## 数据爬取

我爬取了有关“非遗”和“非物质文化遗产”相关的视频以及视频的详细情况。并对其中的信息进行了过滤，只保留下来了与具体的非遗项目相关的视频信息。一共有数据125条，其中过滤调没用的数据，一共有78条。数据的样子如下：

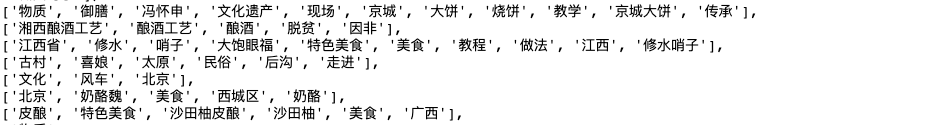


其中核心的字段就是“description”、”keywords”、title”。description字数不是很多，有的长度与title一样大。

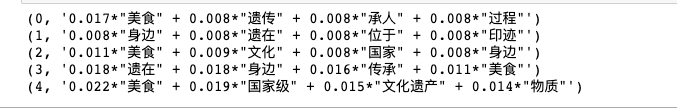
数据的话总的来说的话就是文本数量少，是短文本。

## 数据处理

1. 对其中的description进行分词，去除一些停用词和“非遗”、“非物质文化遗产”这些词之后剩下的词作为关键词（我并没有使用TF-IDF进行关键词提取 因为有的描述就那么两个词，比较少），我在提取关键词的时候又手工进行了一次分词，把里面出现的专有名词和人民都保存了下来，最后跑出来的关键词的样子的话如下图所示



然后进行LDA的训练得到，训练5个主题，每一个主题4个关键词，得到下面的主题-词语分布：



可能是文本比较少的原因吧，现在只是训练出了“美食”这个主题，也是比较符合预期的，因为这个数据集里面美食这个词频数很大，其他的主题 比如“身边”意义基本不大。

## 现在的问题

文本描述过短，LDA效果不明显

## 下一步的工作

从视频发布方入手，找到视频发布方发布的视频，过滤出非遗主题相关的视频，扩充数据集。

因为我发现有的有关非遗的视频在检索中检索不出来，具体原因不知道为什么，比如今天早上更新了一个1年之前发布的视频资源，而我之前通过搜索“非遗”就没有找到。

## 我的想法

对中国所有的非遗资源中的长文本描述进行LDA训练

得到一个文档-主题， 主题-词语分布

比如 马头琴 – 乐器 乐器-古琴

当用户检索“古琴”，则给他推荐到马头琴，然后再去爬取有关西瓜视频中有关马头琴的视频推荐给他