检索系统

1. 网络爬虫获取不低于100篇文档，本地存储

2. 分词，去停用词

3. 放入词袋模型，构建TF-IDF向量

4. 构建倒排索引文件 （某个词语 在哪个文档出现过）

先给每个文档编号，然后对文档里的词条进行编号。以词条创建索引，记录每个词条的所有文档编号

5. 实现 向量空间检索模型匹配算法。

6. 能对检索结果的准确率进行人工评价。

信息抽取系统，以BERT模型为例

1. 在检索系统的基础上，分词时添加特殊标记

2. 对每篇文档进行命名实体识别

3. 关系抽取

4. 能够对抽取结果进行

系统使用流程：

1. 用户查询输入为自然语言字符串

2. 查询结果输出按照相关度从大到小排序，

3. 列出相关度、文档标题、主要匹配内容、URL、文档日期

4. 每个文档的输出结果还要有信息抽取结果，得到里面的实体关系。

5. 用户可以对检索结果和抽取结果进行人工评价。

6. 命令行即可。