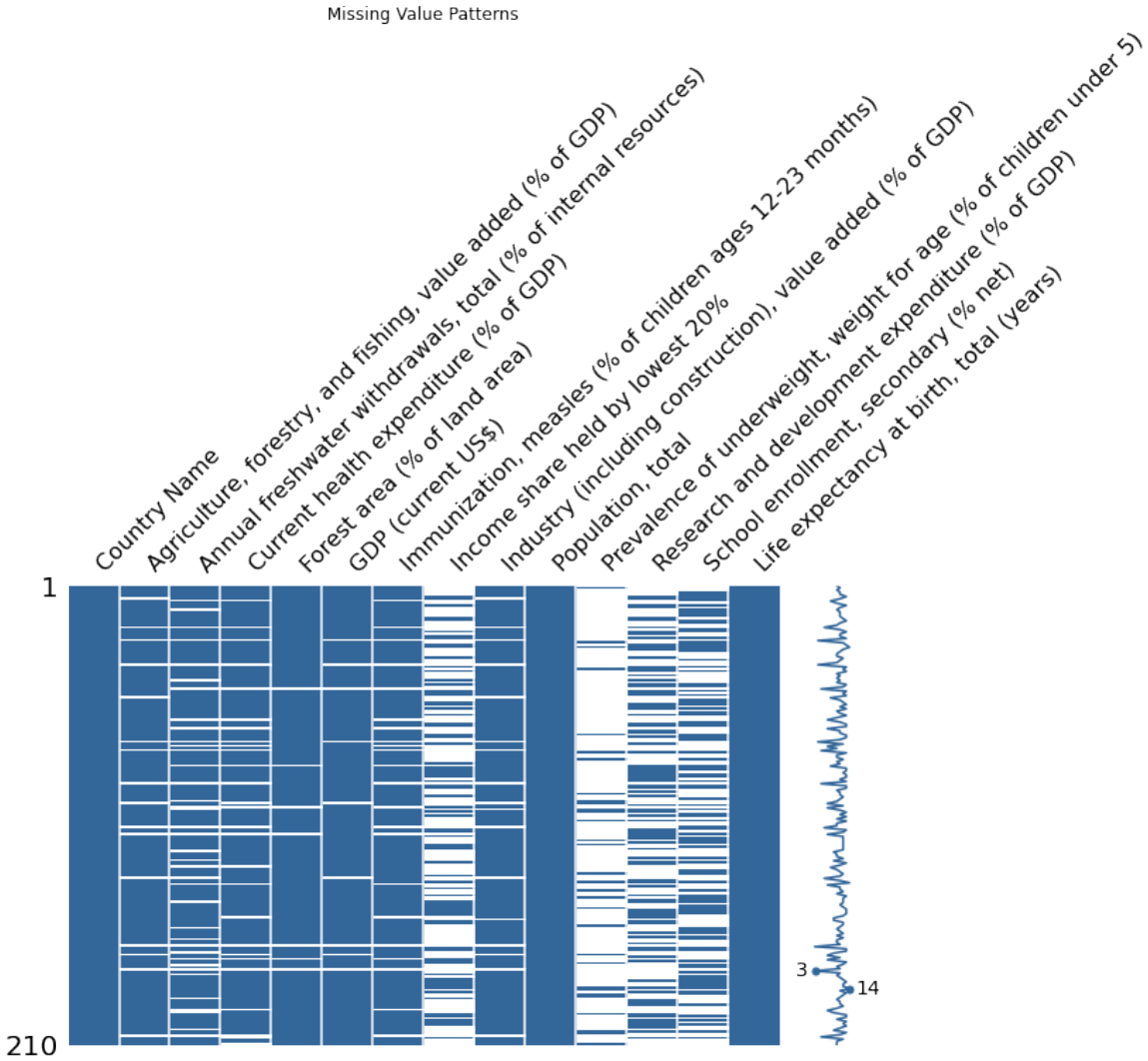
a. Data Understanding

- Look at the distribution of life expectancy at birth to understand its range and variability.



a. Data Understanding

- Deal with any missing data by using different methods and comparing their effectiveness.



a. Data Understanding

- Deal with any missing data by using different methods and comparing their effectiveness.

  - Drop
  - Fillna（前、后、固定值、均值）
  - Interpolate
  - KNN
  - …

对不同的特征可以使用不同的缺失值处理方式

b. Modeling

- Try different models for the prediction task, considering factors like complexity and interpretability.

选择不同复杂度和可解释性的模型，例如：

- **线性模型**：线性回归（基础）、Lasso回归（L1正则化，特征选择）、Ridge回归（L2正则化）。
- **树模型**：随机森林（集成学习，非线性）、XGBoost（梯度提升，高性能）。
- **支持向量机**：SVR（非线性核函数）。
- **神经网络**（可选，需谨慎，小数据集易过拟合）。

数据划分：
训练集：2008-2017年数据
测试集：2018年数据

b. Modeling

- Evaluate the performance of each model and compared them using metrics like MSE and R^2.

1. 均方误差（Mean Squared Error, MSE）

定义

MSE 计算预测值与真实值之间**平方误差的平均值**：

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

其中，y_i 是真实值，$\hat{y}_i$ 是预测值，n 是样本数量。

意义

- **衡量预测的绝对误差**：MSE 直接反映模型预测值与真实值的偏离程度，值越小说明模型越精准。

b. Modeling

- Evaluate the performance of each model and compared them using metrics like MSE and R^2.

2. R方（R-squared, R^2）

定义

R^2 衡量模型**解释目标变量变异性的比例**，范围通常在 $(-\infty, 1]$：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

其中，$\bar{y}$ 是真实值的均值。

意义

- **解释方差比例**：R^2 表示模型相对于简单均值预测的改进程度。例如，R^2=0.8 表示模型解释了目标变量80%的变异性。
- **无量纲性**：R^2 与数据尺度无关，便于跨任务比较。
- **基准对比**：若 R^2 接近1，说明模型拟合良好；若为0，说明模型不优于直接用均值预测；若为负，说明模型比均值预测更差。

b. Modeling

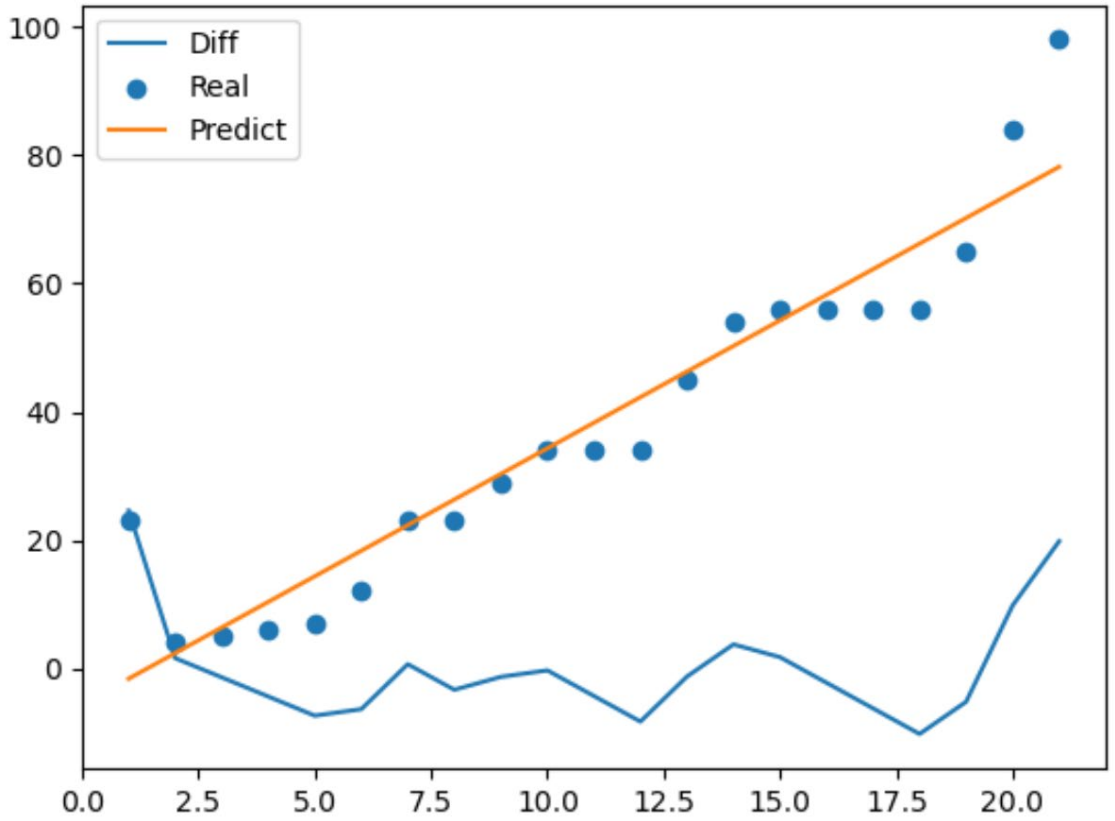
- Identify which features are most important for predicting life expectancy and explain how this determination was made.

  - 线性模型可以直接通过模型权重的绝对值大小直接反映特征对目标变量的影响程度
  - 树模型通过分裂节点时信息增益计算特征重要性
  - 对比不同模型给出的重要特征排名是否一致
  - …

c. Analysis of Predictions

- Visualize the differences between the predicted life expectancy values and the actual values for 2018. Look for any outliers and try to explain why they occurred.

要求对预测值和真实值的差异做出可视化，并对离群值做出解释



c. Analysis of Predictions

- Examine the distribution of prediction errors to see if they follow any patterns or if there are any unexpected trends.

检查预测误差的分布，并观察他们是否存在某种模式

d. Model Improvement

- Try advanced techniques like stepwise forward selection to improve the model's performance.

向前逐步选择法（Forward Stepwise Selection）

1. 什么是向前逐步选择法?

向前逐步 选择法 是一种特征选择（Feature Selection）算法，主要用于模型构建时，从一组候选特征中逐步选择对模型性能影响最大的特征。
通过迭代的方式，逐步向模型中添加特征，直到模型达到预期的性能或满足某些停止准则。

---

2. 目标

- **简化模型**：减少特征数量，提升模型的可解释性。
- **提升性能**：剔除冗余或无关特征，避免过拟合，提高模型的泛化能力。
- **高效计算**：减少特征数量，降低模型计算复杂度。

d. Model Improvement

- Try to create new features that could enhance performance, such as in the health status prediction task, Body Mass Index (BMI) derived from weight and height may be a good high-level indicator.

1. **分析现有特征**：检查数据集中的 12 个原始特征（如 GDP、人口、医疗支出等），寻找可组合或转换的潜在关系。
2. **生成新特征示例**：
   - **人均指标**：如 人均GDP = GDP / 人口。
   - **资源密度**：如 每千人医生数 = 医生数量 / (人口 / 1000)。
   - **复合指标**：如 总健康支出 = 公共健康支出 + 私人健康支出。
   - **比例指标**：如 教育投入占比 = 教育支出 / GDP。
3. **验证新特征**：通过相关性分析或可视化，确认新特征与目标变量（预期寿命）的关系是否显著。

e. Bonus (10 points)

- Is it possible to predict life expectancy for 2025, given the trained model from step d and features (exclude `Life Expectancy at Birth`) ranged from 2008 to 2018?

2 Task2：豆瓣评论情感分析 Part1

Task2: 豆瓣评论情感分析 Part1

Task2. Douban Movie Comment Analysis (60 points)

In this project, we aim to predicte whether a film will be loved by the audience from Douban based on textual reviews. The dataset is named `douban_movie.csv`. The model should take a text input (movie review) and output a specific attitude.

Task2: 豆瓣评论情感分析 Part1

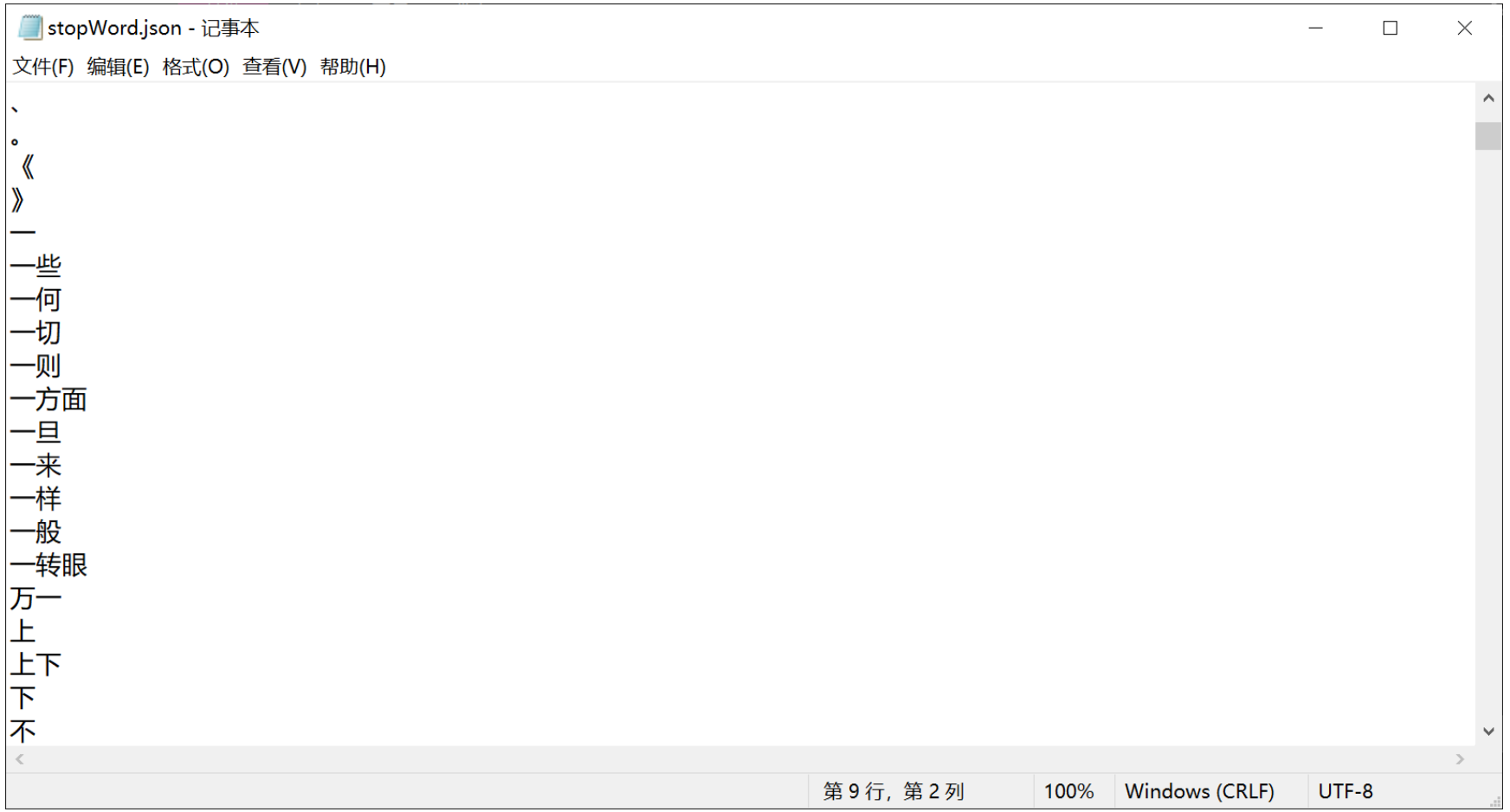
	ID	Movie_Name_EN	Movie_Name_CN	Crawl_Date	Number	Username	Date	Star	Comment	Like
0	0	Avengers Age of Ultron	复仇者联盟2	2017-01-22	1	然潘	2015-05-13	3	连奥创都知道整容要去韩国。	2404
1	10	Avengers Age of Ultron	复仇者联盟2	2017-01-22	11	影志	2015-04-30	4	“一个没有黑暗面的人不值得信任。”第二部剥去冗长的铺垫，开场即高潮、一直到结束，会有人觉...	381
2	20	Avengers Age of Ultron	复仇者联盟2	2017-01-22	21	随时流感	2015-04-28	2	奥创弱爆了弱爆了弱爆了啊！！！！！！	120
3	30	Avengers Age of Ultron	复仇者联盟2	2017-01-22	31	乌鸦火堂	2015-05-08	4	与第一集不同，承上启下，阴郁严肃，但也不会不好看啊，除非本来就不喜欢漫威电影。场面更加宏大...	30
4	40	Avengers Age of Ultron	复仇者联盟2	2017-01-22	41	办公室甜心	2015-05-10	5	看毕，我激动地对友人说，等等奥创要来毁灭台北怎么办厚，她拍了拍我肩膀，没事，反正你买了两份...	16

注意，在这里我们将评分1、2看作是负面评论，评分3、4、5看作是正面评论，只需要通过评论正确预测出正负面评论即可（二分类）

a. Text Preprocessing

- Filter stopwords, low-frequency words, and special symbols.
- Normalize text (e.g., lowercase, stemming/lemmatization).

分词，去掉停用词、低频词、特殊字符



a. Text Preprocessing

- Filter stopwords, low-frequency words, and special symbols.
- Normalize text (e.g., lowercase, stemming/lemmatization).

分词，去掉停用词、低频词、特殊字符

	Comment	Star	comment_processed
0	连奥创都知道整容要去韩国	1	[奥创, 知道, 整容, 韩国]
1	“一个没有黑暗面的人不值得信任” 第二部剥去冗长的铺垫开场即高潮、一直到结束会有人觉得只剩...	1	[一个, 没有, 黑暗面, 值得, 信任, , 第二部, 冗长, 铺垫, 开场, 高潮, ...
2	奥创弱爆了弱爆了弱爆了啊！！！！！！	0	[奥创, 弱, 爆, 弱, 爆, 弱, 爆]
3	与第一集不同承上启下阴郁严肃但也不会不好看啊除非本来就不喜欢漫威电影场面更加宏大单打与团战...	1	[第一集, 不同, 承上启下, 阴郁, 严肃, 不会, 好看, 本来, 喜欢, 漫威, 电影...
4	看毕我激动地对友人说等等奥创要来毁灭台北怎么办厚她拍了拍我肩膀没事反正你买了两份旅行保险惹...	1	[激动, 友人, 说, 奥创, 毁灭, 台北, 厚, 肩膀, 没事, 反正, 买, 两份, ...

b. Text Vectorization

- Convert text to numerical features using:
  - TF-IDF
  - Word2Vec
  - BERT embeddings

(You only need to select one way to do text vectorization. If you compare the impact of different vectorization methods, you will get extra bonus.)

```
[(['今', '天'],
  [array([ 1.26105949e-01,  2.66802788e-01,  8.72311592e-02, -6.03298582e-02,
          -7.75884271e-01, -1.13786355e-01,  3.59727740e-01,  1.85812786e-01,
          -1.47403049e+00, -1.19096741e-01, -7.38845021e-02, -7.87429988e-01,
          -9.21352282e-02,  1.61335021e-01, -8.64771381e-02,  2.56176502e-01,
           8.67898539e-02, -1.32034332e-01, -8.10161680e-02, -1.74522787e-01,
           3.95639017e-02,  7.05541223e-02, -1.02931298e-01, -2.12510973e-01,
           7.06533730e-01, -6.93017840e-02,  8.60072598e-02, -2.62604684e-01,
          -1.59370005e+00, -1.49633154e-01, -1.63491875e-01,  5.32593191e-01,
          -4.69610035e-01, -6.60763383e-02, -1.39506191e-01, -3.14023972e-01,
           7.26258159e-02,  1.34167492e+00, -4.96390201e-02, -3.96719009e-01,
           6.35220185e-02, -1.24420270e-01, -6.83846176e-02,  9.81412828e-03,
          -3.70130911e-02, -3.15944940e-01,  6.09728932e-01,  1.38048425e-01,
          -6.59700036e-02, -9.41008702e-02, -7.49259174e-01,  3.76792192e-01,
           2.29002118e-01, -4.36345071e-01, -2.04388425e-02,  1.27596870e-01,
          -2.80143708e-01,  4.38203327e-02,  1.90831840e-01,  2.05787838e-01,
           1.03129029e+00, -4.77378555e-02, -2.58414745e-01,  1.24642961e-01,
          -5.21564186e-02, -1.89845592e-01,  5.06949201e-02,  7.42469728e-02,
          -1.77084863e-01,  1.19201250e-01,  1.99305415e-01,  2.05253705e-01,
           2.32253790e-01, -3.04696321e-01,  2.71507412e-01,  1.94130927e-01,
           1.54947221e-01, -2.09391028e-01, -1.40534684e-01,  1.79155082e-01,
           1.66332424e-01,  2.83545643e-01, -4.85008389e-01, -1.10102594e-01,
```

c. Model Training & Evaluation

- Train and cross-validate (two both need):
  - Logistic Regression
  - Naive Bayes
- Evaluate accuracy, precision, recall, and F1-score and some reasoning analysis.

数据划分:
训练集: 测试集 = 8 : 2

模型:
逻辑回归、朴素贝叶斯

指标:
accuracy, precision, recall, F1-score, …

3 Task2：豆瓣评论情感分析 Part2

a. Prompt Design & In-Context Learning

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

a. Prompt Design & In-Context Learning (35%)

- Design effective prompts for rating prediction.
- Experiment with few-shot learning (provide examples in the prompt).

Prompt Design示例：
你是一位电影评论分析师。请判断以下评论的情感倾向：
评论：{{review}}
情感倾向（正面/负面）：

few-shot learning示例：
根据示例判断评论情感：
1. 评论：演员表演出色，剧情扣人心弦。
   情感：正面

2. 评论：特效粗糙，节奏拖沓令人失望。
   情感：负面

现在请分析：
评论：{{review}}
情感倾向（正面/负面）：

b. LLM API Testing

- Use **at least two open-source LLM APIs** (e.g., Chatgpt-3.5, deepseek) for prediction.
- Compare their accuracy and robustness.

DeepSeek API 文档

中文（中国） DeepSeek Platform

快速开始
首次调用 API
模型 & 价格
Temperature 设置
Token 用量计算
限速
错误码
新闻
DeepSeek-V3-0324 发布 2025/03/25
DeepSeek-R1 发布 2025/01/20
DeepSeek APP 发布 2025/01/15
DeepSeek-V3 发布 2024/12/26
DeepSeek-V2.5-1210 发布 2024/12/10
DeepSeek-R1-Lite 发布 2024/11/20
DeepSeek-V2.5 发布

快速开始 › 首次调用 API

调用对话 API

首次调用 API

DeepSeek API 使用与 OpenAI 兼容的 API 格式，通过修改配置，您可以使用 OpenAI SDK 来访问 DeepSeek API，或使用与 OpenAI API 兼容的软件。

PARAM	VALUE
base_url *	`https://api.deepseek.com`
api_key	apply for an API key

* 出于与 OpenAI 兼容考虑，您也可以将 `base_url` 设置为 `https://api.deepseek.com/v1` 来使用，但注意，此处 `v1` 与模型版本无关。

* `deepseek-chat` **模型已全面升级为 DeepSeek-V3，接口不变。** 通过指定 `model='deepseek-chat'` 即可调用 DeepSeek-V3。

* `deepseek-reasoner` **是 DeepSeek 最新推出的推理模型 DeepSeek-R1。**通过指定 `model='deepseek-reasoner'`，即可调用 DeepSeek-R1。

调用对话 API

c. Discussion

- Compare ML and LLM results: strengths, limitations, and insights.
  - 指标分析
  - 不一致样本统计
  - 错误案例分析
  - …

Bonus

1. **Advanced Data Analysis (5%)**
   - Generate visualizations (e.g., word clouds, sentiment distribution).
   - Explore correlations between review length/lexicon and ratings.
2. **LLM Prompt Optimization (5%)**
   - Test multi-prompt strategies (e.g., chain-of-thought, role-playing).
   - Analyze how prompt phrasing affects prediction quality.
3. **Fine-Tuning LLMs (10%)**
   - Fine-tune an open LLM (e.g., LLaMA-2) with a small subset of data.
   - Compare performance before/after fine-tuning.

1. 词云、情感分析；探索评论长度/词汇和评分之间的关系
2. Prompting方式优化分析：chain-of-thought、role-playing
3. 开源大语言模型微调

4 问题与讨论

THANK YOU!