使用E-utilities从PubMed批量获取文献摘要

直接用E-utilities搜索

Eutils用法总结 - 宏宇 - 博客园 (cnblogs.com)

Eutils全称是The Entrez Programming Utilities (E-utilities)，是由八个服务器端程序组成的一套编程工具，它提供用于访问NCBI Entrez查询和数据库系统的稳定接口。 这八个工具包括Einfo、ESearch、EPost、ESummary、EFetch、ELink、EGQuery、ESpell。通过这些工具，你可以访问NCBI Entrez所包含的序列、三维结构、文献等所有38个数据库。

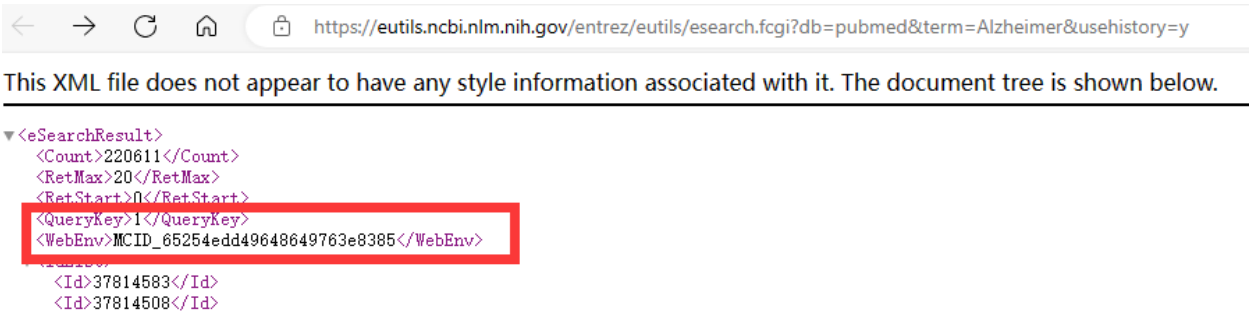
1. **EInfo**：用于获取NCBI数据库的信息，如数据库的名称、描述、记录数等。
2. **ESearch**：用于执行搜索操作，您可以使用它来搜索PubMed文献或其他NCBI数据库中的数据，并获取匹配的记录数和检索的ID列表。
3. **EFetch**：用于获取检索到的记录的详细信息，例如PubMed文献的全文或GenBank序列的序列数据。
4. **ESummary**：用于获取检索结果的摘要信息，它可以提供有关检索结果的概要数据，而无需获取完整的记录。
5. **ELink**：用于查找两个不同NCBI数据库之间的关联，例如查找与特定基因相关的PubMed文献。
6. **EGQuery**：用于执行全局查询，获取NCBI数据库中的记录计数。
7. **EPost**：允许您将一组查询结果（ID列表）提交到NCBI服务器以供后续处理，而无需立即下载。
8. **ECitMatch**：用于查找与PubMed文献相关的DOIs（数字对象标识符）。

- **Esearch示例**

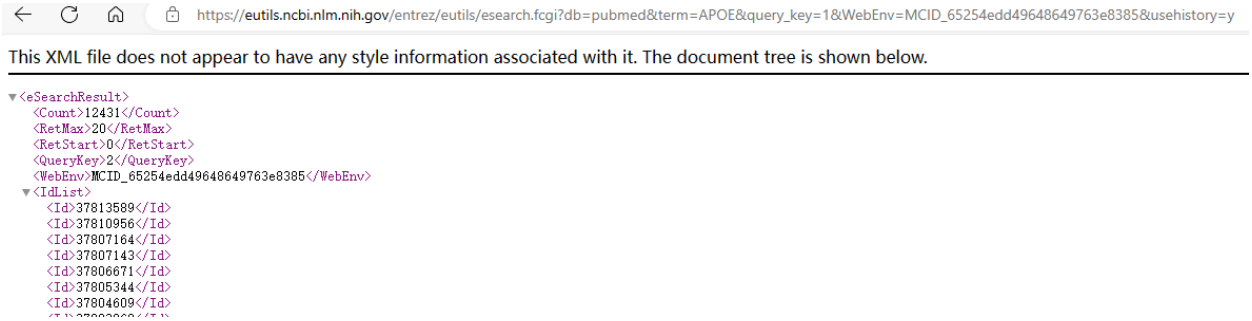
示例1：http://*eutils.ncbi.nlm.nih.gov/entrez/eutils/*esearch.fcgi?db=pubmed&term=Alzheimer

示例2：http://*eutils.ncbi.nlm.nih.gov/entrez/eutils/*esearch.fcgi?term=Alzheimer

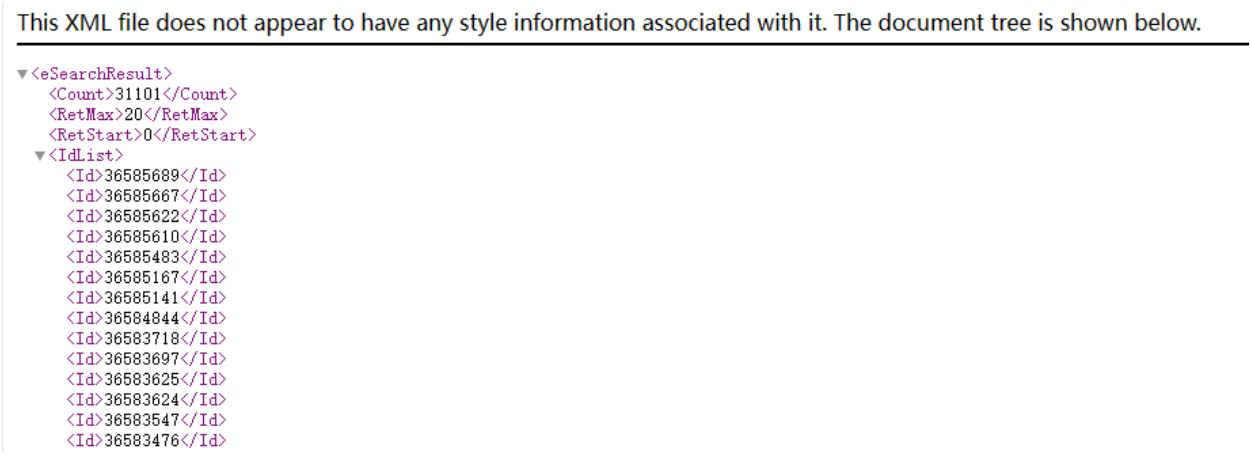
示例3：http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=Alzheimer&usehistory=y



示例4：https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=APOE&query_key=1&WebEnv=MCID_65254edd49648649763e8385&usehistory=y



示例5：https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=Alzheimer&mindate=2021&maxdate=2022



示例1传入必须参数，所查询的数据库名称db和要查询的关键词term；示例2仅传入要查询的关键词term，db的默认值是pubmed，所以示例2将查询以Alzheimer为关键词查询pubmed数据库；

示例3使用可选usehistory参数将在服务器端产生一个查询历史记录，结果中将生成WebEnv和query_key值，下一次使用ESearch、ESummary等操作可利用这些参数在这一次查询结果基础上进行操作，这也是Eutils最强大的地方，可以方便的建立自己的工作流程；

示例4中即使用示例3中的查询结果重新查询以virus为关键词的条目，这等价于将示例3中的“term=Alzheimer”替换为“term=Alzheimer+APOE”的查询结果。

Eutils Name	Entry	Required Parameters	Optional Parameters	Return Format
EInfo	einfo.fcgi		db	xml
ESearch	esearch.fcgi	db term	usehistory WebEnv query_key retstart retmax rettype field datetype reldate mindate, maxdate	xml

1. **usehistory**（可选）：该参数用于指定是否要将搜索结果保存到历史记录中，以便以后进行检索或下载。如果将其设置为 "y"，则结果将保存到历史记录中，您可以使用历史记录中的信息进行后续操作。如果设置为 "n"，则不会保存历史记录。
2. **WebEnv**（可选）：如果使用历史记录（usehistory设置为 "y"），则此参数用于指定历史记录的Web环境标识符，以便检索之前保存的搜索结果。
3. **query_key**（可选）：与WebEnv一起使用，用于指定历史记录中的搜索结果的查询键。这个键可以用来检索之前保存的搜索结果，以进行进一步的操作。
4. **retstart**（可选）：用于指定从搜索结果中返回的记录的起始位置。默认情况下，它是0，表示从搜索结果的第一条记录开始返回。
5. **retmax**（可选）：用于指定要返回的记录的最大数量。可以用来限制返回的结果数量。
6. **rettype**（可选）：用于指定返回的记录的格式类型。可以是不同的格式，例如XML、文本等，具体取决于需求。
7. **field**（可选）：用于指定在哪个字段中执行搜索操作。例如，您可以指定在标题、摘要、作者等字段中搜索。
8. **datetype**（可选）：用于指定日期搜索的类型，例如出版日期、修改日期等。
9. **reldate**（可选）：用于指定与日期相关的搜索条件，例如相对于当前日期的天数。
10. **mindate, maxdate**（可选）：用于指定日期范围搜索的开始日期和结束日期。

获取网页，利用正则表达式获得文献ID列表

curl -s "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=COVID-19&mindate=2021/01/01&maxdate=2021/01/31" | grep -oP '(?<=<Id>)[0-9]+(?=</Id>)' > curl_id.txt

```
muhaha@木哈哈のLaptop:~/NLP_class/test$ curl -s "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&te
rm=COVID-19&mindate=2021/01/01&maxdate=2021/01/31" | grep -oP '(?<=<Id>)[0-9]+(?=</Id>)' > curl_id.txt
muhaha@木哈哈のLaptop:~/NLP_class/test$ ls
alzheimer2021_1.txt  curl_id.txt  idab.txt  idlist.txt
muhaha@木哈哈のLaptop:~/NLP_class/test$ head -10 curl_id.txt
33516166
33516163
33516162
33516153
33516152
33516144
33516131
33516114
33516093
33516085
```

```
ADNI  ADNI_DNA_Methylation_iDAT_files.tar.gz  code  data1  mpi_matrix_point_to_point  mpi_matrix_point_to_point.c  NLP_class  YaWen_data  YaWen_data.zip
(base) [ywliu@localhost ywliu]$ cd NLP_class/
(base) [ywliu@localhost NLP_class]$ curl -X GET "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=COVID-19"
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE eSearchResult PUBLIC "-//NLM//DTD esearch 20060628//EN" "https://eutils.ncbi.nlm.nih.gov/eutils/dtd/20060628/esearch.dtd">
<eSearchResult><Count>390592</Count><RetMax>20</RetMax><RetStart>0</RetStart><IdList>
<Id>37814329</Id>
<Id>37814328</Id>
<Id>37814308</Id>
<Id>37814299</Id>
<Id>37814234</Id>
<Id>37814214</Id>
<Id>37814169</Id>
<Id>37814146</Id>
<Id>37814081</Id>
<Id>37814041</Id>
<Id>37813970</Id>
<Id>37813920</Id>
<Id>37813912</Id>
<Id>37813860</Id>
<Id>37813850</Id>
<Id>37813813</Id>
<Id>37813775</Id>
<Id>37813723</Id>
<Id>37813696</Id>
<Id>37813661</Id>
</IdList><TranslationSet><Translation>     <From>COVID-19</From>     <To>("COVID-19" OR "COVID-19"[MeSH Terms] OR "COVID-19 Vaccines" OR "COVID-19 Vaccines"[MeSH Terms] OR "COVID-19 serotherapy" OR "COVID-19 serotherapy"[Supplementary Concept] OR "COVID-19 Nucleic Acid Testing" OR "covid-19 nucleic acid testing"[MeSH Terms] OR "COVID-19 Serological Testing" OR "covid-19 serological testing"[MeSH Terms] OR "COVID-19 Testing" OR "covid-19 testing"[MeSH Terms] OR "SARS-CoV-2" OR "sars-cov-2"[MeSH Terms] OR "Severe Acute Respiratory Syndrome Coronavirus 2" OR "NCOV" OR "2019 NCOV" OR (("coronavirus"[MeSH Terms] OR "coronavirus" OR "COV") AND 2019/11/01[PDAT] : 3000/12/31[PDAT]))</To>     </Translation></TranslationSet><QueryTranslation>"covid 19"[All Fields] OR "covid 19"[MeSH Terms] OR "covid 19 vaccines"[All Fields] OR "covid 19 vaccines"[MeSH Terms] OR "covid 19 serotherapy"[All Fields] OR "covid 19 nucleic acid testing"[All Fields] OR "covid 19 nucleic acid testing"[MeSH Terms] OR "covid 19 serological testing"[All Fields] OR "covid 19 serological testing"[MeSH Terms] OR "covid 19 testing"[All Fields] OR "covid 19 testing"[MeSH Terms] OR "sars cov 2"[All Fields] OR "sars cov 2"[MeSH Terms] OR "severe acute respiratory syndrome coronavirus 2"[All Fields] OR "ncov"[All Fields] OR "2019 ncov"[All Fields] OR (("coronavirus"[MeSH Terms] OR "coronavirus"[All Fields] OR "cov"[All Fields]) AND 2019/11/01:3000/12/31[Date - Publication])</QueryTranslation></eSearchResult>
(base) [ywliu@localhost NLP_class]$ curl -X GET "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=COVID-19/
```

```
This XML file does not appear to have any style information associated with it. The document tree is shown below.

▼<eSearchResult>
   <ERROR>Search Backend failed: Exception: 'retstart' cannot be larger than 9998. For PubMed, ESearch can only retrieve the first 9,999 records matching the query. To obtain more than 9,999 PubMed records, consider using EDirect that contains additional logic to batch
   PubMed search results automatically so that an arbitrary number can be retrieved. For details see https://www.ncbi.nlm.nih.gov/books/NBK25499/</ERROR>
 </eSearchResult>
```

PMID数量会被限制（9999）

EDirect方法

玩转 edirect — NGS Data Analysis for Pathogens 0.0.1a 文档 (ngs-data-for-pathogen-analysis.readthedocs.io)

EDirect工具集

ESearch命令的参数

EDirect工具集中的ESearch命令用于执行PubMed文献的检索操作，并支持多个参数以定制查询。以下是ESearch命令支持的主要参数：

1. `db DATABASE`: 指定要搜索的数据库名称，例如，`pubmed` 用于 PubMed 文献，`nucleotide` 用于 GenBank 序列记录，`protein` 用于蛋白质数据库，等等。
2. `query "SEARCH_QUERY"`: 指定搜索条件，即您希望在数据库中搜索的关键词、短语或查询字符串。
3. `term "SEARCH_TERM"`: 与 `query` 类似，用于指定搜索条件。在某些情况下，与 `term` 一起使用时可以更精确地匹配特定术语。
4. `usehistory`: 启用搜索历史记录，以便稍后进行操作，如检索、汇总或下载文献记录。
5. `retmax NUMBER`: 限制返回结果的数量。使用此参数可以指定一次检索返回的最大记录数。例如，`retmax 100` 表示最多返回 100 条记录。
6. `sort SORT_ORDER`: 指定返回结果的排序顺序，可选值包括 "relevance"（相关性排序）和 "pub date"（按发布日期排序）等。
7. `mindate DATE` 和 `maxdate DATE`: 指定日期范围，以限制检索结果的发布日期。例如，`mindate 2021` 和 `maxdate 2022` 表示只搜索 2021 年到 2022 年之间的文献。
8. `datetype DATE_TYPE`: 指定日期类型，例如，`datetype PDAT` 表示使用出版日期进行检索，而 `datetype EDAT` 表示使用输入日期进行检索。

EFetch命令的参数

`efetch` 命令是 EDirect 工具集中用于检索文献记录的命令，它可以根据不同的参数设置来获取文献的详细信息。以下是一些常用的 `efetch` 参数：

1. `db DATABASE`：指定要检索的数据库，例如，`pubmed` 用于 PubMed 数据库，`nucleotide` 用于 GenBank，`protein` 用于 Protein 数据库等。
2. `id IDENTIFIERS`：指定要检索的文献或记录的标识符，可以是 PubMed ID (PMID)、GenBank Accession Number、Protein Accession Number 等。
3. `format OUTPUT_FORMAT`：指定输出的格式，您可以选择不同的格式来获取所需的信息，例如 `abstract`（文摘）、`medline`（MEDLINE 格式）、`fasta`（FASTA 格式）等。
4. `mode OUTPUT_MODE`：指定输出的模式，通常可以选择 `text`（文本输出）或 `xml`（XML 格式输出）。
5. `retmode OUTPUT_MODE`：类似于 `mode`，指定输出的模式，通常可以选择 `text`（文本输出）或 `xml`（XML 格式输出）。
6. `rettype OUTPUT_TYPE`：指定要检索的文献记录的类型，例如 `abstract`（文摘）、`full`（完整记录）、`fasta_cds_na`（基因序列）等。
7. `retstart START_INDEX`：用于指定从结果中的第几项开始返回记录。通常与 `retmax` 一起使用，以限制返回结果的数量。
8. `retmax MAX_RECORDS`：用于指定一次检索返回的最大记录数。可以限制返回的记录数量，通常与 `retstart` 一起使用。

9. `query KEYWORD`：可以用于进一步筛选文献记录，类似于 `esearch` 命令中的查询条件。
10. `extra_params`：在一些情况下，您可以使用额外的参数来进一步自定义输出格式或检索条件。

安装

sh -c "$(curl -fsSL https://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/install-edirect.sh)"

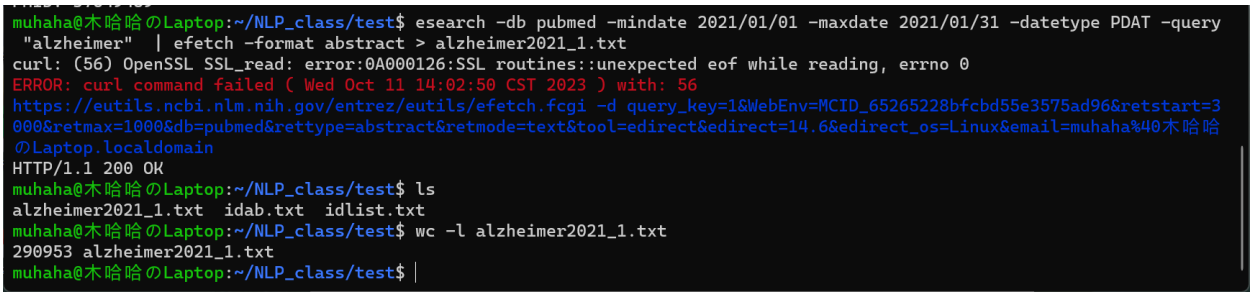
export PATH=${HOME}/edirect : ${PATH}

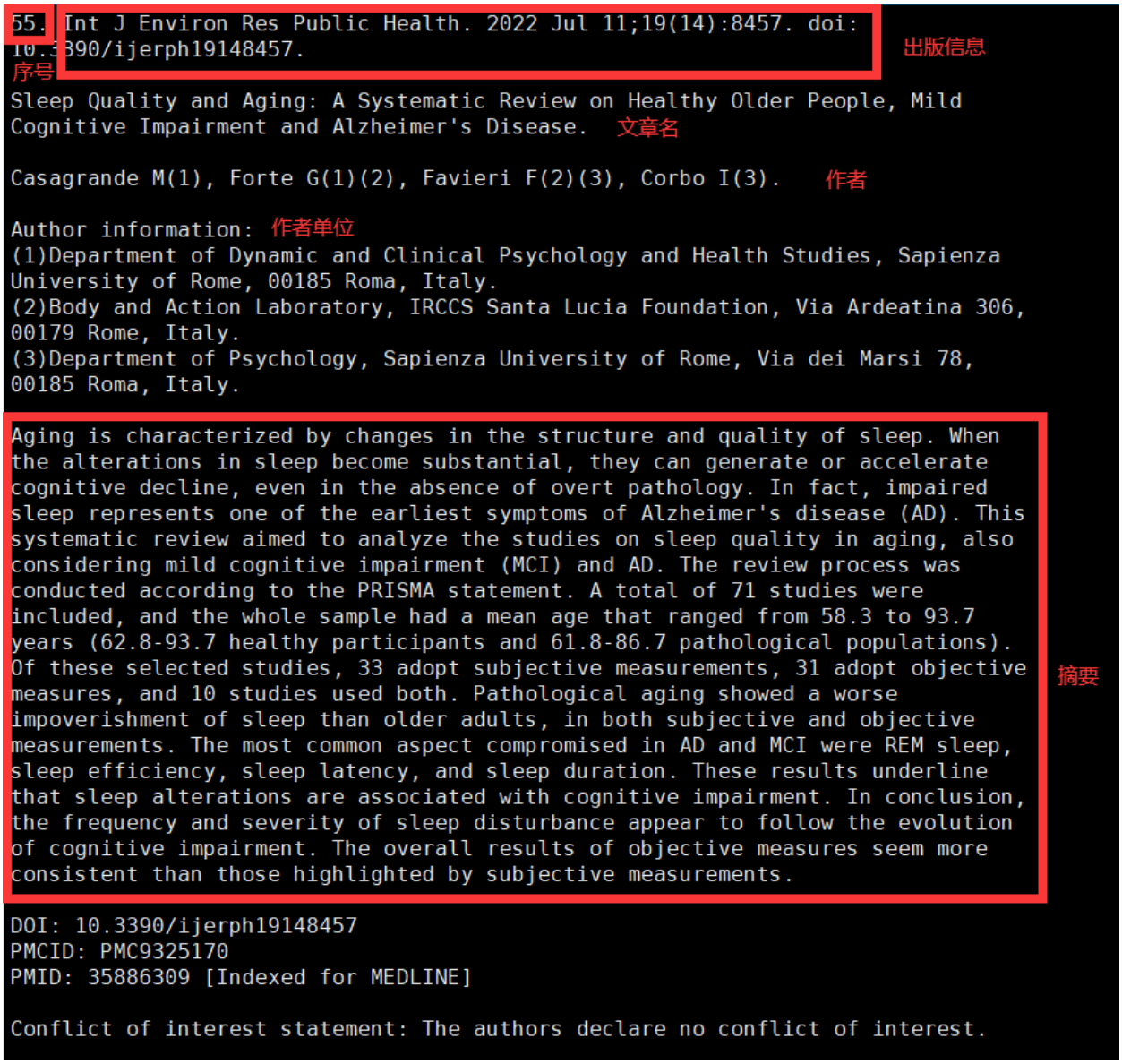
获取2021年1月出版的部分alzheimer相关文献信息

直接获得摘要

nohup esearch -db pubmed -mindate 2021/01/01 -maxdate 2021/01/31 -datetype PDAT -query "alzheimer" | efetch -format abstract > alzheimer2021_01.txt &

(直接用URL在网页上搜索获得1330篇结果，但是用该命令获得的只有684篇)





分步获得

控制每次提交给efetch的PMID数，获得摘要后sleep，防止请求过多被阻止频繁访问。

获取PMID列表（一年）

esearch -db pubmed -mindate 2021/01/01 -maxdate 2021/12/31 -datetype PDAT -query "alzheimer" | efetch -format uid > idlist.txt

根据ID获取摘要

提交一个PMID

efetch -db pubmed -id 37649489 -format abstract > id_abstract.txt

提交两个PMID

efetch -db pubmed -id "33516166,33516163" -format abstract >> id_abstract_.txt

bash脚本

```
#!/bin/bash

# 指定idlist.txt文件路径
idlist_file="idlist.txt"

# 打开文件句柄以读取文件
exec 3<"$idlist_file"

# 循环读取文件中的每个ID
while read -u 3 line; do
  # 提取前2个ID（或更多，如果需要）
  ids=($line)
  id1=${ids[0]}
  id2=${ids[1]}

  # 使用efetch获取摘要信息并保存到文件
  efetch -db pubmed -id "$id1,$id2" -format abstract >> id_abstract.txt

  # 休眠3秒
  sleep 3
done

# 关闭文件句柄
exec 3<&-

echo "摘要获取完成。"
```

提交20个PMID后获得20篇文献的摘要。但是得到的结果id_abstract.txt里每篇摘要的序号都是1，大家可以自行修改该脚本以达到想要效果。