1. 爬虫（诗影）
2. 诗影

诗影是中国诗歌网的一个栏目，它包含了整理好的现代诗及其配图。每个页面都包含一组现代诗及其对应的一张图片，页面结尾处注明了该组诗歌的作者信息。通过分析网站结构可知，每首诗第一行为标题，其后为正文，然后接一张图片。在页面开头有作者姓名和收藏数，收藏数可作为诗影文章受欢迎程度的依据。综上，先爬取原网页，再通过BeautifulSoup提取每首现代诗的标题、作者、正文、配图地址和收藏数这些信息，并将这些信息导出为json文件，方便后续处理

1. 建立索引
2. 现代诗

初步索引

现代诗来源主要是诗歌大全文库和中国诗歌网。首先将两部分来源的数据整合为统一格式，综合了其标题、作者、正文、配图等信息。使用elasticsearch建立索引，首先需要定义索引的映射 (mapping)，及文档中不同字段 (field)的名称和属性。这里使用了两种类型：keyword和text。Keyword指不进行分词，仅能根据精确值进行查找的字段；text指进行分词，可以由不同搜索算法进行搜索的字段。配图地址采用keyword，其他采用text。作者字段可能包含多个作者，之前整合数据时进行过处理，使作者间用空格相隔，所以用whitespace分词器进行索引和搜索。标题和正文各使用两个字段，一个保存原文，另一个保存进行过分词处理的原文。索引分词采用ik\_max\_word分词器，它对文本做最为细粒度的拆分，即穷尽所有可能的组合进行分词。搜索分词采用ik\_smart分词器，进行最为粗粒度地拆分。具体建索引时，使用bulk批量操作，提高导入索引数据的效率。

1. 图片

图片来源主要为veer图库。索引字段为图片存储地址、来源地址、描述文字和联想词。前三者均使用keyword属性，并不进行搜索。联想词由描述文字分词后进行同义词古词联想后得到，用whitespace分词。

1. 诗歌配图
2. 诗图互搜
3. 由诗搜图的过程是由诗的关键词进行同义词扩展后搜索图片的联想词。若由诗的全文分词进行搜索，则查询信息数据量较大，且缺乏有效信息。而关键词通过结巴的tf-idf算法进行抽取，其本身是文章重要信息的提炼，查询有效性提高。
4. 由图搜诗是由图的联想词搜索诗歌的全文。图片的联想词已经进行过同义词古词扩展，搜索全文(分词后的)可以匹配更多的信息，对古诗文部分也有所兼容。
5. 诗配图

根据上述诗图互搜的两种方式，诗配图也有两种方案。

第一种方案是遍历诗歌，通过诗歌查询图片，将评价最高的图片作为其匹配图。第二种是遍历图片，通过图片查询诗歌，在评价得分最靠前的五首诗中，选择未配图的第一首诗作为该图的对应。

数据库中图片有15万张，而诗歌数只有3万。经过测试，第一种方案速度约为1秒5首诗，第二种约为3秒一张图，所以第一种方案速度更快。两种方案均存在诗歌未匹配上图片的情况，后一方案以图片为主体的搜索使这一情况更容易发生，因此还需采用前者对剩余诗歌进行配图，增加了工作量。综合考虑，我们采用了第一种方案。

1. 首页内容

网站首页包含每日推荐和随机推荐两部分。通过从数据库中随机抽取若干首古诗和现代诗生成首页内容，每首诗及其作者分别链接到其对应的诗歌页面和作者页面。对于每日推荐，将当天日期作为随机数种子，定义相应变量保存诗歌id信息，每天更新一次。每日一诗由于位于页首，背景为其配图，限于篇幅，对其正文长度限制在50字以内，对不符合要求的诗歌进行线性搜索。页面最后列出了这些诗歌对应的标签，方便用户点击查询。（此处可加每日一诗及页尾标签部分截图）