

Cluster 0: 房屋, 较大, 城乡, 风险, 抗震

Cluster 1: 发生, 强化, 地震, 意识, 思维

Cluster 2: 群众, 风险, 存量, 房屋, 抗震

Cluster 3: 地震, 救援, 巨灾, 应急, 准备

Cluster 4: 机制, 应急, 预案, 不够, 方面

Cluster 1: 发生, 强化, 地震, 意识, 思维

这个聚类的关键词集中在 地震发生后的意识 和 防范思维，可能涉及如何加强公众和相关部门的地震意识，提升防震减灾的思维方式。

Cluster 2: 群众, 风险, 存量, 房屋, 抗震

这一聚类可能讨论的是 群众房屋的抗震能力，尤其是 房屋的风险存量，可能包含老旧房屋、城乡结合部等抗震能力不足的区域。

Cluster 3: 地震, 救援, 巨灾, 应急, 准备

这个聚类显然与 地震救援 和 应急准备 相关。关键词显示其主题可能集中在如何组织地震救援力量，以及应急准备的策略和规划。

Cluster 4: 机制, 应急, 预案, 不够, 方面

这一聚类的关键词暗示讨论的是 应急预案和应急机制，可能提到了某些地区或机构在应急预案和应对机制方面存在不足，需要改进和完善。

方法简述：

文本预处理：

首先，对文本进行 分词（使用 jieba 进行中文分词），并去除常见的 停用词。这一步是为了将中文文本转化为机器可以处理的格式，并且去除对聚类无关紧要的常见词汇。

TF-IDF 向量化：

使用 TfidfVectorizer 将分词后的文本转换为数值特征。TF-IDF（词频-逆文档频率）是一种衡量单词在文档中重要性的方法，能突出具有区分性的词汇。

每个文本被表示为一个向量，向量的每一维代表一个词在该文本中的重要性。

K-means 聚类：

使用 K-means 算法 对文本向量进行聚类。K-means 试图将文本分成 k 个聚类，每个聚类中心代表一类文本的主题。

K-means 聚类后，每个文本都被分配到了一个聚类（类别）。

关键词提取：

对于每个聚类，提取聚类中心点的前 n 个高权重词汇。这些词汇是该聚类中心最具代表性的关键词，能够反映聚类的主要主题。

这些高权重词汇可以看作是该聚类的多标签，反映了聚类内所有文本的核心内容。

核心步骤概括：

TF-IDF 向量化 用于将文本转化为向量，表示每个词在文本中的重要性。

K-means 聚类 用于将相似的文本分成 k 类，每类代表一个主题。

关键词提取 从每个聚类中提取具有高权重的词汇，作为该类的标签，代表其主题。