**项目思路**

1. 对文档进行分词、去停用词，形成文档集。
2. 把文档集分成测试集和训练集
3. 调用LDA算法生成模型。
4. 用测试集评估模型质量的好坏
5. 根据用户的输入，展现信息。（不清楚展现是怎么样的一个过程）

**疑点**：

使用 Fire.Fox 交互式（？？？）的展现提取的主题信息，可以选择查看具体

的信息情况，完全根据用户的个性信息展现（不懂这个是什么意思，是根据用户输入的文

档？）。

对抽取出的主题及对应关键词做交互式可视化展现。（这也是它官方给的，这个意思是不用管用户？可是上面的文档又给了另外一种说法）

**下面是两个我们认为的展示方式，但不清楚是否是这样的**。

项目业务逻辑1

用户将需要分类的新闻文档通过web上传至分布式系统进行分类，然后将分类的结果打包后发给用户。

项目业务逻辑2

用户将需要分类的新闻文档通过web上传至分布式系统，系统将文档所属的文档-主题以及主题-词项返回给用户