数据清洗与预处理

本节代码来自: https://github.com/Day-yong/Machine_Learning 推荐自学

数据来源：数据采集与爬虫，请仔细阅读01-数据采集部分。

一、常见的数据清洗解决方案

【脏数据的表现形式包括】：

- 数据串行，尤其是长文本情形下
- 数值变量种混有文本/格式混乱
- 各种符号乱入
- 数据记录错误大段缺失（某种意义上不算脏数据）

缺失值处理方法：

- 删除:超过70%以上的缺失
- 填充

In [1]:
```
import pandas as pd
import numpy as np


dates = pd.date_range('20230101', periods=6)
df = pd.DataFrame(np.random.randn(6,4),index=dates,columns=list('ABCD'))
df1 = df.reindex(index=dates[0:4],columns=list(df.columns)+['E'])
df1.loc[dates[0]:dates[1],'E']=1
df1
```

Out[1]:

	A	B	C	D	E
2023-01-01	1.009046	-0.918279	-0.275281	1.179399	1.0
2023-01-02	-0.650371	-0.458422	1.602462	-1.130024	1.0
2023-01-03	2.312536	-0.506544	-0.295946	-1.155632	NaN
2023-01-04	-0.246867	0.537559	-1.304588	-0.614367	NaN

In [2]:
```
#删除包含任何缺失的行
df1.dropna(how='any')
```

Out[2]:

	A	B	C	D	E
2023-01-01	1.009046	-0.918279	-0.275281	1.179399	1.0
2023-01-02	-0.650371	-0.458422	1.602462	-1.130024	1.0

In [3]:
```
#对缺失值进行填充
df1.fillna(value=5)
```

Out[3]:

	A	B	C	D	E
2023-01-01	1.009046	-0.918279	-0.275281	1.179399	1.0
2023-01-02	-0.650371	-0.458422	1.602462	-1.130024	1.0
2023-01-03	2.312536	-0.506544	-0.295946	-1.155632	5.0
2023-01-04	-0.246867	0.537559	-1.304588	-0.614367	5.0

In [4]:
```
df1.fillna(df['A'].mean())
```

Out[4]:

	A	B	C	D	E
2023-01-01	1.009046	-0.918279	-0.275281	1.179399	1.00000
2023-01-02	-0.650371	-0.458422	1.602462	-1.130024	1.00000
2023-01-03	2.312536	-0.506544	-0.295946	-1.155632	0.48251
2023-01-04	-0.246867	0.537559	-1.304588	-0.614367	0.48251

小文本与字符串处理：

- python 字符串处理函数
- 正则表达式

In [5]:
```
# 去除空格
char = '    louwill is a machine learning engineer.    '
char.strip()
```

Out[5]:
```
'louwill is a machine learning engineer.'
```

In [6]:
```
# 字符串分割
char2 = 'louwill,is,a,machine,learning,engineer.'
char2.split(',')
```

Out[6]:
```
['louwill', 'is', 'a', 'machine', 'learning', 'engineer.']
```

In [7]:
```
# 拼接：列表转字符串
char2_2 = char2.split(',')
','.join(char2_2)
```

Out[7]:
```
'louwill,is,a,machine,learning,engineer.'
```

In [8]:
```
# 字符替换
char2.replace(',', ' ')
```

Out[8]:
```
'louwill is a machine learning engineer.'
```

正则表达式

- re模块

对这部分感兴趣，推荐自学字符串和正则表达式。

In [9]:
```
import re

# compile函数：编译正则表达式模式
# re.compile(pattern, flag=0)
text1 = 'lebron is a slight good person, he is cool.'
```

```
rr = re.compile(r'\w*oo\w*')
print(rr.findall(text1))
```

```
['good', 'cool']
```

In [10]:
```
# match函数：从字符串首开始匹配
# re.match(pattern, string, flag=0)
print(re.match('com', 'com.louwill.con').group())
```

```
com
```

In [11]:
```
# search函数：对字符串启动查找模式进行匹配，找到第一个并返回
# re.search(pattern, string, flags=0)
# sub 函数：替换字符串中每一个匹配的子串并返回替换后的字符串
# re.sub(pattern, repl, string, count)
```

二、招聘数据的清洗过程

In [12]:
```
import numpy as np
import pandas as pd
import warnings
warnings.filterwarnings('ignore')

data1 = pd.read_csv('./data_analysis.csv', encoding='utf-8', index_col=0)
data2 = pd.read_csv('./machine_learning.csv', encoding='utf-8', index_col=0)
data3 = pd.read_csv('./data_mining.csv', encoding='utf-8', index_col=0)
data4 = pd.read_csv('./deep_learning.csv', encoding='utf-8', index_col=0)

data = pd.concat((pd.concat((pd.concat((data1, data2)), data3)), data4)).reset_index(drop=True)
data.head(3)
```

Out[12]:

	address	advantage	city	company_name	education	industry	industryLables	label	position_detail	position_name	salary	size	stage
0	['科华北路', '桂溪', '四川大学']	工作氛围好	成都	达疆网络科技（上海）有限公司	本科	O2O	[]	['年底双薪', '绩效奖金', '岗位晋升', '定期体检']	\r职位描述：\r\r工作职责：？\r1、负责新零售业务的数据分析工作，挖掘数据分析需求，制...	数据分析师	10k-20k	2000人以上	D轮及以上
1	NaN	六险二金,晋升通道,独当一面,话语权	北京	贝壳找房（北京）科技有限公司	本科	移动互联网,O2O	['大数据', '商业']	['股票期权', '带薪年假', '绩效奖金', '扁平管理']	\r职位描述：\r\r工作职责:方向一、经营分析/指标体系1.参与公司核心策略的数据分析，...	数据分析类	25k-40k	2000人以上	C轮
2	['中关村', '万泉河', '苏州街']	五险一金	北京	杭州财米科技有限公司	不限	移动互联网,金融	[]	['年底多薪', '岗位晋升', '定期体检', '五险一金']	\r职位描述：\r\r职位描述：\r1、收集、处理用户海量数据，挖掘用户行为特征，为产品、运...	数据分析师(MJ000766)	10k-15k	500-2000人	C轮

In [ ]:
```
data.info()
```

In [14]:
```
data['address'] = data['address'].fillna("['未知']")
data['address'][:5]
```

Out[14]:
```
0    ['科华北路', '桂溪', '四川大学']
1                  ['未知']
2     ['中关村', '万泉河', '苏州街']
3              ['琶洲', '官洲']
4    ['中关村', '北京大学', '颐和园']
Name: address, dtype: object
```

In [15]:
```
for i, j in enumerate(data['address']):
    j = j.replace('[', '').replace(']', '')
    data['address'][i] = j

data['address'][:5]
```

Out[15]:
```
0    '科华北路', '桂溪', '四川大学'
1                  '未知'
2     '中关村', '万泉河', '苏州街'
3              '琶洲', '官洲'
4    '中关村', '北京大学', '颐和园'
Name: address, dtype: object
```

In [16]:
```
for i, j in enumerate(data['industryLables']):
    j = j.replace('[', '').replace(']', '')
    data['industryLables'][i] = j

data['industryLables'][:10]
```

Out[16]:
```
0
1              '大数据', '商业'
2
3    '移动互联网', '社交', '数据运营'
4            '大数据', 'SPSS'
5            '大数据', '数据挖掘'
6            '大数据', '数据挖掘'
7            '大数据', '数据挖掘'
8                    '金融'
9
Name: industryLables, dtype: object
```

In [17]:
```
for i, j in enumerate(data['label']):
    j = j.replace('[', '').replace(']', '')
```

```
    data['label'][i] = j

data['label'][:10]
```

Out[17]:
```
0      '年底双薪', '绩效奖金', '岗位晋升', '定期体检'
1      '股票期权', '带薪年假', '绩效奖金', '扁平管理'
2      '年底多薪', '岗位晋升', '定期体检', '五险一金'
3    '六险一金', '周末双休', '营养工作餐', '暖心下午茶'
4      '绩效奖金', '五险一金', '交通补助', '带薪年假'
5      '专项奖金', '带薪年假', '弹性工作', '管理规范'
6        '丰厚年终', '扁平管理', '追求极致', '本分'
7      '弹性工作', '五险一金', '年度旅游', '年底双薪'
8      '创新开放', '团队牛X', '全员期权', '高速成长'
9      '绩效奖金', '交通补助', '午餐补助', '定期体检'
Name: label, dtype: object
```

In [18]:
```
data['position_detail'][0].replace('\r', '')
```

Out[18]:
```
'职位描述：工作职责：?1、负责新零售业务的数据分析工作，挖掘数据分析需求，制定并实施分析方案，并根据数据分析结果为业务的改进提出合理化建议；?2、通过专题分析，对业务问题进行深入分析，为产品改进、运营决策、营销推广策略提供数据支持，推动业务部门数据驱动业务决策的转化3、通过对公司运营数据研究，提出改善运营质量的方法和建议，搭建BI数据分析体系，为公司决策提供支持。任职资格：?1、统计、数学、信息技术相关专业本科以上学历，互联网2年以上数据分析/挖掘相关经验；2、熟练独立编写商业数据分析报告，及时发现和分析其中隐含的变化和问题；?3、良好的数据敏感度，能从海量数据提炼核心结果；有丰富的数据分析、挖掘、清洗和建模的经验；?4、思维敏捷，具有发散性，能够举一反三，良好的逻辑分析能力及问题解决能力5、良好的跨团队、部门沟通及推动能力，有强烈的主人翁意识，积极发扬团队合作精神6、 熟练掌握SQL7、熟悉或使用过至少一种统计分析/数据挖掘软件（R,?Python等）者优先8、有大数据处理经验，如Hive/Spark/Hadoop等使用经验者优先'
```

In [19]:
```
data['position_detail'] = data['position_detail'].fillna('未知')
```

In [20]:
```
for i, j in enumerate(data['position_detail']):
    j = j.replace('\r', '')
    data['position_detail'][i] = j

data['position_detail'][:3]
```

Out[20]:
```
0    职位描述：工作职责：?1、负责新零售业务的数据分析工作，挖掘数据分析需求，制定并实施分析方案...
1    职位描述：工作职责:方向一、经营分析/指标体系1. 参与公司核心策略的数据分析，基于策略逻辑...
2    职位描述：职位描述：1、收集、处理用户海量数据，挖掘用户行为特征，为产品、运营提供参考依据； ...
Name: position_detail, dtype: object
```

In [21]:
```
data.head(3)
```

Out[21]:

	address	advantage	city	company_name	education	industry	industryLables	label	position_detail	position_name	salary	size	stage
0	'科华北路', '桂溪', '四川大学'	工作氛围好	成都	达疆网络科技（上海）有限公司	本科	O2O		'年底双薪', '绩效奖金', '岗位晋升', '定期体检'	职位描述：工作职责：?1、负责新零售业务的数据分析工作，挖掘数据分析需求，制定并实施分析方案...	数据分析师	10k-20k	2000人以上	D轮及以上
1	'未知'	六险二金,晋升通道,独当一面,话语权	北京	贝壳找房（北京）科技有限公司	本科	移动互联网,O2O	'大数据', '商业'	'股票期权', '带薪年假', '绩效奖金', '扁平管理'	职位描述：工作职责:方向一、经营分析/指标体系1. 参与公司核心策略的数据分析，基于策略逻辑...	数据分析类	25k-40k	2000人以上	C轮
2	'中关村', '万泉河', '苏州街'	五险一金	北京	杭州财米科技有限公司	不限	移动互联网,金融		'年底多薪', '岗位晋升', '定期体检', '五险一金'	职位描述：职位描述：1、收集、处理用户海量数据，挖掘用户行为特征，为产品、运营提供参考依据；...	数据分析师(MJ000766)	10k-15k	500-2000人	C轮

In [ ]:
```
import string
for i in data['salary'][:10]:
    i = i.replace('k', '')
    i1 = int(i.split('-')[0])
    i2 = int(i.split('-')[1])
```

```
    i3 = 1/2 * (i1+i2)
    print(i3)
```

In [23]:
```
data['salary'][0]
```

Out[23]:
```
'10k-20k'
```

In [24]:
```
for i, j in enumerate(data['salary']):
    j = j.replace('k', '').replace('K', '').replace('以上', '-0')
    j1 = int(j.split('-')[0])
    j2 = int(j.split('-')[1])
    j3 = int(1/2 * (j1+j2))
    data['salary'][i] = j3*1000

data['salary'].head(3)
```

Out[24]:
```
0    15000
1    32000
2    12000
Name: salary, dtype: object
```

In [25]:
```
data['size'].value_counts()
```

Out[25]:
```
2000人以上      573
500-2000人    324
150-500人     314
50-150人      269
15-50人       156
少于15人         14
Name: size, dtype: int64
```

In [26]:
```
data['stage'].value_counts()
```

Out[26]:
```
上市公司      337
不需要融资     293
B轮         240
A轮         239
C轮         216
D轮及以上      192
天使轮         69
未融资         64
Name: stage, dtype: int64
```

In [27]:
```
data['work_year'].value_counts()
```

Out[27]:
```
3-5年     730
1-3年     465
不限       221
5-10年    136
应届毕业生     80
1年以下       16
10年以上       2
Name: work_year, dtype: int64
```

In [28]:
```
for i, j in enumerate(data['position_name']):
    if '数据分析' in j:
        j = '数据分析师'
    if '数据挖掘' in j:
        j = '数据挖掘工程师'
    if '机器学习' in j:
        j = '机器学习工程师'
    if '深度学习' in j:
        j = '深度学习工程师'
    data['position_name'][i] = j
data['position_name'][:5]
```

Out[28]:
```
0    数据分析师
1    数据分析师
2    数据分析师
3    数据分析师
4    数据分析师
Name: position_name, dtype: object
```

In [29]:
```
data.head(3)
```

Out[29]:

	address	advantage	city	company_name	education	industry	industryLables	label	position_detail	position_name	salary	size	stage
0	'科华北路', '桂溪', '四川大学'	工作氛围好	成都	达疆网络科技（上海）有限公司	本科	O2O		'年底双薪', '绩效奖金', '岗位晋升', '定期体检'	职位描述：工作职责：?1、负责新零售业务的数据分析工作，挖掘数据分析需求，制定并实施分析方案...	数据分析师	15000	2000人以上	D轮及以上
1	'未知'	六险二金,晋升通道,独当一面,话语权	北京	贝壳找房（北京）科技有限公司	本科	移动互联网,O2O	'大数据', '商业'	'股票期权', '带薪年假', '绩效奖金', '扁平管理'	职位描述：工作职责:方向一、经营分析/指标体系1. 参与公司核心策略的数据分析，基于策略逻辑...	数据分析师	32000	2000人以上	C轮
2	'中关村', '万泉河', '苏州街'	五险一金	北京	杭州财米科技有限公司	不限	移动互联网,金融		'年底多薪', '岗位晋升', '定期体检', '五险一金'	职位描述：职位描述：1、收集、处理用户海量数据，挖掘用户行为特征，为产品、运营提供参考依据；...	数据分析师	12000	500-2000人	C轮

三、清洗过程模块化

In [30]:
```
import numpy as np
import pandas as pd
```

```
import string
import warnings
warnings.filterwarnings('ignore')

class data_clean(object):
    def __init__(self):
        pass

    def get_data(self):
        data1 = pd.read_csv('./data_analysis.csv', encoding='utf-8', index_col=0)
        data2 = pd.read_csv('./machine_learning.csv', encoding='utf-8', index_col=0)
        data3 = pd.read_csv('./data_mining.csv', encoding='utf-8', index_col=0)
        data4 = pd.read_csv('./deep_learning.csv', encoding='utf-8', index_col=0)

        data = pd.concat((pd.concat((pd.concat((data1, data2)), data3)), data4)).reset_index(drop=True)
        return data

    def clean_operation(self):
        data = self.get_data()
        data['address'] = data['address'].fillna("['未知']")
        for i, j in enumerate(data['address']):
            j = j.replace('[', '').replace(']', '')
            data['address'][i] = j

        for i, j in enumerate(data['salary']):
            j = j.replace('k', '').replace('K', '').replace('以上', '-0')
            j1 = int(j.split('-')[0])
            j2 = int(j.split('-')[1])
            j3 = 1/2 * (j1+j2)
            data['salary'][i] = j3*1000

        for i, j in enumerate(data['industryLables']):
            j = j.replace('[', '').replace(']', '')
            data['industryLables'][i] = j

        for i, j in enumerate(data['label']):
            j = j.replace('[', '').replace(']', '')
            data['label'][i] = j

        data['position_detail'] = data['position_detail'].fillna('未知')
        for i, j in enumerate(data['position_detail']):
            j = j.replace('\r', '')
            data['position_detail'][i] = j
```

```
        for i, j in enumerate(data['position_name']):
            if '数据分析' in j:
                j = '数据分析师'
            if '数据挖掘' in j:
                j = '数据挖掘工程师'
            if '机器学习' in j:
                j = '机器学习工程师'
            if '深度学习' in j:
                j = '深度学习工程师'
            data['position_name'][i] = j

        return data


opt = data_clean()
data = opt.clean_operation()
data.head(3)
```

Out[30]:

	address	advantage	city	company_name	education	industry	industryLables	label	position_detail	position_name	salary	size	stage
0	'科华北路', '桂溪', '四川大学'	工作氛围好	成都	达疆网络科技（上海）有限公司	本科	O2O		'年底双薪', '绩效奖金', '岗位晋升', '定期体检'	职位描述：工作职责：?1、负责新零售业务的数据分析工作，挖掘数据分析需求，制定并实施分析方案...	数据分析师	15000.0	2000人以上	D轮及以上
1	'未知'	六险二金,晋升通道,独当一面,话语权	北京	贝壳找房（北京）科技有限公司	本科	移动互联网,O2O	'大数据', '商业'	'股票期权', '带薪年假', '绩效奖金', '扁平管理'	职位描述：工作职责:方向一、经营分析/指标体系1. 参与公司核心策略的数据分析，基于策略逻辑...	数据分析师	32500.0	2000人以上	C轮
2	'中关村', '万泉河', '苏州街'	五险一金	北京	杭州财米科技有限公司	不限	移动互联网,金融		'年底多薪', '岗位晋升', '定期体检', '五险一金'	职位描述：职位描述：1、收集、处理用户海量数据，挖掘用户行为特征，为产品、运营提供参考依据；...	数据分析师	12500.0	500-2000人	C轮

In [31]:
```
data.to_csv('./lagou_preprocessed.csv', encoding='utf-8')
```