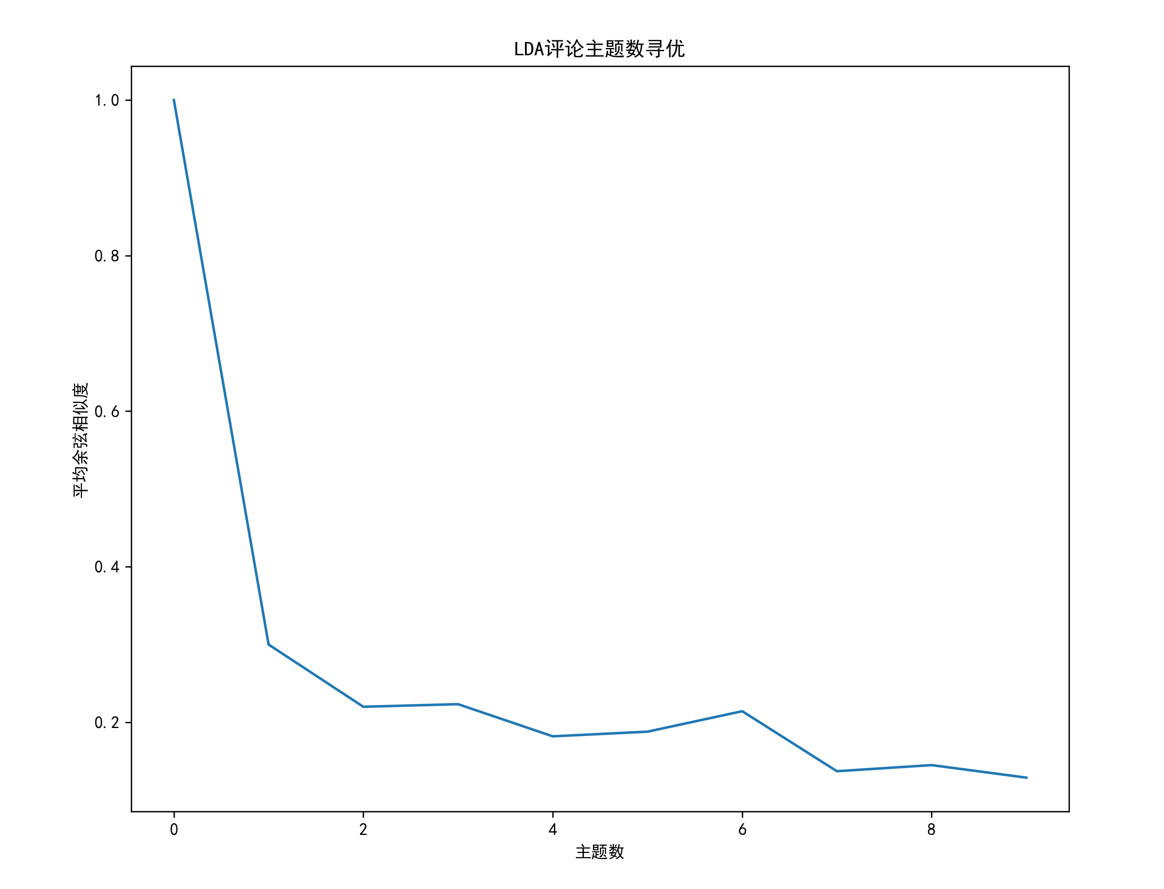
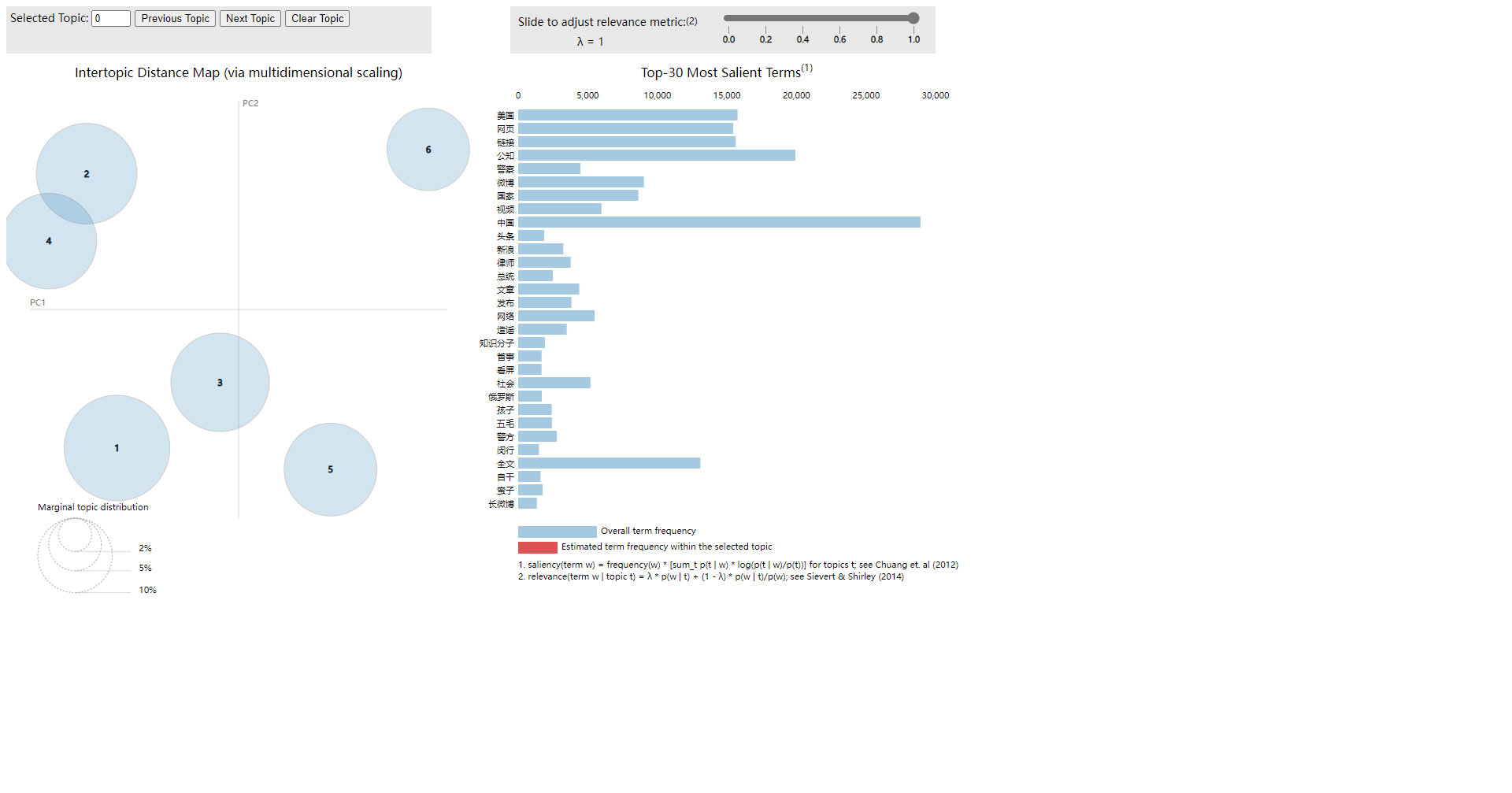
# LDA主题模型分析报告



我们通过数据处理之后，通过平均余弦相似度进行计算，从图中来，从主题数6开始之后，就处于一直下滑的趋势，所以我们的主题数可以选择6



主题相互错开，只有2和4处于交集状态，总体来看，模型主题数选取效果较好，不会出现过拟合的现象

这一步是计算公式：

基于相似度的自适应最优LDA模型选择方法，确定主题数并进行主题分析。实验证明该方法可以在不需要人工调试主题数目的情况下，用相对少的迭代，找到最优的主题结构。具体步骤如下。

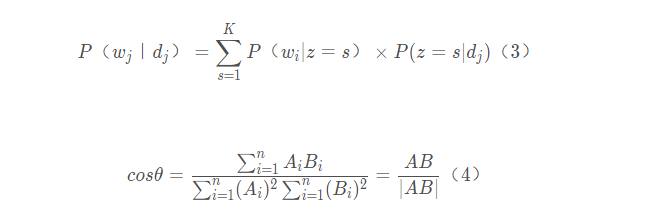
① 取初始主题数k值，得到初始模型，计算各主题之间的相似度（平均余弦距离）。

② 增加或减少k值，重新训练模型，再次计算各主题之间的相似度。

③ 重复步骤②直到得到最优k值。

利用各主题间的余弦相似度来度量主题间的相似程度。从词频入手，计算它们的相似度，用词越相似，则内容越相近。

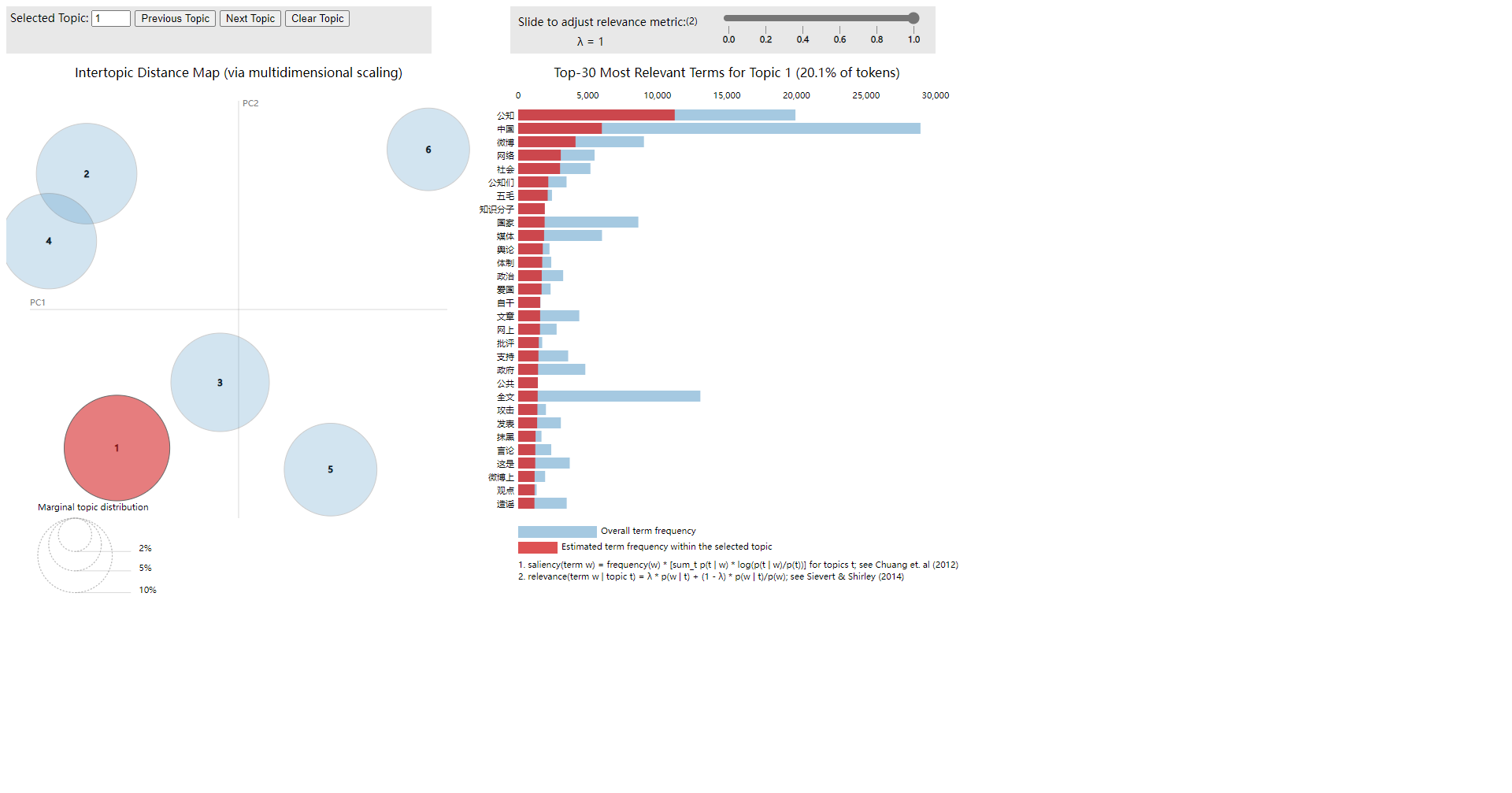
假定A和B是两个n维向量，A是，B是，则A与B的夹角θ的余弦值通过式（4）计算。

​

使用LDA主题模型，找出不同主题数下的主题词；每个模型各取出若干个主题词（比如前100个），合并成一个集合；生成任何两个主题间的词频向量；计算两个向量的余弦相似度，值越大就表示越相似；计算个主题数的平均余弦相似度，寻找最优主题数

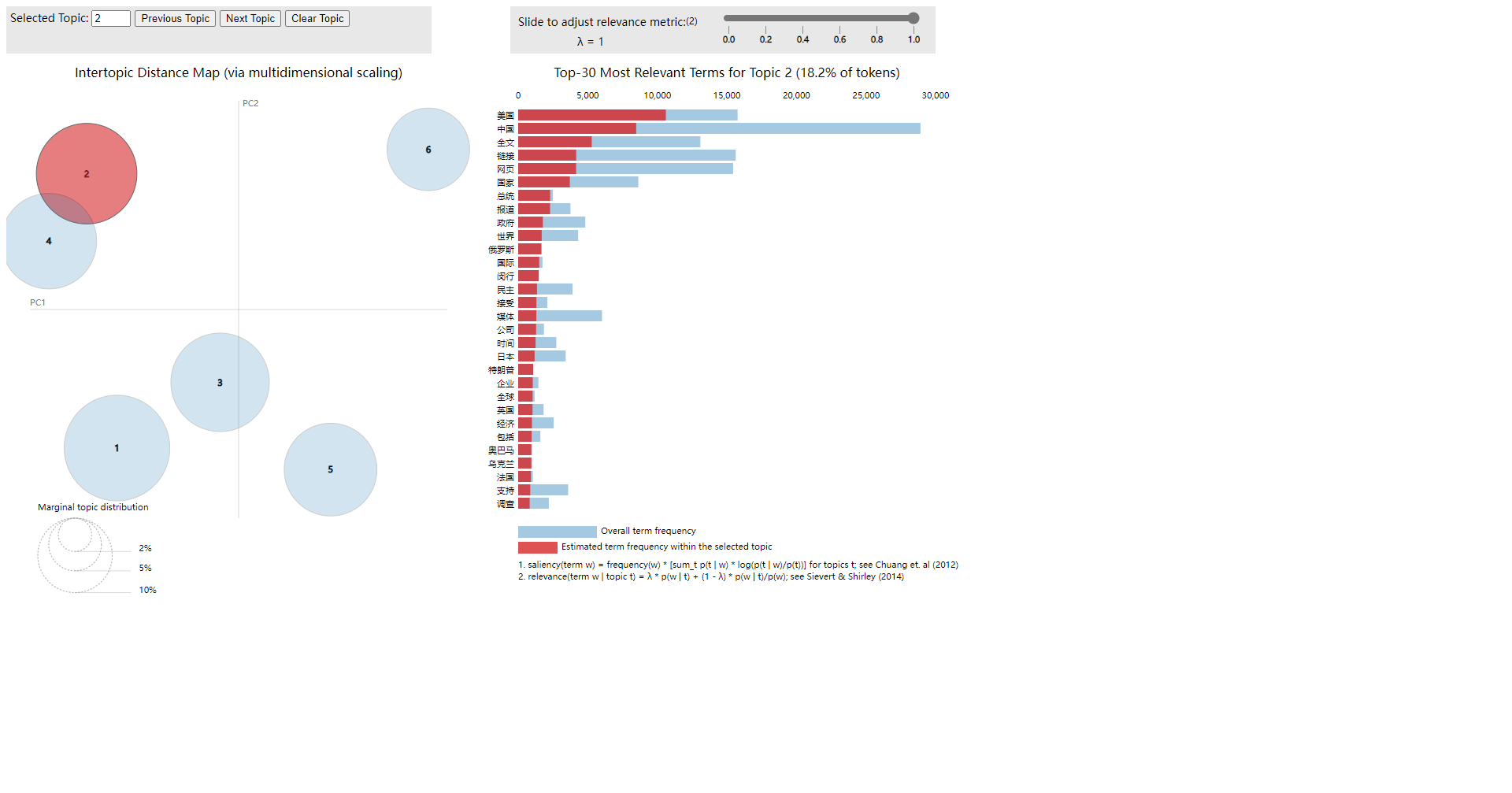
这里一共为6个主题

其中1的主题



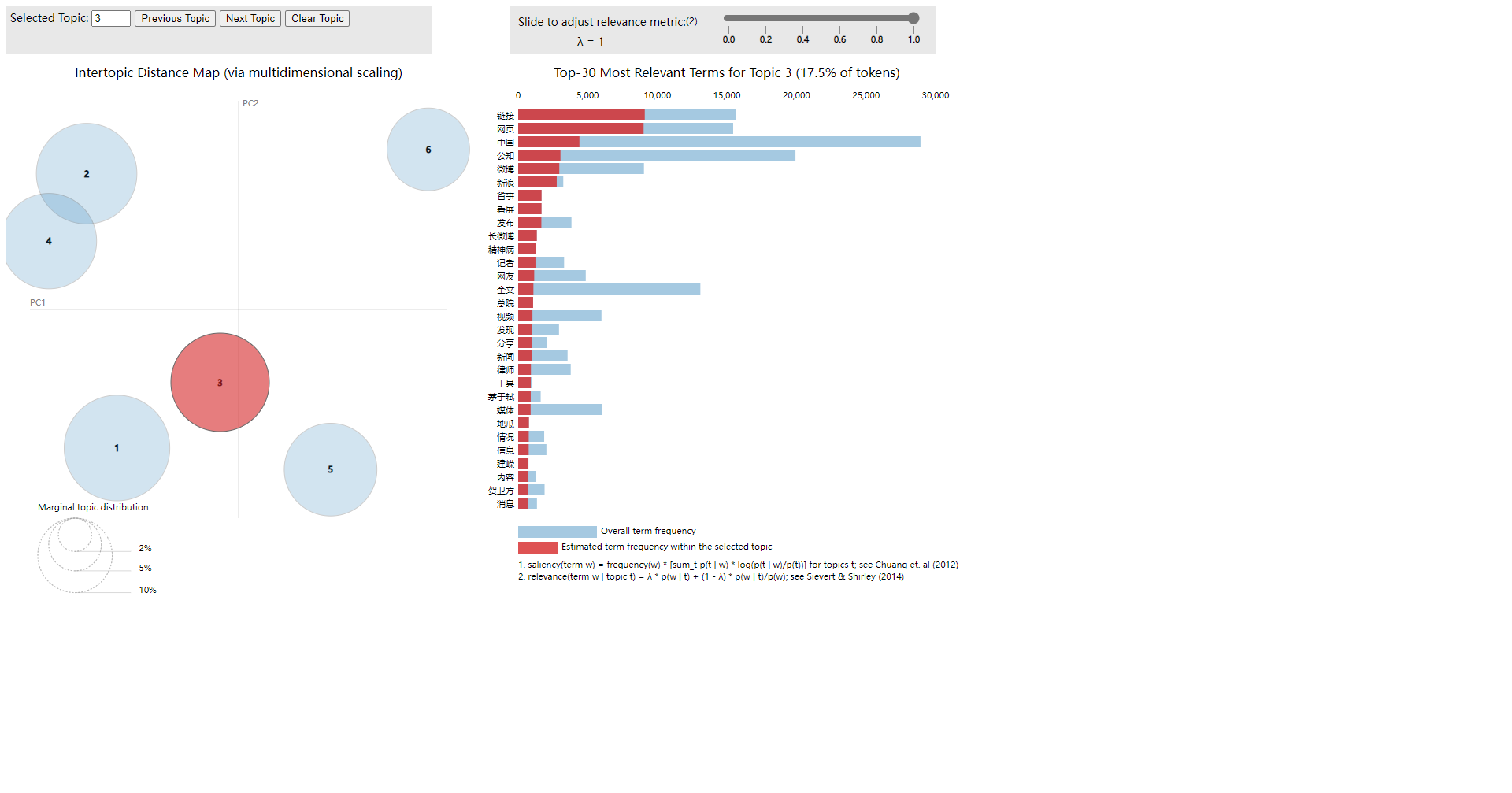
主要是在讲述，公知相关方面的舆论

主题2



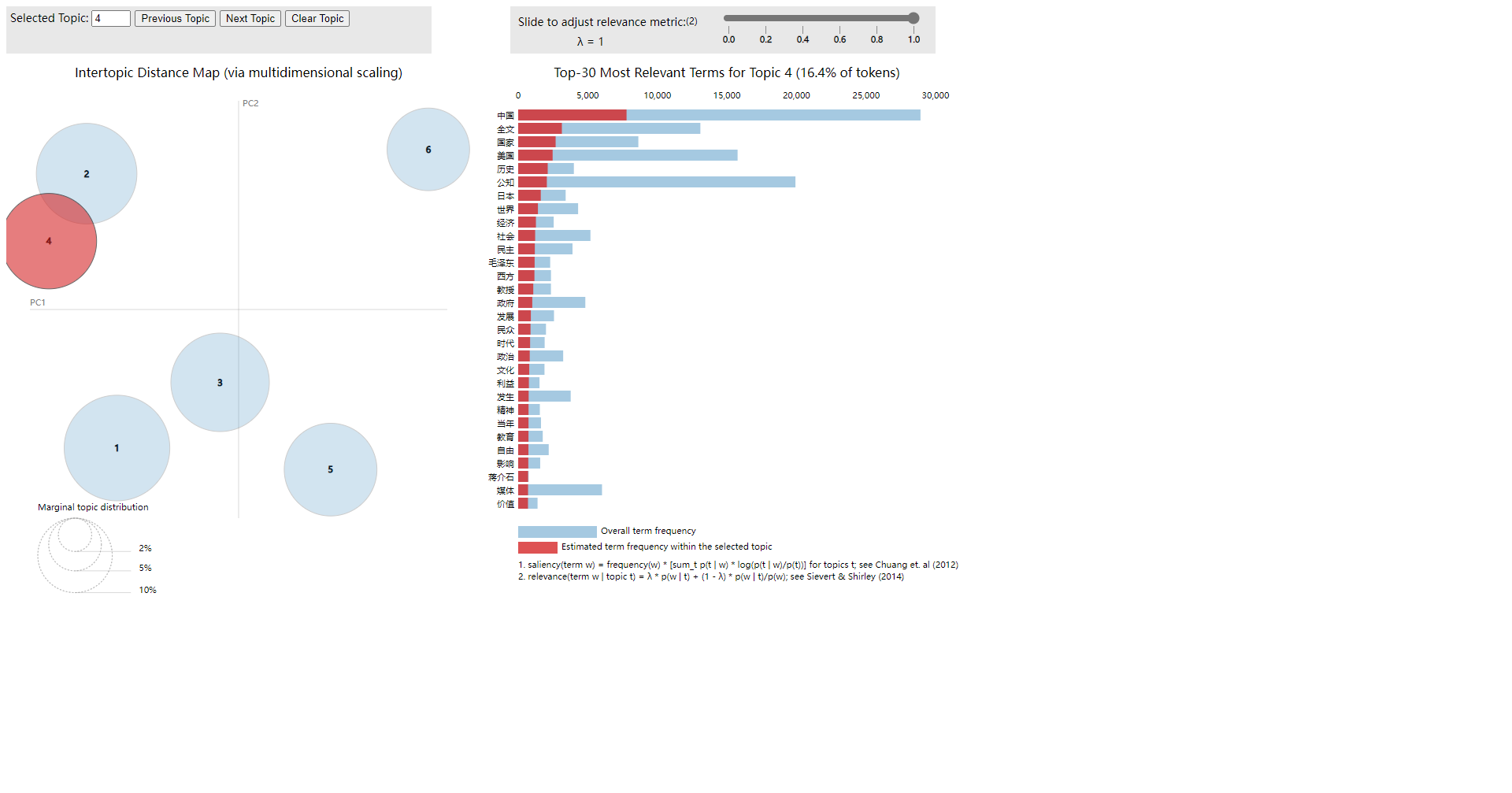
则是在讲述目前国际形势如何，全球的一个整体走势状况

主题3

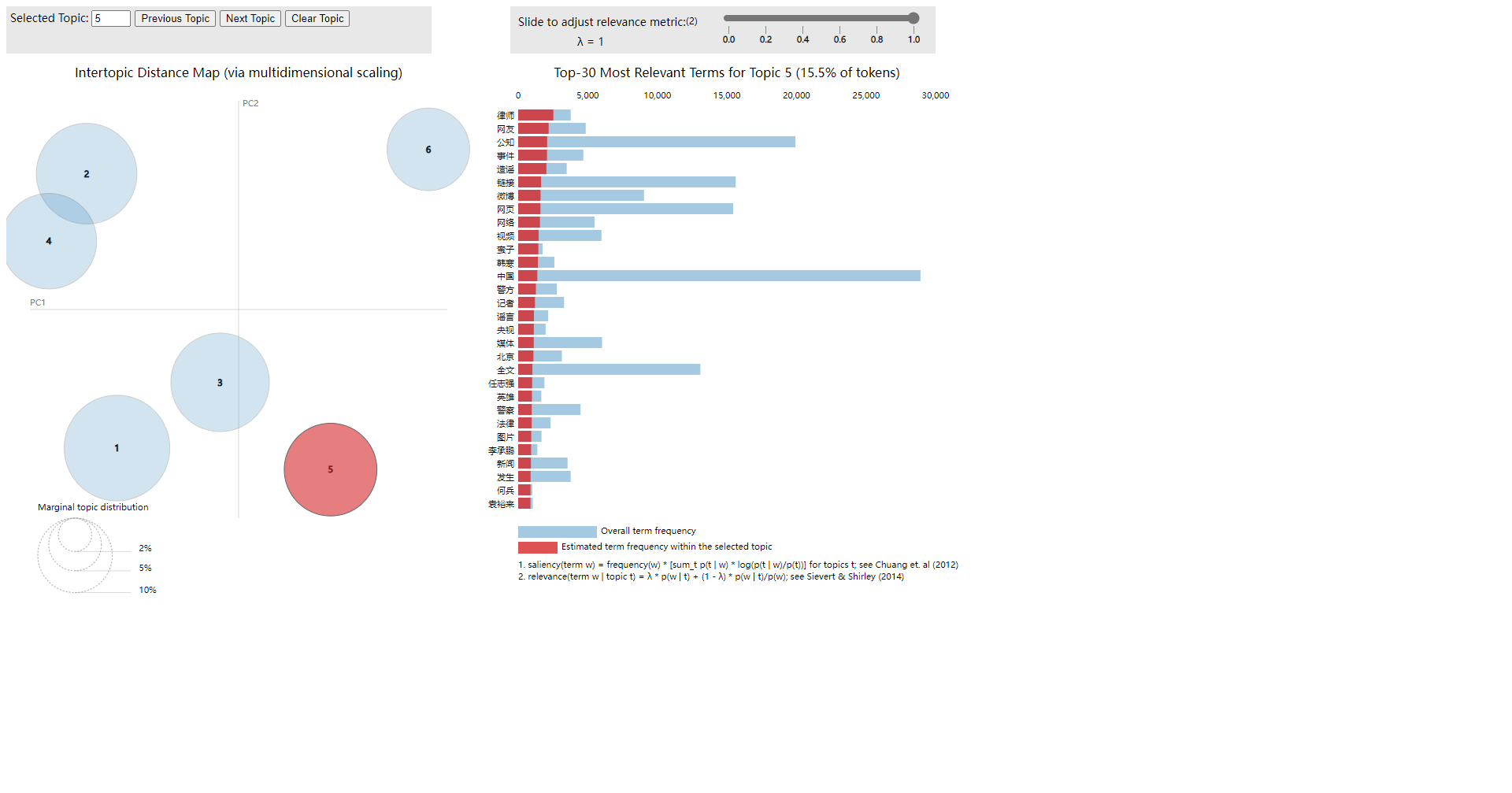


则是在对媒体进行讨论的一个状况

主题4，则是在讲述我国社媒表现的情况，以及我国当前的政治情况



主题5则是在阐述社媒传播的一个路径，通过各界传播的方法



主题6则是美国和中国，公知进行对比的一个具体内容或者事件

