Event Extraction and Knowledge Discovery

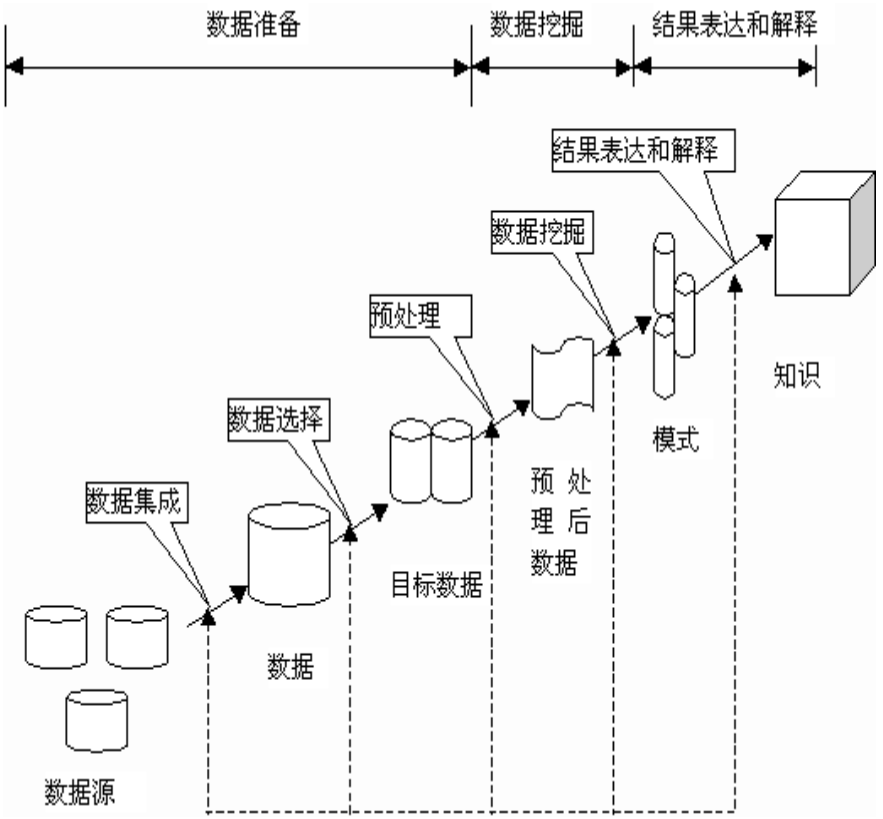
毛盛强

2017/4/27

CONTENTS

Knowledge Discovery in Database

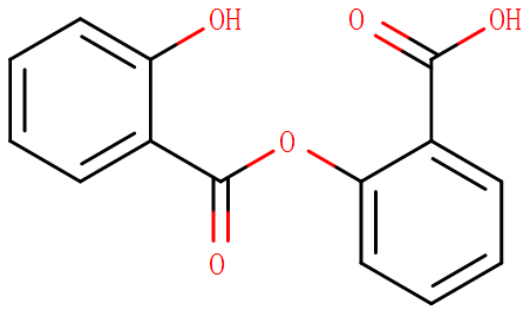
知识发现(KDD),从各种媒体表示的信息中，根据不同的需求获得知识，表示将低层数据转换为高层知识的整个过程。知识发现的目的是向使用者屏蔽原始数据的繁琐细节，从原始数据中确定数据中有效的、新颖的、潜在有用的信息直接向使用者报告。

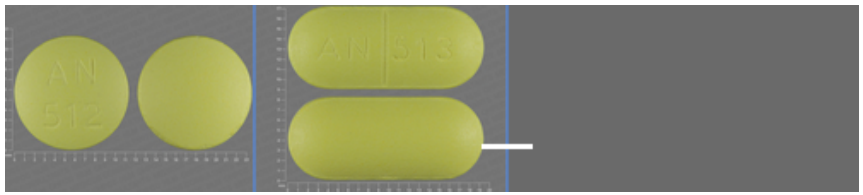


知识发现与数据挖掘的联系

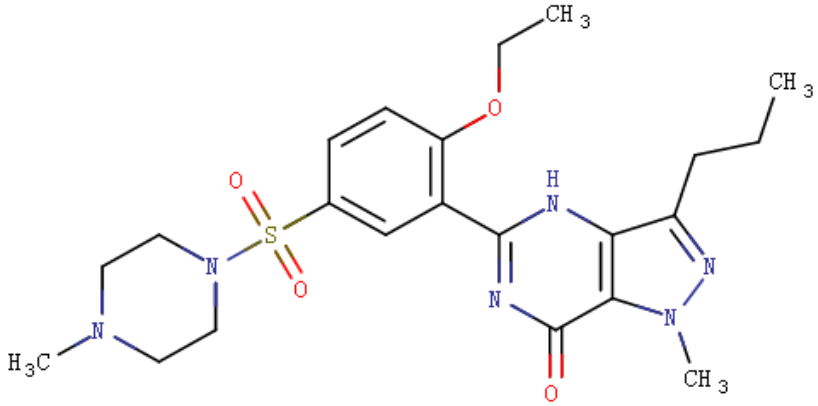
数据挖掘可认为是观察数据中模式或模型的抽取，这是对数据挖掘的一般解释。虽然数据挖掘是知识发现过程的核心，但它通常仅占KDD的一部分(大约是15% 到25%) 。因此数据挖掘仅仅是整个KDD过程的一个步骤。

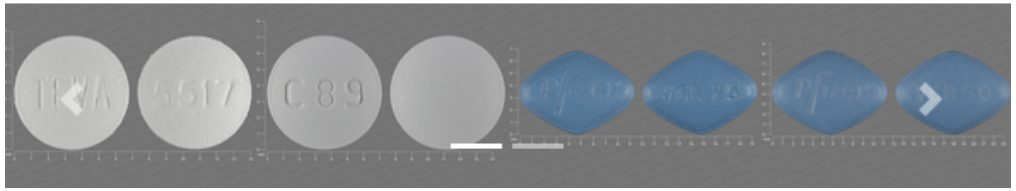
Target drug





Salsalate

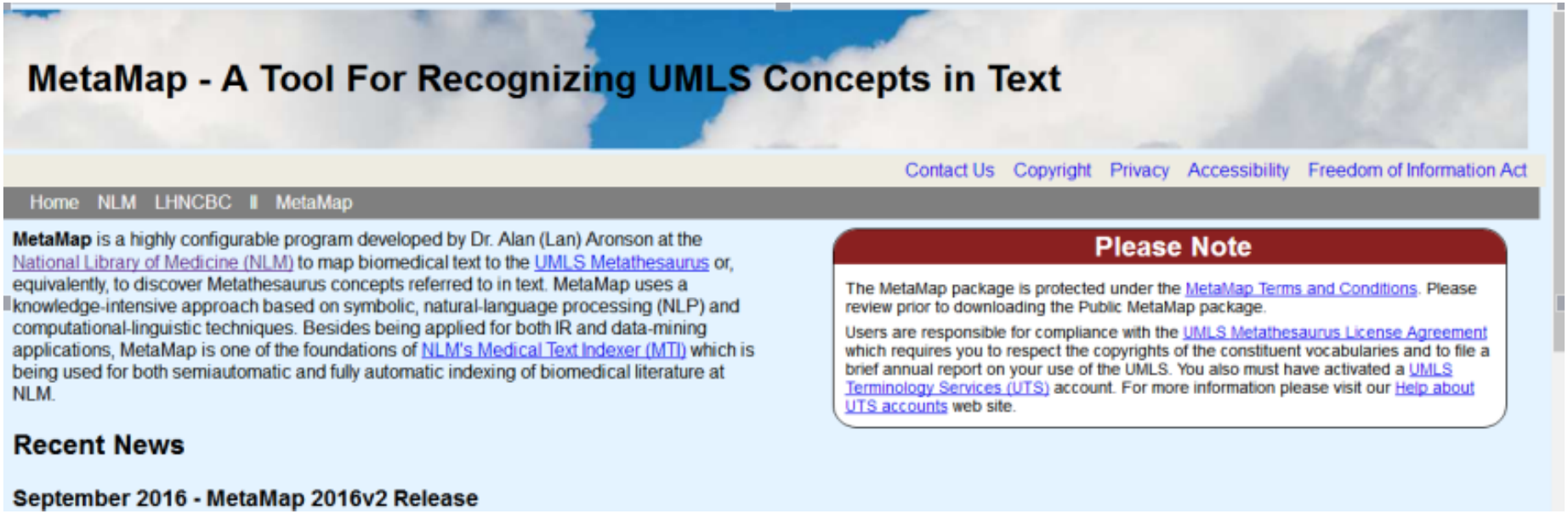




Sildenafil

非甾体类抗炎药，作为抗炎和抗风湿剂的作用方式可能是由于抑制前列腺素的合成和释放。

Metamap



MetaMap是由美国国立医学图书馆开发的高度可配置程序，用于将生物医学文本映射到UMLS（一体化医学语言系统），或者等效地发现文本中涉及到的超级叙词表概念。

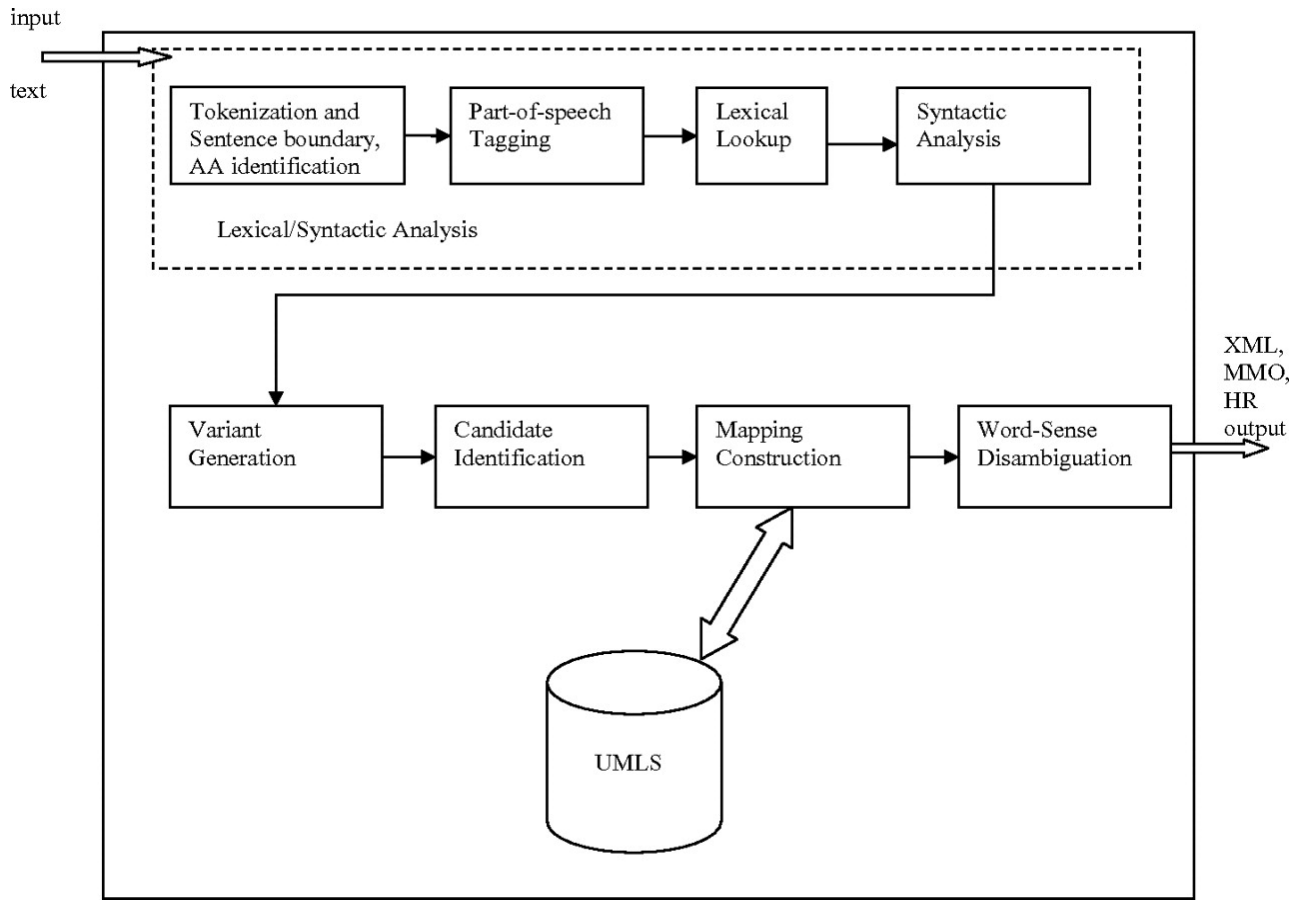
实现过程

语句切分

词性标注

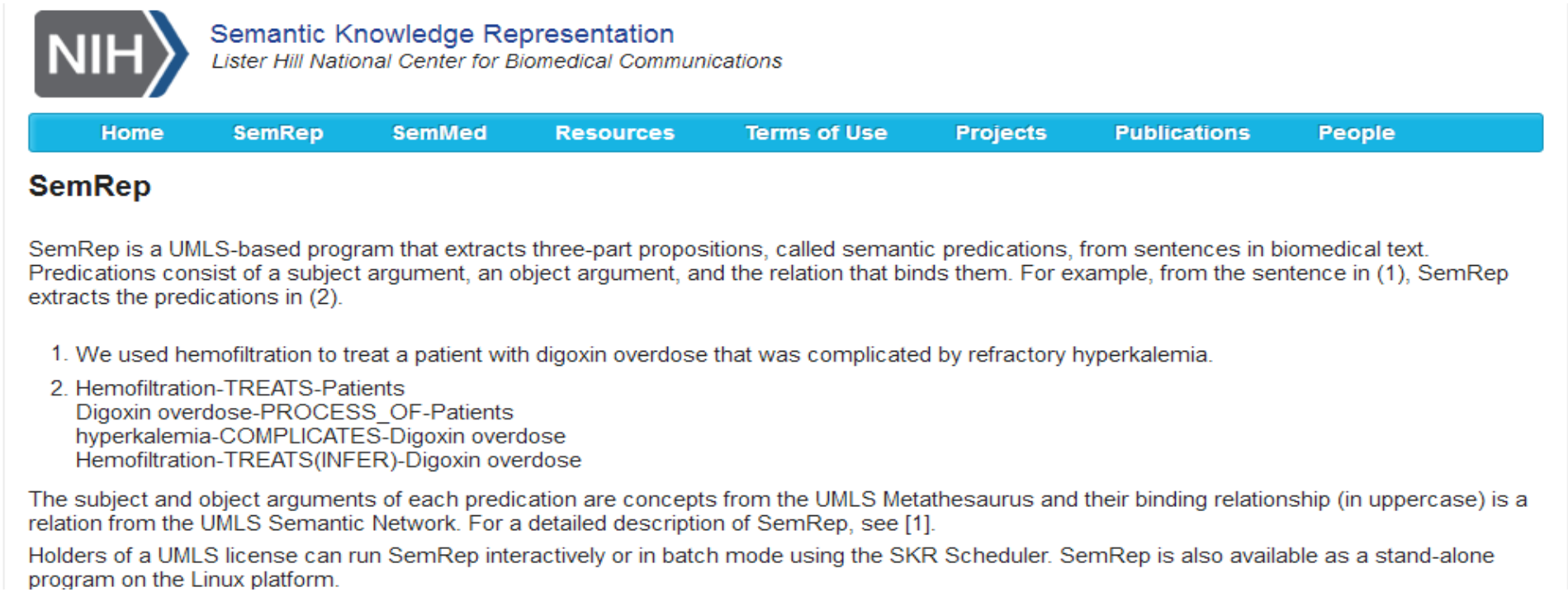
词法查找

语法分析



Journal of the AMAIA，2010,Alan R Aronson

SemRep



Semantic Knowledge Representation
Lister Hill National Center for Biomedical Communications

Home | SemRep | SemMed | Resources | Terms of Use | Projects | Publications | People

SemRep

SemRep is a UMLS-based program that extracts three-part propositions, called semantic predications, from sentences in biomedical text. Predications consist of a subject argument, an object argument, and the relation that binds them. For example, from the sentence in (1), SemRep extracts the predications in (2).

1. We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia.
2. Hemofiltration-TREATS-Patients
   Digoxin overdose-PROCESS_OF-Patients
   hyperkalemia-COMPLICATES-Digoxin overdose
   Hemofiltration-TREATS(INFER)-Digoxin overdose

The subject and object arguments of each predication are concepts from the UMLS Metathesaurus and their binding relationship (in uppercase) is a relation from the UMLS Semantic Network. For a detailed description of SemRep, see [1].

Holders of a UMLS license can run SemRep interactively or in batch mode using the SKR Scheduler. SemRep is also available as a stand-alone program on the Linux platform.

SemRep是从生物医学文本中提取语义预测（主题-关系-对象三元组）的程序。

Modules:

信息提取

统计模块

查询模块

推导模块

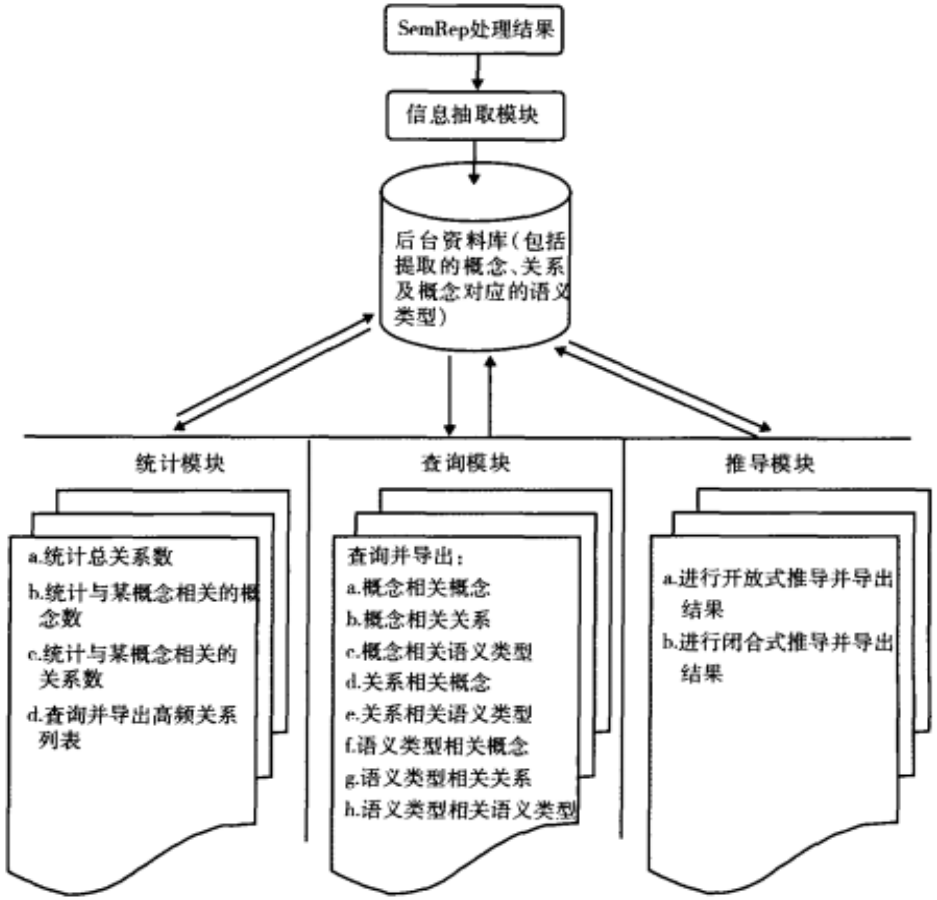


图 1　系统模块

阎雷　*Journal of medicine*　2013

文本预处理

Please see adverse events module for hyperglycemia.
This aim was proposed for the TINSAL-T2D stage 2 trial and is separately reported.
This aim was proposed for the TINSAL-T2D stage 2 trial and is separately reported.
This aim was proposed for the TINSAL-T2D stage 2 trial and is separately reported.
This aim was proposed for the TINSAL-T2D stage 2 trial and is separately reported.

均存在大量的重复文本

Oxygen consumption measurements were taken at peak exercise. Subjects were exercised to maximal volition with an electronically braked cycle ergometer. The protocol consisted of 3 minutes of pedaling in an unloaded state followed by a ramp increase in work rate (watts) to maximal exercise. Metabolic and ventilatory data were obtained throughout the exercise study and for the first 2 minutes of recovery on a breath-by-breath basis with a metabolic cart.
Oxygen consumption measurements were taken at peak exercise. Subjects were exercised to maximal volition with an electronically braked cycle ergometer. The protocol consisted of 3 minutes of pedaling in an unloaded state followed by a ramp increase in work rate (watts) to maximal exercise. Metabolic and ventilatory data were

去重复文本

```python
#!/usr/bin/python
#  -*- coding: utf-8 -*-
def removal (path1):
    import os
    f1=os.listdir(path1)
    for filename in f1:
        file1=open(path1+filename,"r").readlines()
        f1=list(set(file1))
        f1.sort(key=file1.index)
        file2 = open(path1+filename, "w")
        for i in f1:
            file2.write(str(i))
        file2.close()
list_dir_path="C:/Users/Administrator/Desktop/mm/"
removal(list_dir_path)
```

wc结果：行数单词数 字节数 文件名

```
[csxie@localhost test]$ wc viagra.txt
  1207   52059 331059 viagra.txt
[csxie@localhost test]$ wc viagra_del.txt
   962   20135 130395 viagra_del.txt

[csxie@localhost test]$ wc salsalate.txt
  116  5530 34308 salsalate.txt
[csxie@localhost test]$ wc salsalate_del.txt
   68  3593 22363 salsalate_del.txt
```

ERROR

```
Verifying Batch Files: Done
      -- Config File: OK
      -- Text File Exists & Size > 0: OK
      -- Text File Validity Test: FAILED
      -- Single Item Too Large Test: OK
```

Your Batch Job Request contains an erro
Please review the er

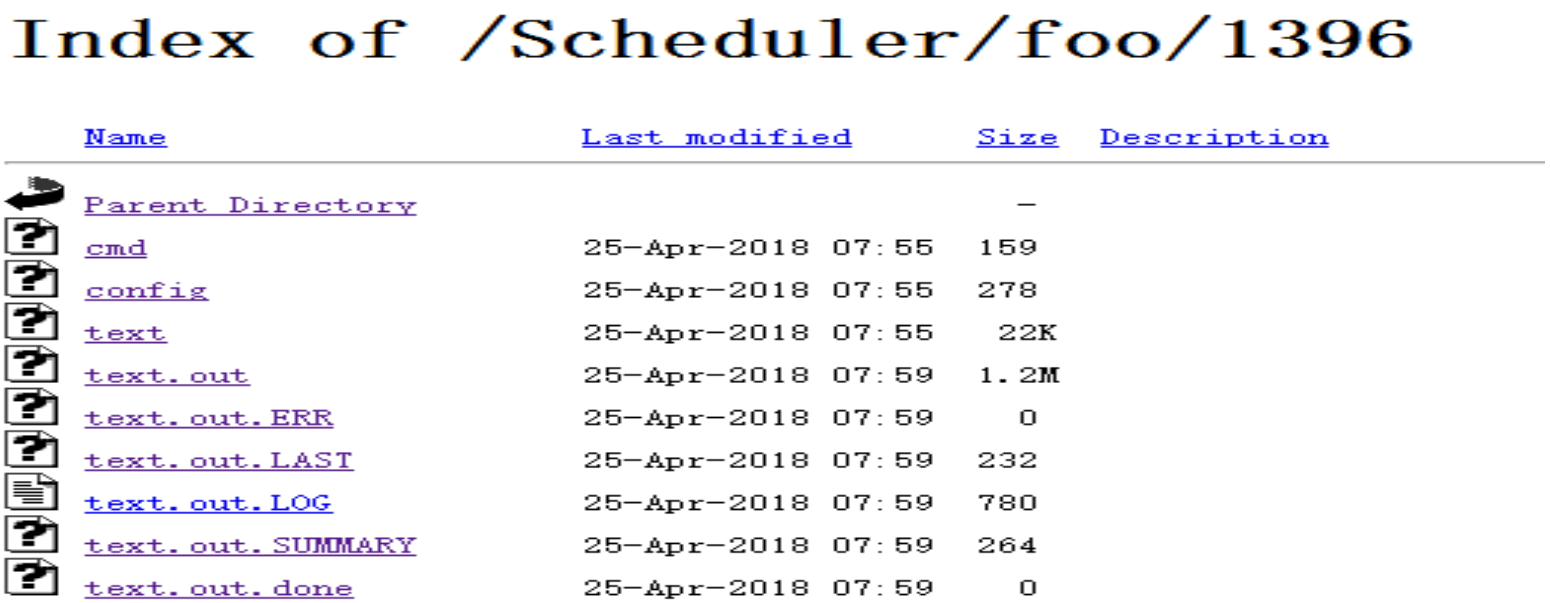
Messages:

```
ERROR: Non-ascii character (£) found on line 7
```

perl -i.bk -pe 's/[^[:ascii:]]//g;' salsalate.txt

```
Verifying Batch Files: Done
      -- Config File: OK
      -- Text File Exists & Size > 0: OK
      -- Text File Validity Test: PASSED
      -- Single Item Too Large Test: OK
```

结果



Index of /Scheduler/foo/1396

Name	Last modified	Size	Description
Parent Directory		-	
cmd	25-Apr-2018 07:55	159	
config	25-Apr-2018 07:55	278	
text	25-Apr-2018 07:55	22K	
text.out	25-Apr-2018 07:59	1.2M	
text.out.ERR	25-Apr-2018 07:59	0	
text.out.LAST	25-Apr-2018 07:59	232	
text.out.LOG	25-Apr-2018 07:59	780	
text.out.SUMMARY	25-Apr-2018 07:59	264	
text.out.done	25-Apr-2018 07:59	0	

查看text.out文件，进行信息提取

awk '/Pathologic Function/ {print}' < salsalate_metamap_out.txt > result.txt

sort result.txt | uniq > result_uniq.txt

Result

rse event module for details HOMA-IR >>>>> Phrase adverse event modu
rse events module for hyperglycemia >>>>> Phrase adverse events module
r reactive hyperemia Reactive hyperemia >>>>> Phrase reactive hyperer
eater rate of insulin resistance Change >>>>> Phrase a greater rate
gh-resolution vascular ultrasound with a 10-MHz linear array transduce
ges in insulin sensitivity >>>>> Phrase changes in insulin sensitivity
ges in markers of inflammation >>>>> Phrase changes in markers of infl
>>>>> Phrase flow mediated dilation <<<<< Phrase >>>>> Mappings Meta
bit >>>>> Phrase inhibit <<<<< Phrase >>>>> Mappings Meta Mapping (94
ers of inflammation >>>>> Phrase markers of inflammation <<<<< Phrase
ured by flow-mediated dilation >>>>> Phrase measured by flow mediated
uding >>>>> Phrase occluding <<<<< Phrase >>>>> Mappings Meta Mapping
rted on the Side Effect Checklist >>>>> Phrase reported on the side e
brachial artery with a blood pressure cuff >>>>> Phrase the brachial
pro-atherogenic inflammatory mediators >>>>> Phrase pro atherogenic i
a manual cuff prior to >>>>> Phrase with a manual cuff prior to <<<<
side effects >>>>> Phrase side effects <<<<< Phrase >>>>> Mappings Me

SemRep **Sildenafil**

Side effects (Adverse effects) [Pathologic Function]
Dilation (Pathological Dilatation) [Pathologic Function]
Dilation (Pathological Dilatation) [Pathologic Function]
Atherogenesis [Pathologic Function]
Cuff (Cuffing (morphologic abnormality)) [Pathologic Function]
Vascular dilation (Vasodilation disorder) [Pathologic Function
Medication side effects (Adverse reaction to drug) [Pathologic
ADVERSE EVENT (Adverse event) [Pathologic Function]
INSULIN RESISTANCE (Insulin Resistance) [Pathologic Function]
INFLAMMATION (Inflammation) [Pathologic Function]
Adverse events (Adverse event) [Pathologic Function]
Insulin Sensitivity [Pathologic Function]
Reactive Hyperemia [Pathologic Function]
Dilation (Pathological Dilatation) [Pathologic Function]
Cuff (Cuffing (morphologic abnormality)) [Pathologic Function]
Dilation (Pathological Dilatation) [Pathologic Function]
INFLAMMATION (Inflammation) [Pathologic Function]
Side Effect (Adverse effects) [Pathologic Function]
Dilation (Pathological Dilatation) [Pathologic Function]

Metamap **Salsalate**

N Hypertension, pulmonary (Pulmonary Hypertension) [Pathologic Function]
N Impairment (Impaired health) [Pathologic Function]
Transformed (metaplastic cell transformation) [Pathologic Function]
Transforming (metaplastic cell transformation) [Pathologic Function]
Dysfunction (Functional disorder) [Pathologic Function]
N Dysfunction (Functional disorder) [Pathologic Function]
Contraction (Contraction (finding)) [Pathologic Function]
HYPERTROPHY (Hypertrophy) [Pathologic Function]
Torsion (Torsion (malposition)) [Pathologic Function]
Transformed (metaplastic cell transformation) [Pathologic Function]
Hypertension, pulmonary (Pulmonary Hypertension) [Pathologic Function]
Contraction (Contraction (finding)) [Pathologic Function]
Ischaemia, NOS (Ischemia) [Pathologic Function]
Hypertension, pulmonary (Pulmonary Hypertension) [Pathologic Function]
Hypertension, pulmonary (Pulmonary Hypertension) [Pathologic Function]
Erectile dysfunction (Erectile Dysfunction Adverse Event) [Pathologic Functic
Hypertension, pulmonary (Pulmonary Hypertension) [Pathologic Function]
N Hypertension, pulmonary (Pulmonary Hypertension) [Pathologic Function]
adverse events (Adverse event) [Pathologic Function]
Erectile dysfunction (Erectile Dysfunction Adverse Event) [Pathologic Functic
Disease Progression [Pathologic Function]

Metamap **Sildenafil**

Side-effect

Salsalate		Viagra	
Metamap	SemRep	Metamap	SemRep
提取文本 释义	提取文本 释义	提取文本 释义	提取文本 释义
Blood disorders 血流阻碍	hyperglycemia 高血糖症	dysfunction 功能障碍	ERROR
insulin resistance 耐受性变敏感	hyperemia 反应性充血	allergy 过敏	
inflation 膨大	insulin resistance 耐受性变敏感	hypertrophy 肥大	
atherogeniC 动脉粥样化	Cardiac Arrest 心脏骤停	anabrosis 溃疡	
hemangiectasis 血管扩张	atherogenic 动脉粥硬化	ischemic necrosis （手指）缺血性坏死	

谢 谢