

第四章 自然语言处理

目标：

- 1. 掌握语自然语言处理的基本概念；
- 2. 知道自然语言处理的典型应用；
- 3. 可以熟练掌握自然语言处理的相关应用；
- 4. 能够利用人工智能开源框架开发自然语言处理相关应用；
- 5. 建立起利用自然语言处理解决生活中所碰到的问题的意识；
- 6. 在日常学习生活中，建立起保护隐私信息的意识，养成主动保护自己及他人的个人隐私数据的习惯。

一、自然语言处理初体验

表 4.1.1 自然语言处理初体验备选工具

环境	工具名称	功能简介	获取方式
浏览器	九歌人工智能诗歌写作系统	九歌人工智能诗歌写作系统。支持绝句、风格绝句、藏头诗、律诗、集句诗和词的自动创作。	访问网址： http://jiuge.thunlp.org/
浏览器	狗屁不通文章生成器	根据关键词或短语生成文章。【请勿用于正式用途】	访问网址： https://suulnnka.github.io/BullshitGenerator/

自然语言处理体验案例：自动创作（九歌人工智能诗歌写作系统、狗屁不通文章生成器）。

1. 九歌作诗系统

应用案例 4-1：九歌人工智能诗歌写作系统

- （1）浏览器打开如表 4.1.1 第一行的九歌系统网址；

- (2) 输入“夏日炎炎”关键词并点击生成诗歌；
- (3) 等待排队后完成诗歌生成，可以对比一下每个人生成的诗歌是否相同？

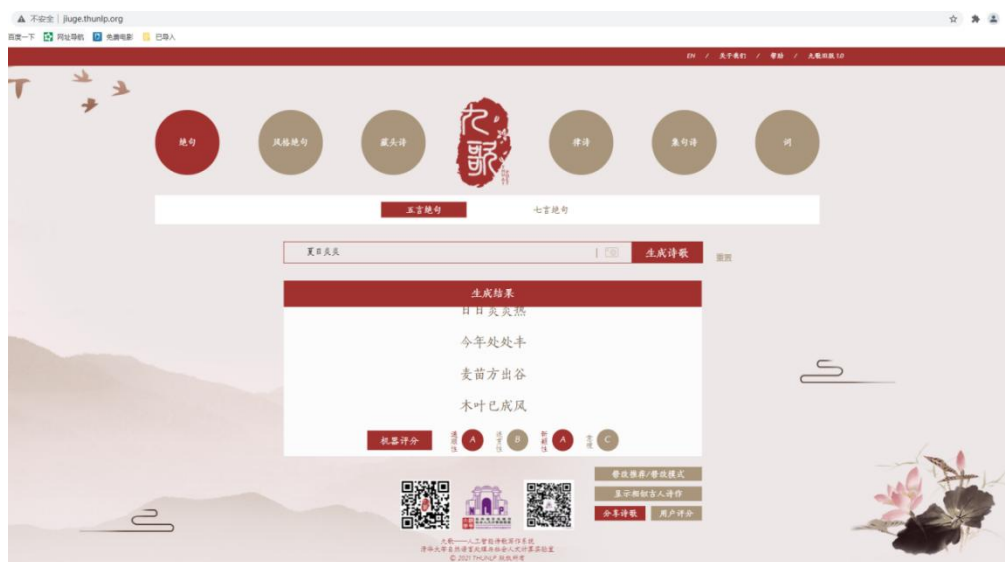


图 4. 1. 1 九歌作诗系统界面

2.狗屁不通文章生成器

应用案例 4-2：狗屁不通文章生成器

- (1) 浏览器打开如表 4. 1. 1 所示第二行链接；
- (2) 在主题区输入短语，点击生成，即生成一篇文章；
- (3) 大家认真查看下该文章，如图 4. 1. 2 所示，你是否发现文章的秘密了？



图 4. 1. 2 狗屁不通文章生成器界面

二、自然语言处理的概念

1. 自然语言处理的概念

《三体》的作者刘慈欣在 2003 年曾经写过一篇小说，名字叫《诗云》，小说中描述制造了一个超级写诗系统，能够包含宇宙中的每一个字。而这样的系统目前已经被清华大学自然语言处理与社会人文计算实验室制作出来了，就是前文描述的九歌系统。这个系统深度应用了自然语言处理技术，按照诗词创作规则自动生成诗词，可以让一个对诗词一窍不通的人也能体验诗词创作的乐趣。这里面应用自然语言理解、自然语言生成的自然语言处理的核心技术，下面我们就来逐步揭开自然语言处理的神秘面纱。

比尔盖茨曾经说过，“语言理解是人工智能领域皇冠上的明珠”。而语言理解是自然语言处理的核心。自然语言处理相当于一个翻译器，在人类与机器之间搭建了一个沟通的桥梁，可以实现人类不同语言之间，人与机器的沟通。

自然语言处理（NLP）是一门融合了计算机科学、人工智能以及语言学的交叉学科，目的是为了让计算机像人一样能够处理、理解、表达语言。¹

2. 自然语言处理的发展水平

自然语言处理分为自然语言理解和自然语言生成，是目前人工智能正在深入研究和迫切需要突破的两个点。

目前制约自然语言处理有五个难点问题，人类的自然语言有时候规律性不明显，且规律不一致，如山东人喜欢用倒装句，不同方言有不同的俚语，因此要把自然语言的规律穷尽本身是一个非常困难的任务；语言可以混合式表达，如可以自由组合，可以多语种混合表达，给机器理解制造了不少“麻烦”；人类在说话时，可以任意“发明创造”，甚至可以创造一种新的表达方式，人类可以很快就适应并理解，但在目前的算法结构下还无法达到和人类一样的水平；自然语言有一定的语境和经验积累，或者说需要一定的上下文，而且有一定的知识依赖，如

¹ 何晗.自然语言处理入门[M].人民邮电出版社，2019 年：1-8.

果不具备相应的知识或者在自然语言处理时没有考虑情境，就会产生理解差异。

目前自然语言处理已经能够支持新闻自动生成、无障碍聊天、自动生成对联、诗文自动生成等。目前在自然语言处理的支持下，语音识别、看图说话等领域已经能够实现日常应用的水平。

3. 自然语言处理的核心问题

自然语言处理是让人工智能从感知智能飞跃到认知智能的关键核心技术，虽然自然语言处理的术语也还在不断地变化过程中，不同学者有不同的表述，但是其内涵大致都是相同的。自然语言处理包括自然语言理解与自然语言生成两个核心任务和内容。

自然语言理解（NLU/NLI），是指实现让机器具备人类的语言理解能力。

自然语言生成（NLG），自然语言生成是人类和机器沟通的一个关键环节，是在响应人类语言并与人类对话的重要环节，可以将非语言格式的数据转换成人类可以理解的自然语言格式，如文章、诗歌、报告等。

4. 自然语言处理的一般过程

自然语言处理经历了传统机器学习方法和深度学习的方法阶段，具体流程如下：

传统机器学习的流程：语料预处理、特征工程、特征提取、特征选择、选择分类器；深度学习的流程：语料预处理、设计模型、模型训练。两者的不同是，深度学习的模型是基于训练生成的。

在自然语言处理的过程中，语料预处理是基础也是非常重要的一个环节。由于中文和英语的差异，预处理过程有所不同具体如下：

英语语料预处理的过程为：分词、词干提取、词性还原、词性标注、命名实体识别和分块。

中文语料预处理的过程为：分词、词性标注、命名实体识别、去除停用词。

16	、
17	。
18	《
19	》
20	—
21	一些
22	一何
23	一切
24	一则
25	一方面
26	一旦
27	一来
28	一样
29	一般
30	一转眼
31	万一
32	上
33	上下
34	下
35	不
36	不仅
37	不但
38	不光
39	不单
40	不只
41	不外乎
42	不如

图 4.2.1 中文停用词库 (<https://github.com/goto456/stopwords>)

三、自然语言处理的应用

自然语言处理的技术包括：分词、词性标注、命名实体标注、信息抽取、文本聚类、文本分类、依存句法分析情感分析等。基于这些技术，自然语言处理的典型应用包括：词云生成、知识图谱、聊天机器人、自动文摘、情感分析、社交媒体审核与分析、翻译、语法检查、智能创作等，其中智能创作在本章开头我们已经进行了初体验。下面我们就以词云、聊天机器人、情感分析、翻译、信息抽取、智能创作为例进行详细了解。

1. 词云

词云是利用自然语言处理的分词技术，去掉停用词之后，按照词的权重，将关键词绘制成大小、颜色不同的图形的一种技术。从技术上来看，词云应该属于分词的可视化，可以直观的让我们看到文本的关键词。其中所抽取的词的大小和颜色取决于词的词频或者权重，一般可以按照词的出现频率来确定，在一些要求比较高的场景中，我们可以采用比较复杂的算法，如 TF-IDF（词频-逆文本频率指数）算法，来计算所抽取词的权重。

表 4.3.1 词云应用

环境	工具名称	功能简介	获取方式
浏览器	图悦	在线词频工具，在线词云图制作。	访问网址： http://www.picdata.cn/picdata/index.php

场景描述：你在写作文时，是否想要知道自己作文所采用的关键词有哪些？还有你是否想要分析下你的作文的中心思想是否符合要求？那么这时候可以借助词云工具来分析。此外，词云工具还可以直观的呈现网上的评论信息，帮助你快速的从海量评论文字中找到关键评论点，从而实现部分舆情趋势分析功能。

体验步骤：

- (1) 浏览器访问如表 4.3.1 网址；
- (2) 左侧文本框输入想要生成词云的文本，点击分析出图，生成词云图。
- (3) 思考一下：词云工具解决了什么问题？是否能够帮到你？你是否有更

好的解决方案？



图 4.3.1 在线词云工具

2.聊天机器人

聊天机器人是人工智能的一种形式，可以通过语音和文本与人类交流。它们中的大多数模仿人类的语言行为，使你觉得在和人类进行交流。聊天机器人心理健康方面可以用于帮助患有抑郁症和焦虑症的人，减缓相关症状；在生活娱乐方面可以用于智能语音助手，使用语音控制智能音箱、家电、家居等；在教育方面，可以实现智能学伴功能，利用语音交互实现小学生的学习辅导，知识问答等。

表 4.3.2 聊天机器人应用

环境	工具名称	功能简介	获取方式
苹果手机	Siri	智能语音助手。	苹果手机自带
华为手机	小艺小艺	智能语音助手。	华为手机自带
小米手机或音箱	小爱同学	智能语音助手。	小米手机或者小爱智能音箱

场景描述：在开车的过程中，想要播放一个音乐，但是开车过程中控制手机是非常危险的，这种情景下如何解决呢？这时候语音助手就非常有用，可以使用语音助手控制音乐播放、导航等。

体验步骤（以苹果手机为例）：

- (1) 苹果手机，确保开启了 Siri，在设置，辅助功能，启用 Siri；
- (2) 用语音唤醒 Siri（嘿 Siri），“播放音乐”；
- (3) 可以尝试和 Siri 进行聊天，如问“你叫什么名字？”、“什么是人工智能”。
- (4) 思考：Siri 解决了什么问题？解决效果如何？是否有改进的点？



图 4.3.2 Siri 对话界面

3.情感分析

情感分析也被称为情感的人工智能或意见挖掘，它是从书面和口头语料库中识别、提取和量化情感和情感状态的过程。情感分析工具用于处理诸如客户评论和社交媒体帖子之类的事情，以理解对特定事物的情绪反应和意见，比如新餐厅的菜品质量。

表 4.3.3 情感分析应用

环境	工具名称	功能简介	获取方式
----	------	------	------

浏览器	对话情绪识别	百度人工智能开放平台，对话情绪识别。	访问网址： https://ai.baidu.com/tech/nlp_apply/emotion_detection
浏览器	情感倾向分析	百度人工智能开放平台，情感倾向分析。	访问网址： https://ai.baidu.com/tech/nlp_apply/sentiment_classify

场景描述：假如你是一个网店的店主，你想知道你的顾客对你的产品的情感倾向，是喜欢还是讨厌？一个评论一个评论的人工去看，费时又费力，这时候怎么办呢？可以借助人工智能情感分析功能来辅助分析。下面我们就体验一下这个功能吧！

体验步骤：

- （1）浏览器打开如表所示对话情绪识别体验网址。
- （2）在功能演示区输入文本：“老师布置的作业太难了，我实在做不出来怎么办？如果不做老师会骂我的，纠结啊！你有什么办法吗？”
- （3）查看文本情绪分析结果。
- （4）自由输入文本查看结果，并与自己的分析做对比。
- （5）思考：机器是如何知道文本的情绪倾向的？这个解决了什么问题？他有哪些应用场景？




图 4.3.3 情绪倾向识别界面

4.翻译

翻译就是实现语种之间的互换，如英文转换为中文，中文转换为英文等。在人工智能能够实用之前，翻译是一个只有人类才能担当的工作。目前，有实时语音翻译软件，实时听写翻译软件，对话翻译软件，实景翻译软件等应用。

表 4.3.4 翻译应用

环境	工具名称	功能简介	获取方式
移动终端	有道翻译官	有道翻译官，支持 107 种语言的随身翻译，支持拍照翻译、对话翻译和同传。	访问网址： http://fanyiguan.youdao.com/ 或移动终端扫描以下二维码下载安装： 

场景描述：学了多年的英语，出国时还是听不懂老外的口音，或者去了其他语言的国家，如韩语、日语等，两眼一抹黑，是否有免费的工具可以帮到你呢？这时候就要用到了翻译软件或产品了。

体验步骤：

- （1）移动终端相应市场中搜索“有道翻译官”，下载并安装。
- （2）在打开的有道翻译官界面中输入要翻译的内容，查看翻译结果。
- （3）体验拍照翻译功能。
- （4）思考：以上翻译功能是否符合你的要求？你是否还想到了其他的翻译场景？

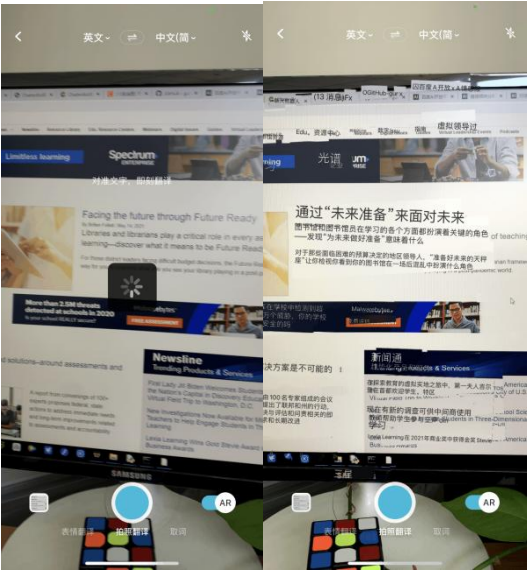


图 4.3.4 拍照翻译

5.Magi 基于机器学习的信息抽取与检索系统

表 4.3.5 信息抽取应用

环境	工具名称	功能简介	获取方式
浏览器	Magi	基于机器学习的信息抽取和检索系统——人工智能知识搜索引擎。	访问网址： https://magi.com/

场景描述：在日常生活中，你是否碰到过一些奇怪的问题，需要使用搜索引擎去搜索，但是发现某些搜索引擎是按照字数截断文字的，这样很可能删掉了你想要的搜索的关键点，这时候需要用到信息抽取与检索的系统，下面来看一下 Magi 基于机器学习的信息抽取与检索系统的功能吧！

体验步骤：

- （1）浏览器打开如表 4.1.2 第二行的 Magi 系统；
- （2）输入人工智能，查看搜索结果；
- （3）对照前面已学内容，复习人工智能的基本知识。
- （4）思考一个问题：Magi 和百度的区别是什么？



图 4.3.5 Magi 系统界面

6.智能创作

智能创作是人工智能的高阶应用，人工智能写新闻、人工智能写书（微软小冰写诗集）、人工智能批改作文等都是相关应用。在本章开篇介绍的九歌和狗屁不通也都属于智能创作的范畴，下面再来体验一个辅助智能创作的应用。

表 4.3.6 智能创作应用

环境	工具名称	功能简介	获取方式
浏览器 微信小程序	get 写作	get 写作，一站式智能写作平台，支持写作时的文献发现，热点发现，标题推荐，智能改写和质量检测。	访问网址： https://getgetai.com 微信小程序搜索“get 写作”

场景描述：写作是一个头疼的任务，即使是专业写手，寻找灵感、寻找金句、写出新意，都是痛点，是否有人工智能助手帮助找灵感，找金句，找文献呢？答案是肯定的，现在就有一些智能创作工具可以提供帮助，下面我们来了解下吧！

体验步骤（以网页版为例）：

- （1）访问上表所示网址；
- （2）注册账号，并完成登录；

- (3) 在文本框中输入你想要撰写的主题；
- (4) 点击开始写作；
- (5) 在左侧文章下面，挑选符合你的写作要求的文章，并点击智能摘要，点击采用，引入自己的文章；
- (6) 可以选中引入的文本，进行改写，找素材和找金句；（注：免费用户只能体验一次）
- (7) 思考：该应用提供了什么功能？这些功能是否符合写作要求？是否有更好的解决方案？



图 4.3.6 get 写作界面

四、自然语言处理的开发

1. 图形化编程

前驱知识准备：图形化编程的界面布局；知道角色背景造型等基本概念；图形化编程的事件（当绿旗被点击）、外观（说**2 秒）、控制（等待 1 秒、重复执行、如果那么）等基本积木，连接运算符；知道图形化编程的基本方法。

硬件准备：带有摄像头和麦克风的笔记本电脑（含 Mac 笔记本）、台式机电脑。注：平板电脑暂不支持机器学习模块。慧编程 mBlock5.3.0 支持 Windows7、Windows10、macOS 10.12+版本的操作系统，推荐 64 位操作系统。

软件工具准备：参照表 2.4.1 所示第一行的获取方式下载并安装慧编程工具。完成账号注册并登陆（可以应用更多功能）。

任务：搭建一个识别你性别并能够识别你说的话的情感倾向的对话机器人



拓展任务：如何实现该机器人一直在识别讲话的情感？

2. Python 编程

（1）词云生成

绘制词云是一个非常专业的任务，在人工智能时代，如何利用相关技术来快速绘制词云呢？Python3 具有很好的解决方案。下面就给大家介绍如何利用 python3 来绘制词云，其中需要导入和借助的库包括：jieba、wordcloud、pillow、

numpy。

库的安装方法如下：

```
pip3 install jieba
```

```
pip3 install wordcloud
```

```
pip3 install pillow
```

```
pip3 install numpy
```

安装好相关库后，就可以编制程序了，具体代码如下：

```
from PIL import Image
import numpy as np
from wordcloud import WordCloud
import jieba
def drawWordCloud(filename,bg_image):
    #绘制词云图
    data=open(filename,'r').read()
    #读取文本
    text=' '.join(jieba.cut(data,cut_all=False))
    #文本分词
    color_mask=np.array(Image.open(bg_image))
    #设置背景叠加图形样式，会绘制成图形的样子
    wc=WordCloud(
        scale=4, #数值越大，分辨率越高
        font_path=r'方正大黑简体.ttf', #默认mac下的配置，windows下为r'c:\windows\fonts\msyh.ttf'
        background_color='black', #设置背景色
        mask=color_mask, #设置蒙版图像
        max_words=800, #最多词
        stopwords=STOPWORDS, # 设置停用词
        max_font_size=160, #最大字号
        random_state=30 #设置多少种随机生成状态，有多少种配色方案
    )
    #设置字体，背景色，背景图，最多词汇数量，最大字体大小
    wc.generate(text) #加载文本
    wc.to_file('draw_wc201905255.png') #将词云绘制为图像
    #绘制词云
if __name__=="__main__":
    drawWordCloud('test_1.txt','bg.png') #调用函数
```

图 3.4.2 词云源代码

(2) 利用 HanLP 为文字生成拼音

安装 HanLP 库：

```
pip3 install HanLP
```

自动生成拼音的代码：

```
import HanLP
```

```
If __name__=="__main__":
```

```
    text="自然语言理解"
```

```
    pinyin_list=HanLP.convertToPinyinlist(text)
```

```
print(pinyin_list)
```

(3) 搭建聊天机器人

方法一：基于在线接口的聊天机器人

```
import json
import urllib.request

api_url = "http://openapi.tuling123.com/openapi/api/v2"
while True:
    text_input = input('我: ')
    if text_input == 'q':
        break

    req = {
        "perception": {
            "inputText": {
                "text": text_input
            },
            "selfInfo": {
                "location": {
                    "city": "上海",
                    "province": "上海",
                    "street": "紫龙路"
                }
            }
        },
        "userInfo": {
            "apiKey": "03826078461f45ee8bb1caec9af8568c",
            "userId": "296075"
        }
    }
    # print(req)
    # 将字典格式的req编码为utf8
    req = json.dumps(req).encode('utf8')
    # print(req)

    http_post = urllib.request.Request(api_url, data=req, headers={'content-type': 'application/json'})
    response = urllib.request.urlopen(http_post)
    response_str = response.read().decode('utf8')
    # print(response_str)
    response_dic = json.loads(response_str)
    # print(response_dic)

    # intent_code = response_dic['intent']['code']
    results_text = response_dic['results'][0]['values']['text']
    print('Turing: ')
    # print('code: ' + str(intent_code))
    print('text: ' + results_text)
```

图 3.4.3 聊天机器人源代码

方法二：搭建一个可训练的聊天机器人

安装库：

```
pip3 install spacy
```

```
pip3 install chatterbot
```

训练代码：


```

1  #!/usr/bin/python
2  # -*- coding: utf-8 -*-
3  from chatterbot import ChatBot
4  from chatterbot.trainers import ListTrainer
5  my_bot = ChatBot("Training demo")
6  my_bot.set_trainer(ListTrainer)
7  my_bot.train([
8      "你叫什么名字?",
9      "我叫ChatterBot。",
10     "今天天气真好",
11     "是啊, 这种天气出去玩再好不过了。",
12     "那你有没有想去玩的地方?",
13     "我想去有山有水的地方。你呢?",
14     "没钱哪都不去",
15     "哈哈, 这就比较尴尬了",
16 ])
17 while True:
18     print(my_bot.get_response(input("user:")))

```

图 3.4.4 聊天机器人训练源代码

测试使用代码:

```

1  #!/usr/bin/python
2  # -*- coding: utf-8 -*-
3  from chatterbot import ChatBot
4  from chatterbot.trainers import ChatterBotCorpusTrainer
5
6  chatbot = ChatBot("myBot")
7  chatbot.set_trainer(ChatterBotCorpusTrainer)
8
9  # 使用中文语料库训练它
10 chatbot.train("chatterbot.corpus.chinese")
11 lineCounter = 1
12 # 开始对话
13 while True:
14     print(chatbot.get_response(input("(" + str(lineCounter) + ") user:")))
15     lineCounter += 1

```

图 3.4.5 聊天机器人源代码

思考与练习

1. 什么是自然语言处理？自然语言处理包含哪两个重要能力？
2. 应用自然语言处理会不会导致个人隐私数据的泄露？如果会，该如何避免？
3. 请使用合适的工具或编程语言，搭建一个给输入汉字文本提供拼音和声调的程序。

