**语义情感分析**

1. **基本步骤**

分类应用

分类算法选择

特征挖掘

文本预处理

训练语料库

1. **文本预处理**

由于一般我们爬下来的文本中有很多htlm的一些标签，所以我们先要将这些标签去除掉。少量的非文本内容可以直接用python的正则（re）表达式删除，复杂的则可以用beautifulsoup（https://www.crummy.com/software/BeautifulSoup/）去除。接下来就可以进行真正的文本预处理。

**中文分词**

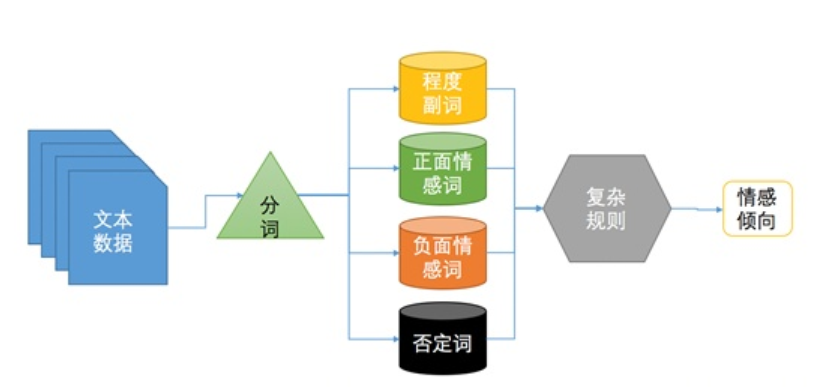
中文分词可以粗略的分为两大类第1类是基于语言学知识的规则方法，如各种形态的最大匹配、最少切分方法、以及综合了最大匹配和最少切分的N-最短路径方法。第2类是基于大规模语料库的机器学习方法，这是目前应用比较广泛、效果较好的解决方案。因此这里我们可以针对我们学要操作的对象选择合适的词库即可。

**极性词挖掘及情感分类的几种不同方式**

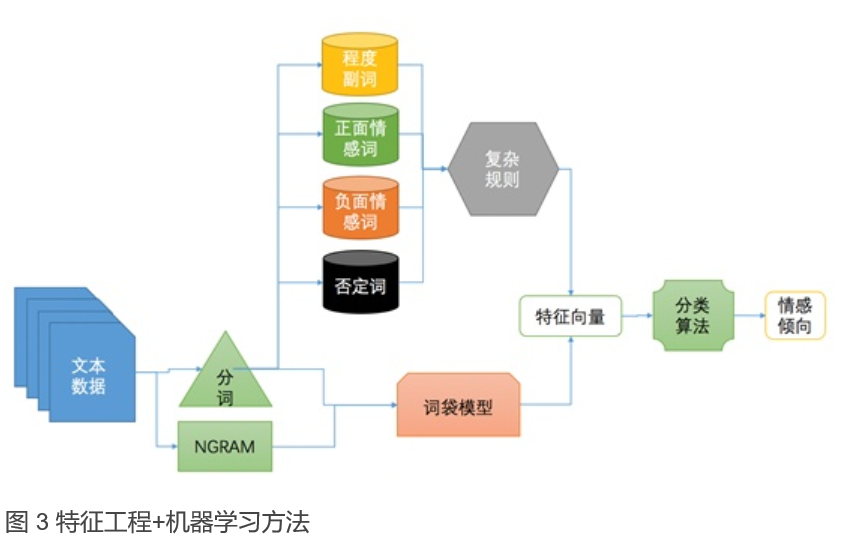
**基于情感词典和规则**

**设置情感词典：**情感词典一般包括五个词典，分别是正面情感词典、负面情感词典、否定词典、程度副词词典和行业情感词典。正面和负面情感词典相对比较容易理解，否定词典例如“不”，会使情感发生逆转。程度副词例如“极其、非常”等，会对情感有增强作用。最终会影响整个句子或者短语的情感倾向。行业情感词典主要在某个特定的行业起作用。

**基于规则匹配：**基于分词和情感词典，可以根据人们平时的语言表达习惯设计一些规则来计算文本的情感倾向。比如每遇到一个正面情感则加一分，遇到负面情感-1分，遇到否定词则乘以-1，遇到程度副词则乘以一个系数然后根据计算出的分数判断情感倾向。



**特征工程加机器学习**

****

**特征提取（极性词挖掘）**

情感分类中“极性词挖掘”，有一种方法是“全词表法”，即将所有词都作为极性词，这样做的好处是所有词语都被保留，但会导致特征维度特别大，计算极其复杂，因此我们采用“极性词表法”，即从文本中挖掘出一些能够表示正负极性（能够表示出评论者的态度）的词或者短语。如“@jjhuang:微信电话本太赞了，能免费打电话，推荐你使用哦”这句话，我们只要挖掘出“赞”、“推荐”这两个极性词。

在极性词挖掘之前可以对分词后的文本爱进行一次预处理，去除掉一些不具有意义的虚词和代词如“的”、“啊”、“我”等，这样便降低了需要处理文件的维度。

在进行完预处理后，就可以进行“极性词挖掘（特征选择）”。特征值挖掘的常用方法有TF-IDF、卡方、互信息、信息增益、X2统计量、交叉熵、Fisher判别式等方法。工业上经常用TF-IDF和信息增益这两种方法。

**TF-IDF**

**TF：词频**，表示特征T在文档D中出现的次数

**DF:** 包含特征T的文档数，DF越高表示特征T衡量文档之间区别作用越低。例如“我”、“的”这样的词的DF一般最高。

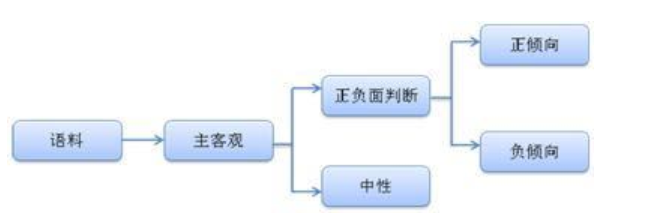
**IDF：**定义为IDF=log|D|/|DF|,|D|为所有文档数。IDF与DF成反比，IDF越高表示特征T对区别文档的意义越大。最终定义**TF-IDF=TF\*IDF**

**信息增益**

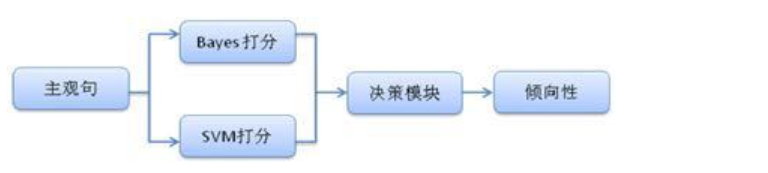
信息增益是工人较好的特征选择方法，它刻画了一个词语在文本中出现与否对情感分类的影响，即一个词语在文本前后出现的信息熵之差

**极性判断**

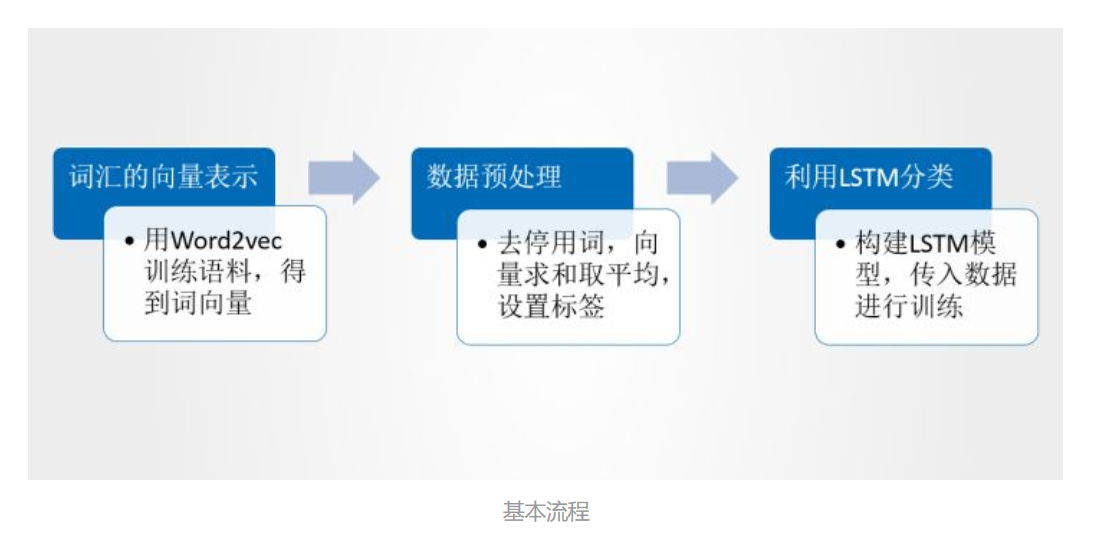
极性判断的任务是判断预料的正、负、中极性这是一个复杂的三分类问题。为了简化问题们可以先对预料做一个主客观判断，客观预料即为中性语句，主观预料在进行正负极性判断。这样就可以讲一个散费雷问题转化为两分类问题。



在分类其中主客观判断可以选择支持向量机模型，在极性判断中可以同时选择支持向量机和朴素贝叶斯模型。朴素贝叶斯使用人工审核后的极性词作为特征，而支持向量机则使用全词表作为特征，两个模型会对输入的语库进行判断，给出正负极性的概率，最后由决策模块给出语料的极性



**基于LSTM的情感分析**



**Word2vec**

