文章编号：2617-6084 (2024) 03-0066-09

基于词频的化学英语可视化分析和教学应用

——以 Python 编程为例

滕英来[1]，蓝　平[1]，蒙永俊[2]，何玉涛[1*]

（1. 暨南大学药学院 先进与应用化学合成研究院，广东 广州 510632；2. 广州都市圈网络科技有限公司，广东 广州 510630）

摘　要：化学英语是高校化学化工类专业教学的重要组成部分。为快速获得重点专业词汇，采用自编的 Python 程序进行化学英语词频统计，分析了两篇不同化学领域的全面性综述和两个化学类高影响因子国际期刊近年刊载的论文标题，并对结果加以图形化展示。所获高频词汇，可帮助教师备课及学生快速学习专业重点词汇，以及辅助有针对性的论文投稿。研究展示了计算机编程辅助词频统计在化学英语教学实践与研究方面的应用潜力。

关键词：专业英语词汇；高频词；Python 编程；数据可视化

中图分类号：G642; G434; H31　　**文献标志码：**A

为了有效掌握当前化学专业英语在教研中的发展动态和趋势，越来越多的院校都开设了专门的化学化工英语课程，有不少教师也将一部分专业英语知识融入到相应的化学专业课教学中。传统的化学英语教学多以命名法为重点，然而，很多学生学完后，依旧对阅读外文文献感到吃力。究其原因，外文文献中除了命名相关词汇外，还有许多专业相关词汇（如：反应、设备名称等）以及非专业直接相关的周边词汇。此外，由于科研进展和学科融合，新的化学专用词汇和“流行词”也在不断产生。笔者在先前的教学中，曾提出利用高影响因子期刊上新近发表的论文尤其是综述类文章，获取最新的专业词汇和词组[1]。但在实践中，学生往往课程紧张，教师也多忙于科研，导致无暇通篇精读论文。因此，如何快速、准确地获取重点词汇，就成了化学英语教研的当务之急。

词汇是语言学习的基础，词频分析是一种快速获取某一领域重点词汇的手段。市面上早已有托福高频词汇书籍，其中的词汇表系通过对托福阅读考试的原文进行词频统计排序而获得[2]。然而，迄今尚未有将词频统计应用在化学英语词汇教学实践中的报道。因此，我们采用流行的 Python 编程语言，编写了专门针对化学英语的词频分析小程序，对化学类综述论文中的词频进行统计分析，获得其中的高频词汇并进行可视化展示；此外，还汇总了化学领域高影响期刊近年来所发表论文的标题，用程序对这些标题进行词频分析，获得了其中的高频词汇。这些对于辅助化学英语的教学，乃至了解最新的科研动态，都起到了很好的效果。

投稿日期：2023-04-22

基金项目：广东省省基础与应用基础研究基金联合基金（2022A1515110367）

作者简介：滕英来（1983-），男（汉族），湖南东安人，科研助理，博士；***通信作者：**何玉涛（1987-），男（汉族），辽宁抚顺人，研究员，博士，主要从事有机合成方法研究，**Tel.** 13600086527，**E-mail** heyutao7@jnu.edu.cn。

1　研究方法

1.1　研究对象

本文分别对化学类论文正文以及化学类高影响期刊所刊载的论文的标题，进行了词频统计分析。论文高频词方面，研究的对象是两篇来自不同化学分支领域的全面性综述（见表 1），它们均发表在 IF > 10 的高影响期刊上。标题高频词方面，选取汇总了美国化学会的《Chemical Reviews》以及英国皇家化学会的《Chemical Society Reviews》在 2019 到 2021 年间刊载的所有综述的标题以供词频分析。

Table 1　Comprehensive review papers used for word frequency analysis in this paper

表 1　本文中用于词频统计的全面性综述论文

编号	标题	领域	篇幅*	刊载期刊、年份与 IF**
论文 1^0	Fermentation for the production of biobased chemicals in a circular economy: a perspective for the period 2022–2050	绿色化学	18874 词	*Green Chemistry*, 2022, 11.034
论文 2^0	Sucrose fatty acid esters: synthesis, emulsifying capacities, biological activities and structure-property profiles	食品化学	11275 词	*Crit. Rev. Food Sci. Nutr.*, 2021, 11.208

* 论文篇幅只计引言（Introduction）、讨论和结论的词数；标题、关键词、摘要、致谢、参考文献等部分均不计入。

** 所列影响因子（IF）为论文均为刊载期刊 2021 年的影响因子。

1.2　文本分析

词频分析应用自行编写的 Python 程序。所用 Python 编程语言的版本为 3.10.6，代码在 Python 自带的集成开发学习环境（IDLE）下编写和调试通过。词频统计的流程如图 1 所示。

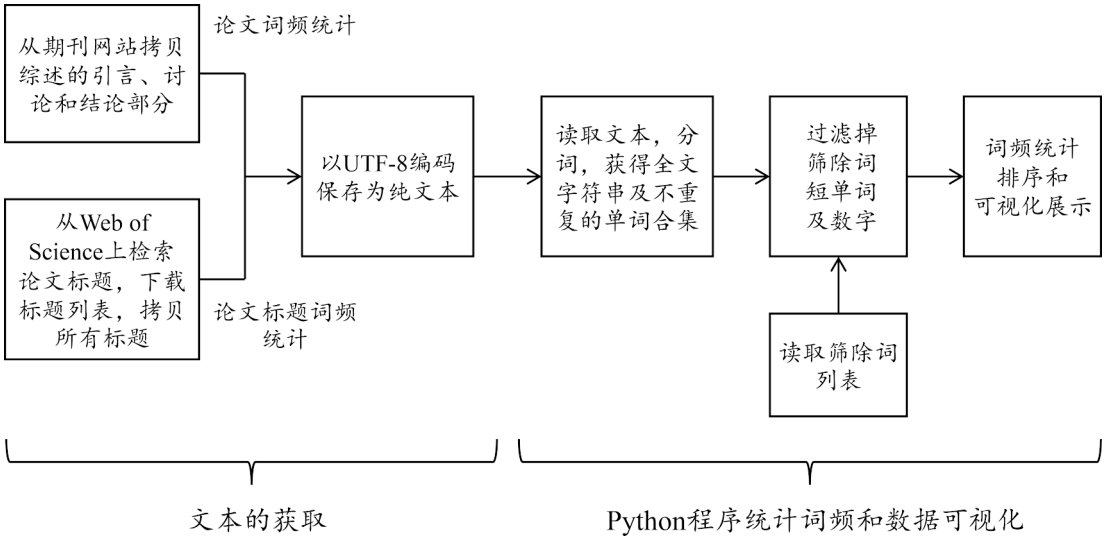

Fig. 1　Flowchart of implementing text word frequency analysis

图 1　实现文本词频率统计的流程图

1.2.1 文本的获取

为获取综述论文的文本，首先，从期刊网站上检索到该文献的网页版（html 格式），再从网页上复制其正文部分到文本编辑器，以 UTF-8 编码保存为纯文本格式。每篇论文的文本仅保留了的正文，即引言、讨论和结论部分。之所以撇除的其他部分，是因为论文的标题、摘要等部分往往会重复一些关键词的频率，而致谢、参考文献等可能带入与论文主题不相关的词汇。

为获取《Chemical Reviews》和《Chemical Society Reviews》在 2019–2021 年中每年刊载的所有综述的标题，先登入 Web of Science 检索这些年份中所刊载的综述的标题，将检索结果导出为 MS Excel 工作表格式（.xls）下载，再将工作表中的所有标题拷贝至文本编辑器中汇总，以供程序读取分析。

1.2.2 词频统计的实现

程序先将待分析的文本以 UTF-8 编码录入为可识别字符串。然后，将字符串中所有字母转换成小写以便于计数。为实现拆分成单词（后文简称为“分词”），先将字符串中起到分词作用的标点和符号（及起到分词作用的标点与符号的特定组合）都转换成空格；再通过 split()方法，以空白字符（包括空格、tab、换行符）为界，将字符串划分为由单词组成的列表。对该列表用 set()函数处理，得到没有重复单词的全文单词合集。对该集合中缺乏统计意义的筛除词（又称为“停用词”）、过短单词和数字，分别使用单独编写的函数加以剔除。最后，利用导入 collections 模块中的 Counter 类对剩下的单词在全文中出现的频次加以计数，辅之以 most_common()方法按词频由高到低的顺序排列，即可获得其中的高频词，并生成相应的结果报表存储在电脑中；尚可通过导入额外的 wordcloud 模块，从词频数据生成相应的词云图形，实现词频结果的可视化。

2 结果与讨论

2.1 词频统计方式的选择

化学化工类学术论文篇幅都较长，尤其是全面性综述（comprehensive review）更是动辄数十甚至上百页之多；即使是本研究中汇总的期刊某一年内所有综述文章标题，其总词数也在 1600 词以上。因此，依靠人力难以实现对这些文本的词频统计和高频词提取。目前，市面上虽有现成词频统计类软件，但这些软件要么需要注册收费，要么只适合分析普通英语文本，难以针对化学符号和命名系统的特殊之处进行功能定制，进而导致结果不准确。因此，本课题选用了具有简洁、高效、功能丰富等优点的 Python 编程语言[3]，编写了词频统计程序，并在程序中考虑到了化学英语的特点。Python 语言本身就带有大量的字符串处理方法和高级功能模块，利于实现快速编程和功能定制，非常适合用于词频统计领域[4]。

2.2　分词和筛词

在对化学论文进行词频统计时，需要充分考虑学术出版和化学词汇的特点。本研究没有直接从 PDF 格式的论文中提取文本分析，是因为在这一格式下的论文含有大量与正文内容不相关的出版信息，以及双栏的排版模式往往会导致位于行末的长单词被连字符打断，影响分词的准确性。就字符编码而言，化学化工中常使用各种非 ASCII 字符（如希腊字母、小型大写字母等），因此，词频分析的文本须采用 UTF-8 编码以兼容这些字符。此外，由于化合物名称可带有逗号、括号和连字符等，故而，在对文本分词时，必须要逐一考虑这些特殊情况，而不能像对普通英语文本那样直接将除字母和数字之外的所有其他字符都作为分词的标志。因此，在本课题所编写的程序中，除了将一些不会在化合物命名中出现的标点（如：感叹号、问号、冒号、分号等）作为分词标志外，还将一些标点和字符组合用于分词，如：逗号和句号后紧跟空白字符的组合以及括号和空格的组合等。

单词筛选是获得有意义词频结果的必要步骤。因为英语有大量起语法作用的词（如冠词、系动词、连词等），这些词通常长度短、出现频率很高，却又在专业英语的教学方面缺乏实际价值。为了将这些词筛除，程序使用了含有这些词的筛除词列表，以及设置了参与词频统计的单词的长度下限（≥4 字符）。最后，化学化工类论文中往往包含了各种数字，包括一些常用常数、序号、引用论文的年份等，这些数字对化学英语教学也缺乏实际意义，所以，程序中还专门设置了筛除数字的功能模块。

2.3　化学论文的高频词分析

本课题从绿色化学和食品化学领域各选取了高影响因子期刊上发表的一篇全面性综述作为高频词的研究对象（表 1），所得高频词结果如表 2 所示。从表中可以看出，高频词统计结果均反映了文章的内容特点。从论文 1 的标题可知，此文是对以发酵制备生物基化学品进行综述和展望，因而发酵（fermentation）、制备（production）、化学品（chemical 和 chemicals）、生物基（biobased）这几个与论文主题相关的词都出现在前 10 大高频词之列是合理的。此外，乙醇作为最重要的发酵产品，在文中被反复提及讨论也属情理之中的。而酸（acid）这个词成为第二高频的词汇，则可归结为许多简单有机酸是重要的发酵产品。论文 2 是一篇关于蔗糖酯类食品乳化剂的文章，因此，相关的化合物名称，如：sucrose fatty acid esters 出现频率较高。而酰基（acyl）的高频出现，则是因为脂肪酸酯类本身就含有这一基团。此外，对某一类化合物进行全面综述，就不可避免地要讨论到它们的制备反应和性质。因此，reaction 和 properties 也成为了高频词。由表 2 可知，词频统计程序能有效地统计出论文中的高频词。只要对相关参数略作调整，就能得到更多的高频词。用这样的词频分析程序处理多篇论文乃至一本著作，就能快捷地提取其中的高频专业词汇，进而使教师能够有目的性地收集这些词汇用于课堂教学，帮助学生快速掌握重要的专业词汇。

Table 2 Word frequency analysis results of review papers in two different chemistry areas in this study

表 2 本研究对于两类不同化学领域综述论文的词频统计结果

词频顺序（由高到低）	论文 1（绿色化学类）[5]		论文 2（食品化学类）[6]	
	单词	词频	单词	词频
1	production	317	sucrose	251
2	acid	271	esters	132
3	fermentation	183	acid	64
4	chemical	86	fatty	44
5	produced	85	used	43
6	annum	78	various	32
7	ethanol	74	reaction	32
8	chemicals	73	properties	31
9	processes	68	using	31
10	biobased	68	acyl	30

Fig. 2 Visualization of the top 200 most frequent words in the papers in the form of word clouds: A) the word cloud generated from paper 1 related to green chemistry; B) the word cloud generated from paper 2 focused on sucrose lipid based food emulsifiers

图 2 以词云的方式将论文的前 200 个高频进行可视化展示：A）由绿色化学相关的论文 1 所生成的词云；B）以蔗糖酯类食品乳化剂为主题的论文 2 所生成的词云

Python 编程环境下有丰富的数据可视化模块可供调用，因此，本文将论文 1 和论文 2 的前 200 个高频词生成了论文相对应的词云。从图 2 中可见，表 3 中的单词都以大的字体展示，这是因为在词云中单词的字体大小和它的词频高低成正比。对这两篇论文分别套用了绿树和黄糖方块图形，这就使得词频和论文主题可以结合在一起进行展示。在图 2A 的右侧中间位置，可以看到以较小字体展示的单词 1,4-bdo，它是由 1,4-BDO（即 1,4-丁二醇的英文缩写）经程序转成小写字母后产生。1,4-丁二醇也是一种可通过发酵制备的重要产品，在论文 1 中有深入的讨论。这从侧面证明：本文所编

写的程序可以应对化合物名词的分词。

2.4　论文标题高频词初探

从表 1 中可以看到，论文标题和关键词往往在论文的正文中也是高频词，那么能否通过分析期刊近年所刊发论文的文章标题，获得其中高频出现的热点词，进而了解该期刊的稿件选取倾向、帮助师生选择合适的期刊投稿？因此，我们收集并分析了化学界具有重大影响的《Chemical Reviews》和《Chemical Society Reviews》在近 3 年刊发的所有综述的标题。这两大期刊在 2021 年影响因子分别高达 72.087 和 60.615，为化学领域排名前二位的期刊。

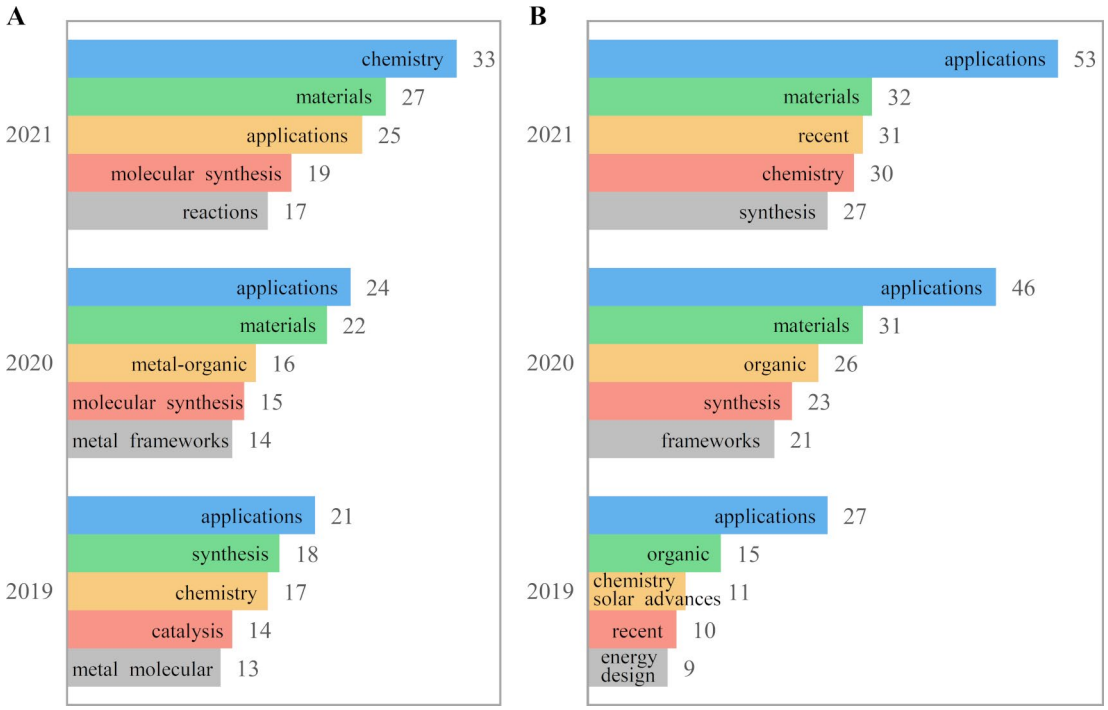

Fig. 3　Top 5 most frequent words in the titles of review papers published in: A *Chemical Reviews*, and B *Chemical Society Reviews* from 2019 to 2021 (words with the same frequency are listed side by side in the bands)

图 3　A《Chemical Reviews》和 B《Chemical Society Reviews》两个化学类期刊在 2019 到 2021 年所刊发综述类论文标题进行词频统计得到的前 5 大标题高频词（具有相同词频的词在条带中并列列出）

图 3 可见，在两大期刊的综述类文章标题中，多个高频词在三年中反复出现。其中，应用（application）一词的频次屡次名列前茅，反映了化学是一门实用性很强的学科。另一个值得注意的高频词是材料（materials）。材料领域的很多合成制备方法和化学直接相关，因此，材料相关研究自然也成为化学领域的热点。此外，化学（chemistry）与合成（synthesis）也在两个期刊的论文标题中频繁出现。前者是因为化学期刊必然要刊载化学方面内容，而后者可以解释为合成工作是化学科研的主要工作。当然，两个期刊在这三年中的标题高频词也有不同的地方。如《Chemical Reviews》中金属（metal）一词在 2019 和 2020 年都经常出现，在 2020 年甚至还以 metal-organic 的复合形容词形式出现，这或许反映了该期刊对金属类化合物题材的偏爱。相较而言，在《Chemical Society Reviews》中，有机（organic）一词多次出现，提示此期刊更倾向于有机化学相关主题；而

形容词“近期”（recent）在 2019 和 2021 年两次上榜，反映了该期刊对近期热点主题的追求。总之，通过对期刊论文标题的高频词进行统计分析，可以从相当程度上反映该期刊的发文倾向性，为师生有针对性的投稿提供参考。

3 启示和局限

词频是单词重要性的指标之一。相对于少见少用的生僻词来说，高频词显然更具有学习和使用价值。因此，通过对化学领域学术论文进行词频分析，得到的高频词可以用来指导学生有目的地学习，达到事半功倍的教学效果。而本文中将词频统计结果转变为可视化词云的方法，既可用在备课上，又可用于生成图形化摘要或学术海报图片。对期刊论文标题汇总的词频统计分析，在一定程度上揭示了这些期刊对来稿的选择倾向性，可以为师生投稿提供一些参考。本文中以程序分析高频词方法，也可以拓展到其他学科的专业英语教研中。如果将词频分析程序结合网络爬虫，还可以让程序自动生成目标论文或期刊的高频词，在运行上更加节省人力。

本文所载词频统计的方法还存在一些局限。如：化学论文中常见一些含有大写字母的缩写词，本文的程序会将这些字母统一转成小写以便词频统计，但在生成词频结果后程序并未将这些小写字母恢复成大写的初始状态。假若文中恰好有和这些缩写词的小写形式一样的单词，就会导致二者并在一起计算词频。对于同一单词的不同形式，如：名词的单复数、动词的各种时态等，程序尚无法实现在一起统计词频。此外，程序尚无法实现对词组的频率统计，而词组尤其是动词词组是英语中重要的表意单元，对学术论文写作有重要意义。要解决这些问题，可能要建立专供程序检索的字典文件、引入语料库甚至是使用人工智能进行词类的识别。

4 结论

本研究通过对两篇近年在高影响化学期刊上发表的全面性综述，以及知名化学类期刊在一定年份中刊发的综述论文标题汇总，以自编的 Python 程序进行词频统计，找出了其中的高频词，并以词云的方式对词频结果结合论文主题进行图形化展示。所获得的高频词汇，可以在化学英语教学方面起到辅助作用，或者应用于追踪论文或期刊的主题。尽管现有程序尚存在一些不足之处，本研究依然展示了编程在辅助专业英语教学方面的潜在应用价值。

参考文献：

[1] 滕英来. 期刊论文在化学专业英语教学中的运用[J]. 海外英语, 2020(17): 60−63.

[2] 李笑来. TOEFL 核心词汇 21 天突破[M]. 北京: 群言出版社, 2010.

[3] MATTHES E. 袁国忠译. Python 编程从入门到实践[M]. 北京: 人民邮电出版社, 2016.

[4] 路丽欢. 基于大数据的雅思词汇可视化分析及应用[J]. 海外英语, 2020(19): 105−107.

[5] EWING T A, NOUSE N, VAN LINT M, et al. Fermentation for the production of biobased chemicals in a circular economy: A perspective for the period 2022–2050[J]. Green Chemistry, 2022,24: 6373−6405.

[6] TENG Y, STEWART S G, HAI Y W, et al. Sucrose fatty acid esters: synthesis, emulsifying capacities, biological activities and structure-property profiles[J]. Critical Reviews in Food Science and Nutrition, 2021,61(19): 3297−3317.

Visualization analysis and teaching applications of Chemistry English based on word frequency using Python programming

TENG Yinglai[1], LAN Ping[1], MENG Yongjun[2], HE Yutao[1*]

(1. *Institute of Advanced and Applied Chemical Syntheses*, *College of Pharmacy*, *Jinan University*, *Guangzhou* 510632, *China*; 2. *Guangzhou O.cn Network Technology Co*., *Ltd*., *Guangzhou* 510630, *China*)

Abstract: Chemistry English is an essential component of the teaching in chemistry and chemical engineering majors at colleges and universities. To quickly acquire key professional vocabulary, a self-developed Python program was used to conduct word frequency analysis in Chemistry English. The analysis was based on comprehensive reviews from two different chemical fields and the titles of papers published in two high-impact international chemical journals in recent years. The results were then graphically displayed. The high-frequency vocabulary obtained can help teachers prepare lessons, assist students learn important professional words as well as aid the in targeted paper submissions. This study demonstrated the potential application of computer programming-assisted word frequency statistics in the teaching practice and research of Chemistry English.

Keywords: professional English vocabulary; high-frequency words; Python programming; data visualization

（上接第 65 页）

Reform of pharmaceutics experiment teaching method based on the combination of “virtual + real” model

LUO Rui[1], LI Sha[1], SUN Pinghua[1], ZHANG Jialin[2], LI Yan[2], XU Jun[1*]

(1. *College of Pharmacy*, *Jinan University*, *Guangzhou* 510632, *China*; 2. *Laboratory and Equipment Management Office, Jinan University*, *Guangzhou* 510632, *China*)

Abstract: The reform strategy of combining traditional experiment teaching (real) and virtual simulation experiment teaching (virtual) is an effective way to break the current limitations of pharmaceutics experiment teaching and to realize the integration of the dual chains of professional course chain and drug development chain. The authors focus on the reform strategy of pharmaceutical lab teaching based on "dual chain integration" and the construction of a "1+1+1" teaching model that combines reality and virtuality. This model involves: one phase of pre-class virtual simulation experimental teaching, one phase of offline teaching for pharmaceutical experiment courses, and one phase of virtual simulation lab teaching of extracurricular quality development, supplemented by multi-dimensional teaching resources and assessment models. This approach aims to cultivate top innovative talents in pharmacy with a solid foundation, strong capabilities, broad vision, innovation ability and high quality.

Keywords: virtual simulation; pharmaceutics experiment; teaching mode