- 贝叶斯分类器的应用：法律裁判文书中的案情要素分类
  - 以“中国裁判文书网”公开的有关婚姻家庭领域的2665条裁判文书为例，基于文书句子文本和每个句子对应的要素标签（多分类），探索朴素贝叶斯分类器在文本分类中的应用
  - 涉及如下两个重要方面：
    - 第一，文本分类预测建模的数据预处理：文本的量化处理
    - 第二，文本和文本标签的组织

  - 文本的量化处理：
    - 分词：将句子分割成若干个词，该过程称为分词
      - “中国 的 发展 是 开放 的 发展”，“中国 经济 发展 的 质量 在 稳步 提升 ， 人民 生活 在 持续 改善”
    - 词的量化：TF-IDF（Term Frequency－Inverse Document Frequency），TF是词频，IDF是逆文本频率，两者结合用于度量词对于某篇文本的重要程度

$$TF_{ji} = \frac{N_{ji}}{\sum_{k=1}^{K} N_{jk}} \qquad IDF_i = \log\left(\frac{\text{总文本数}}{\text{包含词}i\text{的文本数}}\right) \qquad \text{TF} - \text{IDF}_{ji} = TF_{ji} \times IDF_i$$

通常将TF-IDF组织成矩阵形式

	词 1	词 2	……	词 K	文本类别
文本 1	$\text{TF}-\text{IDF}_{11}$	$TF-\text{IDF}_{12}$	……	$\text{TF}-\text{IDF}_{1K}$	2
文本 2	$\text{TF}-\text{IDF}_{21}$	$\text{TF}-\text{IDF}_{22}$	……	$\text{TF}-\text{IDF}_{2K}$	1
……	……	……	……	……	……
文本 N	$\text{TF}-\text{IDF}_{N1}$	$\text{TF}-\text{IDF}_{N2}$	……	$\text{TF}-\text{IDF}_{NK}$	1

- $TF-IDF_{ji}$越大词i的对文本j越重要

- 贝叶斯分类器的应用：法律裁判文书中的案情要素分类
  - 文本的量化处理：

```
documents = ["中国的发展是开放的发展",
    "中国经济发展的质量在稳步提升，人民生活在持续改善",
    "从集市、超市到网购，线上年货成为中国老百姓最便捷的硬核年货",
    "支付体验的优化以及物流配送效率的提升，线上购物变得越来越便利"]
documents = [" ".join(jieba.cut(item)) for item in documents]
print("文本分词结果：\n",documents)
vectorizer = TfidfVectorizer()  #定义TF-IDF对象
X = vectorizer.fit_transform(documents)

words=vectorizer.get_feature_names()
print("特征词表：\n",words)
print("idf:\n",vectorizer.idf_)  #idf
X=X.toarray() #print(X.toarray())  #文本-词的tf-idf矩阵
for i in range(len(X)): ##打印每类文本的tf-idf词语权重，第一个for遍历所有文本，第二个for便利某一类文本下的词语权重
    for j in range(len(words)):
        print(words[j],X[i][j])
```

```
文本分词结果：
 ['中国 的 发展 是 开放 的 发展', '中国 经济 发展 的 质量 在 稳步 提升 ， 人民 生活 在 持续 改善', '从 集市 、 超市 到 网购 ， 线上 年货
成为 中国 老百姓 最 便捷 的 硬核 年货', '支付 体验 的 优化 以及 物流配送 效率 的 提升 ， 线上 购物 变得 越来越 便利']
```

```
特征词表：
 ['中国', '人民', '以及', '优化', '体验', '便利', '便捷', '发展', '变得', '年货', '开放', '成为', '持续', '提升', '支付', '改善', '效率',
'物流配送', '生活', '硬核', '稳步', '线上', '经济', '网购', '老百姓', '质量', '购物', '超市', '越来越', '集市']
idf:
 [1.22314355 1.91629073 1.91629073 1.91629073 1.91629073 1.91629073
 1.91629073 1.51082562 1.91629073 1.91629073 1.91629073 1.91629073
 1.91629073 1.51082562 1.91629073 1.91629073 1.91629073 1.91629073
 1.91629073 1.91629073 1.91629073 1.51082562 1.91629073 1.91629073
 1.91629073 1.91629073 1.91629073 1.91629073 1.91629073 1.91629073]
```

IDF

```
中国 0.32346721385745636
人民 0.0
以及 0.0
优化 0.0
体验 0.0
便利 0.0
便捷 0.0
发展 0.7990927223856119
变得 0.0
年货 0.0
开放 0.5067738969102946
```

TF-IDF

- 贝叶斯分类器的应用：法律裁判文书中的案情要素分类
  - 文本和文本标签的组织：通常采用**JSON**格式组织文本和对应的文本分类标签。**JSON**（**JavaScript Object Notation**）是一种典型的便于数据共享的格式文本，在**Python**中与字典结构相对应
  - Python字典：由键和值构成
    - 例如：{"labels": [], "sentence": "原告林某某诉称：我与被告经人介绍建立恋爱关系，于1995年在菏泽市民政局办理结婚登记手续。"}
  - JSON格式的文本文件：

[{"labels": [], "sentence": "原告林某某诉称：我与被告经人介绍建立恋爱关系，于1995年在菏泽市民政局办理结婚登记手续。"}, {"labels
15日生次女李某丙，2007年11月生一女李某丁。"}, {"labels": [], "sentence": "双方婚后因生活琐事产生矛盾。"}, {"labels": [], "sentenc
[{"labels": [], "sentence": "原告黄某某诉称：婚后，我们未能建立起夫妻感情，被告方某甲脾气暴躁，经常酗酒后殴打辱骂我。"}, {"labe
离婚。"}, {"labels": [], "sentence": "案件受理费100元，由原告黄某某负担。"}]

  - 具体步骤：
    - 读入JSON格式的离婚诉讼文本并以字典对象形式存储。
    - 利用旁置法划分文本数据；
    - 对于训练样本集，进行分词处理，计算TF-IDF并建立朴素贝叶斯分类器
    - 对于测试样本集，进行分词处理，计算TF-IDF并计算模型的测试误差