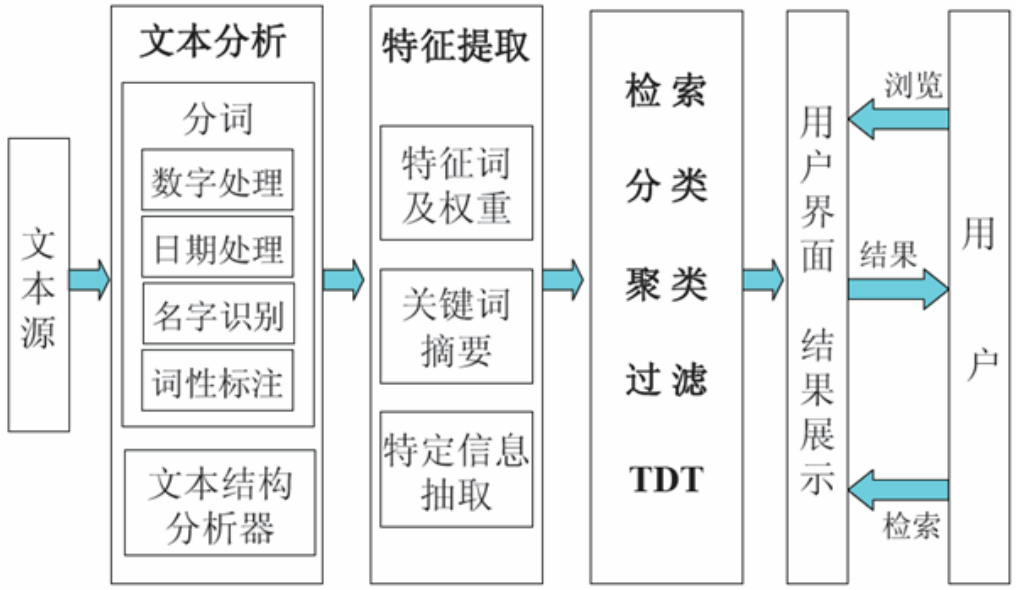
文本特征提取方法

在项目中进行文本挖掘时，需要建立有监督的机器学习模型。下图是一个较为常见的文本挖掘模型结构示意图。



模型是不能直接对文本源进行学习的，我们需要将文本量化为一堆数字，以进行机器学习。从图中我们可以发现，对文本源进行分词，是对文本按照一定的规则进行拆分。而分词是为特征提取服务的。特征提取是文本中的每个词作为特征，使用一定的规则，给每个特征赋予一个或多个特征值。

特征值既可以在送入模型前，对特征数量进行一定减少时，作为可参考的评估值；也可以在送入模型学习时，作为该特征的一个属性被识别。前面两种过程，在不同的论文中都被翻译为不同的名称，在本文中暂且使用下面两组名词进行区分。特征选择：选择值最高的若干个特征。特征提取：构造特征评分函数，把测量空间的数据投影到特征空间。

下面是对目前较为常见的特征提取方法的一些介绍。

（一） 基于评估函数的特征提取方法

基于评估函数的特征提取方法，是在特征独立的假设基础上，对特征集合中的每个特征进行评估，所得值按大小排列，取预定数目的最优结果作为提取结果的特征子集。

下例中，T表示词条，C_i表示类别。

TF-IDF

TF-IDF 由词频(TF)和逆向文件频率(IDF)构成。

$$\mathrm{TF(T,d)} = \frac{n_t}{n}$$

$$\mathrm{IDF(T)} = \log\frac{n_d}{1 + DF(d,T)}$$

$$\mathrm{TF - IDF(T,d)} = \mathrm{TF(T,d)} * \mathrm{IDF(T)}$$

其中n_t是 T 在文档 d 中出现的次数，n 是文档 d 的总次数。 n_d是文档的总个数，$DF(d,T)$是含有 T 的文档个数。在总文本数n_d一定的情况下，词频越高，逆向文本频率越低，说明这个词在不同文本中出现机会越高，相关度越低；反之则说明相关度越高。

词条和类别的互信息

$$\mathrm{MI(T},C_i) = \log(\frac{P(T|C_i)}{P(T)})$$

其中$P(T|C_i)$表示T在C_i中出现的概率，$P(T)$表示 T 在整个训练集中出现的概率。词条和类别的互信息体现了词条和类别的相关程度，互信息越大，相关程度越高。

词条的 CHI 统计

$$\mathrm{CHI(T},C_i) = \sqrt{\frac{(n_{11}+n_{12}+n_{21}+n_{22})(n_{11}\times n_{12} - n_{21}\times n_{22})^2}{(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})}}$$

其中n_{11}表示T在C_i中出现的概率，n_{12}表示T在C_i以外的类别中出现的概率，n_{21}表示T以外的词条在C_i中出现的概率，$n_{22}T$以外的词条在C_i以外的类别中出现的概率。

CHI 统计比较了一个词条对某个类别和其他类别的贡献，同时也比较了该词条和其他词条对分类的影响。

假如$(n_{11}\times n_{12} - n_{21}\times n_{22}) > 0$，说明$T$与$C_i$正相关：假如$T$出现，则$C_i$也可能出现。反之，$(n_{11}\times n_{12} - n_{21}\times n_{22}) < 0$，则说明负相关：假如$T$出现，则$C_i$可能不会出现。

词条的期望交叉熵

$$\mathrm{CE(T)} = \sum_i P(C_i|T) log\frac{P(C_i|T)}{P(C_i)}$$

其中$P(C_i|T)$表示，当词条T出现时，文本属于C_i的概率。该值大，说明T与C_i强相关，反之则弱相关。假如$P(C_i|T)$值很大且$P(C_i)$的值很小，说明该词条对该分类的影响大，则相应的$\mathrm{CE(T)}$的值也会很大，容易被选作为特征值。

* 存在用 TF-IDF 代替$P(T)$的论文

庞剑锋, 卜东波, 白?. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究, 2001, 18(9):23-26.

（二） 基于文本相似度的特征提取方法

向量空间模型 VSM

1. 欧式距离

$$\mathrm{Sim(x,y)} = |\mathrm{x} - \mathrm{y}| = \sqrt{\sum_{k=1}^{t}(x_k - y_k)^2}$$

欧式距离越小，说明两个词语越相似。

2. 向量内积

$$\mathrm{Sim(x,y)} = \mathrm{x} \cdot \mathrm{y} = \sum_{k=1}^{t}(x_k \cdot y_k)$$

3. 向量夹角余弦

$$\mathrm{CosSim(x,y)} = \frac{x \cdot y}{|x| \cdot |y|} = \frac{\sum_{k=1}^{t}(x_k \cdot y_k)}{\sqrt{\sum_{k=1}^{t} {x_k}^2}\sqrt{\sum_{k=1}^{t} {y_k}^2}}$$

向量夹角余弦越大，说明词语与语义维度相似度越高。

例：$D_1 = 2T_1 + 3T_2 + 5T_3$，$D_2 = 3T_1 + 7T_2 + T_3$，$\mathrm{Q} = 0T_1 + 0T_2 + 2T_3$

D 为词向量，T_1为语义维。

1. $\mathrm{Sim}(D_1, D_2) = \sqrt{(1 + 16 + 16)} = 5.74$
2. $\mathrm{Sim}(D_1, \mathrm{Q}) = 2 \times 0 + 3 \times 0 + 5 \times 2 = 10$ $\mathrm{Sim}(D_2, \mathrm{Q}) = 3 \times 0 + 7 \times 0 + 1 \times 2 = 2$
3. $\mathrm{CosSim}(D_1, \mathrm{Q}) = \frac{10}{\sqrt{38 \times 4}} = 0.81$ $\mathrm{CosSim}(D_2, \mathrm{Q}) = \frac{2}{\sqrt{59 \times 4}} = 0.13$

潜在语义索引

VSM 建立在词间相互独立的基础上（正交）

潜在语义分析(LSA)又称为潜在语义索引(LSI)，是一种使用数学和统计的方法对文本中的词语进行抽取，推断它们之间的语义关系，并建立一个语义索引，而将文档组织成语义空间结构的方法。

主要就是降维

$$\mathrm{A} = [a_{ij}], \quad a_{ij} = L(i,j) \times G(i)$$

A 为$\mathrm{m} \times \mathrm{n}$词条-文本矩阵，$L(i,j)$是单词 i 在文本 j 中的局部权重，$G(i)$是单词 i 在文本集中的全局权重，m 为单词书，n 为文本数。对任意矩阵，可以进行以下拆解：

$$\mathrm{A} = \mathrm{U}\Sigma V^T$$

奇异值的计算（右奇异值v_i，奇异值σ_i，左奇异值u_i），：

$$(A^T\mathrm{A})v_i = \lambda_i v_i$$

$$\sigma_i = \sqrt{\lambda_i}$$

$$u_i = \frac{1}{\sigma_i} \mathrm{A} v_i$$

$$\mathrm{A} \approx A_K = \mathrm{U}_K \Sigma_K {V_K}^T$$

其中U为$\mathrm{m} \times \mathrm{m}$矩阵，每一行表示意思相关的一类词，非零元素表示这类词中每个词的重要度，数值越大越相关。Σ为$\mathrm{m} \times \mathrm{n}$矩阵，表示类词与文章类之间的相关度。$V$为$\mathrm{n} \times \mathrm{n}$矩阵，每一列表示同一主题的一类文章，非零元素表示相关度；U和V是正交矩阵，Σ是对角矩阵。

U_K为$\mathrm{m} \times \mathrm{k}$矩阵，$\Sigma_K$为$\mathrm{k} \times \mathrm{k}$，$V_K$为$\mathrm{k} \times \mathrm{n}$（$\mathrm{k} << \min(\mathrm{m}, \mathrm{n})$且$\mathrm{k} < \mathrm{rank}(\mathrm{A})$）。

例：

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

一行表示一个词在哪些 title 中出现了，一列表示一个 title 中有哪些词。对矩阵进行SVD，得到以下矩阵：

book	0.15	-0.27	0.04
dads	0.24	0.38	-0.09
dummies	0.13	-0.17	0.07
estate	0.18	0.19	0.45
guide	0.22	0.09	-0.46
investing	0.74	-0.21	0.21
market	0.18	-0.30	-0.28
real	0.18	0.19	0.45
rich	0.36	0.59	-0.34
stock	0.25	-0.42	-0.28
value	0.12	-0.14	0.23

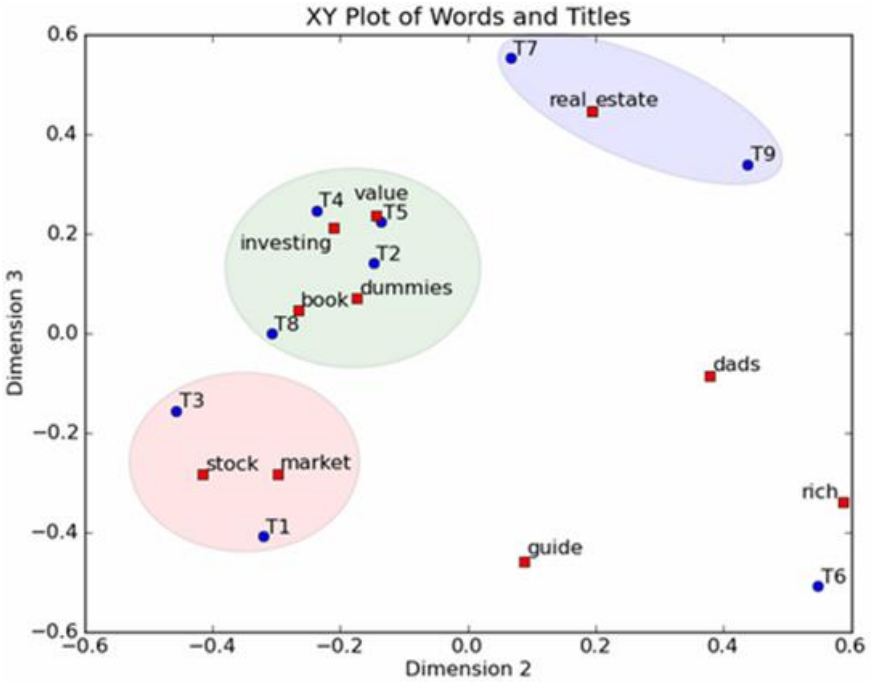
*

3.91	0	0
0	2.61	0
0	0	2.00

*

T1	T2	T3	T4	T5	T6	T7	T8	T9
0.35	0.22	0.34	0.26	0.22	0.49	0.28	0.29	0.44
-0.32	-0.15	-0.46	-0.24	-0.14	0.55	0.07	-0.31	0.44
-0.41	0.14	-0.16	0.25	0.22	-0.51	0.55	0.00	0.34

左奇异向量表示词的一些特性，右奇异向量表示文档的一些特性，中间的奇异值矩阵表示左奇异向量的一行与右奇异向量的一列的重要程序，数字越大越重要。将左奇异向量和右奇异向量都取后 2 维（之前是 3 维的矩阵），投影到一个平面上，可以得到下图。每一个红色的点，都表示一个词，每一个蓝色的点，都表示一篇文档：



其他

除此之外，文本挖掘中比较常用的还有**遗传算法(Genetic Algorithm, GA)**、**N-Gram 算法**等，在这就不作详细介绍了。

（三） 项目中的应用

在项目中，需要进行情感分析，即文本被分类为积极文本和消极文本两类。同时由于训练集的数量和性能的限制，需要对文本特征进行一部分的筛选。在对文本特征提取方法进行调查后，由于 CHI 既代表特征对所属分类文本的相关性，也代表特征对其他文本的相关性。而且在参考既有文献对不同文本特征提取方法用于分类的性能评价时，发现在使用单一文本特征提取方法时，CHI 的性能较优。因此选择使用 CHI 计算用于特征选择的特征值，使用词频计算用于特征提取的特征值。