

## 用 Mathematica 过滤字幕文件只留下生词（可选附中英文释义）

 **LePtC**  
学物理的都好萌 ~

+ 关注他

Yang Hong 等 284 人赞同了该文章

大家有没有碰到过这种两难的情况：看美剧的时候，如果开字幕，就没法好好练听力。如果关掉字幕，就时不时地碰到生词，然后你就得暂停去查生词，影响观片体验。

以前我的解决方案是，提前读一遍字幕，把生词查出来，但这样会被剧透。看剧的话倒不是特别在乎，但看电影的话确实会降低乐趣。

靠人力来识别生词十分消耗时间和精力，于是我突然想到，如果让 matica 酱帮我过滤掉字幕文件中低级的单词，这样既免除了我的手工劳动，又不会被剧透，岂不美哉！

### 低级英文单词表

首先我通过检索从 [based on 450 million word COCA corpus](#) 这里获取了美式英语最高频的 5000 个单词，将纯单词保存为 csv 格式：

A1	:	x	✓	f	the		
	A	B	C	D	E	F	G
1	the						
2	be						
3	and						
4	of						
5	a						
6	in						
7	to						
8	have						
9	to						
10	it						
11	I						
12	that						
13	for						
14	you						
15	he						
4989	stereotype						
4990	sensor						
4991	laundry						
4992	manual						
4993	pistol						
4994	naval						
4995	plaintiff						
4996	kid						
4997	middle-class						
4998	apology						
4999	till						
5000							
5001							

en5000 +

（我删掉了 n't 这一条所以是 4999 个）

然而这些词都是原型，  
Mathematica 的 Word

▲ 赞同 284 ▼

44 条评论

分享

★ 收藏

...

> C: > 用户 > Le > AppData > Roaming > Mathematica > Paclets > Repository >

名称	扩...	大小	修改时间
文件夹 (9)			
WordData_Definitions-10.0.25		3.53 MB	今天 18:23
WordData_InflectedForms-10.0.25		352 KB	今天 16:45
WordData_BinaryIdToSenses-10.0.25		1.98 MB	今天 16:45
WordData_WordTold-10.0.25		1.80 MB	今天 16:44
WordData_Canonicalization-10.0.25		336 KB	今天 16:09

从 InflectedForms 中可以提取单词的所有变形：

```
WordData["fish", "InflectedForms"]
[单词数据]

{{fish, Noun, AquaticVertebrate} -> {fish},
 {fish, Noun, Food} -> {fish}, {fish, Verb, Grab} -> {fished, fishing, fishes},
 {fish, Verb, Search} -> {fished, fishing, fishes}}
```

经过时间一点也不长的 debug 之后，我写出了这个程序，它会读取前面的单词列表，然后为每一个单词添加所有变形形式：



```

SetDirectory[NotebookDirectory[]];
|设置目录      |当前笔记本的目录

WList = Import["en5000.csv"];
|导入

Echo[Length[WList], "Before:"];
|...      |长度

For[i = 1, i ≤ Length[WList], i++,
|For循环      |长度
    Temp = {};
    Dat = WordData[WList[[i, 1], "InflectedForms"];
    |单词数据
    If[! MissingQ[Dat],
    |如果 |丢失判定
        For[j = 1, j ≤ Length[Dat], j++,
        |For循环      |长度
            Temp = Join[Temp, Dat[[j, 2]]
            |连接
        ];
        WList[[i]] = Join[WList[[i]], DeleteDuplicates[Temp]];
        |连接      |删除重复元素
    ]
] // Timing
|计算时间

Export["en5000x.csv", WList];
|导出

WList = Flatten[WList];
|压平

Echo[Length[WList], "After:"];
|...      |长度

» Before: 4999

{2.85938, Null}

» After: 14 509

```

运行之后的效果：

▲ 赞同 284 ▼

● 44 条评论

🔗 分享

★ 收藏

...



## 从字幕中去掉低频词

▲ 赞同 284 ▼ ● 44 条评论 ➤ 分享 ★ 收藏 ...



```

1 1
2 00:00:02,051 --> 00:00:04,315
3 So, what you're eating
4 is not technically yogurt.
5
6 2
7 00:00:04,486 --> 00:00:06,750
8 It doesn't have enough
9 live acidophilus cultures.
10
11 3
12 00:00:06,922 --> 00:00:10,153
13 It's really just ice milk
14 with carrageenan added for thickness.
15
16 4
17 00:00:10,325 --> 00:00:15,160
18 - Well, that's very interesting.
19 - It's also not pink and has no berries.
20

```

程序的思想是这样的：把字幕文件逐行切开，如果每行第一个字符不是数字就认为是要过滤的英文字幕内容了，然后将每行的内容按空格和标点切成单个单词，如果单词存在于前面那个低频表里面就删掉。另外经过实践，我决定把首字母大写的词也删掉，这种词通常是这个剧里面特有的人名，或者如 DNA 这样的首字母缩写词，后者即使你不认识也不影响听力听出来。

```

AllFile = StringSplit[SFile, "\n"];
      [按模式匹配分割字符串]

For[n = 1, n ≤ Length[AllFile], n++,
  [For循环]      [长度]
  Test = AllFile[[n];
  If[Length[ToCharacterCode[Test]] > 0,
    [长度] [将字符转换成字符代码]
    c = ToCharacterCode[Test][[1];
      [将字符转换成字符代码]
    If[c < 48 || c > 57, (* 如果第一个字符不是数字 *)
      [如果]
      Test = StringSplit[DeleteStopwords[Test],
        [按模式匹配...] [删除停用词]
        {Whitespace, ",", ". ", "' ", ":", ";", "!", "-", "(", ")", "[", ""]];
        [空白字符]
      l = Length[Test]; Res = {};
      [长度]
      If[l > 0, For[i = 1, i ≤ l, i++,
        [如果] [For循环]
        If[MemberQ[WList, ToLowerCase[Test[[i]]]] ||
          [成员判定] [转换为小写]
          ToCharacterCode[Test[[i]][[1]] ≤ 90, , (* 如果首字符不是小写字母也不要 *)
            [将字符转换成字符代码]
          Res = Append[Res, Test[[i]]]
            [追加]
        ]];
      AllFile[[n]] = StringJoin[Riffle[Res, " "]];
      [连接字符串] [交互插入]
    ]]]
Export[
  [导出]
  "The.Big.Bang.Theory.S02E01.The.Bad.Fish.Paradigm.BluRay.720p.DTS.x264-CHD.
  words.srt", AllFile, "Text"];

```

过滤后的字幕文件：

▲ 赞同 284 ▼

● 44 条评论

➦ 分享

★ 收藏

...

↑

```
1 1
2 00:00:02,051 --> 00:00:04,315
3
4 technically yogurt
5
6 2
7 00:00:04,486 --> 00:00:06,750
8
9 acidophilus
10
11 3
12 00:00:06,922 --> 00:00:10,153
13
14 carrageenan thickness
15
16 4
17 00:00:10,325 --> 00:00:15,160
18
19 berries
20
```

附英文释义

如果 matica 酱能顺便帮我查单词然后附在字幕里就完美了，然而可惜的是 matica 酱的英译中功能尚不完善：（原来你也有除了生孩子之外不能做的事啊 →\_→）

```
WordTranslation["dog", "Chinese"]
[单词翻译]
{狗}

WordTranslation["acidophilus", "Chinese"]
[单词翻译]
Missing[NotAvailable]

WordData["acidophilus", "Definitions"]
[单词数据]
{{acidophilus, Noun} ->
  a bacterium that is used to make yogurt and to supplement probiotics}
```

想批量查中文的话可以交给谷歌娘，但我不知道怎样让谷歌娘和 matica 酱协作



目前英文解释功能是比较完善的了，下图是补充英文释义后的观片效果：

▲ 赞同 284 ▼

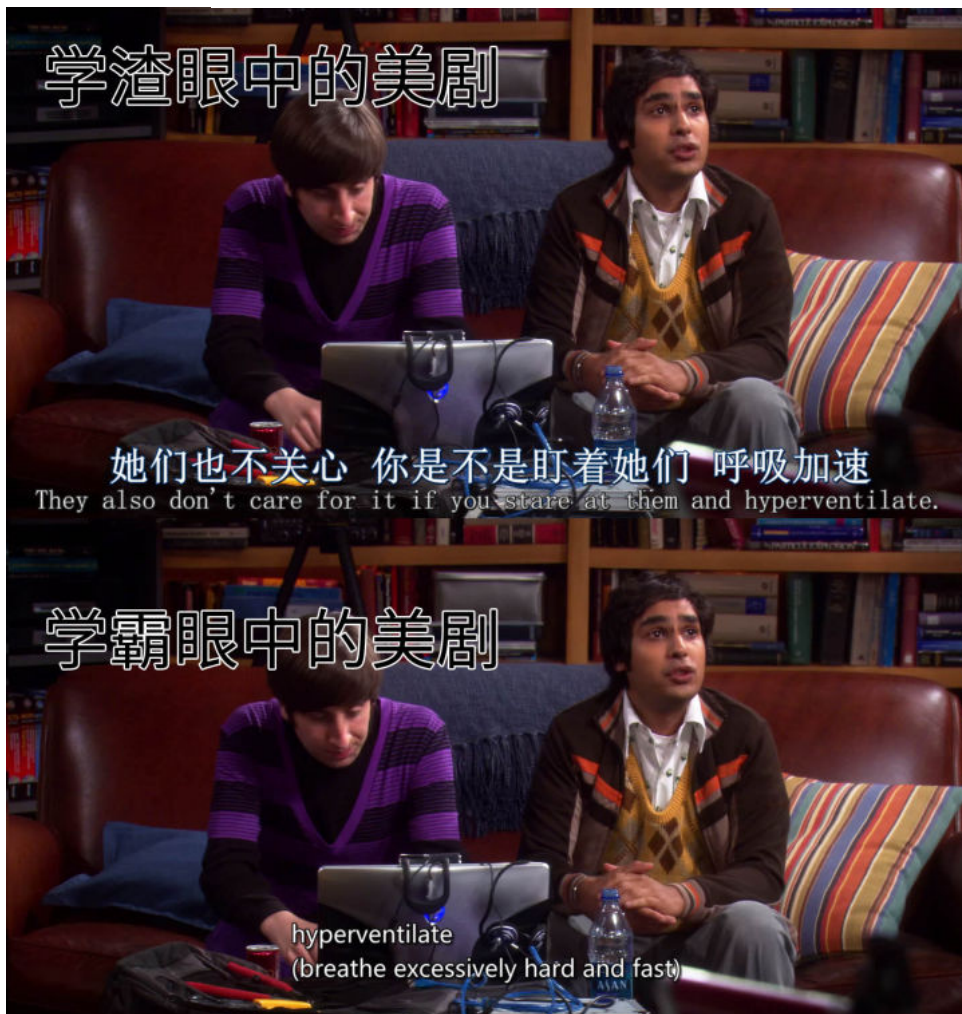
💬 44 条评论

➦ 分享

★ 收藏

...





再也不用手动挑生词吭哧吭哧查单词了！拿程序一跑就能开始看片了！妈妈再也不用担心我看美剧时不学英语了！

## TODO

以上 Mathematica 源代码及 csv, srt 文件均可在我的 GitHub 主页下载：[github.com/LePtC/Matica...](https://github.com/LePtC/Matica...)

欢迎感兴趣的童鞋来协作开发（目前附释义版的程序还有 bug，matica 酱发现词库里没有单词的释义的时候会直接把一滩代码吐在字幕里 = =）

目前过滤五千的标准对于我来说还是偏弱了点，以后还会做一个过滤八千到一万词汇量的，然而那个网站一万词汇是要收费的.....所以我还得想想从哪找词源，如果网友有 Excel 版的单词表欢迎共享出来（~△~）~

## Update 20170319

感谢 @王竞先 提供的指引，matica 酱现在能获取中文释义了

▲ 赞同 284 ▼

● 44 条评论

➦ 分享

★ 收藏

...



```
MyTextTranslation[en_] :=
StringCases[URLRead["http://dict.youdao.com/search?q=" <> en, "Body",
|字符串匹配 |读取URL响应
FollowRedirects → True, CharacterEncoding → None],
|遵循重定向 |真 |字符编码 |无
"<div class=\"trans-container\">" ~~ Shortest[a__] ~~ "<li>" ~~
|最短
Shortest[c__] ~~ "</li>" → c][1]
|最短

MyTextTranslation["hyperventilate"]

vi. 强力呼吸；换气过度
```

（在国内的网络环境下跑一集 TBBT 字幕大概是不到一分钟的时间）

matica 酱的顺序是先上有道娘，有道没有再上谷歌娘，如果谷歌娘也没有的话返回空（解决了吐一滩代码的 bug ...）

```
CNTrans[en_] :=
(ans = StringCases[URLRead["http://dict.youdao.com/search?q=" <> en,
|字符串匹配 |读取URL响应
"Body", FollowRedirects → True, CharacterEncoding → None],
|遵循重定向 |真 |字符编码 |无
"<div class=\"trans-container\">" ~~ Shortest[a__] ~~ "<li>" ~~
|最短
Shortest[c__] ~~ "</li>" → c];
|最短
If[Length[ans] == 0,
|... |长度
ans =
StringCases[
|字符串匹配
Quiet@
|不输出任何消息
URLRead[
|读取URL响应
"https://translate.google.cn/m?hl=en&sl=en&tl=zh-CN&ie=UTF-8&\prev=_m&q
=" <> en, "Body", FollowRedirects → True,
|遵循重定向 |真
CharacterEncoding → None],
|字符编码 |无
"<div dir=\"ltr\" class=\"t0\">" ~~ Shortest[c__] ~~ "</div>" → c]
|最短
];
If[Length[ans] > 0, ans[[1]], ""]
|... |长度
)

(* 自动查询有道中文释义，若没查到再查谷歌娘 *)
```

另外我在过滤程序里补充了单字母也去掉的规则，手动在低级词表加入 gonna 之类的口语词（所以这个程序的改进是需要收集大量实战经验的，请大家多多反馈啊）

观片体验视频：[【Mathematica】如何像一个学霸一样看美剧\\_野生技术协会\\_科技\\_bilibili\\_哔哩哔哩](#)

代码还是在前面那个地址下载，请认准 En filter v2.nb (￣△￣)~

## 常见问题统一回复 20170320

Q：有学英语需求的人很多，但装有 Mathematica 的人太少，用不了你这个程序

解决方案 ①：将程序

▲ 赞同 284 ▼

● 44 条评论

➦ 分享

★ 收藏

...



@薛定谔的喵 已经写了一个 Python 版的：[Celthi/meltSubtitles](#)  
（我本人不用 Python，所以也无法提供评测）

解决方案 ②：公共服务

我可以帮你们把最热门的几部美剧的生词版字幕给跑出来，然后挂在 GitHub 或者字幕网站上供你们下载

③ 源头解决方案：其实这种事由字幕组来做是最合适的不是吗？他们已经制作出双语字幕了，顺手就可以再改出一个“5k生词版”“1w生词版”的字幕，释义的准确度也会是最高的。希望大家帮忙传播，让字幕组的成员看到这个建议\_(:3」∠)\_

Q：只看生词还是跟不上，能不能保留全部英文字幕然后突出显示生词

A：Update：我发现 srt 字幕是可以做特效的.....



Q：有没有其它语言的

A：很难，Mathematica 只对英文单词有变形数据库，此外，汉语和日语都会碰到难以断词的问题


```
WordData["naïf", "InflectedForms", "List"]
[单词数据]

Missing[NotAvailable]
```

编辑于 2018-01-29

英语学习 美剧 Wolfram Mathematica

文章被以下专栏收录



**Mathematica 还能这样玩**  
Mathematica（mma、麦酱），宇宙第一计算姬（钦定的），投喂 CPU 时间就能把答...

已关注

▲ 赞同 284 ▼ 44 条评论 分享 ★ 收藏 ...

推荐阅读



美剧《老友记》中一共包含多少单词？

用户不存在



追了十年美剧，推荐五部给你学英语

Claire



为什么很多英语学习者都提到要看老友记？

姜小白 发表于姜小白的随...



用这个漂迅速转换

Linux...

44 条评论 切换为时间排序

写下你的评论...

精选评论 (2)

 list 2 年前

用Mathematica干这事儿是因为你最熟悉这个工具吧？否则其它随便哪个语言都比它方便吧？

3

查看回复

 LePtC (作者) 回复 list 2 年前

给出单词的所有变形肯定是 matica 最方便（一行代码），字幕过滤的话别的语言也不一定比 matica 的行数少。我尽量从善意的角度理解你发这条评论的意图，你可能是想说装有 mathematica 的人太少影响程序的传播和使用。我的回答是我欢迎你们把我的代码改写成更有利于传播的语言，我自己目前暂时没有时间做这种事

21

查看回复

评论 (44)

 wjxway 2 年前

[mathematica.stackexchange.com...](#)

1

 wjxway 2 年前

里面给出了两种种翻译的方法.....

赞

 LePtC (作者) 回复 wjxway 2 年前

第一种方法要收费？我选择谷歌娘.....

赞

 list 2 年前

用Mathematica干这事儿是因为你最熟悉这个工具吧？否则其它随便哪个语言都比它方便吧？

3

 LePtC (作者) 回复 list 2 年前

给出单词的所有变形肯定是 matica 最方便（一行代码），字幕过滤的话别的语言也不一定比 matica 的行数少。我尽量从善意的角度理解你发这条评论的意图，你可能是想说装有 mathematica 的人太少影响程序的传播和使用。我的回答是我欢迎你们把我的代码改写成更有利于传播的语言，我自己目前暂时没有时间做这种事

21

 Dan Wu 2 年前

这个stern

赞同 284

44 条评论

分享

收藏



查看全部 6 条回复



wjxway

2 年前

稍微仔细地看了下代码.....恕我直言.....很多地方写的并不mma，稍加改进效果会好很多.....  
(速度上和代码读起来容易程度上)



2



LePtC (作者) 回复 wjxway

2 年前

是，我还是小白水平 ٩ ( ̄ ̄ ) 感谢指点



赞



wjxway

2 年前

至少For和大量的CharacterCode不太应该在mma代码里出现.....显然更多的使用  
Map,StringSplit,StringCases会使代码更好。



4



大铀子

2 年前

真棒，有没有解决日语的？



赞



LePtC (作者) 回复 大铀子

2 年前

很难，一个是日语单词的变形没有现成的数据库，二个是断词不像英语那么好处理，三  
个是日语有很多含义是通过语法来表达的...



赞



薛定谔的喵 回复 大铀子

2 年前

日语找到可以分词的api，应该就可以弄了，



赞



椰叶

2 年前

厉害了，一直苦恼于这个问题却没有想到过这个方案。感谢楼主



赞



飞呀飞

2 年前

这个好！必须赞



1



SorrowCancer

2 年前

没有。【全文完】【真的没有】【逃



赞



吴月

2 年前

厉害



赞



St Jason

2 年前

谢谢，之前想过用matlab处理一些简单的字幕乱码，你这个太强大了，想问问：1.你的编辑器  
是什么啊？是用苹果系统吗？2.你是mathematica 11？



赞



LePtC (作者) 回复 St Jason

2 年前

Sublime，win10（装了个黑色主题），mma 11.0



1



薛定谔的喵

2 年前

▲ 赞同 284



44 条评论

分享

★ 收藏



公共服务可以是做一个网站，用户只需上传字幕，等待，下载字幕。（这个就看对这个需求的人多不多，人多弄比较有意义，或者是自己练手弄（我有时间可能会练一下手））。

字幕可以保留生词字幕所在的对话（这个我的repo加上了，用户可以选择保留生词还是和生词一起的对话）

日语找到可以分词的api,应该也可以弄，

👍 1

 Steven 回复 薛定谔的喵 2 年前

讲道理，这个点子真的超级赞。期待网站。//就是比较懒--

👍 赞

 LePtC (作者) 回复 薛定谔的喵 2 年前

这种服务器一般是要花钱的吧，当然如果有人愿意做确实不错

👍 赞

[查看全部 6 条回复](#)

 FunnyFanny 2 年前

想法很棒

👍 赞

 何从 2 年前

[zhuanlan.zhihu.com/p/25...](https://zhuanlan.zhihu.com/p/25...) 写了个网页版的，欢迎试用

👍 1

 酷暑一夏1 2 年前

感觉这个常用表某些词还是挺生的。。

👍 赞

 LionKiss 2 年前


小白问一下：怎么实用啊？不会弄啊

👍 赞

 方文 2 年前

好神器呀，感觉不是理科生工科生都没法学好英语了呢。

👍 赞

 王维 2 年前

厉害了，我有个组建自己生词库的工具，[vitamin/subtitle](#)，干活的几个聚聚搞搞？

👍 1

 LePtC (作者) 回复 王维 2 年前

我不会用 Python😂 如果你需要词库我倒可以帮忙

👍 赞