实验报告

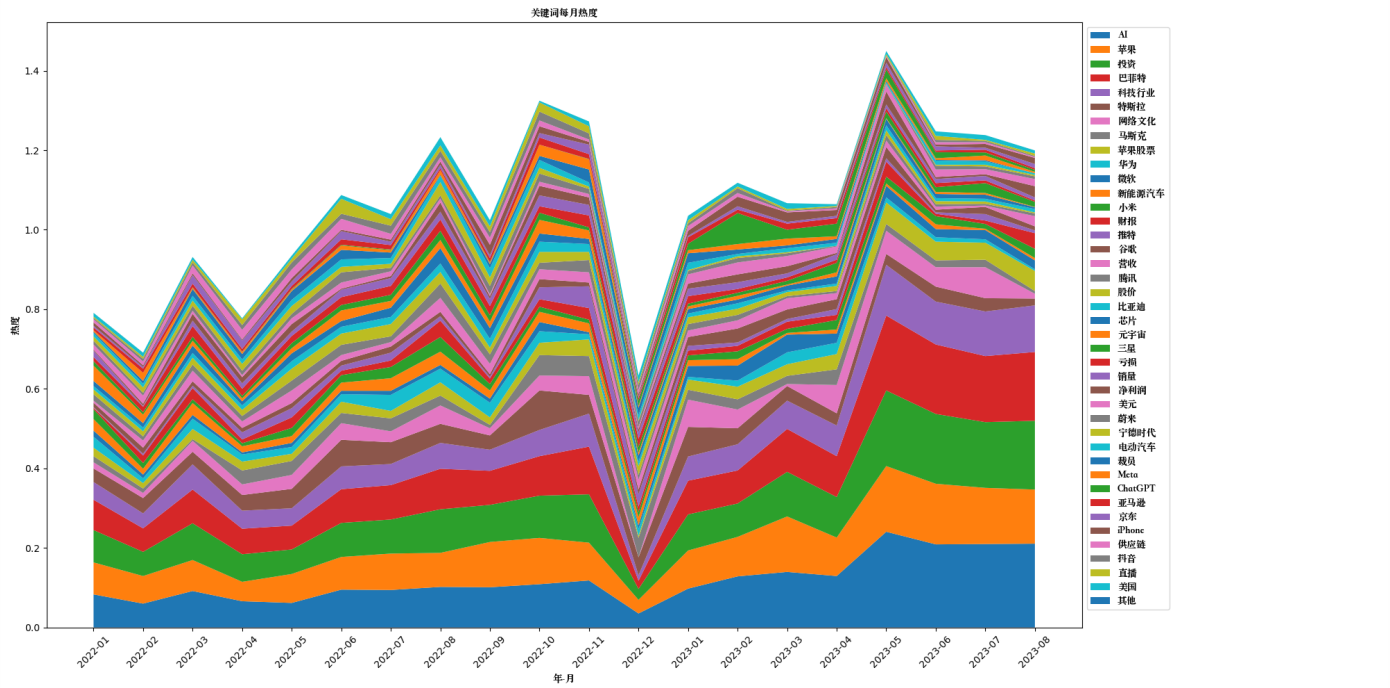
经22-计28 吕博涵 2022011547

1. *通过绘制从2022年1月到2023年8月的关键词热度图得到公众聚焦点的变化*

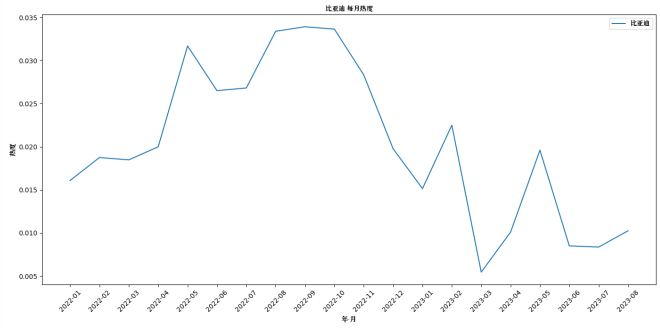
关键词热度的定义是本月包含该关键词的文章数目/本月所有文章数目。因为每个月的绝对文章数目可能不同，因此新闻种类对应的绝对新闻数目不一定有意义，但是这个比例有意义。

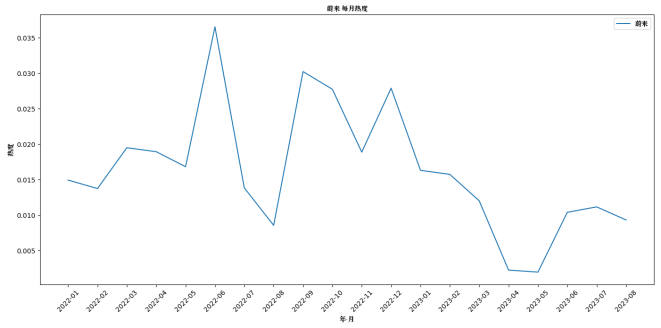
我分别为挑选出来的40个关键词绘制了各自的折线图，也绘制了这些关键词的累积图。

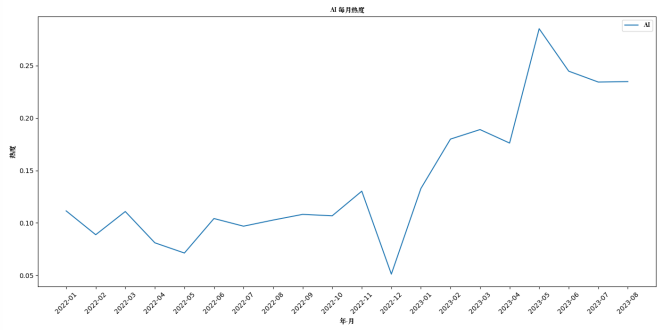
累积图：

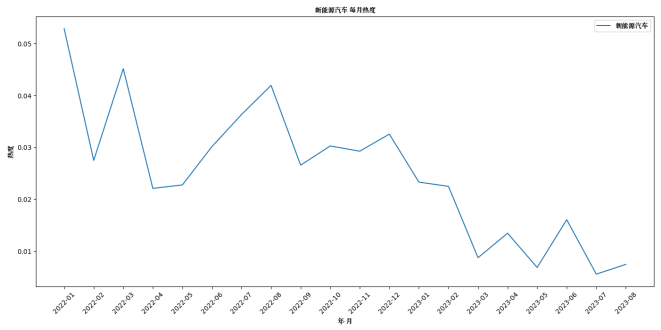


折线图举例：









第一行从左到右分别是：关键词比亚迪、关键词蔚来

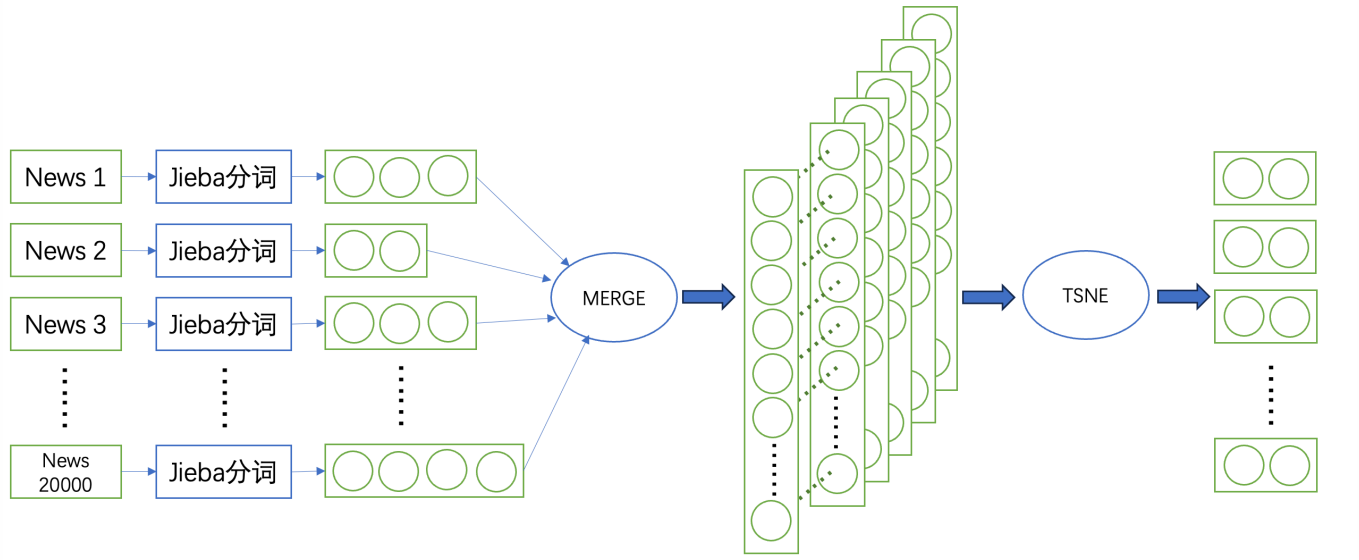
第二行从左到右分别是：关键词AI，关键词新能源汽车

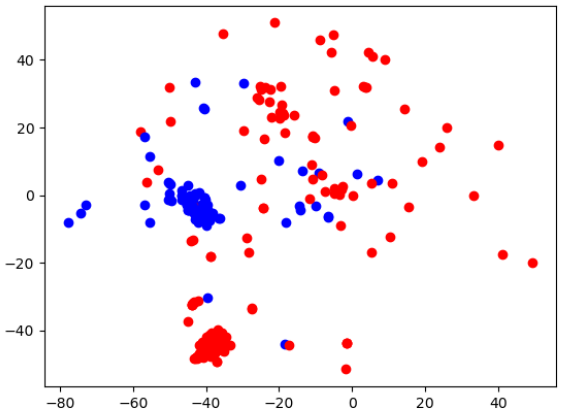
结论1: 这两年来AI的关注度在逐渐升高，而新能源汽车的关注度到了2023年呈下降趋势。

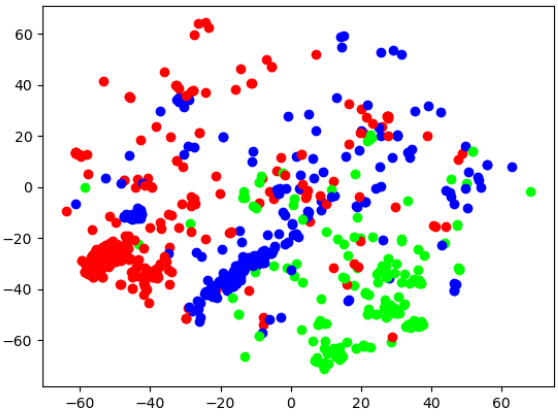
结论2: 关系相近的关键词的热度趋势相似，如上面的比亚迪、蔚来都是新能源汽车概念关键词，其变化趋势与新能源汽车关键词的变化趋势相似。上面没有列举但是同样得到了图片的特斯拉、宁德时代两个关键词也能反映出相同的趋势（宁德时代是新能源汽车电池的制造商），这也反映了相近行业的趋势变化是相近的。

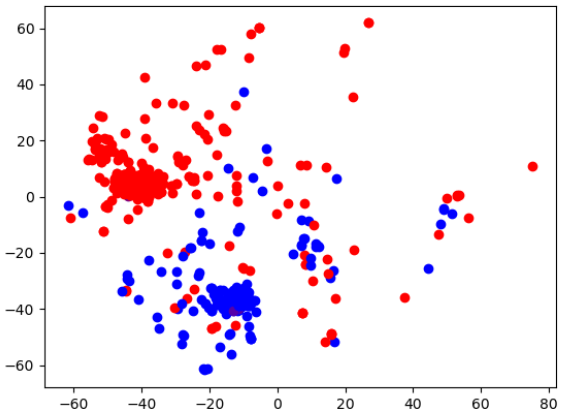
2. *Doc2Vec: 分词+TSNE*

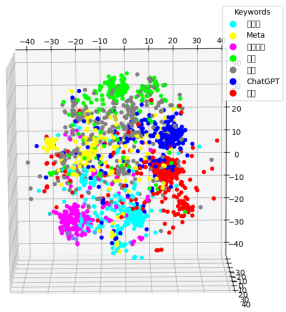
Doc2Vec示意图：

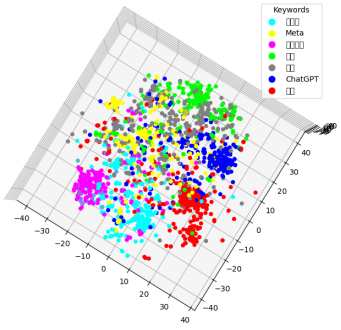


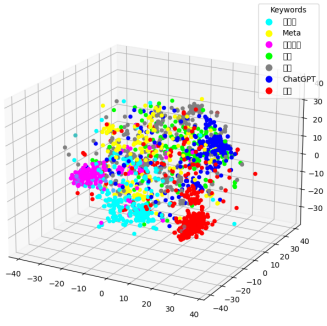


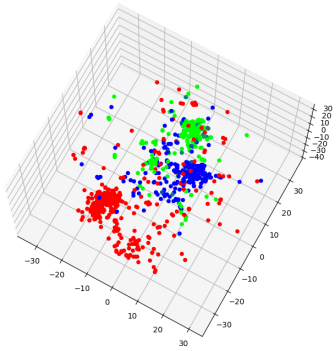


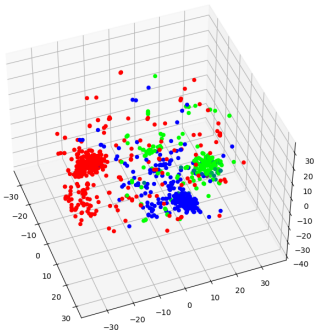



Keywords
Meta
ChatGPT


Keywords
Meta
ChatGPT


Keywords
Meta
ChatGPT



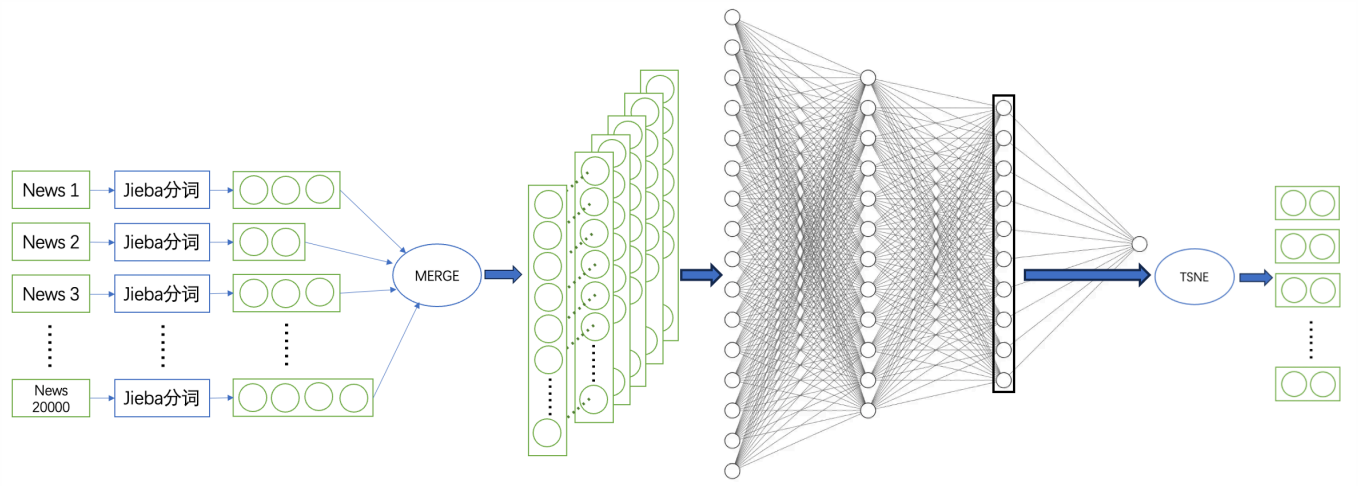


实现方法：我将每篇文章根据分词结果处理成20万维左右的向量（每个分词）对应一个维度，每个文章这个维度的值就是这个文章这个分词的数目。然后根据这些数据使用TSNE绘制成二维或三维的数据并进行可视化，最后用关键词对每个文章对应的点进行颜色标注。从结果来看，不同关键词对应的文章对应的点出现了明显的聚集。

其中部分图中的红色点是“小米”关键词，蓝色点是“ChatGPT”关键词。这两个关键词基本没有重叠的文章，因此出现了明显的划分。

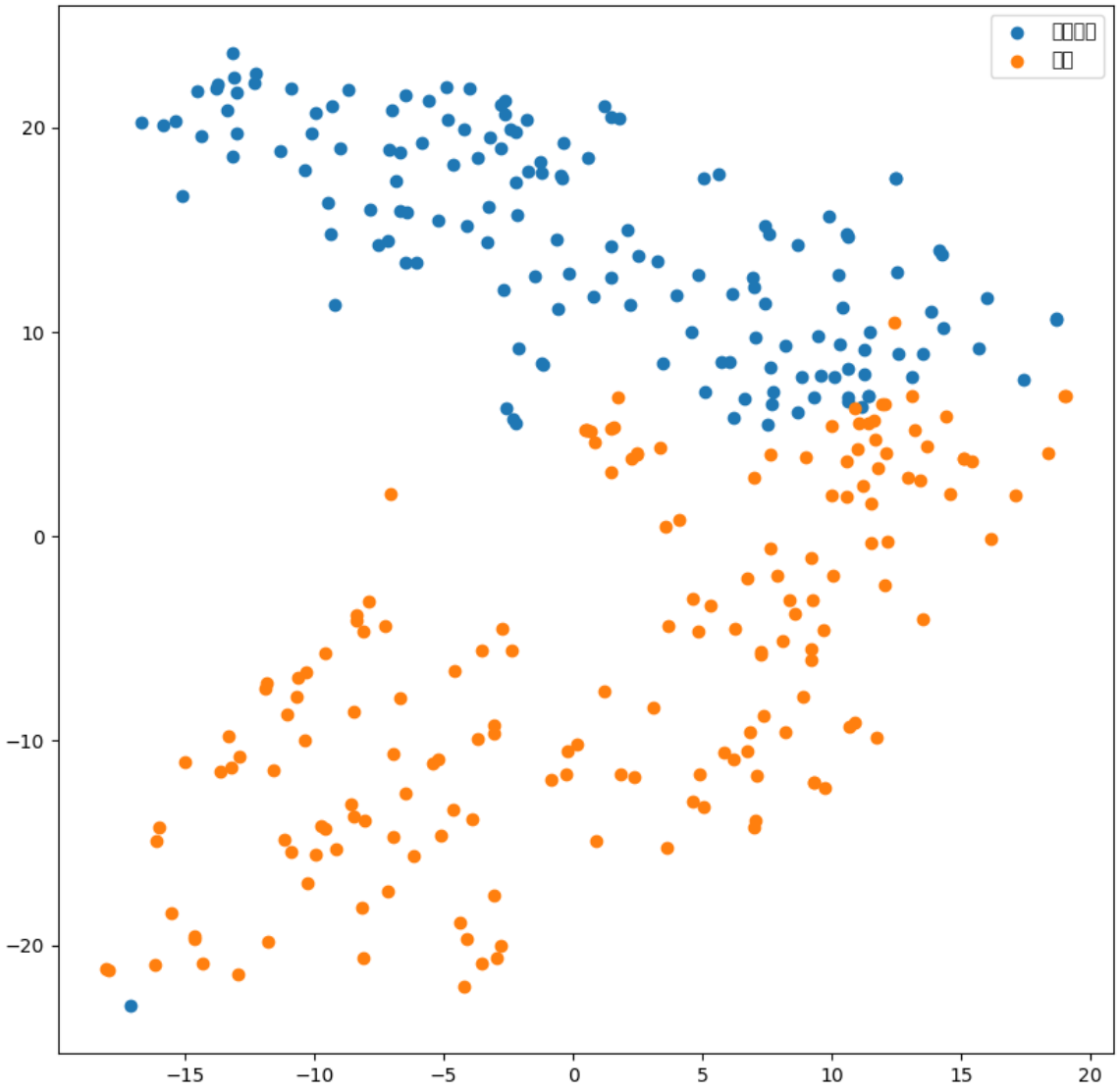
结论3: 文章的分词结果包含了文章的信息。通过用TSNE处理分词向量能对文章进行较好的分类。

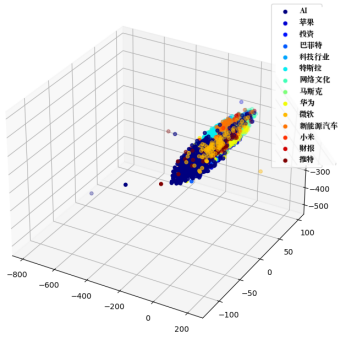
2. *Doc2Vec Plus: 分词+MLP+TSNE*

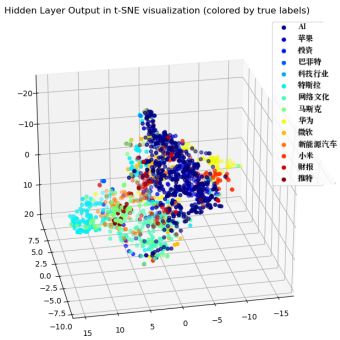


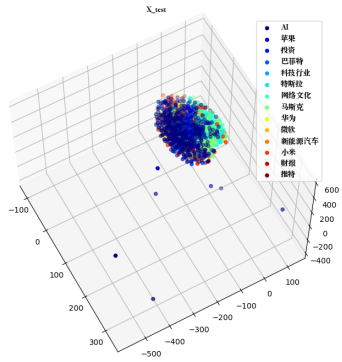
即使上面的结果已经很好了，但是20万维的分词向量包含了很多无用信息。我找到一个方法，在分词向量和TSNE之间添加一个深层感知神经网络（MLP），通过训练用MLP执行分类任务来对其进行训练，然后取每个分析向量到MLP最后一个隐藏层的向量作为TSNE的输入向量。

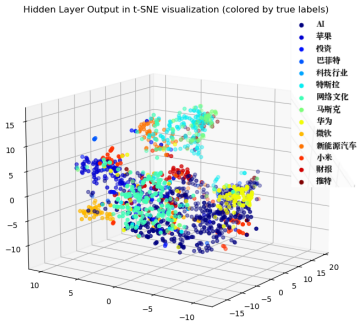
下图是在一个对于两个关键词分类任务的结果（TSNE处理到二维）：（蓝色和黄色分别是宁德时代和小米）


□□□□
□□
20
10
0
−10
−20
−15
−10
−5
0
5
10
15
20









上面这些是添加了更多关键词标注的结果。左边的两个是直接用测试集合进行绘制，右面两个是用测试集数据到最后一个隐藏层的结果绘制。可以看出，虽然左边的两个图不同的颜色（关键词）也有聚集，但是不是非常明显，而右面两个图则非常明显。

结论4: 通过MLP调整分词各个维度的权重可以更好让TSNE更好地进行可视化和分类，也能更好地体现出文章的特征。