第三天笔记-数据处理

- 1.环境：
  - 库文件：
    - sys
    - pickle（把特有的数据类型转成python的数据类型，并且可以保存模型）
    - re 正则， 清洗数据，把数据里面的空格，标点符号，英文大小写等
    - tqdm 进度条，耗时
-
- 2.语料收集
  - 聊天记录
  - 电影对话
  - 台词片段语
-
- 3.语料清洗
  - 多余的空格
  - 不正规的符号
  - 多余的字符、英文
  - 清洗的方法
    - 正则化
    - 切分
    - 好坏语句判断
-
- 4.句子向量的编码化 - 计算机无法直接识别句子中文的含义，转成计算机识别的词向量
  - 原始文本不可以直接训练
  - 将句子转成向量的形式
  - 将向量转成句子
-
- 5.语料问答对的构建
  - 问答对的处理和拆分
    - 问： 早上好 - 答 ： 早安
-
- 6.语料模型的保存
  - 使用pickle来保存模型
  - 生成pkl格式进行语料的训练
    - 再进行深度模型训练，打包成restful的形式
-
-
-
-

-
-
-
-
-
-
-
-
-
-
-
-