自然语言处理技术基础

网络空间安全与计算机学院

教材和教参

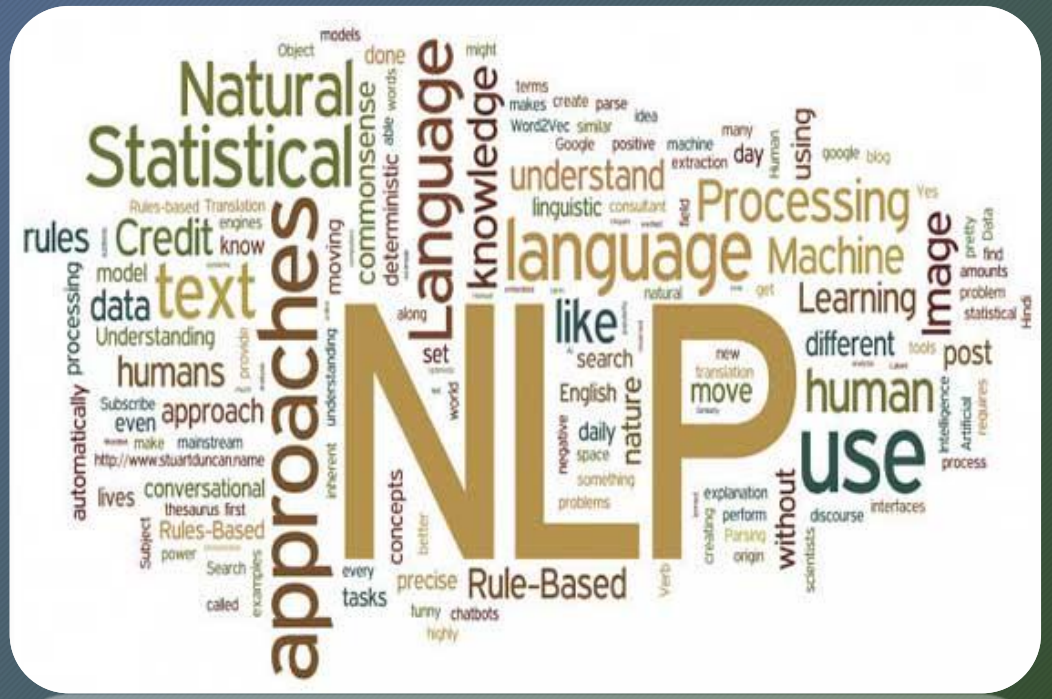
教材：

自然语言处理基本理论和方法（第2版）陈鄞[yín]

教参：

1. 数学之美（第3版）吴军
2. 统计自然语言处理（第2版） 宗成庆
3. 自然语言处理综论（第3版） Dan Jurafsky
4. 自然语言处理入门 何晗
5. 统计学习方法（第2版） 李航

第1章 绪论



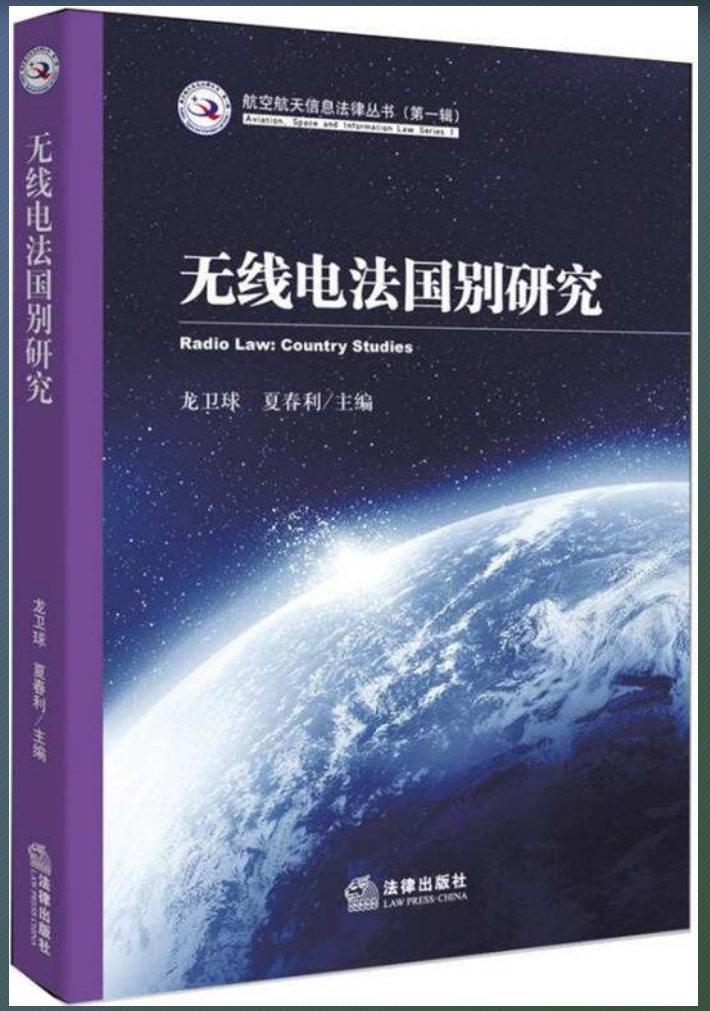
1.1 什么是自然语言处理

自然语言

- 人类使用的语言
- 人类交流和思维的主要工具
- 书面形式：文字
- 口头形式：语音
- 举例：汉语、英语等

#自然语言理解太难了#


三餐二楼欢迎新老师生前来就餐
@北冥乘海生


航空航天信息法律丛书（第一辑）
Aviation, Space and Information Law Series 1
无线电法国别研究
Radio Law: Country Studies
龙卫球　夏春利/主编
法律出版社
LAW PRESS·CHINA

#自然语言理解太难了# 三楼二餐欢迎新老师生前来就餐

http://www.sogou.com/labs/webservice/

搜狗实验室

https://www.xfyun.cn/services/lexicalAnalysis

科大讯飞

#自然语言理解太难了#

要你管和不要你管
气死他和气不死他
掉地上和掉地下
了得和了不得
结婚前和没结婚前
大败和大胜

说错了话=
说了错话
说错了话
说话错了
话说错了

能穿多少就穿多少和能穿多少就穿多少
喜欢一个人和喜欢一个人
爱上一个人和爱上一个人
谁都打不过和谁都打不过

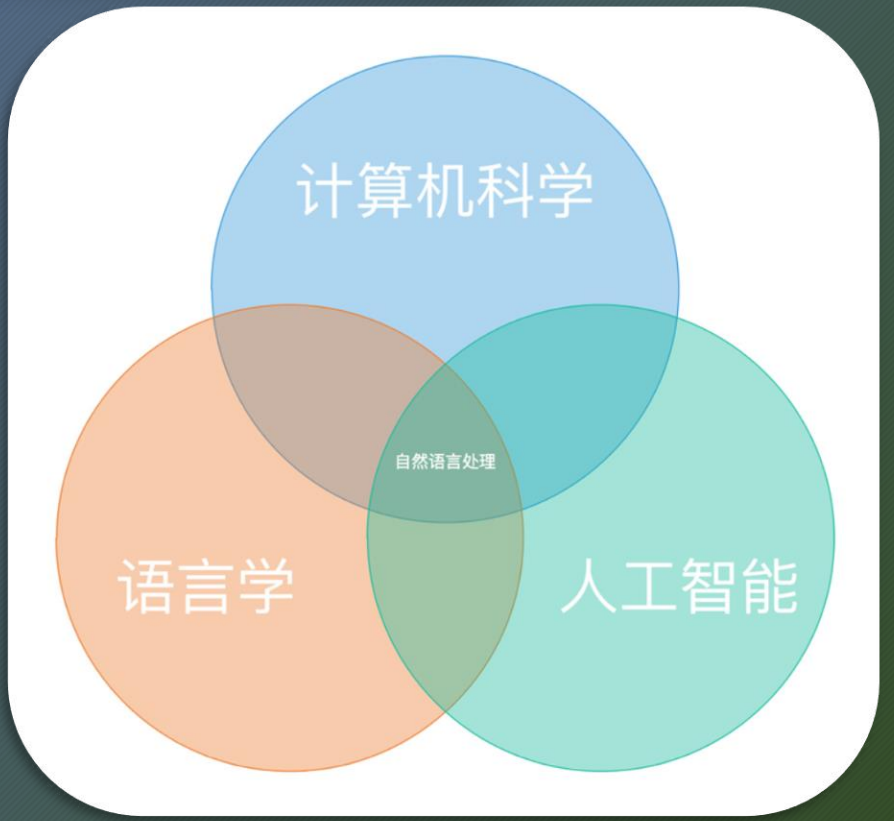
1.1 什么是自然语言处理

自然语言处理

Natural Language Processing，NLP

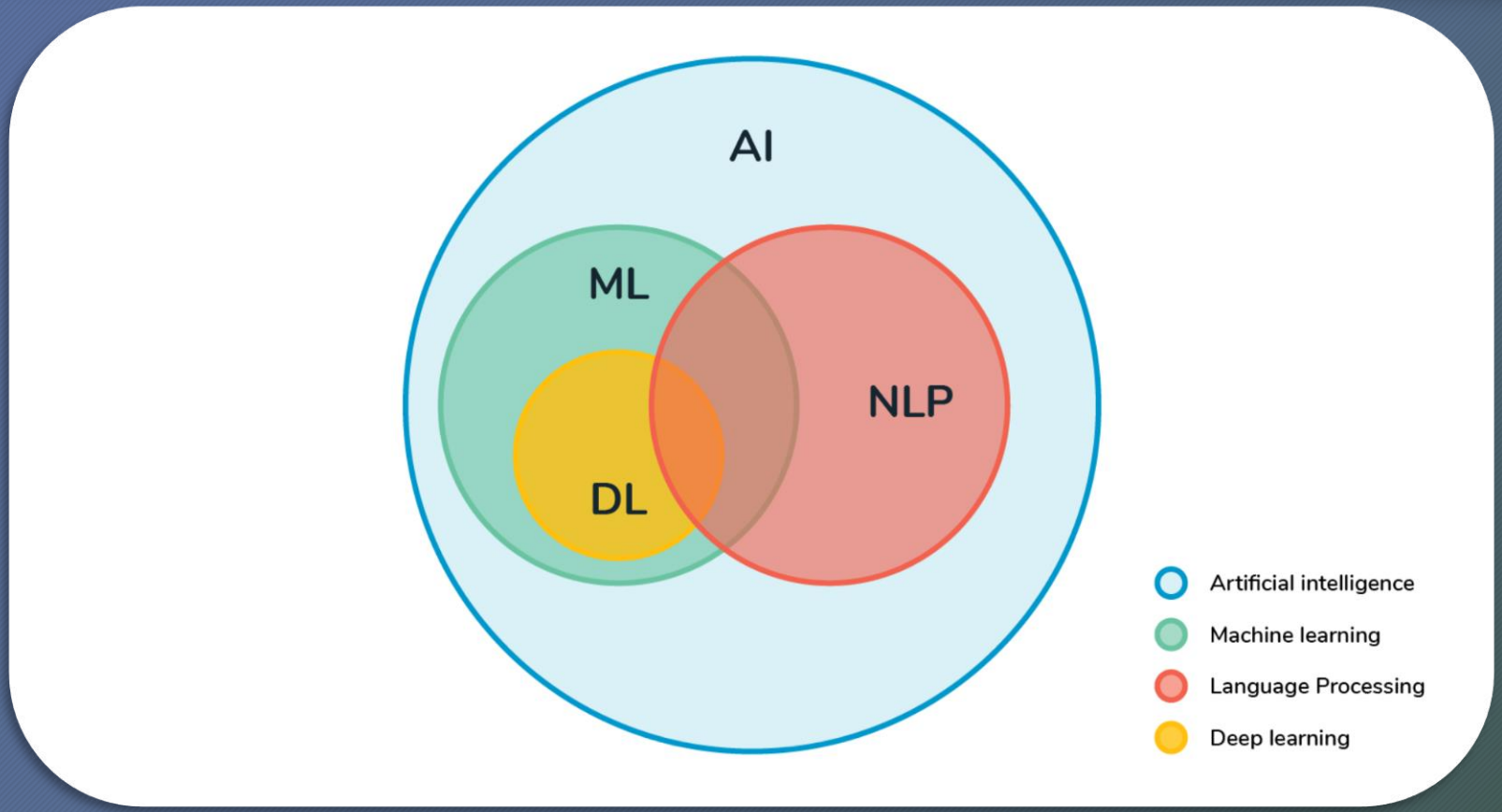
- 利用计算机为工具对自然语言的信息进行处理和加工
- 计算机科学领域与人工智能领域中的一个重要方向
- 研究人与计算机之间有效通信的各种理论和方法
- 涉及学科：语言学、计算机科学、数学

REF：冯志伟《自然语言的计算机处理》



比尔·盖茨：「语言理解是人工智能皇冠上的明珠」

AI ML DL 与 NLP



AI：Artificial intelligence

ML：Machine learning

DL：Deep learning

LP：Language Processing

NLP：Natural Language Processing

Ref：《自然语言处理入门》何晗 著 https://www.ituring.com.cn/book/tupubarticle/28291

DL 与 NLP

	深度学习算法	NLP 使用
1	神经网络	词性标记 词语切分（Tokenization） 实体命名识别 目的提取
2	循环神经网络	机器翻译 问答系统 图像描述
3	递归神经网络	句子解析 情感分析 释义检测（Paraphrase detection） 关系分类 物体识别
4	卷积神经网络	句子/文本分类 关系提取和分类 垃圾邮件检测 搜索词条的归类 语义关系提取

1.2 自然语言处理的主要过程

自然语言理解 NLU

Natural Language Understanding

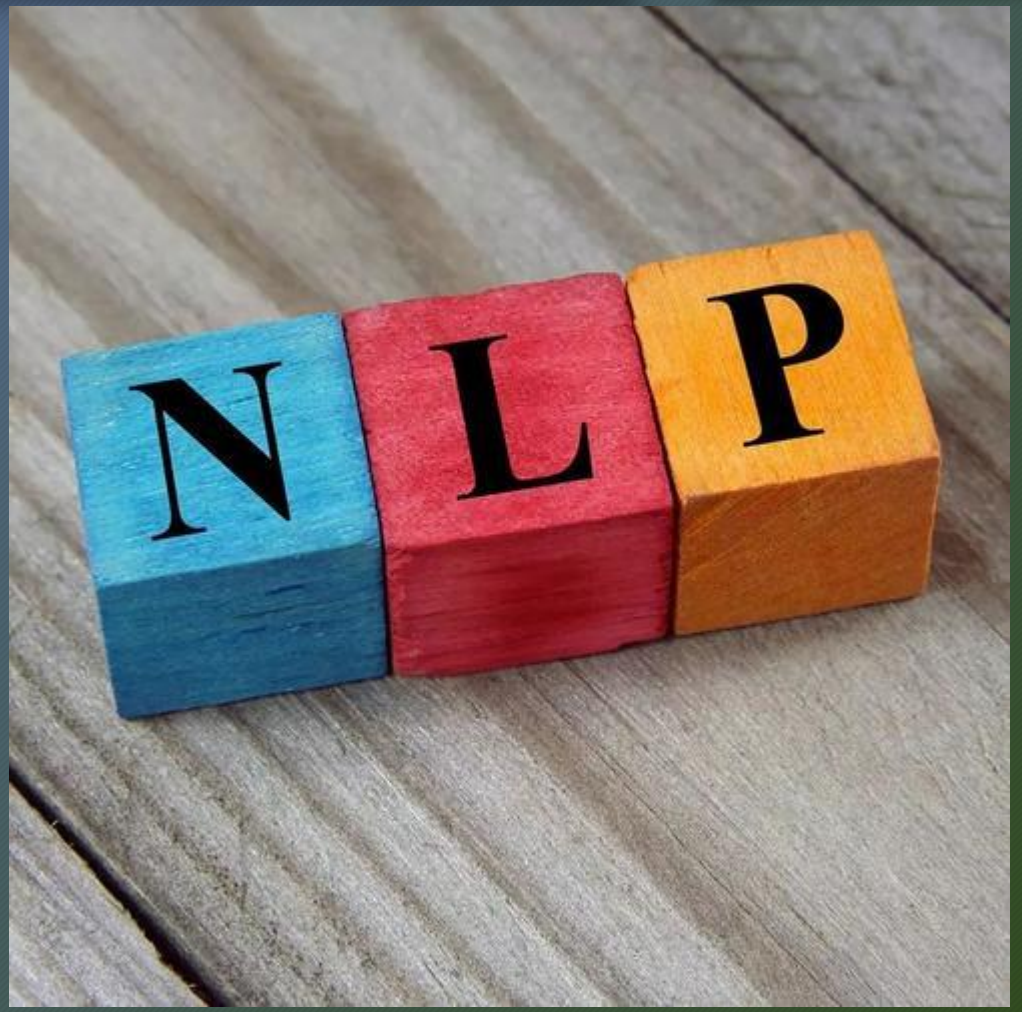
- 语言信息录入:
  - 语音录入
  - 文字录入
- 文本理解:
  - 词法分析
  - 句法分析
  - 语义分析

自然语言生成 NLG

Natural Language Generation

- 语音合成
- 自动作诗
- 机器翻译
- 摘要生成
- 文本更正

主要研究内容



NLP研究内容分为以下几个层次:

基础研究

字符集编码体系	语言计算模型	语料库、知识库

应用基础研究

英语形态分析	中文自动分词	词性标注	短语切分	命名实体识别	句法分析	语义分析	篇章理解

应用研究

机器翻译	信息检索	自动问答	文本校对	自动摘要	辅助写作	语音识别	语音合成

1.3 自然语言处理的应用领域

文化教育

- 图书分类
- 语音识别
- 自动判卷
- 机器翻译
- 智能解说
- 新闻定制

医疗

- 聊天机器人
- 残疾人辅助

政务

- 自动咨询
- 汇总决策

内容安全

- 垃圾邮件过滤
- 内容监控

商务

- 自动呼叫中心
- 投诉分类汇总

公共设施

- 天气预报
- 餐饮查询

社交网络

- 舆情分析
- 热点发现

1.4 自然语言处理中用到的知识

语音学	• 语言学的一个分支。 • 主要研究语言的发音机制，语音特性和在言谈中的变化规律。
词法学	• 构词学，又称形态学，是语言学的一个分支。 • 研究单词的内部结构和其形成方式。
句法学	• 研究语言的句子结构。
语义学	• 研究自然语言的意义。
语用学	• 语言学、哲学和心理学的一个分支学科。 • 研究脉络如何影响人运用和理解语言。

1.5 自然语言处理面临的困难

- 歧义现象的处理
  - 语法层面
  - 句法层面
  - 语义层面
  - 语用层面
  - 语音层面
- 未知语言现象的处理

  灌水、盖楼、沙发、童鞋、盆友、驴友

  喜大普奔

  Awsl、 xswl、 ssfd

他说：“她这个人真有意思funny。”

她说：“他这个人也怪有意思的funny。”

人们以为他俩有了意思wish，就让他向她意思意思express。

他急了：“我根本没那个意思thought！”

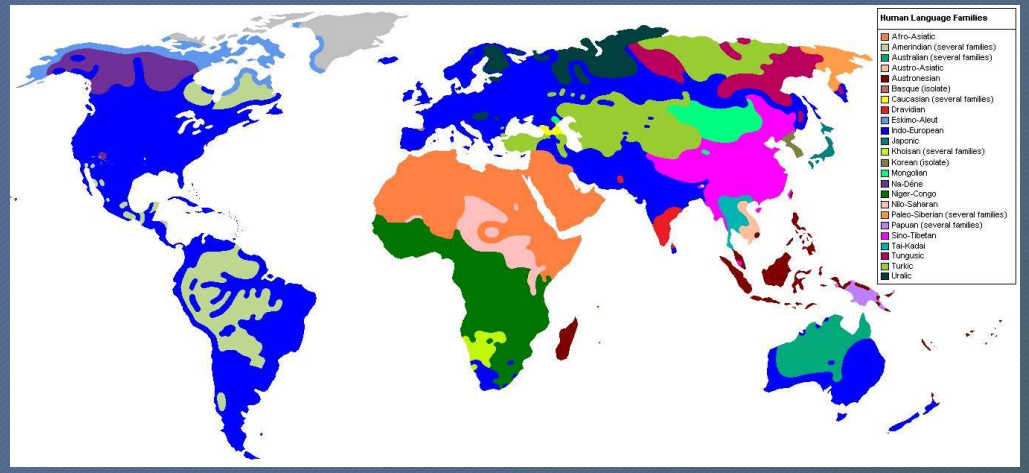
她也生气了：“你们这么说是什么意思intention？！”

有人觉得这个段子很有意思funny，但是也有人觉得这个段子并没有意思sense。

-《生活报》 1994.11.13 第六版

1.5 自然语言处理面临的困难

- 语言是没有规律的，或者说规律是错综复杂的。
- 语言是可以自由组合的，可以组合复杂的语言表达。
- 语言是一个开放集合，我们可以任意的发明创造一些新的表达方式。
- 语言需要联系到实践知识，有一定的知识依赖。
- 语言的使用要基于环境和上下文。



1.6 自然语言处理的基本方法及其发展

- 理性主义方法：基于规则的方法
- 经验主义方法：基于统计的方法

1.7 学科现状

有些问题得到了基本解决

- 词性标注
- 命名实体识别
- 垃圾邮件识别

有些问题取得长足进展

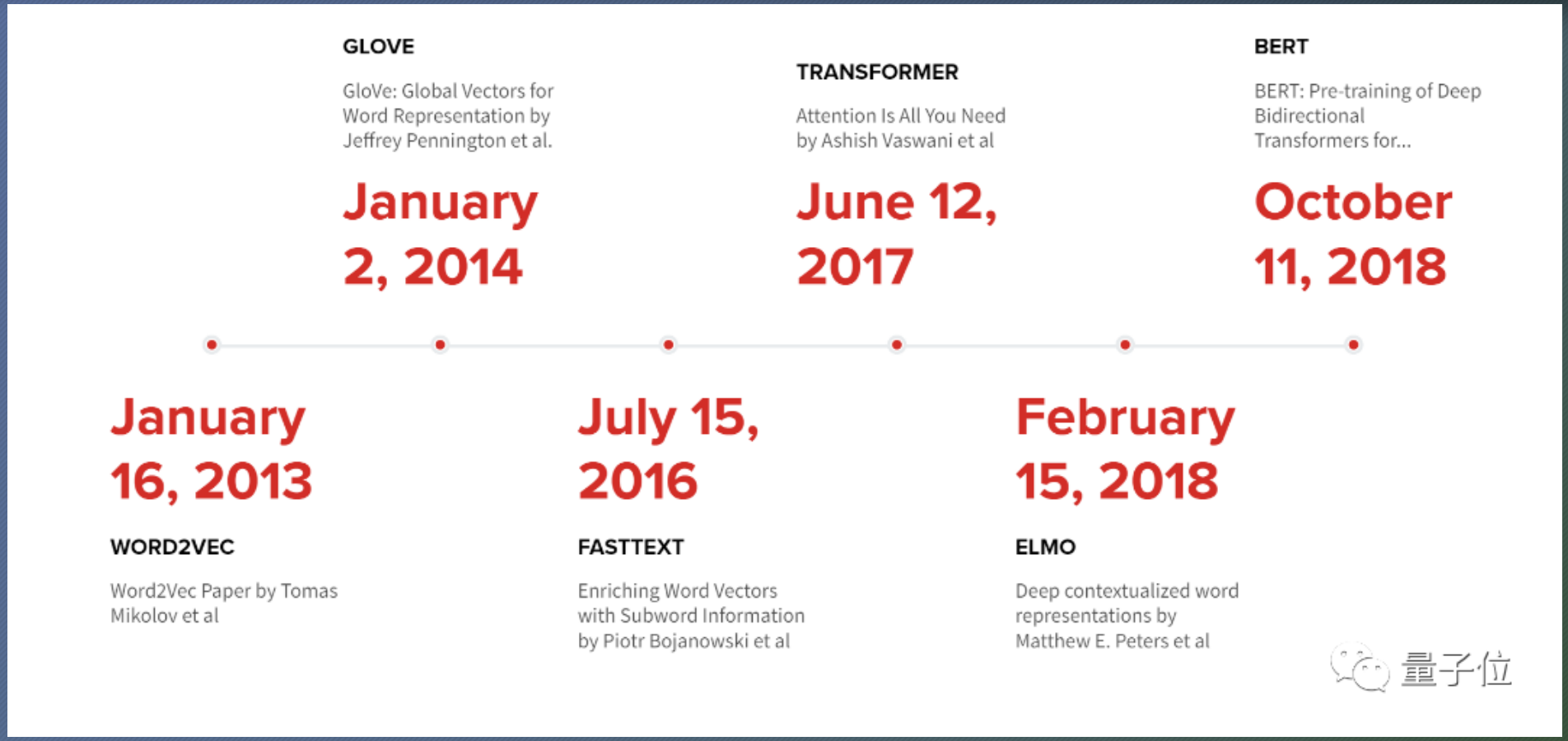
- 情感分析
- 共指消解
- 词义消歧
- 句法分析
- 机器翻译
- 信息抽取

有些问题依然充满挑战

- 自动问答
- 复述
- 文摘提取
- 会话机器人

Stafford，Dan Jurafsky：https://blog.csdn.net/longxinchen_ml/article/details/50543337

前沿研究



https://cloud.tencent.com/developer/article/1581666

前沿研究

NLP领域国际会议

- **ACL**
- EMNLP
- NAACL
- COLING

Association for Computational Linguistics

中国中文信息学会 Chinese Information Processing Society of China

http://www.cipsc.org.cn/

中文信息学报 JOURNAL OF CHINESE INFORMATION PROCESSING

http://jcip.cipsc.org.cn/CN/volumn/home.shtml

其中 ACL、EMNLP、NAACL均由 ACL 举办。
ACL 是 CCF 推荐A类国际学术会议，
EMNLP 和 COLING 是B类，
NAACL 则是C类。

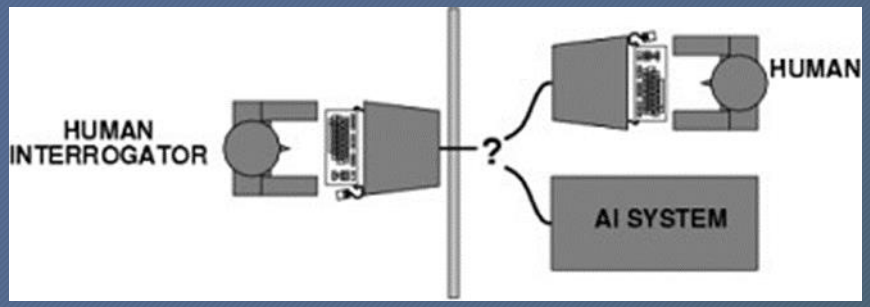
https://blog.csdn.net/weixin_34613450/article/details/86679630

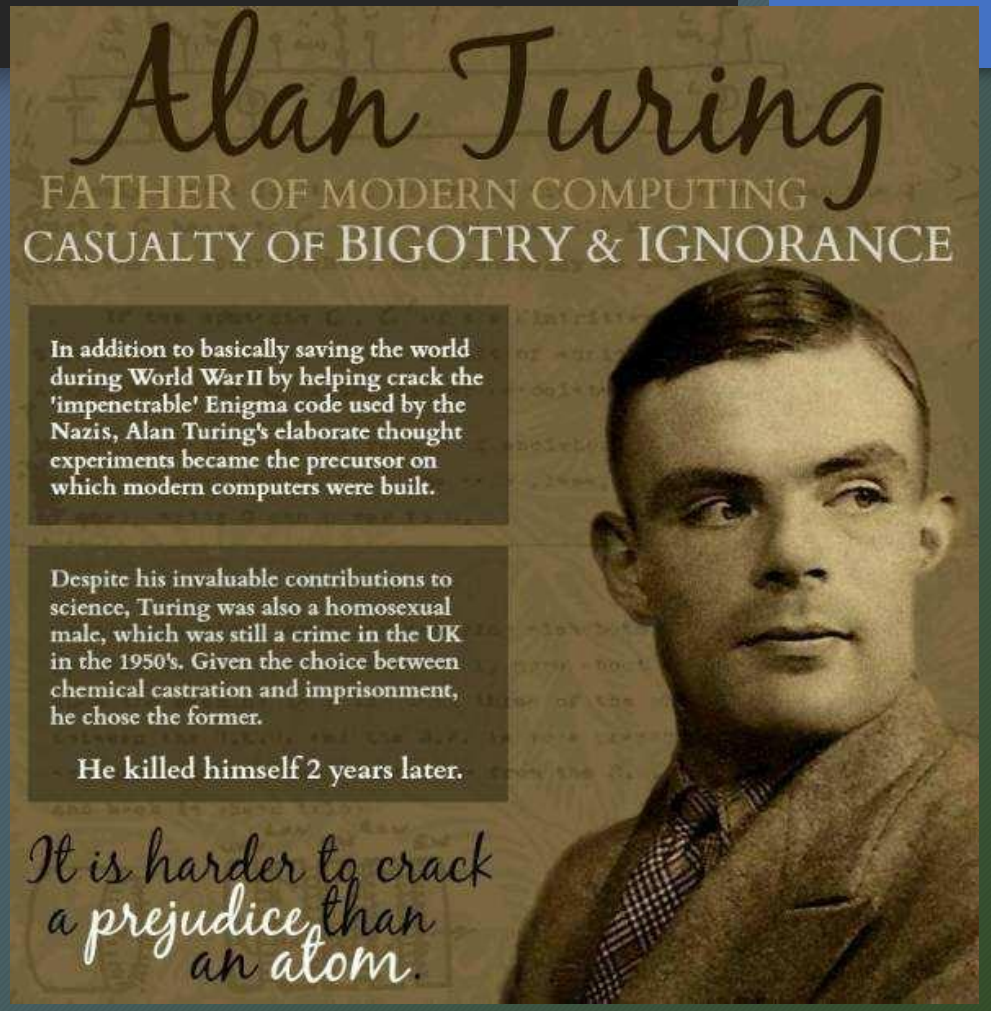
1.8 语言、思维和理解

机器是否具有智能？

图灵测试（Turing，1950）

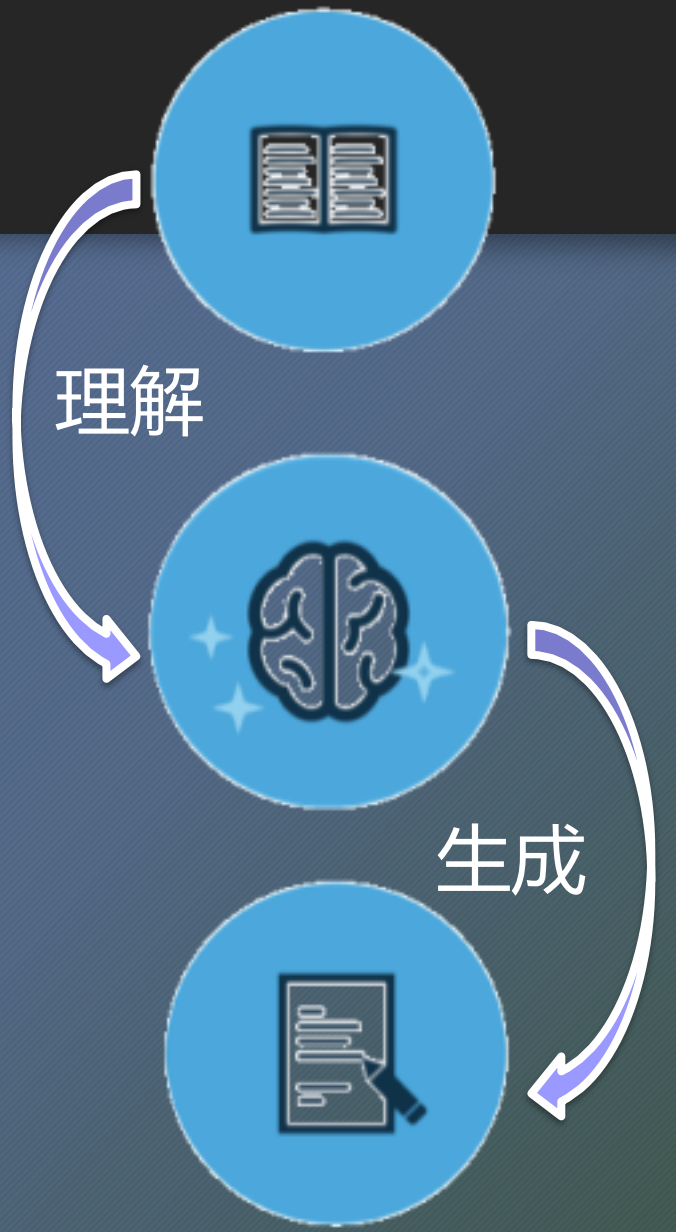
把一个人和一台机器隔开，人可以随意向电脑提问，进行多次测试后，如果人不能确定被测试者是人还是机器，那么这台机器就通过了测试，被认为具有人工智能。





自然语言处理任务

词法分析
句法分析
篇章分析



语言模型
机器翻译
问答系统

1.9 本书结构

基础知识：1-6章

- 语言学资源建设：第2章
- 语言计算模型：第3、4、5章
- 字符集编码体系：第6章

基本技术：7-9章

- 词法分析：第7章
- 句法分析：第8章
- 语义分析：第9章

自然语言处理工具包

- 结巴分词 jieba
- **HanLP**
- SnowNLP
- NLPIR
- 哈工大 LTP
- 中科院 **ICTCLAS**
- 清华大学 THULAC
- 复旦大学 FudanNLP

- NLTK
- Genism
- TextBlob
- Stanford NLP
- Spacy

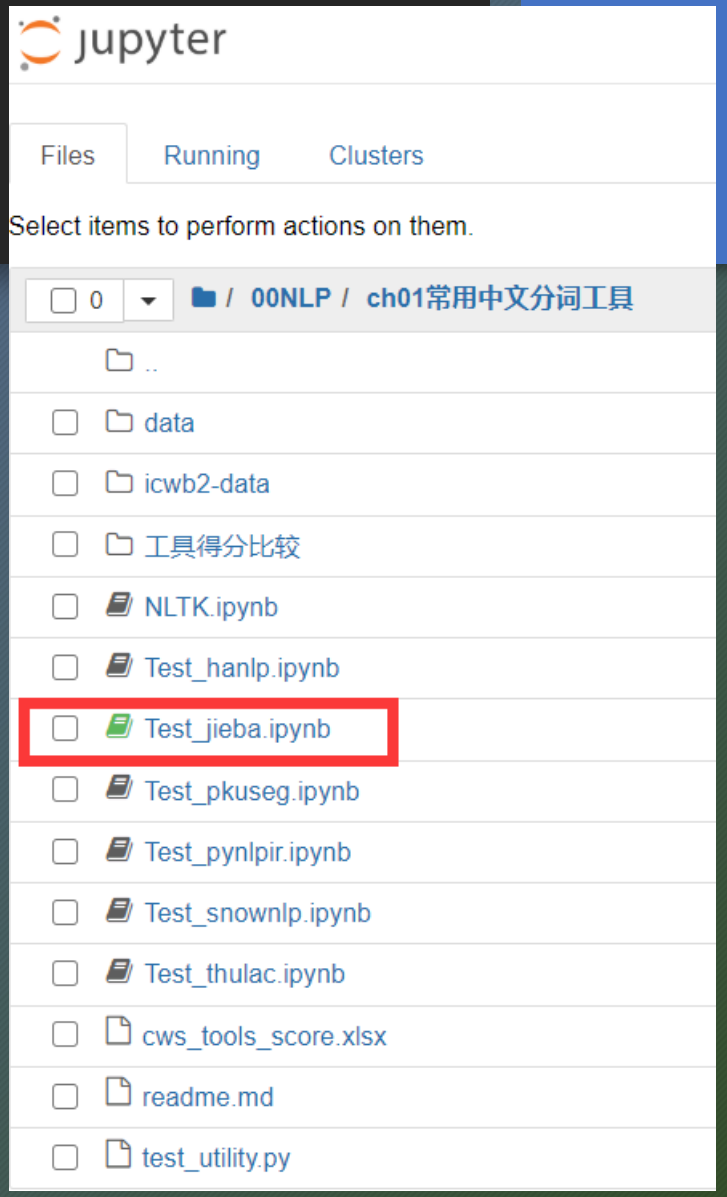
https://github.com/HBU/Jupyter/tree/master/00NLP/ch01%E5%B8%B8%E7%94%A8%E4%B8%AD%E6%96%87%E5%88%86%E8%AF%8D%E5%B7%A5%E5%85%B7

Jieba分词

```
import jieba
```

```
seg_list = jieba.cut("我来到北京清华大学", cut_all=True)# 全模式
print("\nFull Mode: " + "/ ".join(seg_list))

seg_list = jieba.cut("我来到北京清华大学", cut_all=False) # 精确模式
print("\nDefault Mode: " + "/ ".join(seg_list))

seg_list = jieba.cut("我来到北京清华大学")  # 默认是精确模式
print('\n默认模式:'+", ".join(seg_list))
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\David\AppData\Local\Temp\jieba.cache
Loading model cost 0.648 seconds.
Prefix dict has been built successfully.


Full Mode: 我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学

Default Mode: 我/ 来到/ 北京/ 清华大学

默认模式:我, 来到, 北京, 清华大学
```



Pkuseg 北京大学

```
import pkuseg
sentence = '萨哈夫说，伊拉克将同联合国销毁伊拉克大规模杀伤性武器特别委员会继续保持合作。'
seg = pkuseg.pkuseg(postag=False)  # 以默认配置加载模型，不进行词性分析
sentence = seg.cut(sentence)  # 进行分词
print(' '.join(sentence))
```

```
萨哈夫 说 ， 伊拉克 将 同 联合国 销毁 伊拉克 大规模 杀伤性 武器 特别 委员会 继续 保持 合作 。
```

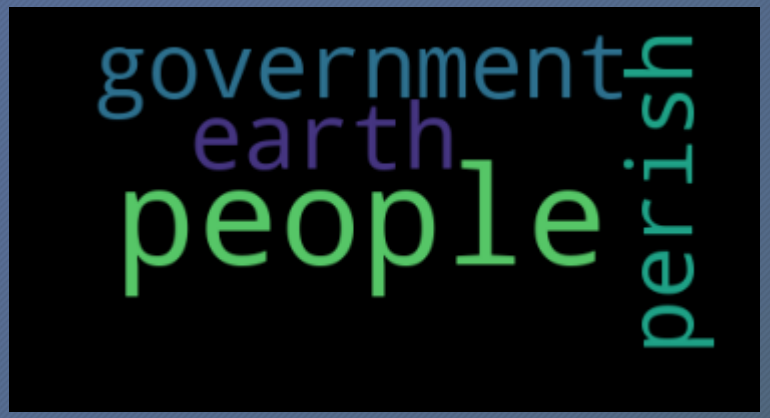
HanLP

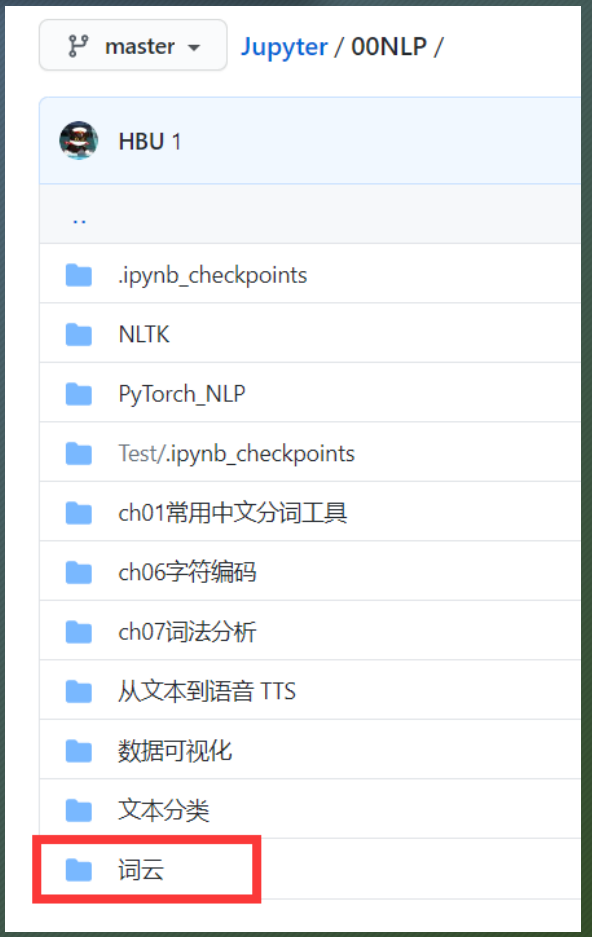
```
from pyhanlp import *

content = "萨哈夫说，伊拉克将同联合国销毁伊拉克大规模杀伤性武器特别委员会继续保持合作。"
print(HanLP.segment(content))
```

```
[萨哈夫/nrf, 说/v, ，/w, 伊拉克/nsf, 将/d, 同/p, 联合国/nt, 销毁/v, 伊拉克/nsf, 大规模/b, 杀伤性/n, 武器/n, 特别/d, 委员会/nis, 继续/v, 保持/v, 合作/vn, 。/w]
```

词云 wordcloud

```
import wordcloud
w = wordcloud.WordCloud()
w.generate('and that government of the people, by the people, for the people, shall not perish from the earth.')
w.to_file('wordclouds/output1.png')
```





词云 wordcloud

TTS-pyttsx3 普通语速，读文章

```python
import pyttsx3
# 普通语速，读文章

f = open('Spring.txt',encoding='utf-8')
msgf = f.read()
print(msgf)
teacher = pyttsx3.init()
teacher.say(msgf)
teacher.runAndWait()
```

盼望着，盼望着，东风来了，春天的脚步近了。
一切都像刚睡醒的样子，欣欣然张开了眼。山朗润起来了，水涨起来了，太阳的脸红起来了。
小草偷偷地从土里钻出来，嫩嫩的，绿绿的。园子里，田野里，瞧去，一大片一大片满是的。坐着，躺着，打两个滚，踢几脚球，赛几趟跑，捉几回迷藏。风轻悄悄的，草软绵绵的。
桃树、杏树、梨树，你不让我，我不让你，都开满了花赶趟儿。红的像火，粉的像霞，白的像雪。花里带着甜味儿，闭了眼，树上仿佛已经满是桃儿、杏儿、梨儿。花下成千成百的蜜蜂嗡嗡地闹着，大小的蝴蝶飞来飞去。野花遍地是：杂样儿，有名字的，没名字的，散在花丛里，像眼睛，像星星，还眨呀眨的。
“吹面不寒杨柳风”，不错的，像母亲的手抚摸着你。风里带来些新翻的泥土的气息，混着青草味儿，还有各种花的香，都在微微润湿的空气里酝酿。鸟儿将巢安在繁花嫩叶当中，高兴起来了，呼朋引伴地卖弄清脆的喉咙，唱出宛转的曲子，跟轻风流水应和着。牛背上牧童的短笛，这时候也成天在嘹亮地响着。
雨是最寻常的，一下就是三两天。可别恼。看，像牛毛，像花针，像细丝，密密地斜织着，人家屋顶上全笼着一层薄烟。树叶儿却绿得发亮，小草也青得逼你的眼。傍晚时候，上灯了，一点点黄晕的光，烘托出一片这安静而和平的夜。在乡下，小路上，石桥边，有撑起伞慢慢走着的人；还有地里工作的农民，披着蓑戴着笠。他们的草屋，稀稀疏疏的，在雨里静默着。
天上风筝渐渐多了，地上孩子也多了。城里乡下，家家户户，老老小小，也赶趟儿似的，一个个都出来了。舒活舒活筋骨，抖擞抖擞精神，各做各的一份儿事去，“一年之计在于春”；刚起头儿，有的是工夫，有的是希望。
春天像刚落地的娃娃，从头到脚都是新的，它生长着。
春天像小姑娘，花枝招展的，笑着，走着。
春天像健壮的青年，有铁一般的胳膊和腰脚，他领着我们上前去。

配置编程环境

- Anaconda
- Jupyter / pycharm / VS code
- Python 3

THE END