1. **（英文）文本预处理**
2. 去掉html里面的标签
3. 去掉语音语调
4. 把缩写展开
5. 去掉特殊符号
6. 把动词原型找出来
7. 去掉停用词
8. **词袋下的特征向量**
9. 全部单词做为词袋的0和1模型
10. N-Grams模型
11. TF-IDF模型
12. **文章聚类**

窗体顶端

你好窗体底端

1. 成对文档相似度矩阵（余弦相似度） & Hierarchical Clustering
2. LDA模型 & K-means