In [1]:

```
import pandas as pd
```

In [2]:

```
train=pd.read_csv('s.csv')
```

In [3]:

```
train=train.iloc[:,1:]
train
```

Out[3]:

	0	1	2	3	4	5	6	7	8	9	...	3991	3992	3993	3994	3995	3996	3
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
13272	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
13273	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
13274	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
13275	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
13276	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	

13277 rows × 4001 columns

In [4]:

```
train_x=train.iloc[:,0:4000]
train_y=train.iloc[:,4000]
```

In [5]:

```
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import accuracy_score
```

In [6]:

```
df=pd.read_csv('Total_data.csv')
import re
pattern = re.compile(u'\s|\n|<[^>]*>|&.*;|\\(.*?\\)', re.S)
for i in range(df.shape[0]):
    df.content[i] = pattern.sub('', df.content[i])
    df.content[i] = df.content[i].replace('”“','')
df.category.unique()
df=df[['category','content','src','time','title']]

import jieba
for i in range(df.shape[0]):
    df.content[i]=jieba.lcut(df.content[i])

df_content=df[['category','content']]

stopwords=pd.read_csv("stopwords.txt",index_col=False,sep="\t",quoting=3,names=['sto
def drop_stopwords(contents,stopwords):
    contents_clean = []
    all_words = []
    for line in contents:
        line_clean = []
        for word in line:
            if word in stopwords:
                continue
            line_clean.append(word)
            all_words.append(str(word))
        contents_clean.append(line_clean)
    return contents_clean,all_words

contents = df_content.content.values.tolist()
stopwords = stopwords.stopword.values.tolist()
contents_clean,all_words = drop_stopwords(contents,stopwords)
df_content=pd.DataFrame({'contents_clean':contents_clean,'label':df_content.category
```

```
/Users/apple/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.
py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (h
ttp://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#re
turning-a-view-versus-a-copy)
  """
/Users/apple/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.
py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (h
ttp://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#re
turning-a-view-versus-a-copy)

Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/y1/hn1g_j8d3dld698c0g4g97qm0000g
n/T/jieba.cache
Loading model cost 0.791 seconds.
Prefix dict has been built succesfully.
```

In [7]:

```
df_content
```

Out[7]:

	contents_clean	label
0	[精彩, 弹幕, 客户端, 近日, 广东, 深圳, 凌晨, 挂, 高楼, 外, 墙上, 喊救...	news
1	[近日, 湖南省, 宁远县, 人民法院, 公布, 一名, 肢解, 自杀, 妻子, 抛, 尸,...	news
2	[希望, 这起, 带有, 第一案, 光环, 案件, 司法, 力量, 划定, 人脸识别, 技术...	news
3	[兰乔, 圣菲, 小区, 多名, 业主, 进, 华商报, 新闻热线, 小区, 物业, 贴, ...	news
4	[华商报, 讯, 虚假, 招聘, 诱饵, 骗, 全国, 大学生, 前来, 应聘, 暴力手段,...	news
...	...	...
13995	[近日, 山西, 原平, 市民, 政府, 官网, 市长, 信箱, 举报, 称, 一名, 小学...	baby
13996	[妈妈, 爸爸, 家, 玩, 段时间, 这句, 话, 贵州, 消防员, 朋友, 圈里, 刷,...	ntes
13997	[俄罗斯, 卫星, 网, 报道, 俄罗斯内务部, 官网, 公布, 预防, 违法犯罪, 政府,...	ntes
13998	[面对, 择校, 升学, 压力, 中国, 家长, 不惜, 花, 重金, 孩子, 英国, 补习...	baby
13999	[岁, 藏族, 小女孩, 格日, 草, 病房, 里望, 窗外, 起床, 头, 读物, 玛, ...	baby

14000 rows × 2 columns

In [8]:

```
df_content_pre = df_content[df_content.isnull().T.any().T]
```

In [9]:

```
df_content_pre
```

Out[9]:

	contents_clean	label
10002	[古人云, 赐子, 千金, 教子, 一艺, 教子, 一艺, 赐子, 好名, 起名, 改名, ...	NaN
10019	[水逆, 当成, 倒霉, 代名词, 水逆, 怎么回事, 简单, 分析, 水逆, 水逆, 下次...	NaN
11001	[一生, 花, 四季, 我开, 绿, 枯, 秃, 永远, 猜, 不到, 明天, 绿帽, 先,...	NaN
11031	[霜降, 过后, 深秋, 小编, 整理, 一首, 二胡, 晚秋, 舒服, 句, 话, 送给,...	NaN
11047	[一问, 疲惫, 生活, 日夜, 奔波, 省吃俭用, 钱, 攒, 几个, 烙下个, 病, 身...	NaN
...	...	...
12994	[上班族, 颈椎病, 这话, 一点, 夸张, 常年, 伏案, 工作, 颈椎, 肩膀, 僵硬,...	NaN
12995	[封面, 新闻记者, 李雨心, 现代医学, 发达, 做, 一场, 手术, 难事, 躺, 无菌...	NaN
12996	[生活, 越来越, 大鱼大肉, 肥胖, 本草, 厨房, 带来, 一杯, 茶品, 三七, 龙眼...	NaN
12997	[伽人, YO, 酱, 常说, 拉伸, 本质, 肌肉, 一种, 牵拉, 增加, 关节, 活动...	NaN
12998	[说, 芦荟, 第一, 美容, 芦荟, 美容, 帮, 血管, 减龄, 降低, 三高, 修复,...	NaN

571 rows × 2 columns

In [10]:

```
words_content = []
for line_index in range(df_content_pre.shape[0]):
    try:
        #x_train[line_index][word_index] = str(x_train[line_index][word_index])
        words_content.append(' '.join(df_content_pre['contents_clean'].values[line_i
    except:
        print (line_index)
```

In [11]:

```
words_content[0]
```

Out[11]:

'古人云 赐子 千金 教子 一艺 教子 一艺 赐子 好名 起名 改名 慎之又慎 忽视 名 姓 关系 字形 字义 搭配 吴姓 起名 吴德 可想而知 一生 样子 胡姓 胡伟 胡作非为 怪事 位叫 王栓柱 拴住 被判 无期徒刑 永远 拴住 林彪 名字 林中 一只 花斑 猛虎 伤人 彪 字 带有 三撇 刑伤 有位 女孩 张雪 推断 婚姻 不顺 妇科 调 怕冷 事实上 结婚 时间 不长 离婚 孩子 朋友 感悟 名字 名字 杨跃东 羊要 越冬 艰辛 绿草 清泉 挨过 冬天 生机盎然 春天 等待 这位 朋友 一生 操劳 难关 羊 啃 干 草皮 精神 才行 张姓 弓 不宜 后接 动物 名字 如渔 震 默 家 祥 有位 朋友家 孩子 张翀 羽毛 鸟 中 字 好像 一只 鸟 射中 感觉 肿胀 象 事实上 男孩 前不久 肝癌 离开 临走 肿 刚刚 岁 令人 痛惜 字 五行 吉凶 关乎 一生 朋友 运势 不佳 情感 事业 财运 健康 不好 建议 长 识别 二维码 测测 名字 朋友 测试 担心 名字 起好 初衷 写给 起名 改名 朋友 成年人 不太 改 担心 起个 笔名 网名 补救 心态 调整'

In [12]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(analyzer='word', max_features=4000,  lowercase = False)
vectorizer.fit(words_content)
```

Out[12]:

```
TfidfVectorizer(analyzer='word', binary=False, decode_error='strict',
        dtype=<class 'numpy.float64'>, encoding='utf-8', input='conten
t',
        lowercase=False, max_df=1.0, max_features=4000, min_df=1,
        ngram_range=(1, 1), norm='l2', preprocessor=None, smooth_idf=T
rue,
        stop_words=None, strip_accents=None, sublinear_tf=False,
        token_pattern='(?u)\\b\\w\\w+\\b', tokenizer=None, use_idf=Tru
e,
        vocabulary=None)
```

In [13]:

```
params = {'n_estimators':100, 'max_depth':5, 'min_samples_split':20, 'learning_rate
GB = GradientBoostingClassifier(**params)
GB.fit(train_x,train_y)
predict_GB_trn = GB.predict(train_x)
accuracy_score(train_y,predict_GB_trn)
```

Out[13]:

```
0.9975144987572494
```

In [14]:

```
y_predict=GB.predict(vectorizer.transform(words_content))
```

In [15]:

```
len(y_predict)
```

Out[15]:

```
571
```

In [32]:

```
y_predict=pd.DataFrame(y_predict)
df_content_pre=df_content_pre.reset_index(drop=True)
predict_data=pd.concat([df_content_pre,y_predict],axis=1,ignore_index=True)
```

In [34]:

```
predict_data
```

Out[34]:

	0	1	2
0	[古人云, 赐子, 千金, 教子, 一艺, 教子, 一艺, 赐子, 好名, 起名, 改名, ...	NaN	20
1	[水逆, 当成, 倒霉, 代名词, 水逆, 怎么回事, 简单, 分析, 水逆, 水逆, 下次...	NaN	20
2	[一生, 花, 四季, 我开, 绿, 枯, 秃, 永远, 猜, 不到, 明天, 绿帽, 先,...	NaN	20
3	[霜降, 过后, 深秋, 小编, 整理, 一首, 二胡, 晚秋, 舒服, 句, 话, 送给,...	NaN	10
4	[一问, 疲惫, 生活, 日夜, 奔波, 省吃俭用, 钱, 攒, 几个, 烙下个, 病, 身...	NaN	20
...	...	...	...
566	[上班族, 颈椎病, 这话, 一点, 夸张, 常年, 伏案, 工作, 颈椎, 肩膀, 僵硬,...	NaN	14
567	[封面, 新闻记者, 李雨心, 现代医学, 发达, 做, 一场, 手术, 难事, 躺, 无菌...	NaN	14
568	[生活, 越来越, 大鱼大肉, 肥胖, 本草, 厨房, 带来, 一杯, 茶品, 三七, 龙眼...	NaN	14
569	[伽人, YO, 酱, 常说, 拉伸, 本质, 肌肉, 一种, 牵拉, 增加, 关节, 活动...	NaN	20
570	[说, 芦荟, 第一, 美容, 芦荟, 美容, 帮, 血管, 减龄, 降低, 三高, 修复,...	NaN	14

571 rows × 3 columns

In [16]:

```
data=pd.read_csv('Total_data.csv')
import re
pattern = re.compile(u'\s|\n|<[^>]*>|&.*;|\\(.*?\\)', re.S)
for i in range(data.shape[0]):
    data.content[i] = pattern.sub('', data.content[i])
    data.content[i] = data.content[i].replace('”“','')
#data.category.unique()
data=data[['category','content','src','time','title']]
```

```
/Users/apple/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.
py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (h
ttp://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#re
turning-a-view-versus-a-copy)
  """
/Users/apple/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.
py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (h
ttp://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#re
turning-a-view-versus-a-copy)
```

In [17]:

```
data=data[['category','content']]
```

In [18]:

```
data_pre = data[data.isnull().T.any().T]
```

In [19]:

```
data_pre=data_pre.content
data_pre
```

Out[19]:

```
10002    古人云：赐子千金，不如教子一艺；教子一艺，不如赐子好名。因此无论起名或改名，一定要慎之又慎，...
10019    现在很多人都会把水逆当成“倒霉”的代名词，然而水逆究竟是怎么回事呢？今天我们就来简单分析一下...
11001    人的一生就像花的四季“我开了……我绿了……我枯了……我秃了……”你永远也猜不到明天和“绿帽”...
11031    霜降过后，进入了深秋，今天小编就整理了一首二胡《晚秋》，以及让人非常舒服的5句话，要送给大家...
11047    一问自己活得疲惫吗？为了生活日夜奔波，省吃俭用钱也没攒几个，最后烙下个病身子，给座金山也没福...
                                                  ...
12994    10个上班族9个颈椎病，这话一点都不夸张，我就是其中一个，常年伏案工作，颈椎和肩膀都僵硬了，...
12995    封面新闻记者李雨心在现代医学发达的当下，做一场安全的手术，并不是什么难事。躺在无菌的手术室，...
12996    现在大家生活越来越好，大鱼大肉很容易让自己肥胖，今天《本草厨房》就给大家带来一杯茶品—“三...
12997    哈喽，各位伽人，我是你们的YO酱！我们常说的拉伸，它的本质就是对肌肉的一种牵拉，与此同时还能...
12998    说到芦荟，我们的第一反应是“美容”，但其实芦荟除了美容，它还能够帮我们的血管减龄，降低“三高...
Name: content, Length: 571, dtype: object
```

In [20]:

```
data_pre=pd.DataFrame(data_pre)
data_pre
```

Out[20]:

	content
10002	古人云：赐子千金，不如教子一艺；教子一艺，不如赐子好名。因此无论起名或改名，一定要慎之又慎，...
10019	现在很多人都会把水逆当成“倒霉”的代名词，然而水逆究竟是怎么回事呢？今天我们就来简单分析一下...
11001	人的一生就像花的四季“我开了……我绿了……我枯了……我秃了……”你永远也猜不到明天和“绿帽”...
11031	霜降过后，进入了深秋，今天小编就整理了一首二胡《晚秋》，以及让人非常舒服的5句话，要送给大家...
11047	一问自己活得疲惫吗？为了生活日夜奔波，省吃俭用钱也没攒几个，最后烙下个病身子，给座金山也没福...
...	...
12994	10个上班族9个颈椎病，这话一点都不夸张，我就是其中一个，常年伏案工作，颈椎和肩膀都僵硬了，...
12995	封面新闻记者李雨心在现代医学发达的当下，做一场安全的手术，并不是什么难事。躺在无菌的手术室，...
12996	现在大家生活越来越好，大鱼大肉很容易让自己肥胖，今天《本草厨房》就给大家带来一杯茶品——“三...
12997	哈喽，各位伽人，我是你们的YO酱！我们常说的拉伸，它的本质就是对肌肉的一种牵拉，与此同时还能...
12998	说到芦荟，我们的第一反应是“美容”，但其实芦荟除了美容，它还能够帮我们的血管减龄，降低“三高...

571 rows × 1 columns

In [28]:

```
y_predict=pd.DataFrame(y_predict)
data_pre=data_pre.reset_index(drop=True)
predict_data=pd.concat([data_pre,y_predict],axis=1,ignore_index=True)
```

In [29]:

```
predict_data
```

Out[29]:

	0	1
0	古人云：赐子千金，不如教子一艺；教子一艺，不如赐子好名。因此无论起名或改名，一定要慎之又慎，...	1
1	现在很多人都会把水逆当成“倒霉”的代名词，然而水逆究竟是怎么回事呢？今天我们就来简单分析一下...	14
2	人的一生就像花的四季“我开了……我绿了……我枯了……我秃了……”你永远也猜不到明天和“绿帽”...	10
3	霜降过后，进入了深秋，今天小编就整理了一首二胡《晚秋》，以及让人非常舒服的5句话，要送给大家...	10
4	一问自己活得疲惫吗？为了生活日夜奔波，省吃俭用钱也没攒几个，最后烙下个病身子，给座金山也没福...	14
...	...	...
566	10个上班族9个颈椎病，这话一点都不夸张，我就是其中一个，常年伏案工作，颈椎和肩膀都僵硬了，...	14
567	封面新闻记者李雨心在现代医学发达的当下，做一场安全的手术，并不是什么难事。躺在无菌的手术室，...	14
568	现在大家生活越来越好，大鱼大肉很容易让自己肥胖，今天《本草厨房》就给大家带来一杯茶品——“三...	18
569	哈喽，各位伽人，我是你们的YO酱！我们常说的拉伸，它的本质就是对肌肉的一种牵拉，与此同时还能...	14
570	说到芦荟，我们的第一反应是“美容”，但其实芦荟除了美容，它还能够帮我们的血管减龄，降低“三高...	14

571 rows × 2 columns

In [23]:

```
data
```

Out[23]:

	category	content
0	news	精彩弹幕，尽在客户端近日，广东深圳，一男子凌晨被挂高楼外墙上，大喊救命。目击者称，男子拉着一...
1	news	近日，湖南省宁远县人民法院公布了当地一名男子肢解自杀妻子并抛尸一案的判决书。判决书显示，20...
2	news	希望这起带有“第一案”光环的案件，能用司法力量划定人脸识别技术的应用边界。近日，发生在杭州的...
3	news	11月1日，兰乔圣菲小区多名业主打进华商报新闻热线反映，小区物业贴通知称，供热公司要求物业按...
4	news	华商报讯以虚假招聘为诱饵，骗全国各地的大学生前来应聘，然后用暴力手段限制人身自由，胁迫被害人...
...	...	...
13995	baby	近日，有山西原平市民在政府官网市长信箱举报称，当地一名小学代课教师在校外私自补课，“课上不讲...
13996	ntes	“妈妈，爸爸什么时候才来我们家玩？”这段时间，这句话在贵州消防员的朋友圈里刷了屏。当这条朋友...
13997	ntes	据俄罗斯卫星网10日报道，俄罗斯内务部官网公布了预防违法犯罪政府委员会议纪要，其中称若未成年...
13998	baby	面对择校和升学的压力，一些中国家长不惜花重金让孩子来英国补习。和担忧子女学业的中国家长相似，...
13999	baby	8岁藏族小女孩格日草从病房里望着窗外，有时又拿起床头读物《玛蒂娜》翻看。夜晚，护士稍闲下来，...

14000 rows × 2 columns

In [26]:

```
df_content_use = data.dropna(axis=0,how='any')
df_content_use = df_content_use[df_content_use.category!='video']
from sklearn import preprocessing
le = preprocessing.LabelEncoder()
le.fit(df_content_use.category.tolist())
```

Out[26]:

```
LabelEncoder()
```

In [27]:

```
len(le.classes_)
```

Out[27]:

```
21
```

In [30]:

```
cate=predict_data.iloc[:,1].tolist()
cate
```

Out[30]:

```
[1,
 14,
 10,
 10,
 14,
 14,
 15,
 14,
 14,
 14,
 14,
 10,
 10,
 14,
 5,
 14,
 14,
 10,
```

In [31]:

```
predict_label=list(le.inverse_transform(cate))
```

In [32]:

```
len(predict_label)
```

Out[32]:

```
571
```

In [33]:

```
content=predict_data.iloc[:,0]
```

In [34]:

```
predict_label=pd.DataFrame(predict_label)
```

In [35]:

```
final_data=pd.concat([content,predict_label],axis=1,ignore_index=True)
```

In [36]:

```
final_data
```

Out[36]:

	0	1
0	古人云：赐子千金，不如教子一艺；教子一艺，不如赐子好名。因此无论起名或改名，一定要慎之又慎，...	baby
1	现在很多人都会把水逆当成“倒霉”的代名词，然而水逆究竟是怎么回事呢？今天我们就来简单分析一下...	news
2	人的一生就像花的四季“我开了……我绿了……我枯了……我秃了……”你永远也猜不到明天和“绿帽”...	mil
3	霜降过后，进入了深秋，今天小编就整理了一首二胡《晚秋》，以及让人非常舒服的5句话，要送给大家...	mil
4	一问自己活得疲惫吗？为了生活日夜奔波，省吃俭用钱也没攒几个，最后烙下个病身子，给座金山也没福...	news
...	...	...
566	10个上班族9个颈椎病，这话一点都不夸张，我就是其中一个，常年伏案工作，颈椎和肩膀都僵硬了，...	news
567	封面新闻记者李雨心在现代医学发达的当下，做一场安全的手术，并不是什么难事。躺在无菌的手术室，...	news
568	现在大家生活越来越好，大鱼大肉很容易让自己肥胖，今天《本草厨房》就给大家带来一杯茶品——“三...	tech
569	哈喽，各位伽人，我是你们的YO酱！我们常说的拉伸，它的本质就是对肌肉的一种牵拉，与此同时还能...	news
570	说到芦荟，我们的第一反应是“美容”，但其实芦荟除了美容，它还能够帮我们的血管减龄，降低“三高...	news

571 rows × 2 columns

In [37]:

```
final_data.to_csv('final_data_2.csv')
```

In [ ]:

```

```

In [ ]:

```

```