



NLP

自然语言处理技术基础

Natural Language Processing, NLP

网络空间安全与计算机学院

第2章 语料库与词汇知识库

2.1 语料库

- 2.1.1 基本概念

- 2.1.2 语料库类型

- 2.1.3 典型语料库介绍

- 2.1.4 语料处理的基本问题

2.2 词汇知识库

- 2.2.1 WordNet

- 2.2.2 知网

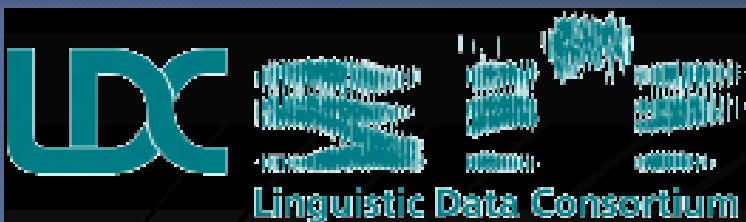
2.1 语料库 Corpus (Corpora 复数)

语料库：存放**语言材料**的数据库

- ①存放真实出现过的语言材料
- ②计算机为载体承载语言知识
- ③分析处理后的真实语言材料

ID	语句
1	进城/v 后任/n 天津/ns 女中/j 教导/n 主任/n ， /w 以后/nt 到/v 河北大学/ni 、 /w 沧洲师院/ni 、 /w 天津师院/ni 任教/v 。 /w #语料来源 [样本名称]:N/A;[作者]:N/A;[写作时间]:N/A;[出版时间]:1982-4-3;[书刊名称]:天津日报;[编著者]:N/A;[出版社]:天津日报社
2	在/p 河北大学/ni 执教/v 的/u 我国/n 著名/a 历史学家/n 、 /w 教授/n 漆侠/nh 和/c 人大代表/j 、 /w 副教授/n 张亚丽/nh 等/u ， /w 七月/nt 一/m 日/nt 加入/v 中国共产党/ni 。 /w #语料来源 [样本名称]:N/A;[作者]:N/A;[写作时间]:N/A;[出版时间]:1984-7-2;[书刊名称]:河北日报;[编著者]:N/A;[出版社]:河北日报社

主要语料库



<https://www.ldc.upenn.edu/>

LDC, 全名Linguistic Data Consortium, 语言数据联盟

由大学、图书馆、企业、政府、研究机构共同合办的联合企业，
成立于1992年，
目前由宾夕法尼亚大学负责主要运营。

Pay the membership fee

- Not-for-profit organizations and US Government entities
 - Standard membership: \$2,400
 - Subscription membership: \$3,850
- For-Profit organizations
 - Standard membership: \$24,000
 - Subscription membership: \$27,500



货币兑换

2400美元=16240.56人民币

1. LDC 中文树库

LDC 中文树库(Chinese Tree Bank, CTB)^①是由美国宾夕法尼亚大学(UPenn)负责开发,并通过语言数据联盟(LDC)发布的中文句法树库,该树库收集的语料取材于新华社和香港新闻等媒体,目前该语料库已经发展成为第7版,由2400个文本文件构成。含45000个句子,110万个词,165万个汉字。文件由GBK和UTF-8两种编码格式存储。



<http://icame.uib.no/>



<https://ota.bodleian.ox.ac.uk/repository/xmlui/>

国内语料库

教育部语言文字应用研究所计算语言学研究室

<http://corpus.zhonghuayuwen.org/index.aspx>

北京大学计算语言学研究

<https://klcl.pku.edu.cn/zygx/zyxz/index.htm>

哈工大信息检索研究中心

http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

现代汉语语料库检索

http://corpus.zhonghuayuwen.org/CnCindex.aspx



- 首页
- 语料库检索
- 语料分析处理
- 语料字词检索

现代汉语语料库检索

查询条件格式

词类标记代码

查询条件: 河北大学

每页句数: 100

检索

模式: ☐ 整词匹配 ☒ 模糊匹配 ☐ 全文检索 输出: ☒ 生语料 ☐ 标注语料 显示: ☒ 语料出处 ☐ 关键词居中

第1到2条, 共查询到2条符合要求的例句!

下载语料

ID	语句	
1	在河北大学执教的我国著名历史学家、教授漆侠和人大代表、副教授张亚丽等, 七月一日加入中国共产党。 #语料来源 [样本名称]:N/A; [作者]:N/A; [写作时间]:N/A; [出版时间]:1984-7-2; [书刊名称]:河北日报; [编著者]:N/A; [出版社]:河北日报社	
2	进城后任天津女中教导主任, 以后到河北大学、沧洲师院、天津师院任教。 #语料来源 [样本名称]:N/A; [作者]:N/A; [写作时间]:N/A; [出版时间]:1982-4-3; [书刊名称]:天津日报; [编著者]:N/A; [出版社]:天津日报社	

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

使用工具为免费文本资源加标记



语料库语言学 (Corpus Linguistics)

基于语料库进行语言学研究的一门学科

研究自然语言电子文本的采集、存储、标注、检索、统计等方法的一门学问

目的是通过对客观存在的大规模真实文本中的语言事实进行定量分析，为语言学研究或自然语言处理系统开发提供支持。

- 当代语言学与计算机科学交叉
- 用计算机对巨量的语料库进行高速检索、统计和展示
- 揭示真实语言使用的倾向性规律及其所传递的意义、功能乃至思想意识



2.1.2 语料库类型

- 按用途：通用语料库 专用语料库
- 按语种：单语语料库 多语语料库 双语/平行语料库
- 按时效性：共时语料库 历时语料库
- 按是否被标注：生语料库 熟语料库

2.1.3 典型语料库介绍

- Brown语料库
- LOB (Lancaster Oslo Bergen)
- Penn TreeBank
- PropBank
- NomBank
- FrameNet
- The Canadian Hansards
- LC-STAR
- C-STAR
- **北京大学语料库**
- LDC中文树库CTB

北大《人民日报》语料库

北京大学对1998年全年《人民日报》分词、词性标注
1999.4-2002.4

【免费资源】现代汉语切分、标注、注音语料库-1998年1月份样例与规范

发布日期：2012-02-15

现代汉语切分、标注、注音语料库（目前已有1998、2000两年加工好的新闻语料，视规模需另签委托加工协议）

（发布者：网站管理员）

附件：[12973779684875000004-现代汉语切分、标注、注音语料库-1998年1月份样例与规范20110330.rar](#)

北京大学版权所有

联系方式：北京大学计算语言学教育部重点实验室 邮编：100871

使用以下浏览器可获得最佳效果1024*768 Internet Explorer 6.0 及更高版本

<https://klcl.pku.edu.cn/zygx/zyxz/index.htm>

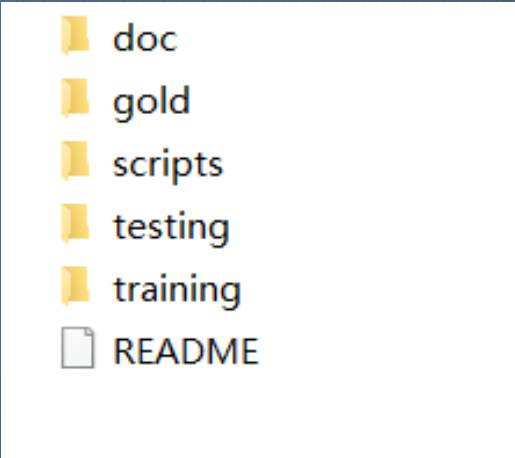
icwb2-data 中文分词数据集

2005.11联合发布:

- 北京大学
- 香港城市大学
- 台湾 CKIP, Academia Sinica
- 中国微软研究所

AS 和 CityU 为繁体中文数据集
PK 和 MSR 为简体中文数据集

Corpus	Encoding Types	Word	Words Types	Character	Characters
Academia Sinica	Big Five Plus	141,340	5,449,698	6,117	8,368,050
CityU HKSCS	Big Five	69,085	1,455,629	4,923	2,403,355
Peking University	CP936	55,303	1,109,947	4,698	1,826,448
Microsoft Research	CP936	88,119	2,368,391	5,167	4,050,469



- doc
- gold
- scripts
- testing
- training
- README

Second International Chinese Word Segmentation Bakeoff

第二届国际中文分词测评

<http://sighan.cs.uchicago.edu/bakeoff2005/>

SIGHAN - 汉字特别兴趣小组 国际计算语言学会 (ACL)

Special Interest Group for Chinese Language Processing of the Association for Computational Linguistics



国际中文分词评测

International Chinese Word Segmentation Bakeoff

- 第一届2003年, 日本札幌举行 (Bakeoff 2003)
- 第二届2005年, 韩国济州岛举行 (Bakeoff 2005)
- 第三届2006年, 澳大利亚悉尼举行 (Bakeoff 2006) 加入了中文命名实体识别评测
- SIGHAN-4
- SIGHAN-5
- SIGHAN-6
- SIGHAN-7 2013 Nagoya名古屋, Japan
- SIGHAN-8 2015 Beijing北京, China
- SIGHAN-9 2017 Taipei台北, China

2010年第一届CIPS-SIGHAN联合会议 北京

2012年第二届CIPS-SIGHAN联合会议 天津

2014年第三届CIPS-SIGHAN联合会议 武汉

2.1.4 语料处理的基本问题

汉语预处理:

- 分词 (第六章详细介绍)

英语预处理:

- 空格围起了多个词
 - ① 词+标点: etc.
 - ② 词+单撇号: isn't
 - ③ 连接字符连接多个单词: 26-years-old
- 空格不是分界标志
 - 例如: 138 0312 8888 、 New York

2.2 词汇知识库

2.2.1 WordNet

2.2.2 知网

教参《统计自然语言处理》（第二版）宗成庆

4.2	语言知识库.....	67
4.2.1	WordNet	68
4.2.2	FrameNet	69
4.2.3	EDR	70
4.2.4	北京大学综合型语言知识库	71
4.2.5	知网	73
4.2.6	概念层次网络	77

2.2 语言知识库

“语言知识库”比“语料库”包含更广泛的内容

包括：词汇知识库、句法规则库、语法信息库、语义概念库等


2.2.1 WordNet

基于认知语言学的英语词典。

不仅按单词以字母顺序排列，
而且按照单词的意义组成一个“单词的网络”。

包含描述：

- 概念含义
- 一义多词
- 一词多义
- 类别归属
- 近义
- 反义
-

 PRINCETON UNIVERSITY


WordNet

A Lexical Database for English

What is WordNet

People

News

Use Wordnet Online 

Download

Citing WordNet

License and Commercial Use

Related Projects

Documentation

Publications


Frequently Asked Questions

What is WordNet?

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the creators of WordNet and do not necessarily reflect the views of any funding agency or Princeton University.

When writing a paper or producing a software application, tool, or interface based on WordNet, it is necessary to properly **cite the source**. Citation figures are critical to WordNet funding.

About WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the **browser** . WordNet is also freely and publicly available for **download**. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

- <https://wordnet.princeton.edu/>

Wordnet 使用

- wordnet是nltk的一个组件，因此需要先下载nltk

```
from nltk.corpus import wordnet as wn

print(wn.synsets('published')) # 打印publish的多个词义
[Synset('print.v.01'), Synset('publish.v.02'), Synset('publish.v.03'), Synset('published.a.01'), Synset('promulgated.s.01')]

dog = wn.synset('dog.n.01') # 狗的概念
print(dog)
Synset('dog.n.01')

print(dog.hypernyms()) # 狗的父亲(上位词)
[Synset('canine.n.02'), Synset('domestic_animal.n.01')]

print(dog.hyponyms()) # 狗的子类(下位词)
[Synset('basenji.n.01'), Synset('corgi.n.01'), Synset('cur.n.01'), Synset('dalmatian.n.02'), Synset('great_pyrenees.n.01'), Synset('g riffon.n.02'), Synset('hunting_dog.n.01'), Synset('lapdog.n.01'), Synset('leonberg.n.01'), Synset('mexican_hairless.n.01'), Synset('n ewfoundland.n.01'), Synset('pooch.n.01'), Synset('poodle.n.01'), Synset('pug.n.01'), Synset('puppy.n.01'), Synset('spitz.n.01'), Synset('toy_dog.n.01'), Synset('working_dog.n.01')]

Test1 = wn.synset('motorcar.n.01')
print(Test1)
Synset('car.n.01')

print(Test1.lemma_names()) # lemma_names() # 同义
['car', 'auto', 'automobile', 'machine', 'motorcar']

Test2 = wn.synset('good.a.01')
print(Test2)
Synset('good.a.01')

print(Test2.lemmas()[0].antonyms()) # antonyms() # 反义
[Lemma('bad.a.01.bad')]
```

(1) 上位词/下位词

- 1 | `hypernyms()` # 上位(父类)
- 2 | `hyponyms()` # 下位(子类)

(2) 同义词/反义词

- 1 | `lemma_names()` # 同义
- 2 | `antonyms()` # 反义

(3) 蕴涵关系

`entailments()`

(4) 整体与部位

- 1 | `part_meronyms()` # 部分
- 2 | `substance_meronyms()` # 实质
- 3 | `member_holonyms()` # 成员

(5) 计算概念之间距离

- 1 | `path_similarity()` # 相似度
- 2 | `lowest_common_hypernyms()` # 在何种层面相似

2.2.2 知网 HowNet

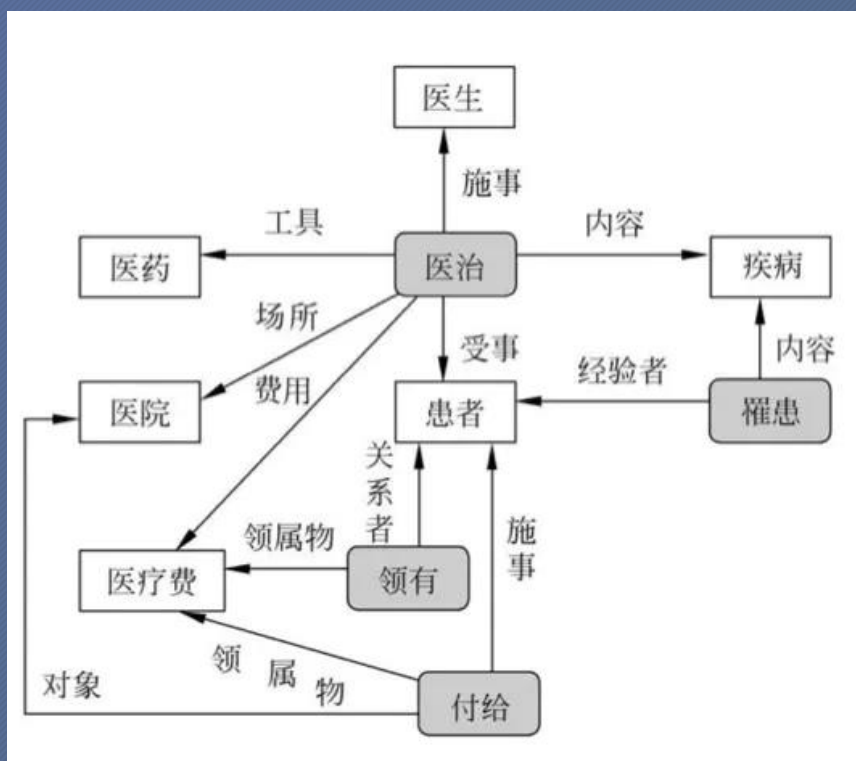
- 此知网(HowNet)非彼知网 (CNKI)

- <http://www.keenage.com/>
- <https://www.cnki.net/>

- 知网是董振东和董强经过多年努力创建的语言知识库
- 是一个以汉语和英语的词语所代表的概念为描述对象



2.2.2 知网



常识知识库：
揭示概念与概念之间，
以及概念所具有的属性之间的关系

国外语料库

- BBC语料库: <http://bcc.blcu.edu.cn/>
- BNC-英国国家语料库 (British National Corpus) : <http://www.natcorp.ox.ac.uk/>
- BOE-柯林斯英语语料库 (the Bank of English) : <http://www.collinslanguage.com/wordbanks/>
- 联合国文件数据库 (提供80万份六种语言平行文档) <http://documents.un.org/simple.asp>
- ANC-美国国家语料库 (American National Corpus) :<http://www.anc.org/>
- 兰开斯特汉语语料库 (LCMC) <http://ota.oucs.ox.ac.uk/scripts/download.php?otaid=2474>
- OLAC语言开发典藏社群 (Open Language Archives Community) <http://search.language-archives.org/index.html>
- COCA-美国当代英语语料库(Corpus of Contemporary American English)<http://www.americanacorp.org/>
- COHA-美国近当代英语语料库 (Corpus of Historical American English) : <http://corpus.byu.edu/coha/>
- SKETCHENGINE多语言语料库: www.sketchengine.co.uk
- BASE-英国学术口语语料库 (British Academic Spoken English Corpus) : <http://www2.warwick.ac.uk/fac/soc/celte/research/base/>
- Leeds: <http://corpus.leeds.ac.uk/internet.html>
- JustTheWord: <http://193.133.140.102/JustTheWord/index.html>
- Lextutor: <http://www.lex tutor.ca/>
- Web Concordancer: www.edict.com.hk

国内语料库 1

- 语料库: <http://yulk.org/>
- 语料库在线: <http://www.cncorpus.org/>
- 北京大学中国语言学研究中心: <http://ccl.pku.edu.cn/corpus.asp>
- 国家语委现代汉语语料库: <http://www.cncorpus.org/>
- 北外语料库语言学: <http://www.bfsu-corpus.org/>
- 古代汉语语料库: <http://www.cncorpus.org/login.aspx>
- 语料库语言学在线: <http://ccl.pku.edu.cn/corpus.asp>
- 《人民日报》标注语料库: http://www.icl.pku.edu.cn/icl_res/
- 汉语国际教育技术研发中心: HSK动态作文语料库<http://202.112.195.192:8060/hsk/login.asp>
- 语言研究所: 北京口语语料查询系统 (BJKY)
http://www.blcu.edu.cn/yys/6_beijing/6_beijing_chaxun.asp
- 现代汉语平衡语料库: <http://www.sinica.edu.tw/SinicaCorpus/>
- 古汉语语料库: <http://www.sinica.edu.tw/ftms-bin/ftmsw>
- 近代汉语标记语料库: http://www.sinica.edu.tw/Early_Mandarin/

国内语料库 2

- 树图数据库: <http://treebank.sinica.edu.tw/>
- 中英双语知识本体词网<http://bow.sinica.edu.tw/>
- 搜文解字: <http://words.sinica.edu.tw/>
- 文国寻宝记: <http://www.sinica.edu.tw/wen/>
- 唐诗三百首: <http://cls.admin.yzu.edu.tw/300/>
- 汉籍电子文献: <http://www.sinica.edu.tw/~tdbproj/handy1/>
- 红楼梦网络教学研究数据中心: <http://cls.hs.yzu.edu.tw/HLM/home.htm>
- 中国传媒大学文本语料库检索系统: <http://ling.cuc.edu.cn/RawPub/>
- 哈工大信息检索研究室对外共享语料库资源: http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm
- 香港教育学院语言资讯科学中心及其语料库实验室: <http://www.livac.org/index.php?lang=sc>
- 中文语言资源联盟: <http://www.chineseldc.org/>
- 杨百翰大学语料库: <http://view.byu.edu/>

第三周，上课带笔记本电脑

课上计划完成以下目标：

- ① 分词 WS
- ② 词云 WordCloud
- ③ 文本转语音 TTS
- ④ 使用wordnet



THE END