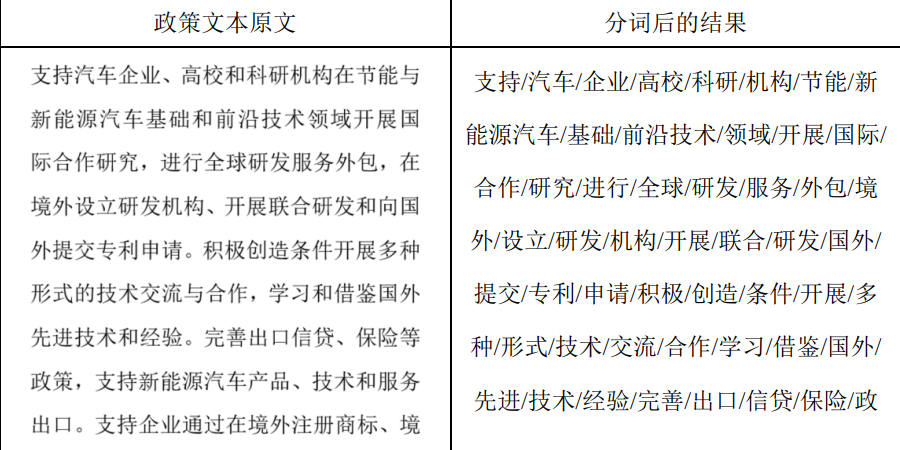
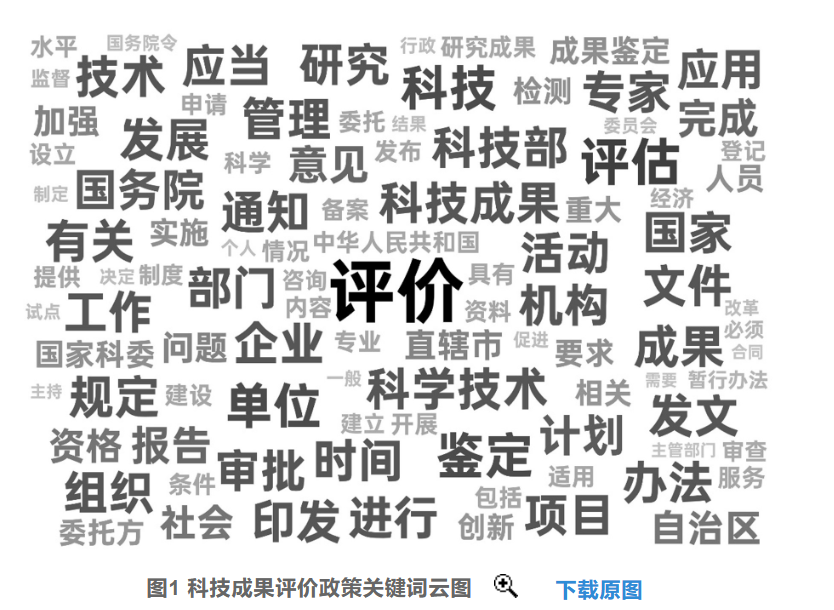
目的：我提供的文本分为四个阶段，做一个主题演化，分析演化过程。（数据还在整理中，具体如何分阶段还在考虑中，但是应该是分四个阶段，想问一下如果前面阶段的文本比较少，然后用四个阶段的困惑度和一致性的变化曲线去确定主题数目，会不会非常影响结果）

Tips：以下哪些不能做可以直接告知，没关系的，大致思路是这样，有问题告诉我可以灵活改变。另外需要有简单的代码注释，尤其是比较复杂的地方，有不懂之处可以帮忙解答或者分享给我相关的博客、论文等我自己去看。

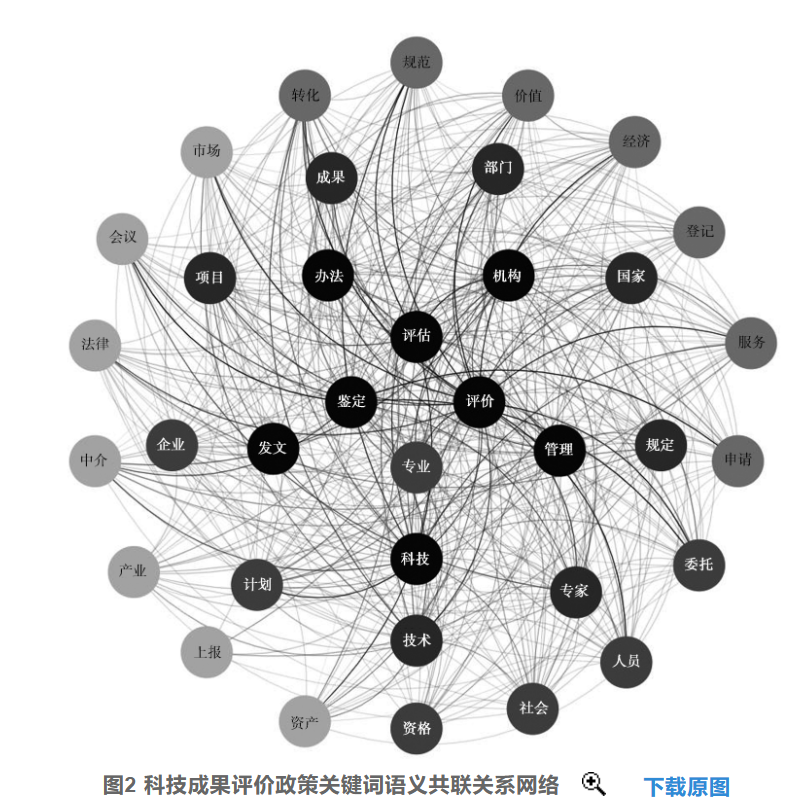
数据：excel，第一列是文本，第二列是时间

1去停用词去标点去空格等，分词前后展示（告诉我如何将专业词汇添加到分词表中帮助分词）

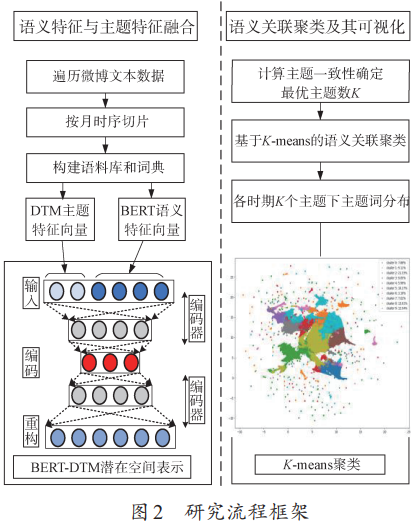
  
2高频词展示，不分阶段

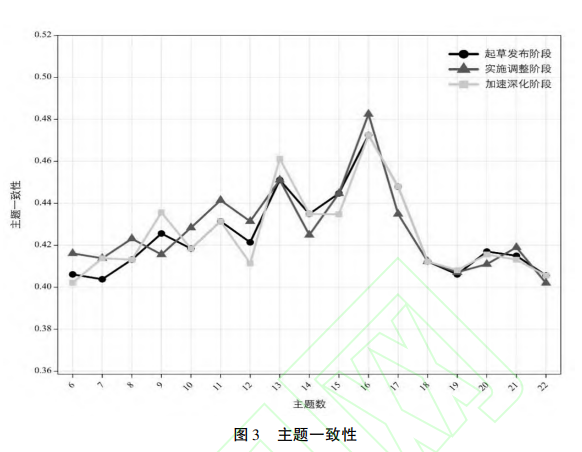


3共词网络图（选取词频统计前80的关键词构建共现矩阵不分阶段）

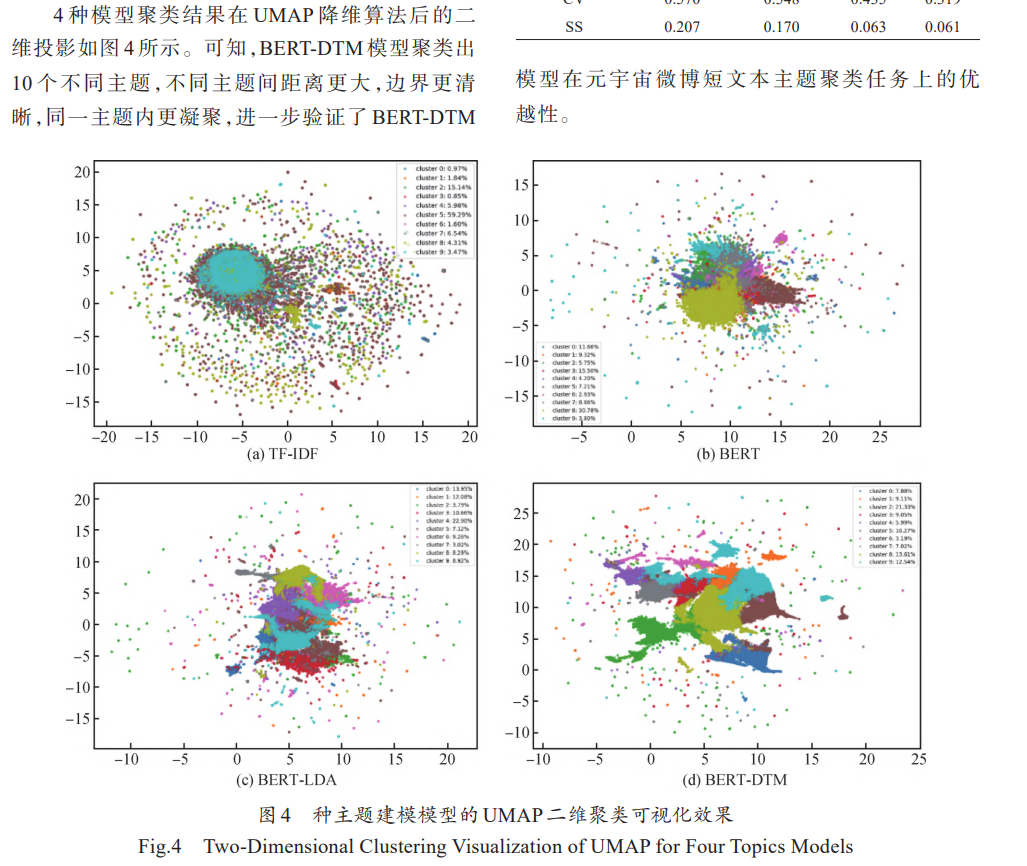


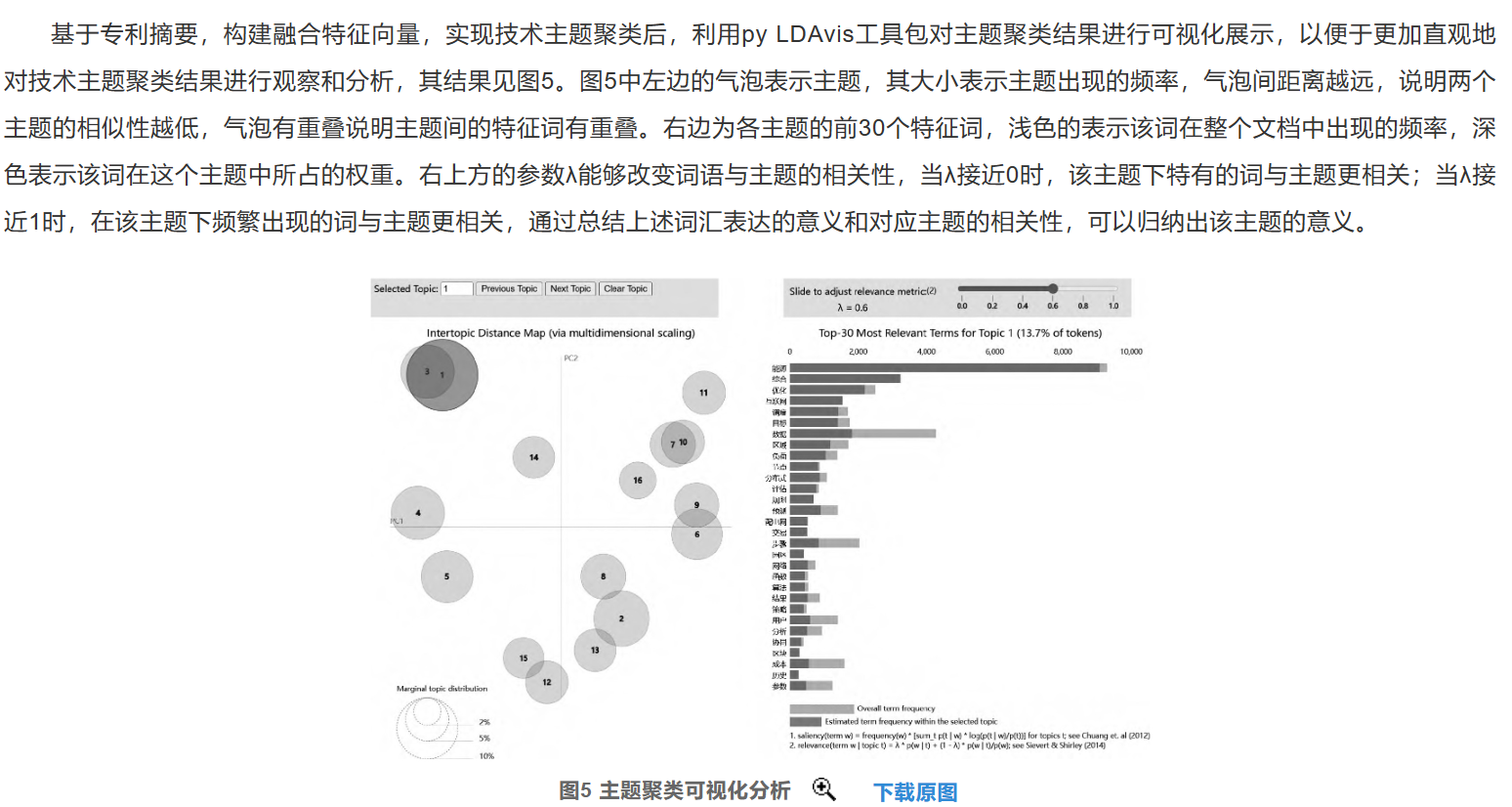
4根据提供的论文做bert+dtm的处理（哪里和论文一样，哪里有改动我需要知道，所以编写的时候希望有一些代码注释，让人知道代码的大概意思）



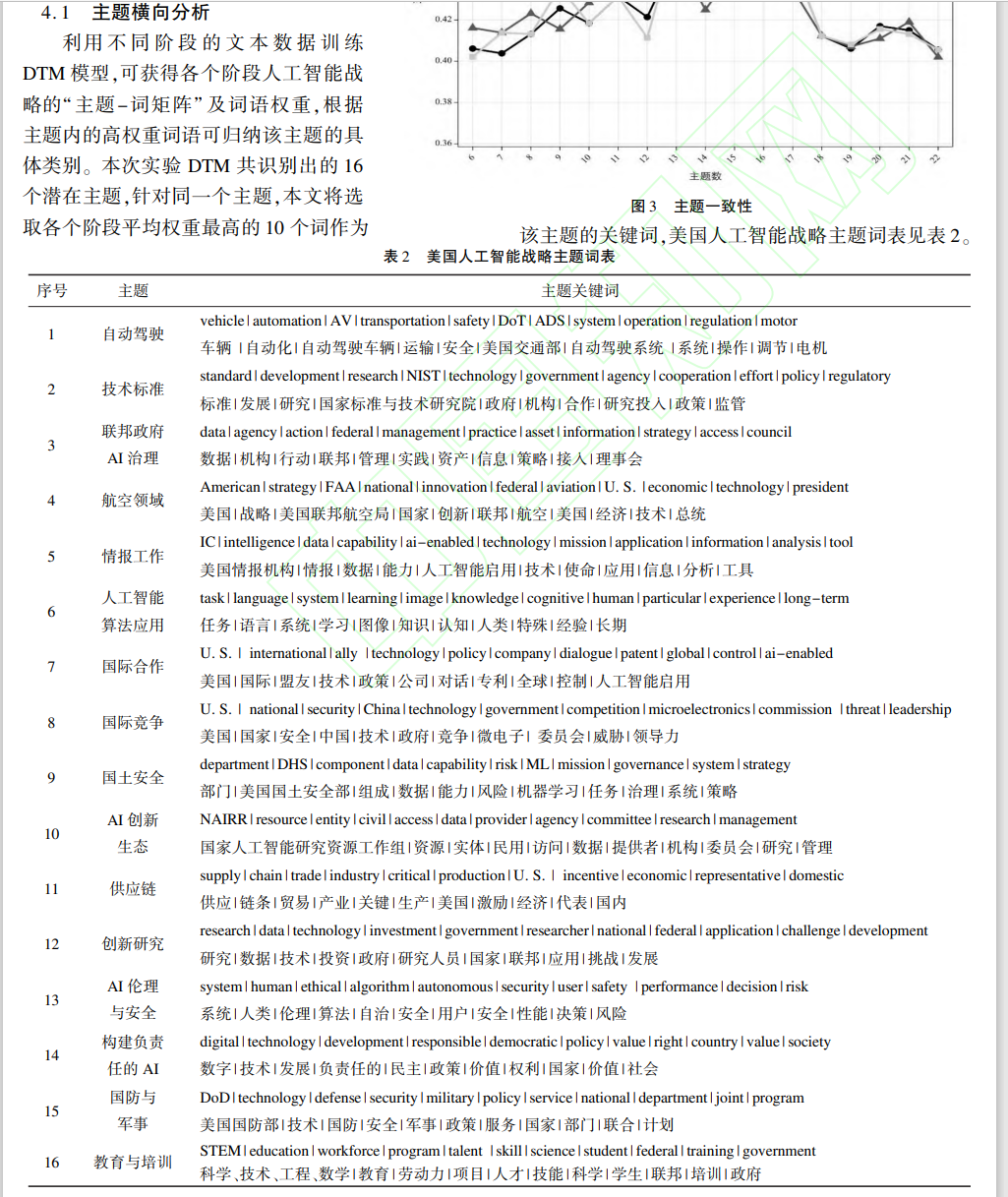
5主题数量确定（困惑度图+一致性图，分阶段，四个阶段在一个图上，多加一个困惑度的）

6聚类效果图（在UMAP降维算法后的二维投影图，只需要展示我们这种方法的投影图+pyldavis气泡图）

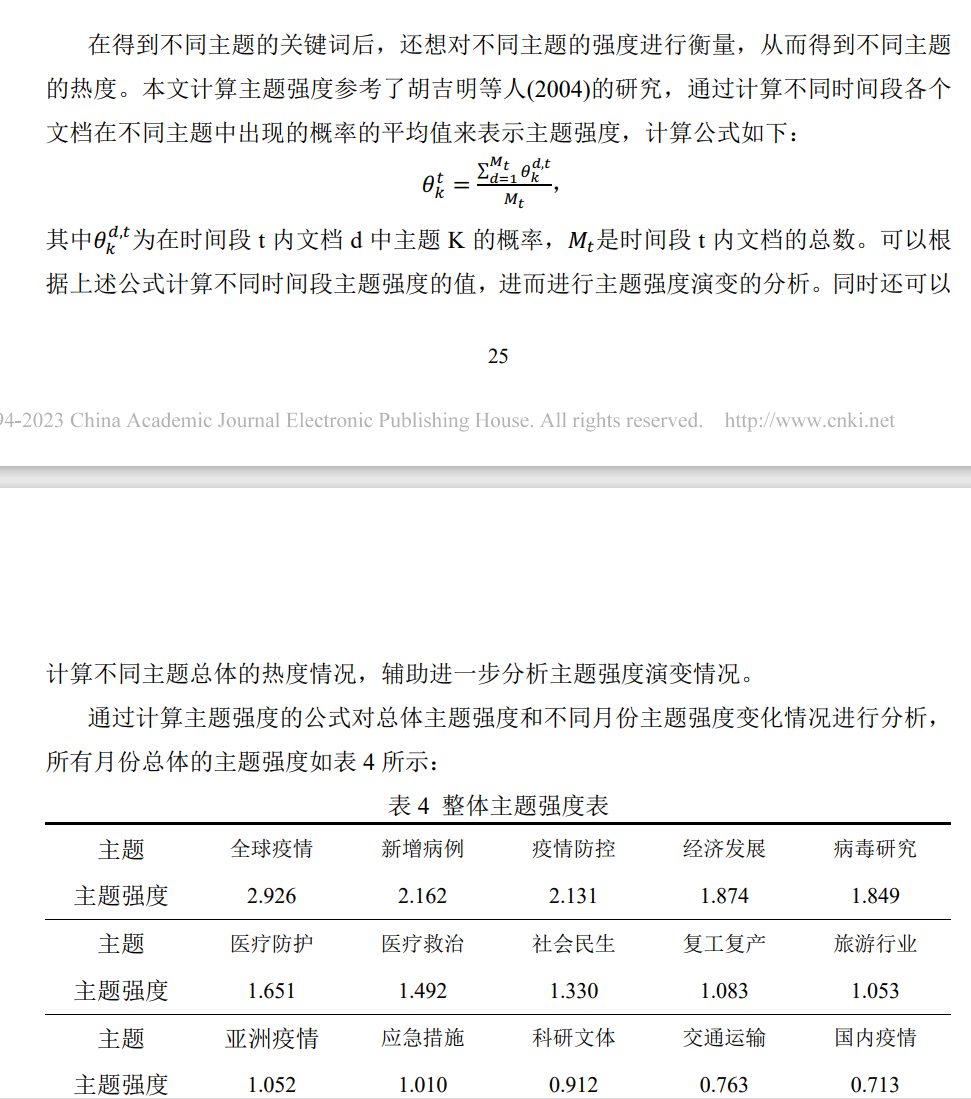




7主题词表



8主题强度计算，得到图表（分四个阶段，可以画折线图看强度变化？）



9主题热度计算和桑基图（和下图一样只不过分四个阶段）

