解题思路方案说明：

整体思路：数据导入🡪数据预处理🡪通过规则判断数据是否为关键词🡪存储判断得到的关键词

数据预处理：

1. 导入数据分成训练集和测试集
2. 将训练集的题目和文章中的数字去除
3. 用jieba分词的extract\_tags提取题目中的5个关键词，并限定关键词的词性为名词和动词，人名，地名，名动词。
4. 用jieba分词的textrank提取文章中的5个关键词，并限定关键词的词性为名词和动词，人名，地名，名动词。
5. 用hanlp分割标题，得到分割以后的标题和标题中的词的词性。
6. 提取题目中《》中的内容。

通过规则判断数据是否为关键词:

规则判断流程图如下图所示，其中图中的题目关键词是通过jieba的提取关键词函数得到，提取了前5个比较重要的词，文章关键词也是通过jieba的提取关键词函数得到，提取了前5个比较重要的词。

程序按《》内容，人名，专业名词，题目关键词，文章关键词这样的优先级来定规则，并在程序中剔除重复的词（比如人名和题目关键词有可能重复，这样就换另一个题目关键词）。

最后将得到的两个关键词存到要提交的表中。



最后，感谢DC竞赛的平台以及神策数据公司提供的数据，让我在这次比赛中有所成长和收获，希望DC竞赛平台和神策数据公司越来越好。