原始网页信息

http://thuocl.thunlp.org/

# 词库

## 词条信息

| **Name** | **Entity Szie** | **Date** | **Link** |
| --- | --- | --- | --- |
| IT | 16000条 | 2016-12-24 | [download](http://thuocl.thunlp.org/source/THUOCL_it.txt) |
| 财经 | 3830条 | 2016-12-24 | [download](http://thuocl.thunlp.org/source/THUOCL_caijing.txt) |
| 成语 | 8519条 | 2016-12-24 | [download](http://thuocl.thunlp.org/source/THUOCL_chengyu.txt) |
| 地名 | 44805条 | 2017-06-01 | [download](http://thuocl.thunlp.org/source/THUOCL_diming.txt) |
| 历史名人 | 13658条 | 2016-12-24 | [download](http://thuocl.thunlp.org/source/THUOCL_lishimingren.txt) |
| 诗词 | 13703条 | 2017-01-20 | [download](http://thuocl.thunlp.org/source/THUOCL_poem.txt) |
| 医学 | 18749条 | 2017-01-20 | [download](http://thuocl.thunlp.org/source/THUOCL_medical.txt) |
| 饮食 | 8974条 | 2017-04-20 | [download](http://thuocl.thunlp.org/source/THUOCL_food.txt) |
| 法律 | 9896条 | 2017-04-28 | [download](http://thuocl.thunlp.org/source/THUOCL_law.txt) |
| 汽车 | 1752条 | 2017-05-15 | [download](http://thuocl.thunlp.org/source/THUOCL_car.txt) |
| 动物 | 17287条 | 2017-06-01 | [download](http://thuocl.thunlp.org/source/THUOCL_animal.txt) |

# THUOCL：清华大学开放中文词库

## 目录

* [词库简介](http://thuocl.thunlp.org/#%E9%A1%B9%E7%9B%AE%E4%BB%8B%E7%BB%8D)
* [词库格式及词频统计语料库](http://thuocl.thunlp.org/#%E7%9B%B8%E5%85%B3%E8%AE%BA%E6%96%87)
* [词库清单](http://thuocl.thunlp.org/#%E5%9C%A8%E7%BA%BF%E6%BC%94%E7%A4%BA)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [IT](http://thuocl.thunlp.org/#IT) | [财经](http://thuocl.thunlp.org/#caijing) | [成语](http://thuocl.thunlp.org/#chengyu) | [地名](http://thuocl.thunlp.org/#diming) | [历史名人](http://thuocl.thunlp.org/#lishimingren) | [诗词](http://thuocl.thunlp.org/#shici) | [医学](http://thuocl.thunlp.org/#yixue) | [饮食](http://thuocl.thunlp.org/#food) | [法律](http://thuocl.thunlp.org/#law) | [汽车](http://thuocl.thunlp.org/#car) | [动物](http://thuocl.thunlp.org/#animal) |

* [开源协议](http://thuocl.thunlp.org/#%E5%BC%80%E6%BA%90%E5%8D%8F%E8%AE%AE)
* [作者](http://thuocl.thunlp.org/#%E4%BD%9C%E8%80%85)

## 词库简介

THUOCL（THU Open Chinese Lexicon）是由清华大学自然语言处理与社会人文计算实验室整理推出的一套高质量的中文词库，词表来自主流网站的社会标签、搜索热词、输入法词库等。THUOCL具有以下特点：

1. 包含词频统计信息DF值（Document Frequency），方便用户个性化选择使用。
2. 词库经过多轮人工筛选，保证词库收录的准确性。
3. 开放更新，将不断更新现有词表，并推出更多类别词表。欢迎专业人士加入，协作建设开放词库，有意者请致信[thunlp@gmail.com](mailto:thunlp@gmail.com)。

该词库可以用于中文自动分词，提升中文分词效果。建议搭配本组研制开发的[THULAC工具包](http://thulac.thunlp.org/)使用，提升特定领域中文分词的效果。

## 词库格式及词频统计语料库

词库每一行由两部分组成，分别是词和DF值（存在此单词的文档个数），中间由Tab间隔。

词频统计语料库：

1. CSDN博客 时间：2014.07-2016.07 文档数：3785976
2. 新浪新闻 时间：2008.01-2016.11 文档数：8421097
3. 搜狗语料 文档数：729008561

## 词库清单

IT

**词表简介：**本词表包含了大量IT类词汇。

**词条样例：**文件备份、虚拟地址、C++编程、事务调度、强连通缩点。

**词条数量：**16000条

**词频统计语料库：**CSDN博客

**更新时间：**2016-12-24

**贡献者：**马云山、韩世依、张钰晖

**下载链接：**[点此下载](http://thuocl.thunlp.org/message)

财经

**词表简介：**本词表包含了大量财经类词汇。

**词条样例：**年期、调整方案、全面收购、差价、萎缩。

**词条数量：**3830条

**词频统计语料库：**新浪新闻

**更新时间：**2016-12-24

**贡献者：**韩世依、张钰晖、马云山

**下载链接：**[点此下载](http://thuocl.thunlp.org/message)

成语

**词表简介：**本词表包含了大量成语词汇。

**词条样例：**故作高深、有理有据、用之不竭、人微言轻、因地制宜、求贤若渴。

**词条数量：**8519条

**词频统计语料库：**新浪新闻

**更新时间：**2016-12-24

**贡献者：**韩世依、张钰晖、马云山

**下载链接：**[点此下载](http://thuocl.thunlp.org/message)

地名

**词表简介：**本词表包含了大量地名词汇。

**词条样例：**浙江、上海、澳大利亚、珠穆朗玛峰、湘潭县、大甲镇。

**词条数量：**44805条

**词频统计语料库：**搜狗语料

**更新时间：**2017-06-01

**贡献者：**韩世依、张钰晖、马云山

**下载链接：**[点此下载](http://thuocl.thunlp.org/message)

历史名人

**词表简介：**本词表包含了大量历史名人类词汇。

**词条样例：**陆游、荀彧、诸葛亮、孙权、张伯伦。

**词条数量：**13658条

**词频统计语料库：**新浪新闻

**更新时间：**2016-12-24

**贡献者：**韩世依、张钰晖、马云山

**下载链接：**[点此下载](http://thuocl.thunlp.org/message)

诗词

**词表简介：**本词表包含了大量诗词名句。

**词条样例：**更上一层楼、犹抱琵琶半遮面、路漫漫其修远兮、任尔东西南北风。

**词条数量：**13703条

**词频统计语料库：**新浪新闻

**更新时间：**2017-01-20

**贡献者：**张钰晖、韩世依、马云山

**下载链接：**[点此下载](http://thuocl.thunlp.org/message)

医学

**词表简介：**本词表包含了大量医学类词汇。

**词条样例：**患者、充血、皮疹、冬虫夏草。

**词条数量：**18749条

**词频统计语料库：**新浪新闻

**更新时间：**2017-01-20

**贡献者：**张钰晖、韩世依、马云山

**下载链接：**[点此下载](http://thuocl.thunlp.org/message)

饮食

**词库简介：**本词库包含了大部分饮食类词汇。

**词条样例：**土豆、火锅、意大利面、果佳、猴头菇。

**词条数量：**8974条

**词频统计语料库：**搜狗语料

**更新时间：**2017-04-20

**贡献者：**王盟源、吴佼玉、黄伟杰，林永天

**下载链接：**[点此下载](http://thuocl.thunlp.org/message)

法律

**词库简介：**本词库包含了大部分法律类词汇。

**词条样例：**版权、有关部门、有限责任公司、土地审裁处法官、日本庄园制度。

**词条数量：**9896条

**词频统计语料库：**搜狗语料

**更新时间：**2017-04-28

**贡献者：**王盟源、吴佼玉、黄伟杰，林永天

**下载链接：**[点此下载](http://thuocl.thunlp.org/message)

汽车

**词库简介：**本词库包含了大部分汽车类词汇。

**词条样例：**轿车、车展、东风本田、前挡风玻璃、四川丰田。

**词条数量：**1752条

**词频统计语料库：**搜狗语料

**更新时间：**2017-05-15

**贡献者：**王盟源、吴佼玉、黄伟杰，林永天

**下载链接：**[点此下载](http://thuocl.thunlp.org/message)

动物

**词库简介：**本词库包含了大部分动物类词汇。

**词条样例：**信鸽、梅花鹿、街鸽、四方藤、斑尾林鸽。

**词条数量：**17287条

**词频统计语料库：**搜狗语料

**更新时间：**2017-06-01

**贡献者：**王盟源、吴佼玉、黄伟杰，林永天

**下载链接：**[点此下载](http://thuocl.thunlp.org/message)

## 开源协议

1. THUOCL面向国内外大学、研究所、企业、机构以及个人免费开放，可用于研究与商业。
2. 欢迎对该工具包提出任何宝贵意见和建议。请发邮件至thunlp@gmail.com。
3. 如果您在THUOCL基础上发表论文或取得科研成果，请您在发表论文和申报成果时声明“使用了清华大学开放中文词库”，并按如下格式引用：
   * **中文： 韩世依, 张钰晖, 马云山, 涂存超, 郭志芃, 刘知远, 孙茂松. THUOCL：清华大学开放中文词库. 2016.**
   * **英文： Shiyi Han, Yuhui Zhang, Yunshan Ma, Cunchao Tu, Zhipeng Guo, Zhiyuan Liu, Maosong Sun. THUOCL: Tsinghua Open Chinese Lexicon. 2016.**