◁ BIT ▷

新词发现

答辩人：张阔 董文轩 唐卓然 牛子儒

指导教师：张华平

时间：2022/03/22

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

德以明理　学以精工

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目 录

CONTENTS

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

PART ONE

概述

新词发现是 NLP 的基础任务之一，通过对已有语料进行挖掘，从中识别出新词。新词发现也可称为未登录词识别，但未登录词不一定是新词。严格来讲，新词是指随时代发展而新出现或旧词新用的词语。而未登录词是指分词处理系统无法识别的词汇或已有词库中未出现过的词汇。

“发现”：

依据某种手段或方法，从文本中挖掘词语，组成新词表；

“新词”：

借助挖掘得到的新词表，和之前已有的旧词表进行比对，不在旧词表中的词语即可认为是新词。

新词发现的发展历史

孕育初期

出现了少量研究新词新意现象和新词术语等方面的研究学者。

20世纪80年代之前

发展期

学者们对汉语新词语进行了较多的研究，这种研究呈现出了多方位，多角度，多层次和立体化的趋势。

20世纪80年代

增长期

在吕叔湘先生的呼吁下，大部分人开始关注新词并对新词的产生、构成语义等方面进行了研究。

20世纪90年代

猛烈增长期

新词被更多人接纳，并随着计算机算力迅猛发展，有监督的机器学习开始兴起，研究视角、方法、技术手段也不断丰富。

21世纪初

持续增长期

进入互联网时代，尤其是移动互联网时代，随着互联网，手机，人工智能等技术的发展，新词发现也在不断完善。

今天

分类方法

1.新造词语，如“打假”、“扶贫”等。
2.旧词新用，如“下课”、“高就”等。
3.方言词汇，如“套磁”、“搞笑”等。
4.外来词，如“的士”、“欧佩克”等。
5.简略词，如“博导”、“超市”等。
6.修辞用法稳定下来构成的新词语，如“豆腐渣工程”、“空姐”等。
7.专有术语，如“黄牌”、“套牢”等。
8.字母词，如“TOFEL”、“CEO”等。

2006年崔世起（中文新词检测与分析）通过对大量的语料分析，得到如下新词构词模式：

模式	1+1、1+1+1和1+1+1+1	2+1和3+1	其他
比例	61.4%	31.2%	7.4%

新词发现的难点

1.由于新词定义的模糊性，构词灵活多变，导致很难找到统一且通用的方法进行识别工作。

2.由于数据稀疏，专业词汇和圈内词汇的低频新词流通性不强，导致新词识别难度增大。

3.无论是从词典参照角度还是时间参照角度来看，都很难发现新词。

4.新词尤其是非命名实体，在构成方面没有普遍的规律，无法直接预测。

1.基于统计

2.基于规则

3.基于统计和规则

4.基于深度学习

新词发现的出现是为了解决什么问题？

新词发现的意义

机器翻译

提高新词，旧词新意翻译的准确率，增强翻译效果。

挖掘实体

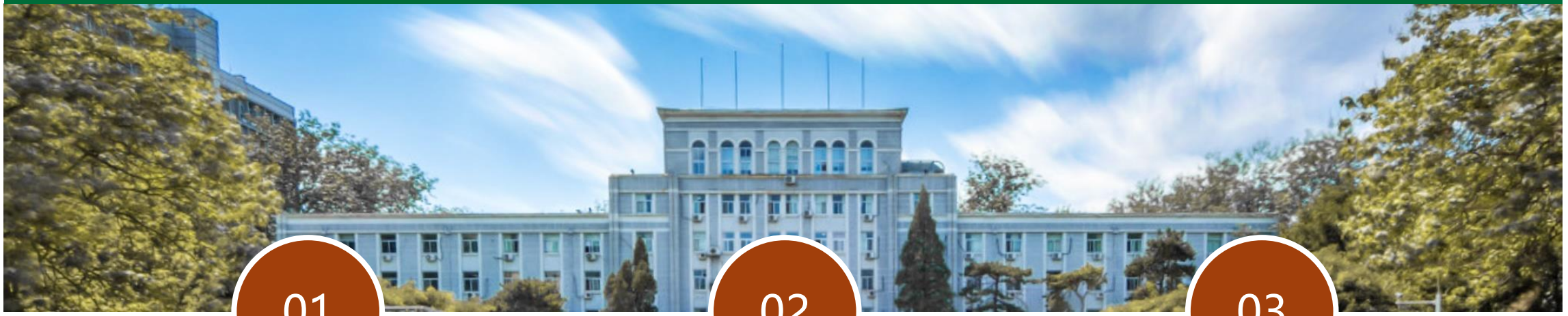
从文本语料库挖掘尽可能多的高质量词汇。

构建词库

运用无监督方法把所有抽取得到的词和已有的词库进行比较，就能得到新词。

分词

有效识别新词，提高分词准确率。



01

精确率

新词发现结果中正确识别的词数占识别为新词总数的比例。

02

召回率

新词发现结果中正确识别的新词数占实际新词数的比例。

03

F数

精确率和召回率的调和值，即为综合考虑的效果。

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

PART TWO

新词发现方法

根据语言学构词规则，比如中文里名词、动词、形容词的组合规律建立专用的规则库，再根据规则库从文章里抽取新词。

- 优势：针对性强，新词发现的效果好。
- 劣势：规则库的建立与领域或应用耦合较深，通用性差，维护耗费大量人力。

计算大量语料数据，根据词的共同特征或前后词语关系，利用相关度、左右熵等计算策略识别新词

- 优势：通用性强，不受领域限制，易于移植和维护。
- 劣势：效率不是很高，复杂度和准确率主要取决于算法的优劣。

CRF分词

CRF模型把分词任务当成一个序列标注问题

例：我爱北京天安门

CRF标注后：我/S　爱/S　北/B　京/E　天/B　安/M　门/E

分词结果：我/爱/北京/天安门

优点：CRF模型的分词考虑了上下文语境，因此对于未登录词的识别也有很好的效果

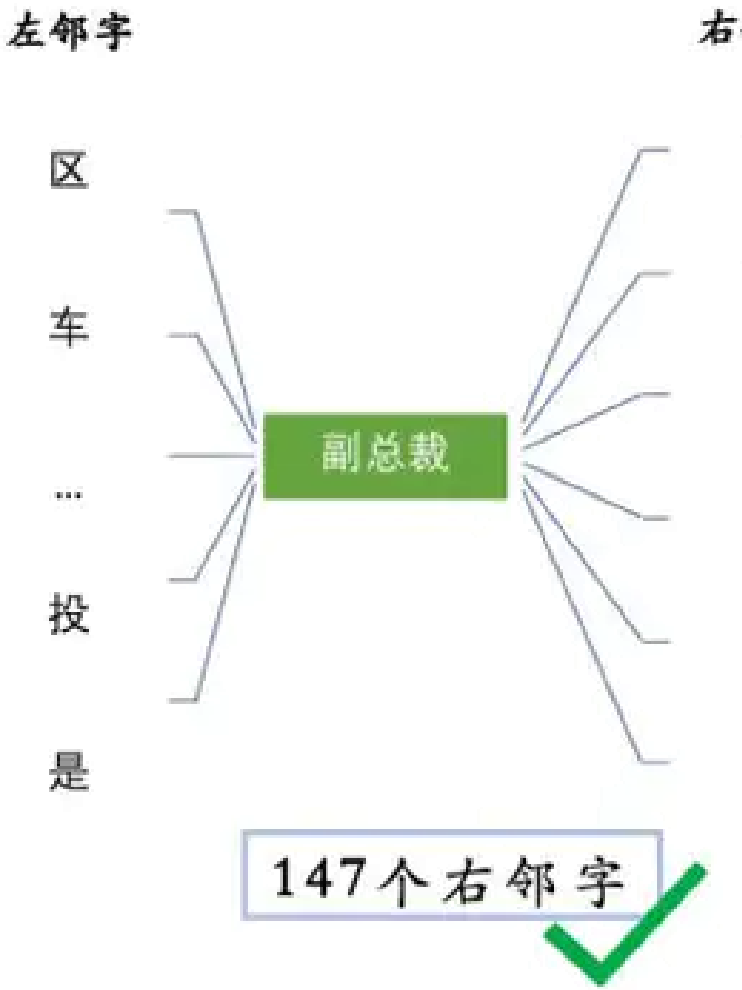
- 基于统计和规则相结合的方法融合了统计学的特征计算和规则准确率高的优点来提高系统的性能，是近几年主流的方法。
- 基于深度学习的方法通用性增强，对于低频新词的识别率更好，利用上下文信息的能力更强。

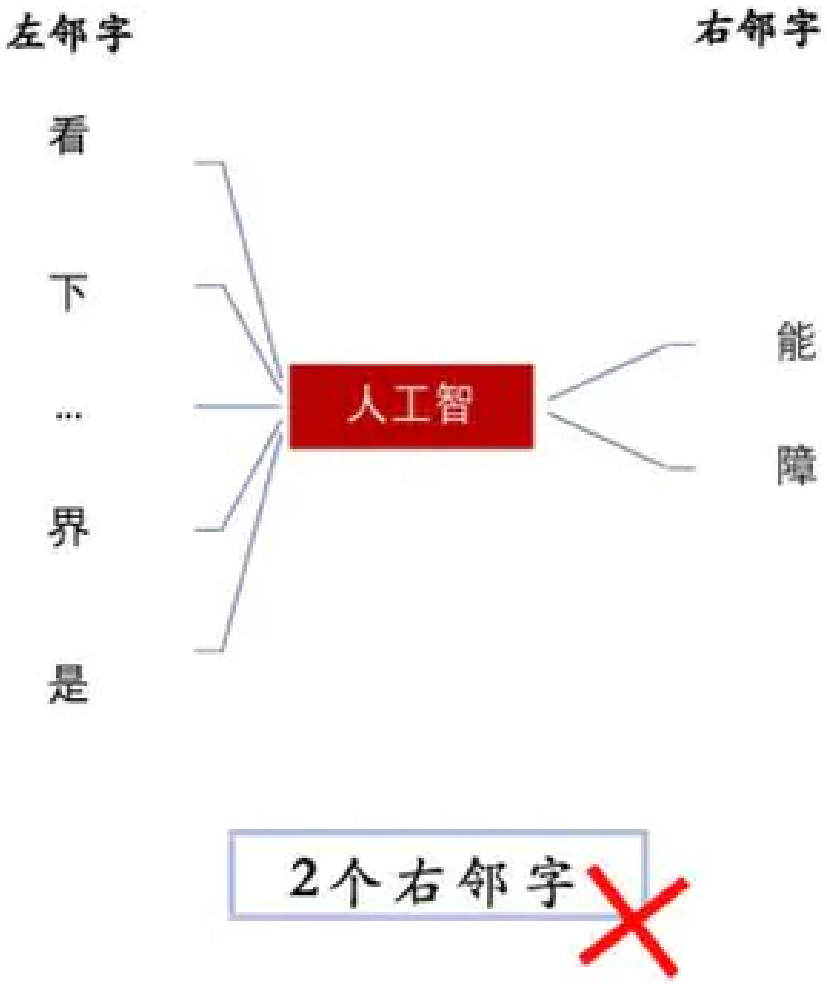
左邻接熵 $H_{left}(W) = -\sum_{w_{left} \in C_{left}} p(w_{left}|w) \log_2 p(w_{left}|w)$

$w_{left} \in C_{left}$：候选词左侧词的集合

$p(w_{left}|w)$：候选词为w条件下，左侧词的w_{left}的条件概率

- 熵值越大，表示当前字能够与其他字组成的词越多，它本身是个完整词的概率就越大。


左邻字
右邻字
区
车
…
投
是
副总裁
张
王
周
…
说
曾
147个右邻字


左邻字
右邻字
看
下
…
界
是
人工智
能
障
2个右邻字

- 互信息度量两个对象之间的相互性。字与字之间的相关性越大，字与字能组成词的概率也就越大。

$$I(X;Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$

x,y共同出现的概率

x,y：语料库中的字样本

x,y单独出现的概率

缺陷：互信息处理多字词效果不好，而新词发现任务中有大量的三字或四字词需要识别。

基于二元语法的互信息改进算法

语言模型：把句子表示成单词列表，用以估计整个句子出现的概率

$$p(w)=p(w_1w_2\cdots w_k) = p(w_1|w_0) \times p(w_2|w_0w_1) \times \cdots \times p(w_{k+1} | w_0w_1w_2\cdots w_k)$$

二元语法模型

$$p(w)=p(w_1w_2\cdots w_k)=p(w_1|w_0)\times p(w_2|w_1)\times\cdots\times p(w_{k+1}|w_k)$$

$$\mathrm{M}(\mathrm{w}_1,\ldots,\mathrm{wn}) = \log \frac{P(w_1, w_2, \ldots, wn)}{P(w_1)\, P(w_2|w_1)\ \ldots P(wn|wn_{-1})}$$

$$\mathrm{M}(\mathrm{w}) = \min(\mathrm{Mw}_{1,\ \mathrm{w2},\ldots\ \mathrm{wn}})$$

张华平1,2, 商建云3. 面向社会媒体的开放领域新词发现[J]. 中文信息学报, 2017, (3):55-61.

- 文章采用了CRF模型进行分词，通过对命名实体过滤，对词频、左右熵、互信息低于阈值的候选词过滤来实现新词发现。

- 文章亮点有：

1、对命名实体进行过滤，非常严谨。

2、CRF模型对分词的优越性

3、弥补了互信息的缺陷

4、复杂度低

1、语料库预处理

2、用CRF模型分词并统计词频，过滤词频低于阈值的候选词

3、过滤命名实体

4、用二元语法模型对候选词进行再次切分，同时计算互信息和左右熵

5、过滤互信息和左右熵低于阈值的词

6、计算剩下候选词的权重，取权重值前30%-40%的词作为新词。

$$\mathrm{Weight}_w = \alpha \frac{H_{\mathrm{ADJ}}(w)}{H_{\mathrm{ADJ}}^{\max}} + (1-\alpha)\frac{M(w)}{M^{\max}}$$

算法评价

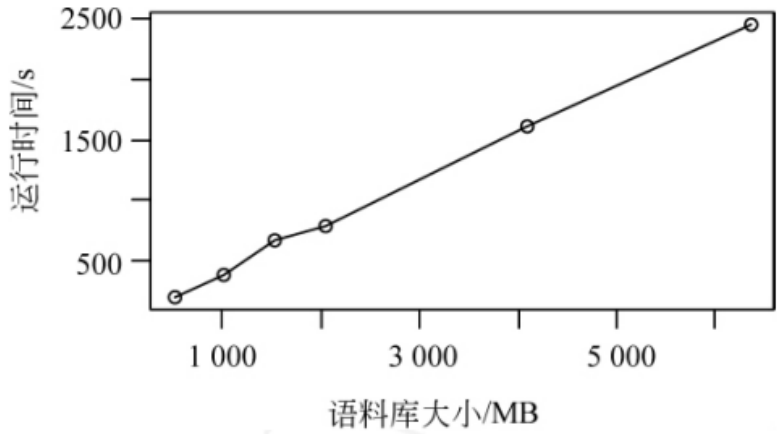


图 1　语料库大小与运行时间的关系

表 4　不同算法的召回以及精度

算法	召回数量	$P^1_{TOP1000}$	$P^2_{TOP1000}$
Peng04	5 062	0.63	0.60
文献[16]	2 779	0.72	0.76
本文	2 463	0.81	0.80

优势：

1、算法及实验严谨性好，包括对候选词过滤的严谨性和对参数不同取值的严谨性。

2、时间复杂度为O(n)，处理大批量数据具有很大优越性

3、经过多次实验和对比，文章中的算法具有很高的召回和精度。

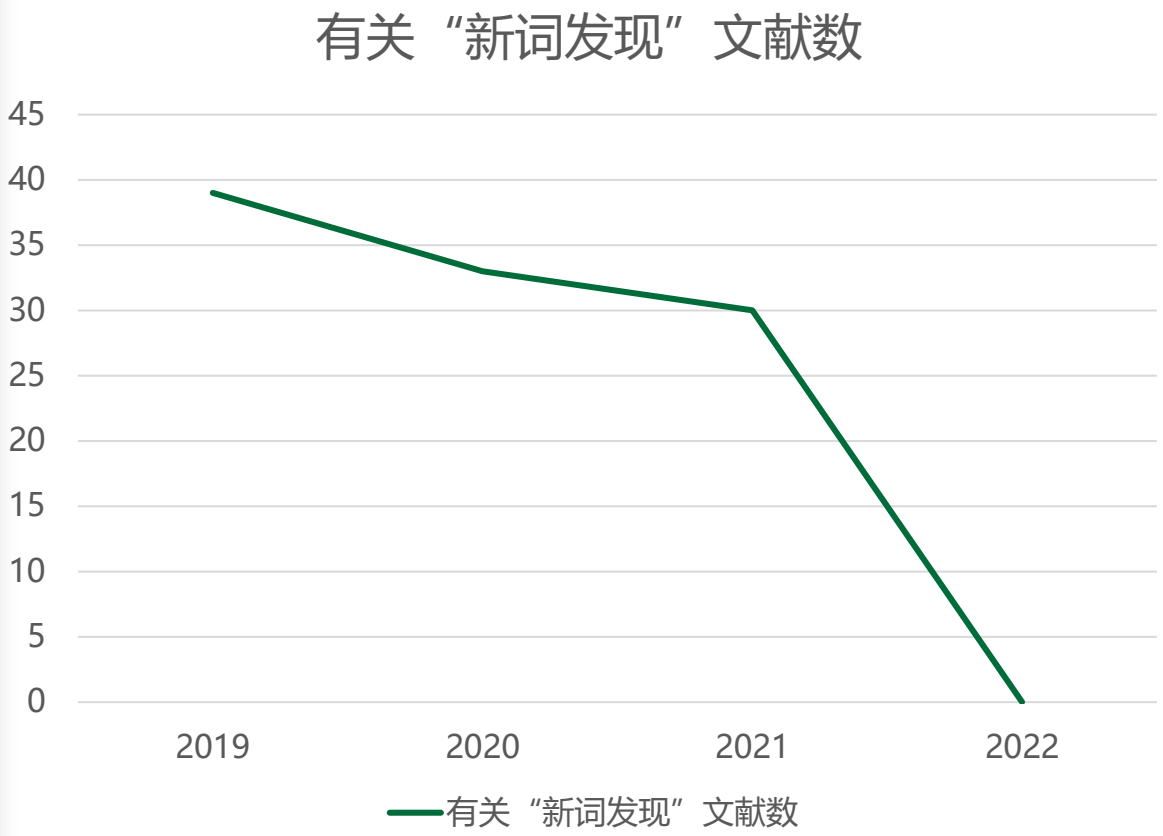
北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

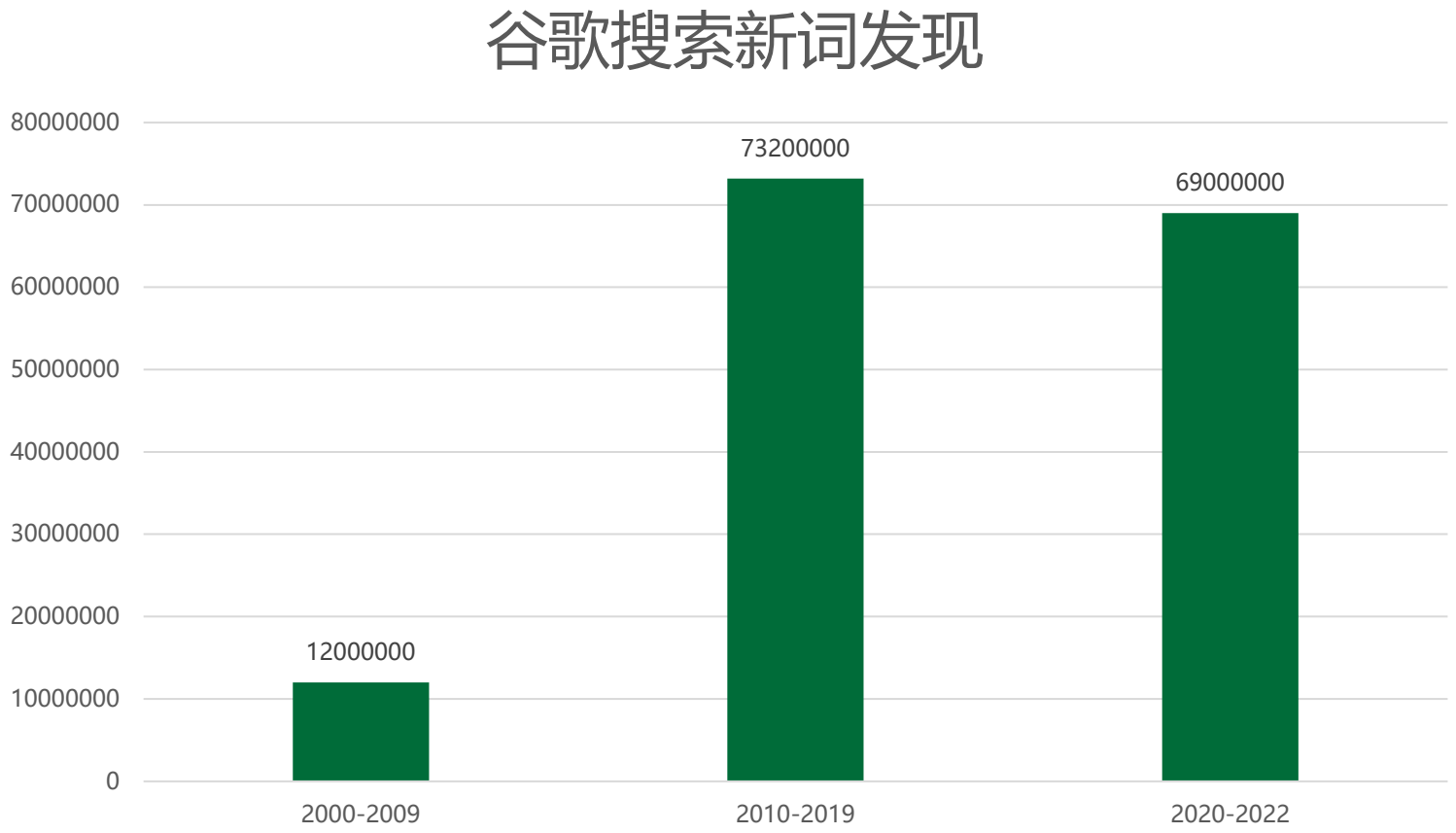
PART THREE

现状和前沿研究

截至2022年3月19日，在知网以“新词发现”为关键词搜索，总找到298条结果。

在2019至2021这三年，分别有39、33、30条结果。




谷歌搜索新词发现
80000000
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
12000000
73200000
69000000
2000-2009
2010-2019
2020-2022

将分词和新词发现相结合

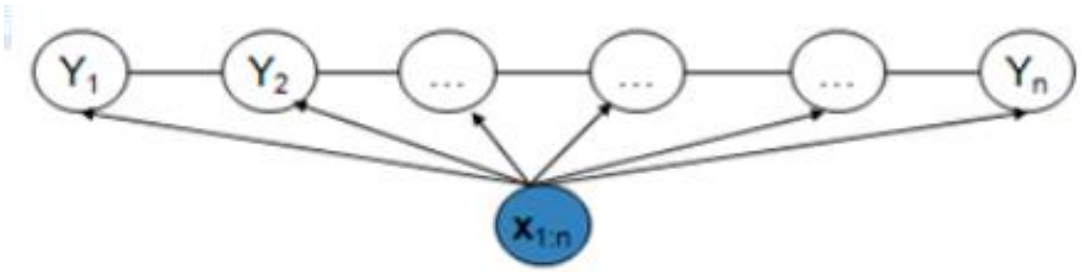
使用专家总结的语法规则和相关知识来匹配新词

将用户行为数据考虑在内来检测新词

使用统计特征来发现新词

1.将分词和新词发现相结合

使用CRF模型进行分词任务，将置信度高但不被认为是词语的词作为新词加入词典。



2.使用专家总结的语法规则和相关知识来匹配新词

- 根据词语原子的构词法将一个或多个词语子组合到一起来检测未知词语的模型。

- 将一些种子词语放入字典中来总结这些中自此的词法模板，通过模板匹配文本，在迭代的过程中将新词加入到字典中。

3.将用户行为数据考虑在内来检测新词

根据特定领域经常使用的术语，提取领域新词。

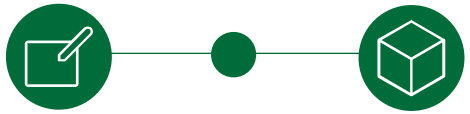
4.使用统计特征来发现新词

- 利用改进的点互信息(PMI)结合一些基本规则来识别网络新词的无监督方法。
- 利用增强互信息(EMI)算法对旅游领域的候选新词进行了过滤。
- 提出了一种基于领域词字典模型(D-WDM)的领域新词提取方法，该方法采用了领域词字典模型(DWDM)来代替传统的 WDM 模型，并利用一个领域得分函数来区分一个词语是来自于普通词字典还是领域词字典。

4.使用统计特征来发现新词

- 使用 SVM 来提取新词。
- 通过使用 LSTM 网络来提取古汉语中的新词。
- 使用 TF-IDF 算法结合关键词抽取方式来提取新词的方法。
- 基于改进的 PMI 算法和熵相结合的无监督 OOV 识别方法来完成新词发现任务。

针对旅游领域

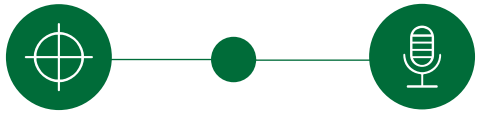


研究古汉语

SVM

LSTM

针对食品安全



针对金融知识

CRF

条件随机场(CRF)于2001年提出，结合了最大熵模型和隐马尔可夫模型的特点，是一种无向图模型，近年来在分词、词性标注和命名实体识别等序列标注任务中取得了很好的效果。

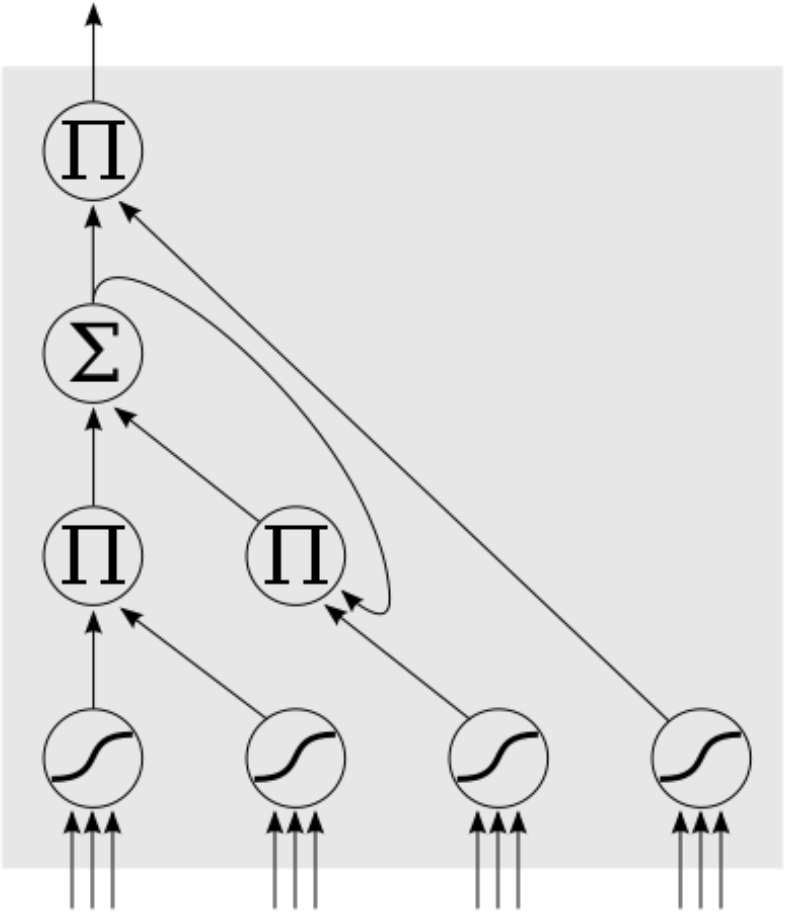
VS

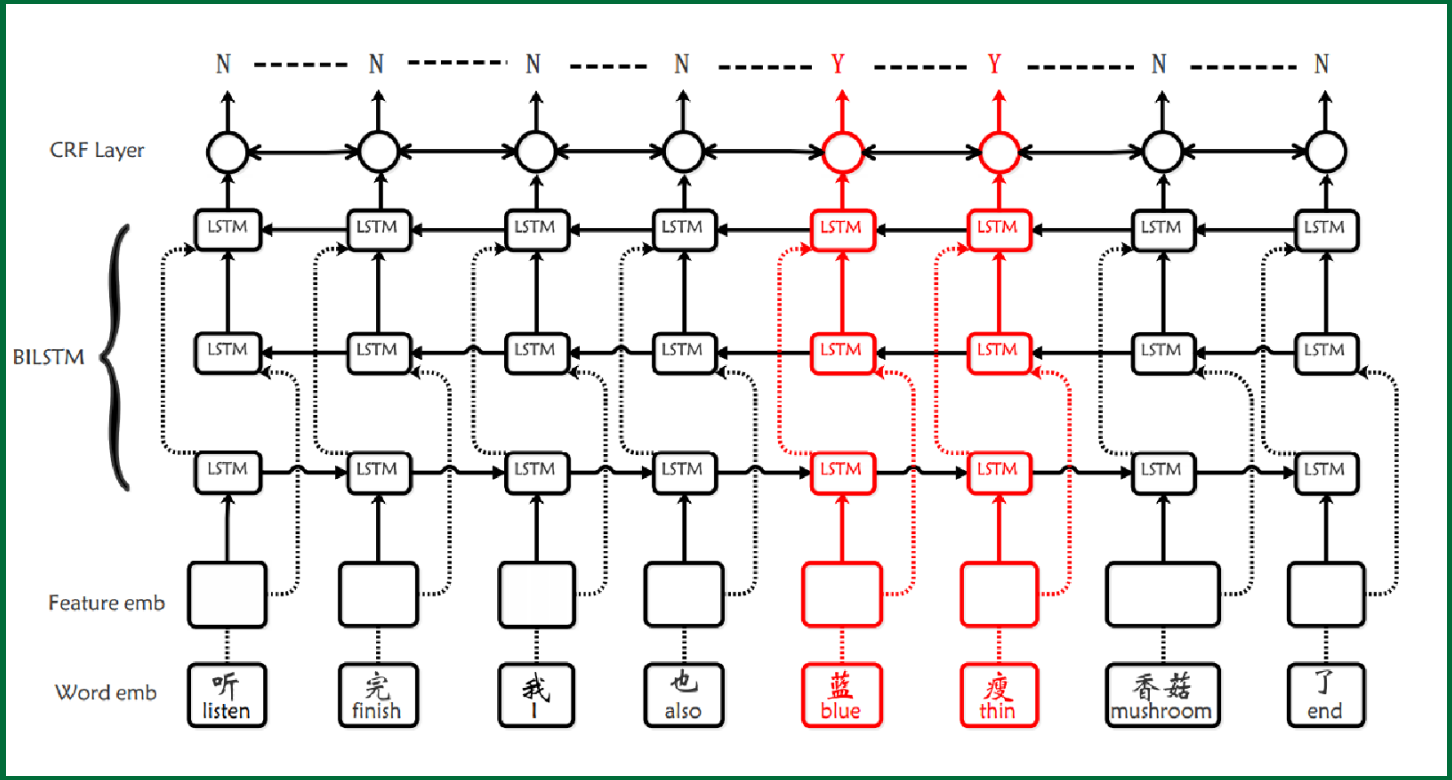
LSTM

长短期记忆网络（Long-Short Term Memory,LSTM）论文首次发表于1997年。由于独特的设计结构，LSTM适合于处理和预测时间序列中间隔和延迟非常长的重要事件。

LSTM的结构

LSTM是一种含有LSTM区块（blocks）或其他的一种类神经网络，文献或其他资料中LSTM区块可能被描述成智能网络单元，因为它可以记忆不定时间长度的数值，区块中有一个gate能够决定input是否重要到能被记住及能不能被输出output。




N
N
N
N
Y
Y
N
N
CRF Layer
BILSTM
LSTM
Feature emb
Word emb
听 listen
完 finish
我 I
也 also
蓝 blue
瘦 thin
香菇 mushroom
了 end

听完我也蓝瘦香菇了

融合特征的Bi-LSTM与CRF模型对比

Model	pre	recall	F1
CRF	15.06	34.08	20.87
Bi-LSTM+CRF	18.59	30.98	23.24
Bi-LSTM+CRF+Features	**35.19**	**52.36**	**42.09**

未融合特征

基于统计方法的CRF受制于词语的内部和外部特征，识别效果不佳

融合特征后

各项得分远超CRF模型得分

融合特征的Bi-LSTM与nlpir模型对比

Model	pre	recall	F1
nlpir	15.04	46.69	22.74
Bi-LSTM+CRF	18.59	30.98	23.24
Bi-LSTM+CRF+Features	**35.19**	**52.36**	**42.09**

未融合特征

准确率高于nlpir模型，但召回率远低于nlpir模型

融合特征后

各项得分远超nlpir模型得分，三者中最佳

各模型提取新词情况

结合图表可看出，融合特征后的Bi-LSTM+CRF模型能预测出数量最多的新词

新词	nlpir	CRF	Bi	Bi+Feature
蓝瘦香菇	✓	×	✓	✓
沙雕	×	×	×	✓
喜大普奔	×	×	×	×
淘旅行	✓	×	×	✓
压力山大	×	×	✓	✓
金拱门	×	✓	✓	✓

图1　NLPIR全链条大数据语义智能分析平台

包含了精准采集，文档格式转换、新词发现、批量分词、语言统计、文本聚类、文本分类、摘要实体、智能过滤、情感分析、文档去重、全文检索和编码转换十三项独立功能

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

PART FOUR

Demo展示

中文分词方法介绍

基于字符串匹配的分词方法

基于字符串匹配的分词方法又称**机械分词方法**，它是按照**一定的策略**将**待分析的汉字串**与一个“**充分大的**”**机器词典**中的词条进行配，**若在词典中找到某个字符串**，则**匹配成功**（识别出一个词）。

基于理解的分词方法

基于理解的分词方法是**通过让计算机模拟人对句子的理解**，达到识别词的效果。其**基本思想**就是**在分词的同时进行句法、语义分析**，利用**句法信息**和**语义信息**来**处理歧义**现象。这种分词方法**需要使用大量的语言知识和信息**。由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前还处在**试验阶段**。

基于统计的分词方法

基于统计的分词方法是在**给定大量已经分词的文本**的前提下，**利用统计机器学习模型学习词语切分的规律**（称为**训练**），从而实现**对未知文本的切分。基于统计的中文分词方法是如今的主流方法。**

jieba分词简介

jieba分词是国内使用人数最多的**中文分词工具。jieba分词**支持**三种模式**：
（1）**精确模式**：试图将句子最精确地切开，**适合文本分析**；
（2）**全模式**：把句子中所有的可以成词的词语都扫描出来，速度非常快，但是**不能解决歧义**；
（3）**搜索引擎模式**：在精确模式的基础上，**对长词再次切分，提高召回率，适合用于搜索引擎分词。**

jieba分词过程中主要涉及如下几种算法：
（1）基于**前缀词典**实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)；
（2）采用了**动态规划**查找最大概率路径，找出**基于词频的最大切分组合**；
（3）对于**未登录词**，采用了基于汉字成词能力的 **HMM 模型**，采用**Viterbi 算法**进行计算；
（4）基于**Viterbi**算法做**词性标注**；
（5）基于**tf-idf**和**textrank**模型**抽取关键词。**

测试


```python
import jieba

#全模式
test1 = jieba.cut("初夏方好，小荷初露，一起来读那些美不胜收的咏荷经典诗词。", cut_all=True)
print("全模式: " + "| ".join(test1))

#精确模式
test2 = jieba.cut("初夏方好，小荷初露，一起来读那些美不胜收的咏荷经典诗词。", cut_all=False)
print("精确模式: " + "| ".join(test2))

#搜索引擎模式
test3= jieba.cut_for_search("初夏方好，小荷初露，一起来读那些美不胜收的咏荷经典诗词。")
print("搜索引擎模式:" + "| ".join(test3))
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\49876\AppData\Local\Temp\jieba.cache
Loading model cost 0.569 seconds.
Prefix dict has been built successfully.
全模式: 初夏| 方| 好| ，| 小| 荷| 初露| ，| 一起| 起来| 读| 那些| 美不胜收| 不胜| 收| 的| 咏| 荷| 经典| 诗词| 。
精确模式: 初夏| 方好| ，| 小荷| 初露| ，| 一| 起来| 读| 那些| 美不胜收| 的| 咏| 荷| 经典| 诗词| 。
搜索引擎模式:初夏| 方好| ，| 小荷| 初露| ，| 一| 起来| 读| 那些| 不胜| 美不胜收| 的| 咏| 荷| 经典| 诗词| 。

进程已结束，退出代码为 0
```

程序简介


```
import warnings
import regex as re
import numpy as np
import math
import pandas as pd
import jieba
from tqdm import tqdm
```

库的引用

```
def __init__(self, content='', topK=-1, tfreq=10, tDOA=0, tDOF=0, is_jieba=False, mode=[1]):
```

主函数介绍：
content: 待成词的文本
topK: 返回的词数量
tfreq: 频数阈值
tDOA: 聚合度阈值
tDOF: 自由度阈值
Mode：jieba中的分词方式

处理流程

读取

01

用收集到的数据（头条标题）读取该文件导入程序。

分词

02

利用jieba库的相应功能先对文本进行初步的分词。

生成词语

03

去掉标点符号以及相关的屏蔽词。

计算信息熵

04

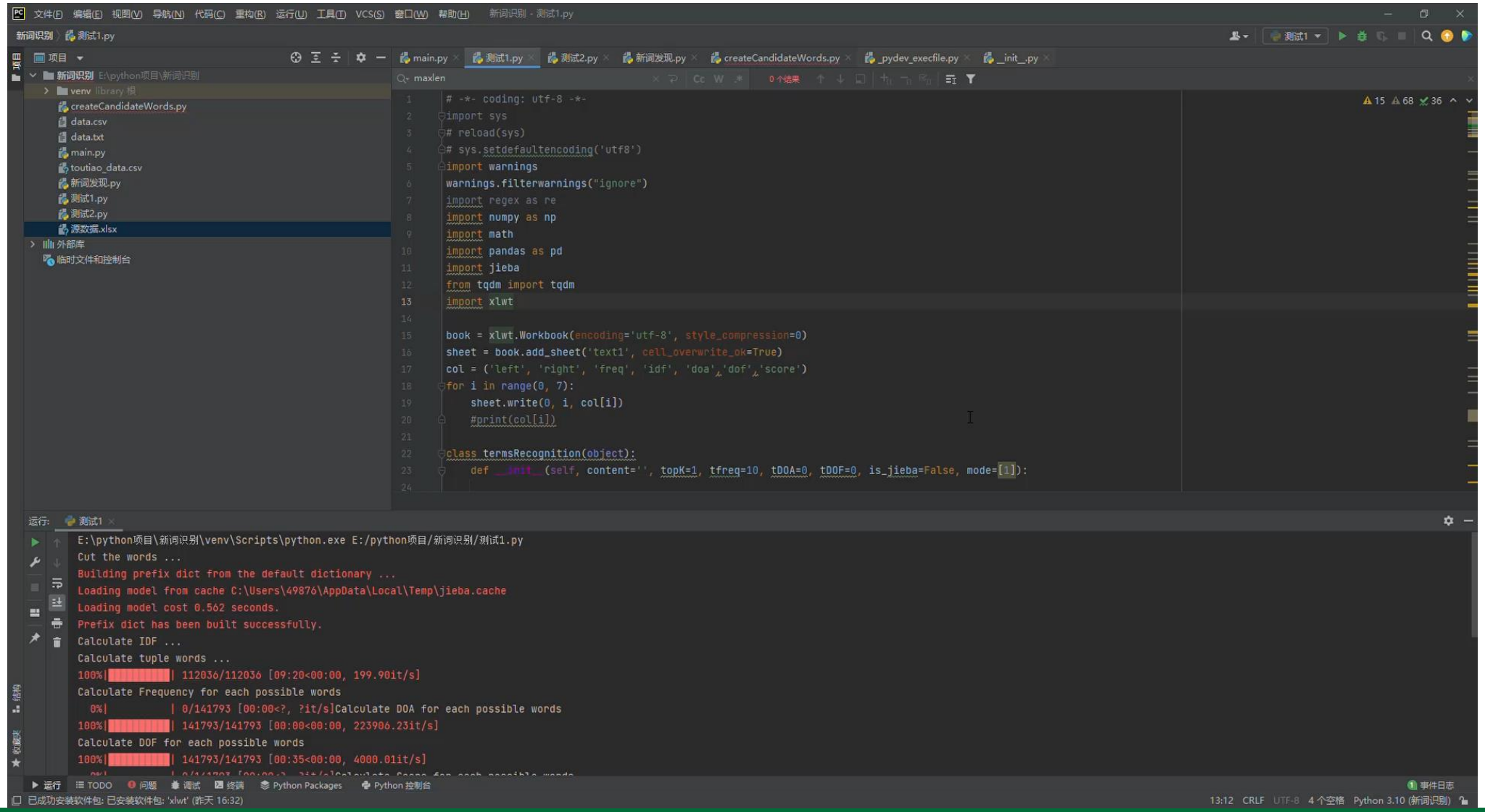
根据词语的自由度以及凝固度（参数）将简单词合成为复合词语。

得到数据

05

对于得到的词语进行计数并得到统计数据。

Demo演示


新词识别 - 测试1.py
新词识别 E:\python项目\新词识别
venv library 根
createCandidateWords.py
data.csv
data.txt
main.py
toutiao_data.csv
新词发现.py
测试1.py
测试2.py
源数据.xlsx
外部库
临时文件和控制台
main.py
测试1.py
测试2.py
新词发现.py
createCandidateWords.py
_pydev_execfile.py
__init__.py
maxlen
# -*- coding: utf-8 -*-
import sys
# reload(sys)
# sys.setdefaultencoding('utf8')
import warnings
warnings.filterwarnings("ignore")
import regex as re
import numpy as np
import math
import pandas as pd
import jieba
from tqdm import tqdm
import xlwt
book = xlwt.Workbook(encoding='utf-8', style_compression=0)
sheet = book.add_sheet('text1', cell_overwrite_ok=True)
col = ('left', 'right', 'freq', 'idf', 'doa','dof','score')
for i in range(0, 7):
sheet.write(0, i, col[i])
#print(col[i])
class termsRecognition(object):
def __init__(self, content='', topK=1, tfreq=10, tDOA=0, tDOF=0, is_jieba=False, mode=[1]):
运行: 测试1
E:\python项目\新词识别\venv\Scripts\python.exe E:/python项目/新词识别/测试1.py
Cut the words ...
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\49876\AppData\Local\Temp\jieba.cache
Loading model cost 0.562 seconds.
Prefix dict has been built successfully.
Calculate IDF ...
Calculate tuple words ...
100%| | 112036/112036 [09:20<00:00, 199.90it/s]
Calculate Frequency for each possible words
0%| | 0/141793 [00:00<?, ?it/s]Calculate DOA for each possible words
100%| | 141793/141793 [00:00<00:00, 223906.23it/s]
Calculate DOF for each possible words
100%| | 141793/141793 [00:35<00:00, 4000.01it/s]
运行 TODO 问题 调试 终端 Python Packages Python 控制台
已成功安装软件包: 已安装软件包: 'xlwt' (昨天 16:32)
13:12 CRLF UTF-8 4 个空格 Python 3.10 (新词识别)

新词汇总

季	司令部	大幅	克罗	绿盲	应援	高低	村委	奶
冬冬	千人	最低	打响	亿吨	齐祖	内容	精装	集字
仅次	万人	纷纷	帝王	教育部	孵化	厌恶	奔驰	导演
球虐	招商	上古	文化	渐唱	丝路	推掉	盖伦	伤痕
全	显示	十六	细节	有时	荔枝	空投	演唱会	罗冲
何以	流量	多种	师徒	太空游	张	无条件	花	把握
最优	畏难情绪	合圈	莱加	万世	三位	规模	泰达	组织
荣威	沙比	雕工	予	子弹	五四	幽灵	市场化	转化
念错	女朋友	天目	诚寻	暗自	凸肚	淞沪	校区	至
直入	唯一	前任	您	中海	侠	工业	氛围	京
到达	罗	篆刻	万美金	杭州	资产	窥探	近	太极拳
状态	镀金	宿务	碎片	月共迎	六大	运到	痛点	选手
第一场	笔	尧	骂	高端	同款	完美	被迫	哈登
一句	晒	内部	其	巍巍	超有	全场	神话	马拉松
东方	擒	传闻	车辆	启程	宝贝	乔屡失	步行者	收购
雷克萨斯	传播	登陆	理想	窒息	平方	达	争夺战	书香
画像	预期	推手	白色	喝酒	肆意	永不	早	销售量
原生态	零	三节	四年级	市场主体	回馈	发力	外海	道奇
生猪	彩排	继詹皇	赚	因为	下周	新手	美景	集中
解析	山体	巴西	楼下	曝出	首次	级	民族	应对
老头	政企	皇马	艺术家	最帅	效应	行动	傻了眼	增删
夏窗	变	心理	内地	保平争	监管	醒脑	剧本	艺术节
卓尔	幻想	表演	发音	仍然	办法	枉然	专车	男友
难	迷宫	应用	哈尼	热情	看似	不再	研究	降级
高尔夫	回归	批	技巧	锋芒	言论	图表	宣布	优
赛场	年轻	仍	外高内	冒雨	危房改造	静脉	非法经营	律韵
电脑	吉尼斯	需求	享受	京东	蒜薹	十年	这位	火热
饰品	闪崩	上下游	十	排除	很漂亮	扶贫	完败	熊黛林
国际	数据	企业家	正在	小镇	泰腐剧	九大	莫斯科	表示
马栏	供	不同	送别	洗	混迹	季节	珠海	落子
太子妃	超英	同比	书荒	希望	小伙伴	店门口	普吉	碰
斩落	满	总体方案	兄弟	近距离	城市更新	科技	月球	钢琴曲
预告	机会	重磅	撩妹	女儿	西游记	图文	这套	全景
双拼	面临	四月	优雅	市场需求	去留	违规行为	拳台	
别去	度假	巴勒斯坦人	放手	传动	锅多盖少	投入	微笑	
这句	雄伟	无奇不有	彪马	服务	风光	入学	融资	
支付宝	术后	接着	小学	领投	伽途	优惠	大型	
转账	大马	喜庆	运载火箭	私屠滥	年销	如歌	烟雨	
炒作	量身	勇士	不准			体罚	爆炸	
						宣传		

改进与不足

程序对于多线程的优化方面略有不足，面对38万+的数据集运行时长约30分钟而CPU的占用率不高。

重复词的过滤还需完善。

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

谢谢观看
敬请各位老师同学批评指正

答辩人：张　阔　董文轩
唐卓然　牛子儒
指导教师：　张华平