特征选择.

1. TF-IDF 原理.

2. 文本矩阵化. 使用 Bag of word 模型. 以 TF-IDF 特征值为权重. (the library TfidfTransformer in Python)

3. 互信息的原理.

4. 使用第二步生成的特征矩阵, 利用互信息进行特征筛选.

TF-IDF 原理.

(b) Normalised Term Frequency.

$$tf_{ij} = \frac{f_{ij}}{\max_k(f_{kj})}$$

measure of how common or rare a particular term is.

1

$$idf_i = \log \frac{|D|}{|\{d_j \in D : term_i \in d_j\}|}$$

(c)

Tf-idf

$$tf_{ij} = tf_{ij} \times idf_i$$

文本相似

点互信息和互信息.

点互信息 (Pointwise Mutual Information).

$$PMI(x;y) = \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}$$

表示x与y的相关程度.

互信息 (Mutual Information).

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

两随机变量间的关联性，一个随机变量中包含的关于另一个随机变量的信息量.

X和Y的所有可能的取值情况的互信息，即二者的加权和。