YiCorpus 多语种单语语料库使用手册

2022.6

上海一者信息科技有限公司

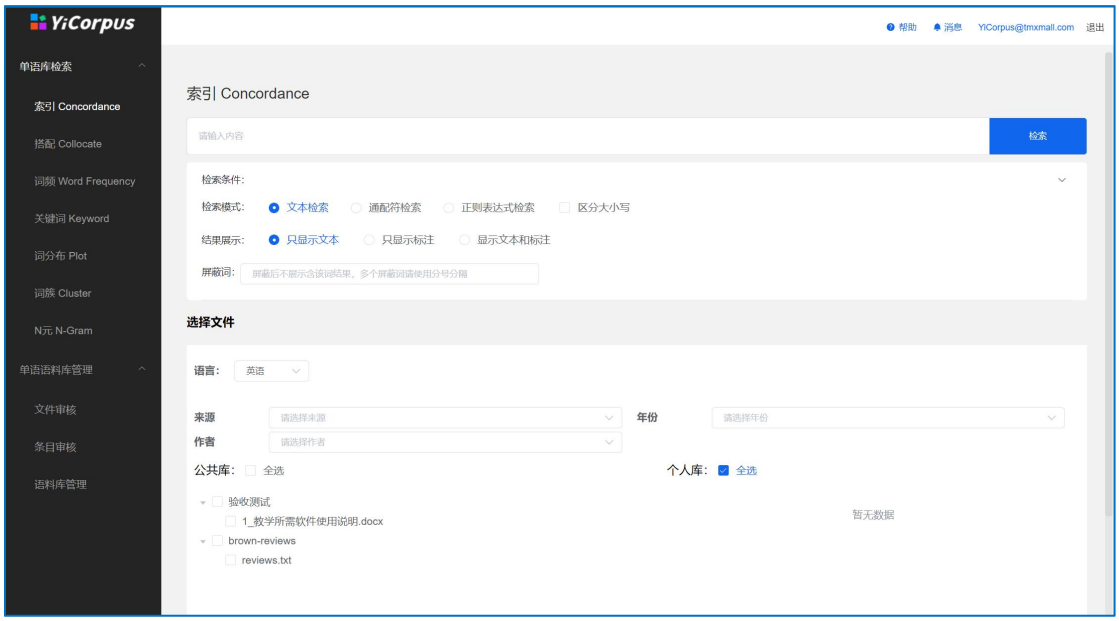
目录

YiCorpus

1. YiCorpus 多语种单语语料库简介

1.1 产品简介

YiCorpus 多语种单语语料库是一款专业的单语语料检索及资源管理平台。平台支持多语种、多格式文件导入、提供自动分词及词性标注功能、支持多模式语料检索，并支持个人及公共语料资源独立储存管理，可满足多场景语料查询及分析需求，为教学科研及翻译实践提供辅助支持。



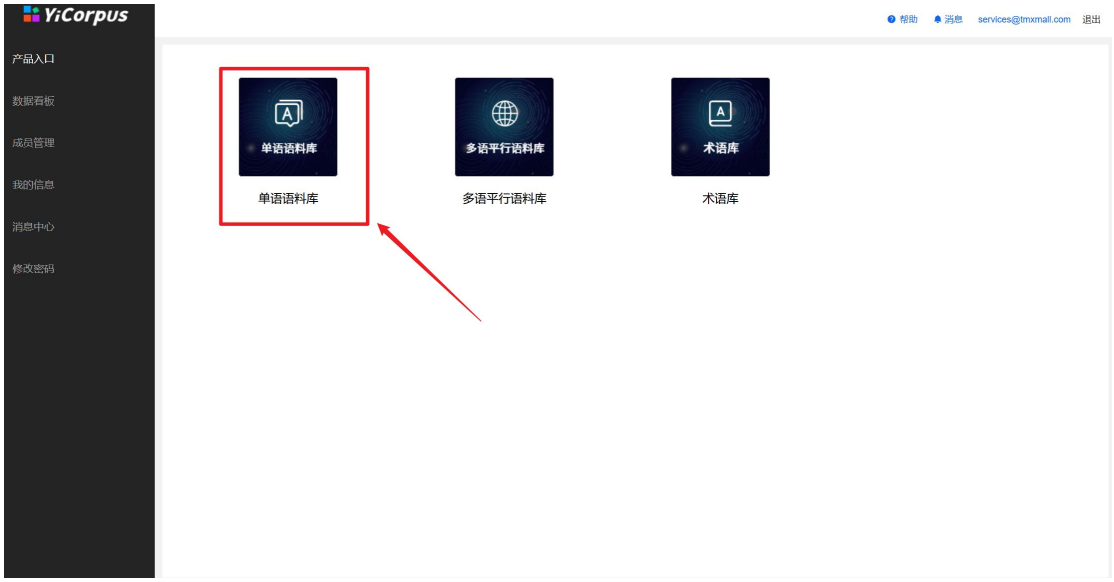
1.2 功能特点

1) 多种检索模式：平台支持索引（Concordance）、搭配（Collocate）、词频（Word Frequency）、关键词（Keyword）、词分布（Plot）、词簇（Cluster）、N 元（N-Gram）七大主流检索模式。
2) 高级检索设置：平台支持使用通配符及正则表达式查询，并可设置是否区分大小写、添加忽略词等。支持用户通过基本信息过滤库、跨库检索等。
3) 多维检索结果：各模式下检索结果包含统计概况、频次、统计数据等多维信息，支持按各主要维度排序，并可切换标准/详细数据模式方便查看。

4) 多类统计方法：以“搭配”模块为例，支持生成 MI、MI3、T-SCORE、Z-SCORE、Log-ratio、Log-likelihood、Dice 等多种统计结果。
5) 多重可视化功能：支持生成词云图、检索词分布图、数据看板（管理员可访问）等。
6) 语料区分个人/公共库：用户可上传语料至个人库独立使用，亦可将语料上传至公共库，待管理员审核通过后与所有团队成员共享。
7) 语料实时更新：普通用户可实时更新上传个人语料，且可对公共库中需修改、编辑的条目提出修改建议，待管理员审核通过后可更新语料。

2. 语料检索功能

输入平台网址，登录账号，点击产品入口中的“单语语料库”模块，即可进入 YiCorpus 多语种单语语料库。

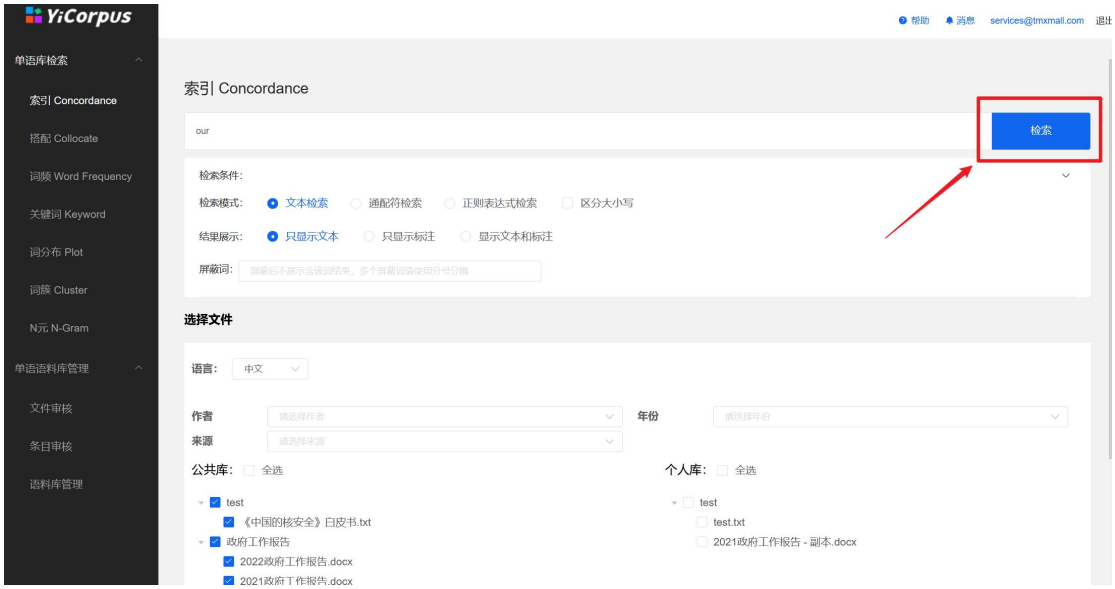


2.1 索引（Concordance）

索引功能可用于查询检索词所在语境内容。

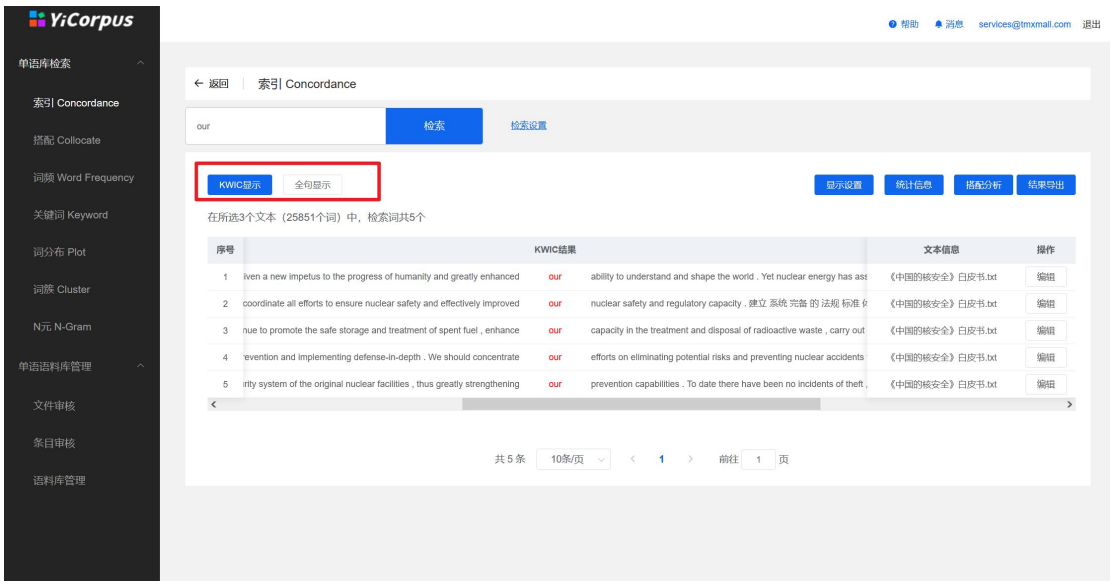
如下图所示，请在输入框内输入需查询的检索词（文本/通配符/正则表达式），并在检索栏下方选择对应检索模式。根据实际情况选择结果展示方式、填充屏蔽词（停用词）等。

在下方“选择文件”区域筛选出对应语种、类型的语料库并勾选，点击“检索”按钮，即可跳转结果展示页。



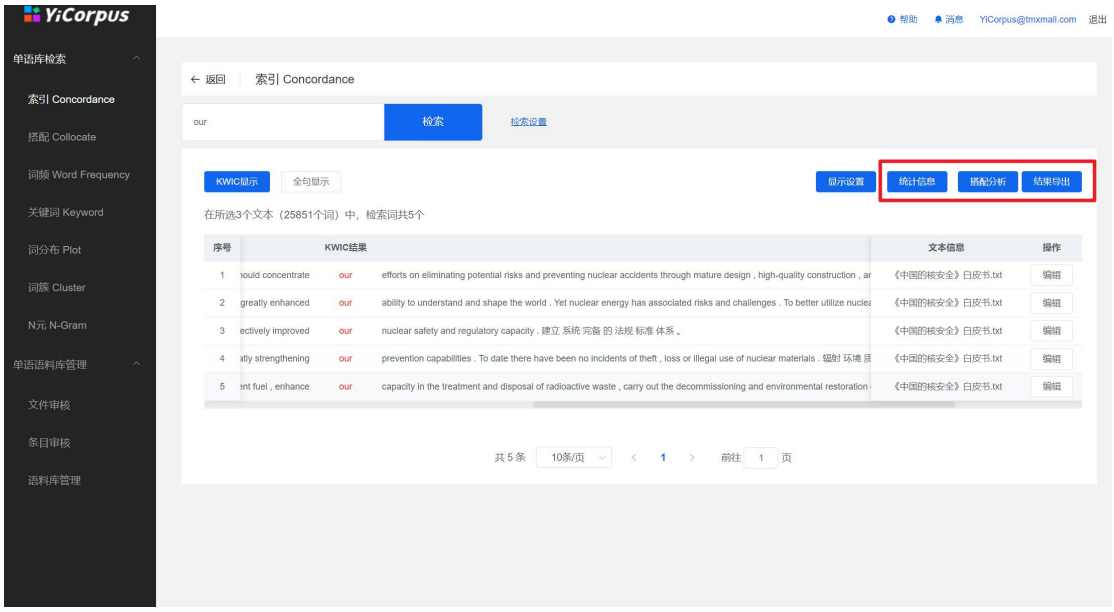
如需进一步了解文件信息，可点击文件名称，查看文件详情。

索引结果可按“KWIC（Key word in context，中心词检索显示方式）”或“全句展示”两种模式展现，点击按钮可切换模式。



点击右侧区域“显示设置”按钮，可设置结果排序、展示方式等。

点击“统计信息”、“搭配分析”、“结果导出”还可分别查看检索词分布统计情况、跳转搭配（Collocate）页面、导出当前页面检索结果等。

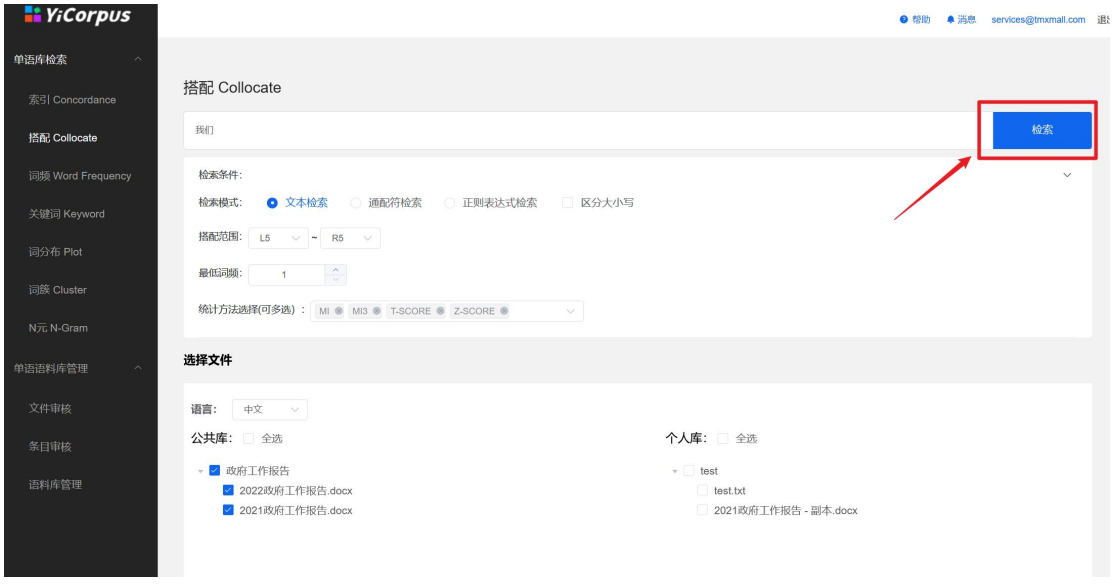


2.2 搭配（Collocate）

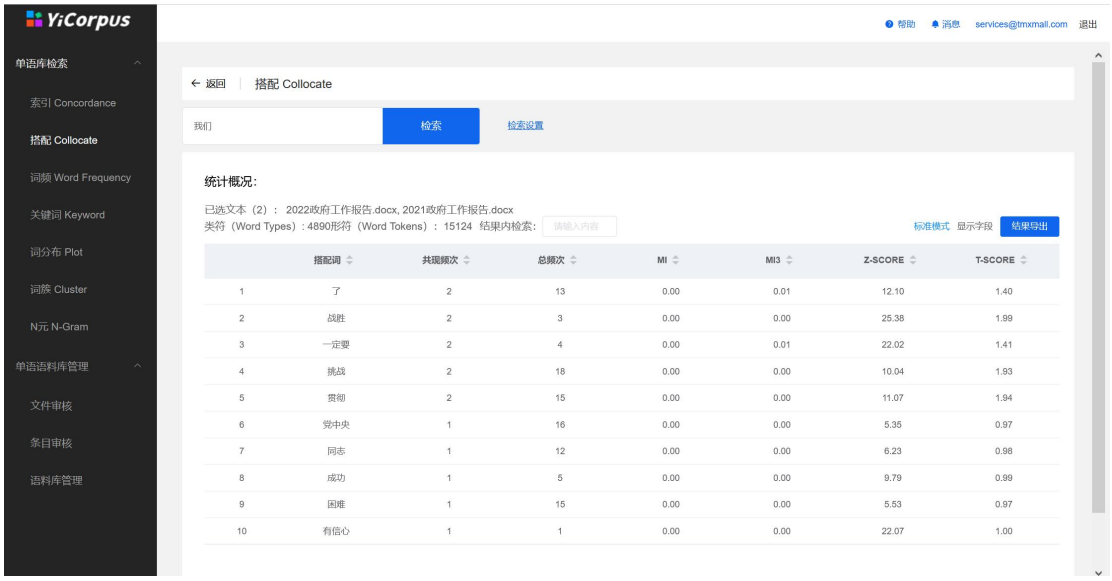
搭配功能可用于查询与检索词搭配使用的单词。

如下图所示，请在输入框内输入需查询的检索词（文本/通配符/正则表达式），并在检索栏下方选择对应检索模式。根据实际情况选择搭配范围、最低次品、统计方法（可多选）等。

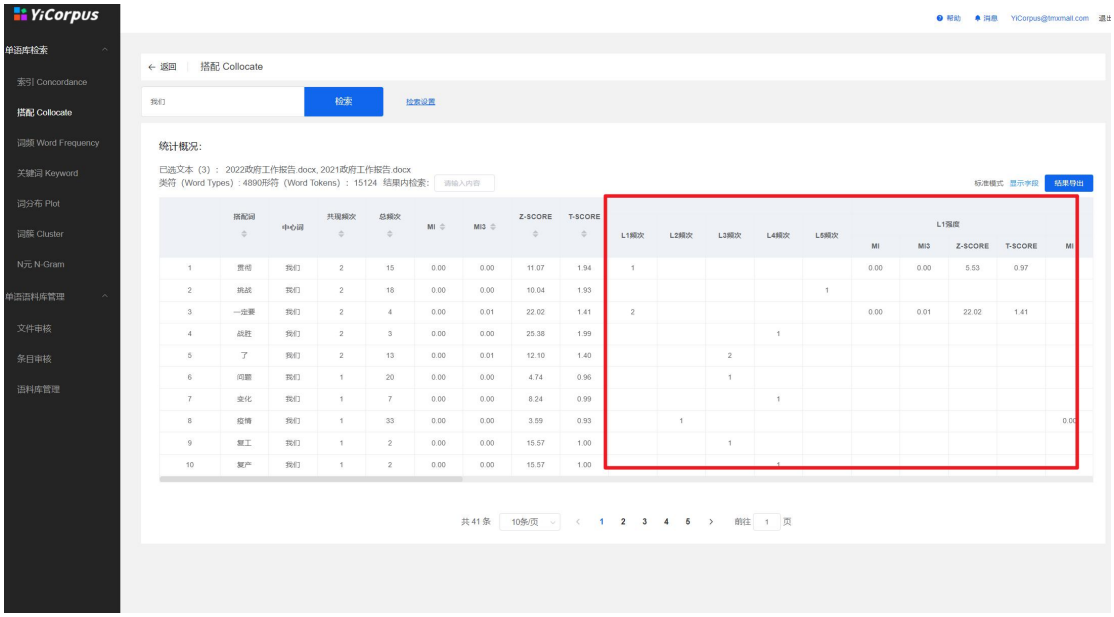
在下方“选择文件”区域筛选出对应语种、类型的语料库并勾选，点击“检索”按钮，即可跳转结果展示页。



在检索结果界面，可以看到与检索词搭配使用的单词列表，点击表头指标右侧的 图标，可按该维度切换结果排列顺序。



点击切换“详细数据”模式，还可查看搭配词在检索词附近分布的位置、相应频次和统计结果等。

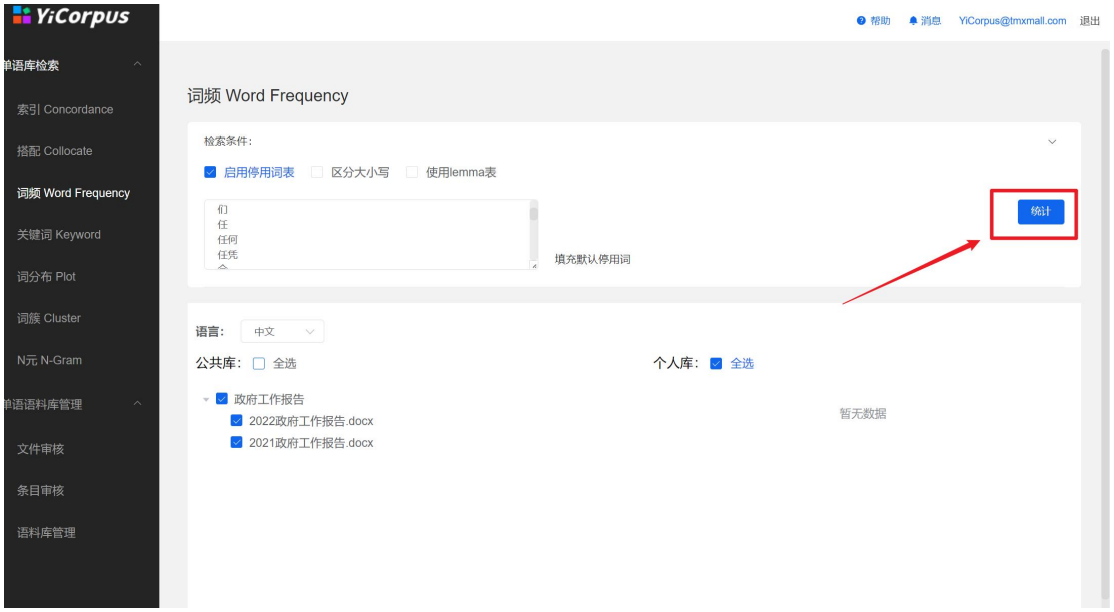


2.3 词频（Word Frequency）

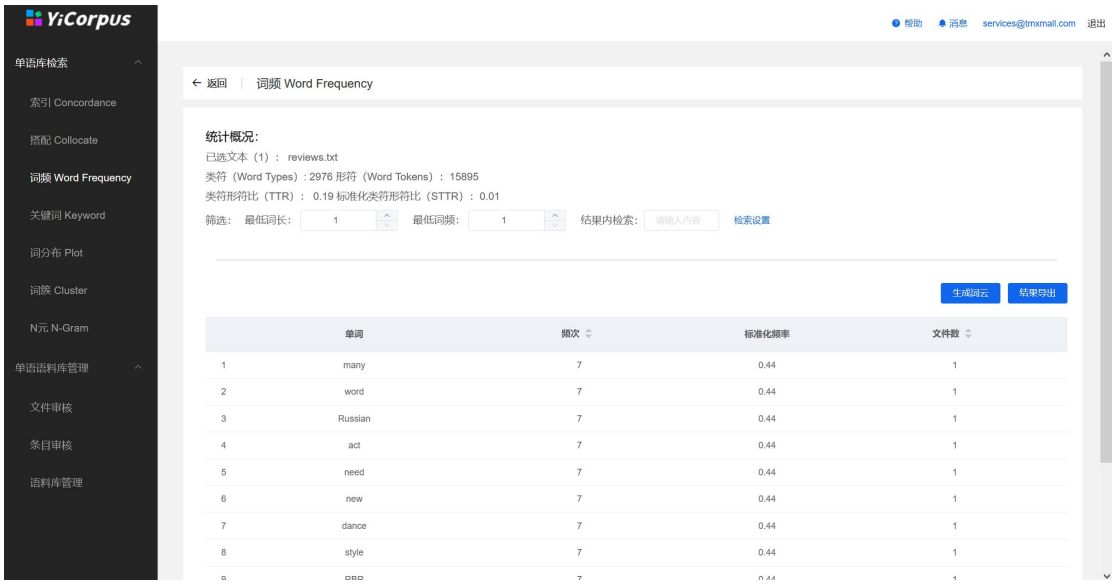
词频功能可用于统计语料库中单词出现的频率。

如下图所示，请设置是否启用停用词表、是否区分大小写、是否使用 Lemma 表等。本平台内置停用词表，并支持所有语种，如不符合需求，可直接在输入框内手动添加。内置 Lemma 表支持德语、英语、西班牙语、法语、意大利语、葡萄牙语和俄语。

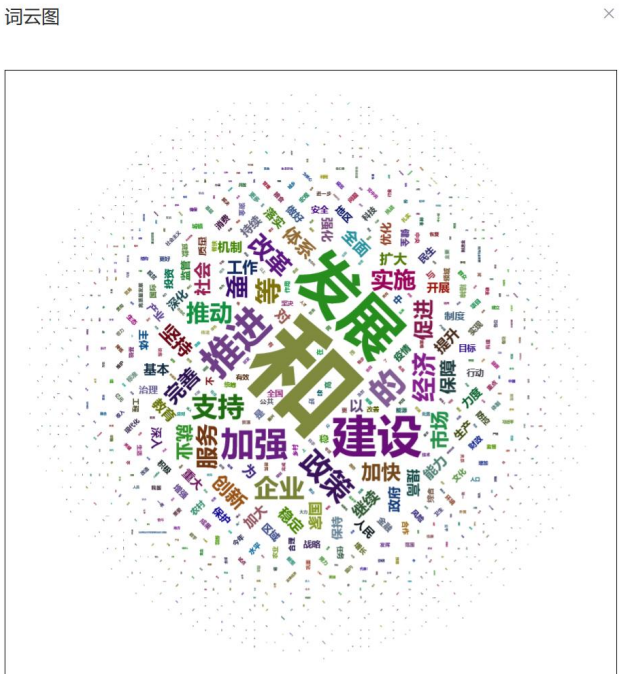
设置完成后，在下方“选择文件”区域筛选出对应语种、类型的语料库并勾选，点击“统计”按钮，即可跳转结果展示页。



在检索结果界面，可以看到语料库中各单词出现的频率，点击表头指标右侧的 图标，可按该维度切换结果排列顺序。



点击“生成词云”、“结果导出”按钮，可分别生成可下载的词云图、导出检索结果等。



2.4 关键词（Keyword）

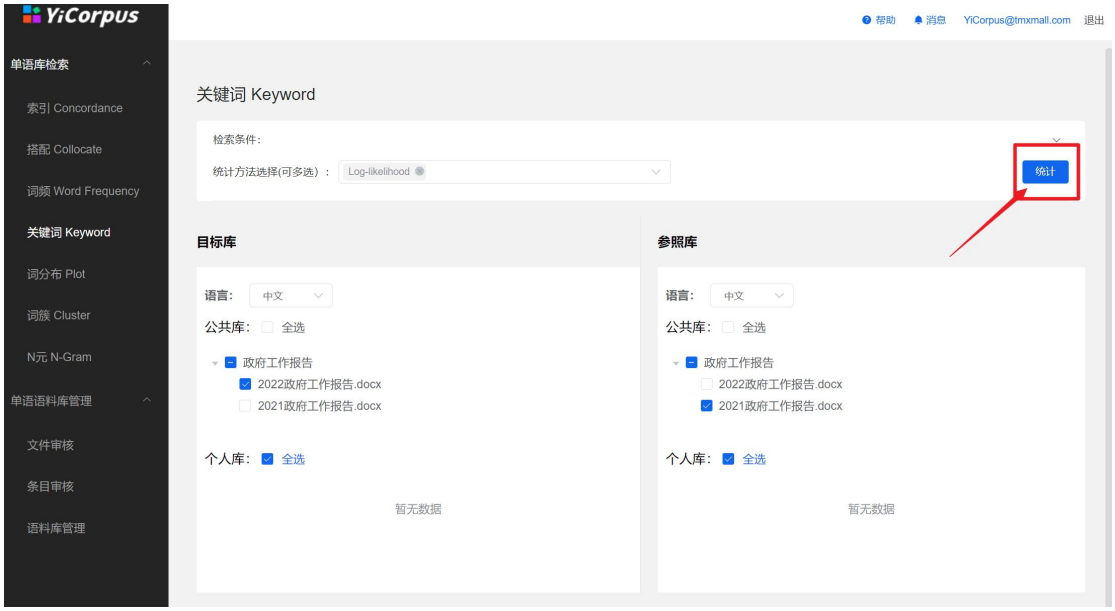
关键词功能可用于得出所选目标库中的关键词语。计算过程需对比参照库。

关键词按照关键性抽取。高频词仅取决于词频大小，因而不需要参照库；而关键词的关键性需要比对不同的两个库才能得出。

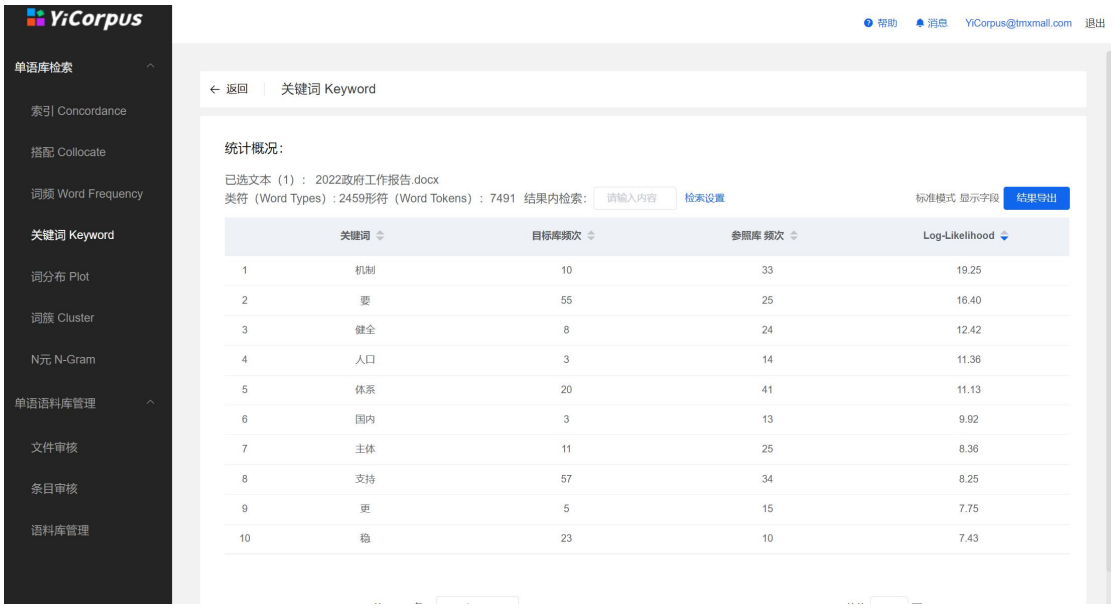
语料库语言学视角下，关键词的含义为：“在某个语篇、语料库中出现频率显著高于其在另一语料库中出现频率的词语列表”，这些关键词“具有语篇的本质属性“。

关键词的应用场景：舆情研究、教学辅助等。

如下图所示，请选择合适的统计方法。并在下方区域筛选出对应语种、类型的语料库并勾选，点击“统计”按钮，即可跳转结果展示页。



在检索结果界面，可以看到语料库中关键词列表、出现频次、关键性等，点击表头指标右侧的  图标，可按该维度切换结果排列顺序。

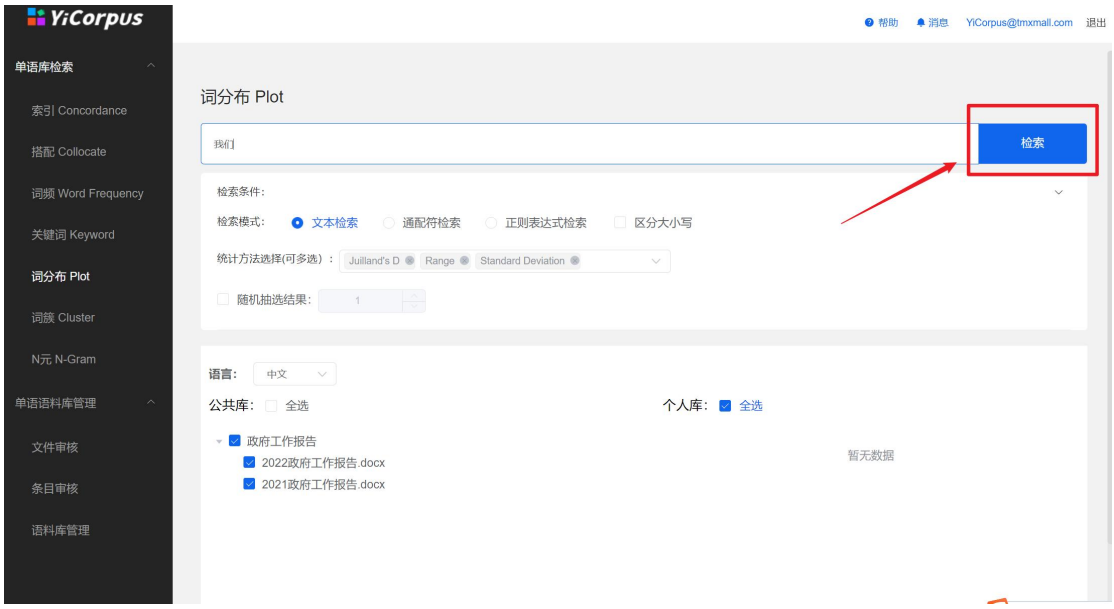


2.5 词分布（Plot）

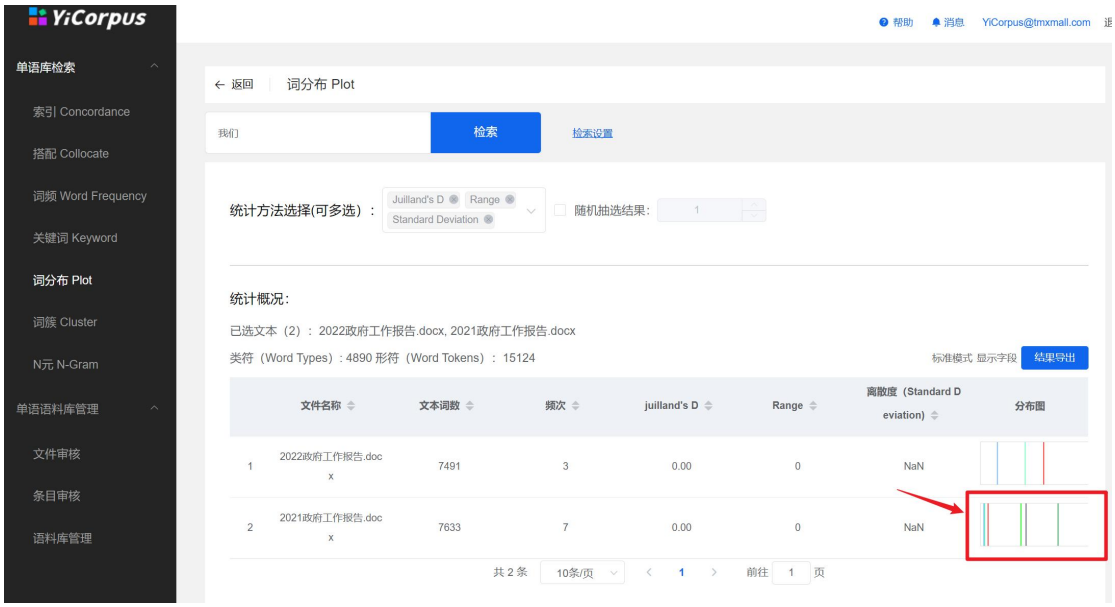
词分布功能可查看检索词在语料文件中的分布情况。

如下图所示，请在输入框内输入需查询的检索词（文本/通配符/正则表达式），并根据实际情况选择统计方法（可多选）。

设置完成后，在下方“选择文件”区域筛选出对应语种、类型的语料库并勾选，点击“检索”按钮，即可跳转结果展示页。



在结果展示页面，可查看检索词所分布的文本、文本次数及各统计指标得出的离散度。点击放大分布图，可查看检索词在文本中分散的具体位置。



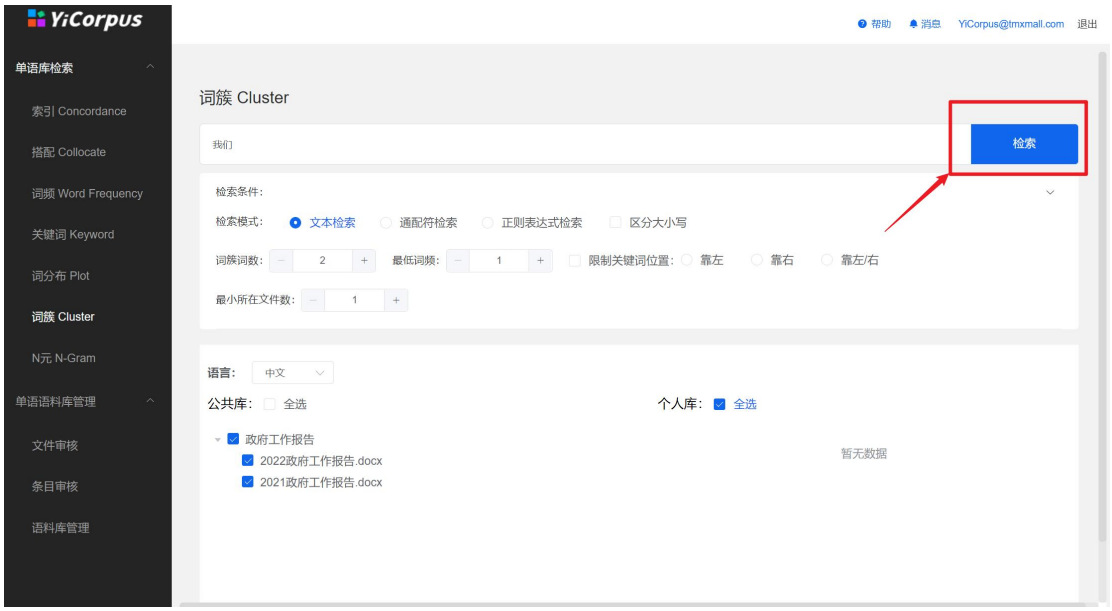
如下图所示，点击相应竖条，可查看检索词在该位置的上下文。

2.6 词簇（Cluster）

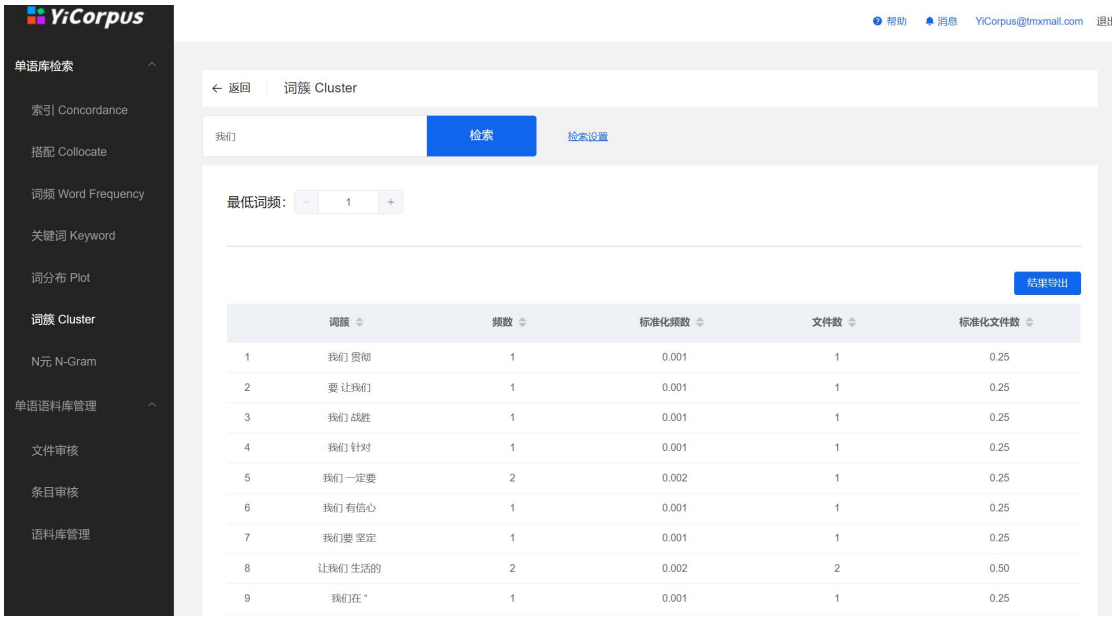
词簇功能可用于检索共同出现的多词序列。

如下图所示，请在输入框内输入需查询的检索词（文本/通配符/正则表达式），并在检索栏下方选择对应检索模式。根据实际情况选择词簇词数、最低词频、关键词位置（检索词在检索结果中的位置）、最小所在文件数等等。

在下方“选择文件”区域筛选出对应语种、类型的语料库并勾选，点击“检索”按钮，即可跳转结果展示页。



在检索结果页面，可查看词簇列表、频数、标准化频数（每千词出现的概率）、文件数、标准化文件数（命中文件数占总所选文件数的比例）等。点击表头指标右侧的  图标，可按该维度切换结果排列顺序。

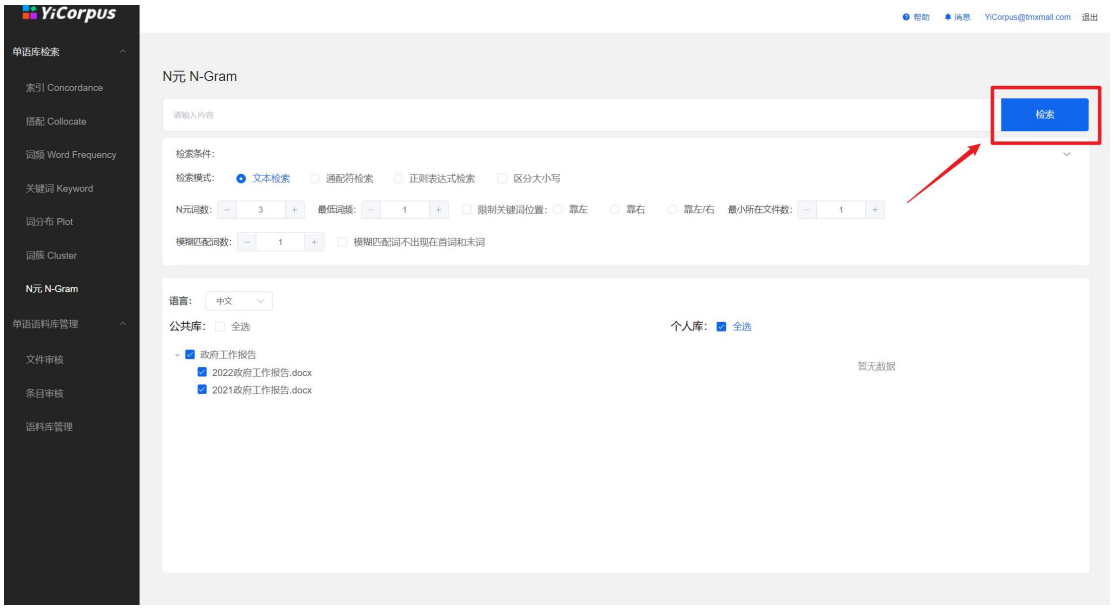


2.7 N 元（N-Gram）

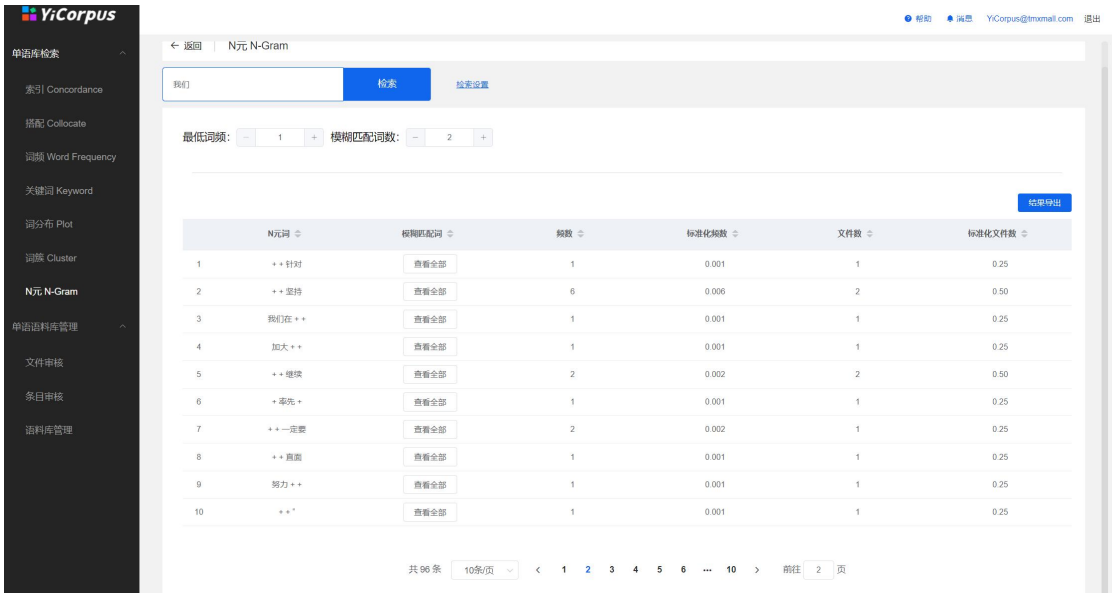
N 元功能可用于检索特定元数的 N 元语法结构。

如下图所示，请在输入框内输入需查询的检索词（文本/通配符/正则表达式，也可不填），并在检索栏下方选择对应检索模式。根据实际情况选择 N 元词数、最低词频、关键词位置（检索词在检索结果中的位置）、最小所在文件数、模糊匹配词数（归类同语法结构）等等

在下方“选择文件”区域筛选出对应语种、类型的语料库并勾选，点击“检索”按钮，即可跳转结果展示页。



在检索结果页面，可查看 N 元词列表、模糊匹配词、频数、标准化频数（每千词出现的概率）、文件数、标准化文件数（命中文件数占总所选文件数的比例）等。点击表头指标右侧的 图标，可按该维度切换结果排列顺序。



3. 语料管理功能

3.1 文件审核功能

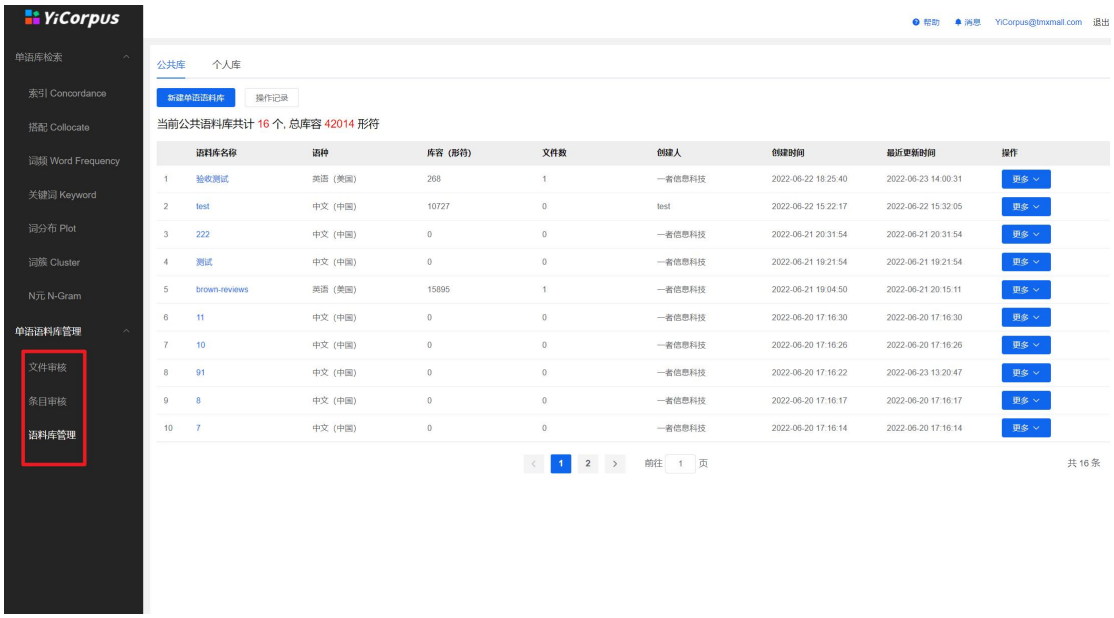
当有成员向公共库中上传语料时，需管理员在“单语语料库管理-文件审核”中审核。任一管理员审核通过后，该语料即可由团队成员共享。审核记录会在系统中保留。

3.2 条目审核功能

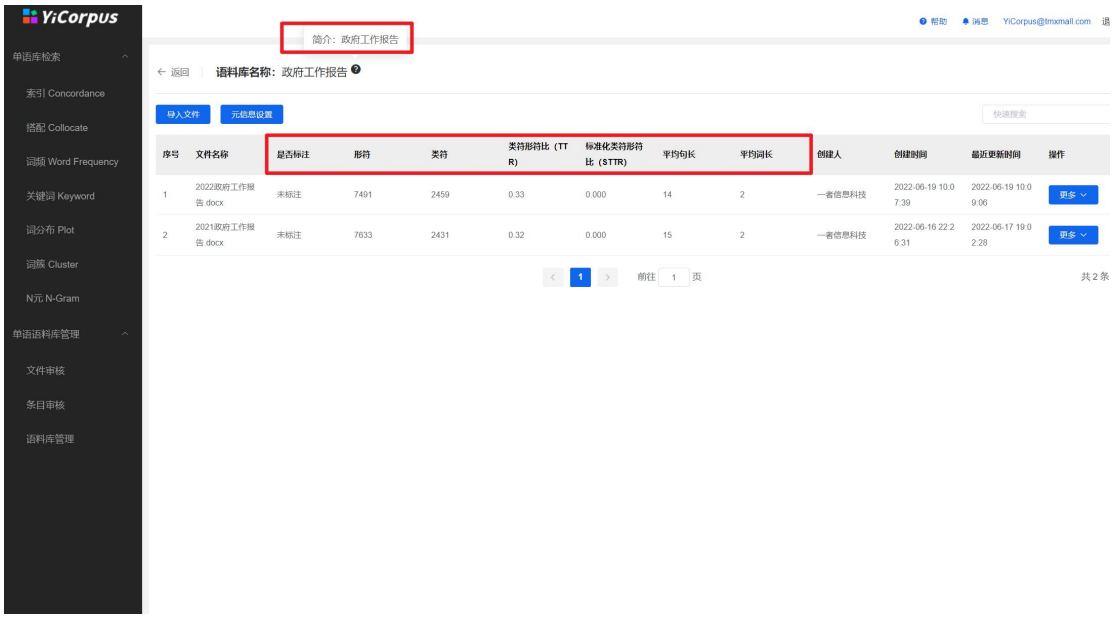
当有成员在检索结果中申请编辑条目时，需管理员在“单语语料库管理-条目审核”中审核。任一管理员审核通过后，该修改即可生效，对应语料内容将被更新。

3.3 语料管理功能

成员可在该模块中对公共库/个人库进行新增、修改、删除、查询、导入、导出、查询操作记录等，其中普通成员对公共库仅有上传权限，且需管理员审核。

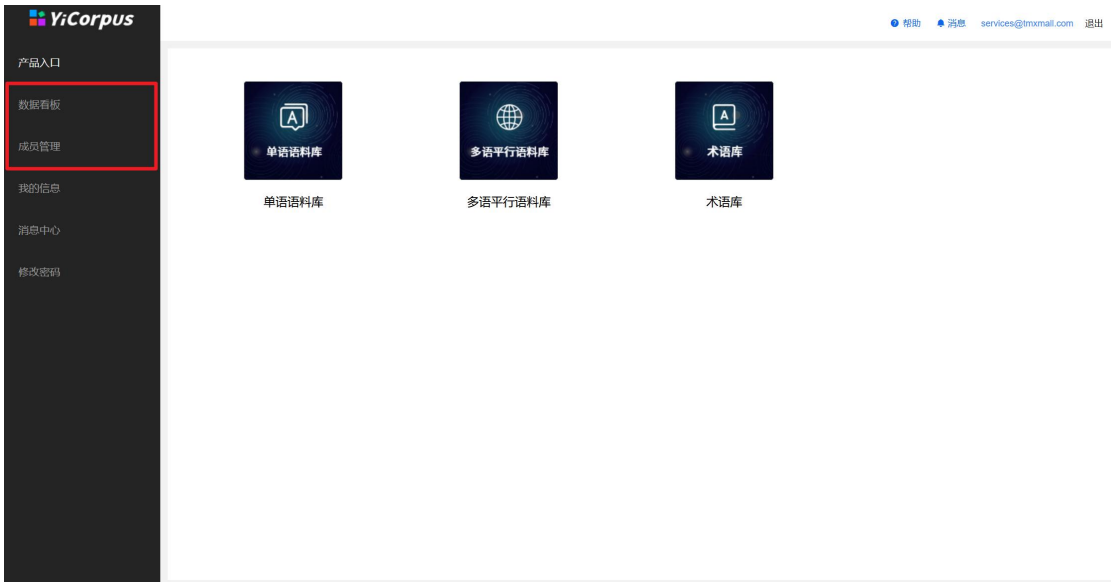


点击语料库名称，可查看语料库所包含文件详细信息。如文件简介、是否标准、形符、类符、类符形符比（TTR）、标准化类符形符比（STTR）、平均词长、平均句长等。

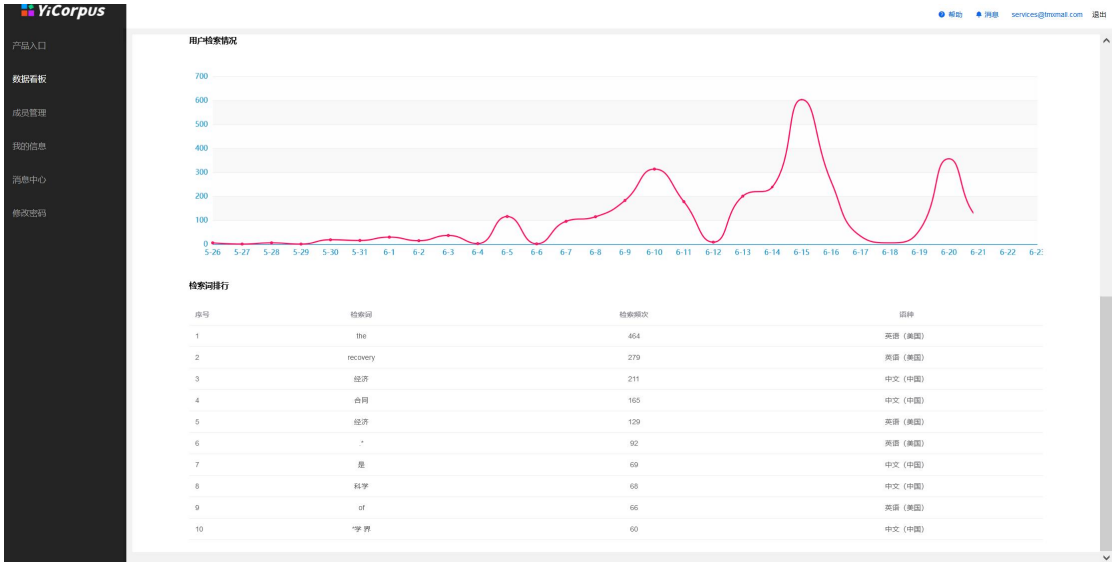

YiCorpus
单语库检索
索引 Concordance
搭配 Collocate
词频 Word Frequency
关键词 Keyword
词分布 Plot
词簇 Cluster
N元 N-Gram
单语语料库管理
文件审核
条目审核
语料库管理
简介：政府工作报告
帮助
消息
YiCorpus@tmxmall.com
退
返回
语料库名称：政府工作报告
导入文件
元信息设置
快速搜索
序号
文件名称
是否标注
形符
类符
类符形符比（TTR）
标准化类符形符比（STTR）
平均句长
平均词长
创建人
创建时间
最近更新时间
操作
1
2022政府工作报告.docx
未标注
7491
2459
0.33
0.000
14
2
一者信息科技
2022-06-19 10:07:39
2022-06-19 10:09:06
更多
2
2021政府工作报告.docx
未标注
7633
2431
0.32
0.000
15
2
一者信息科技
2022-06-16 22:26:31
2022-06-17 19:02:28
更多
前往 1 页
共 2 条

4. 团队管理功能

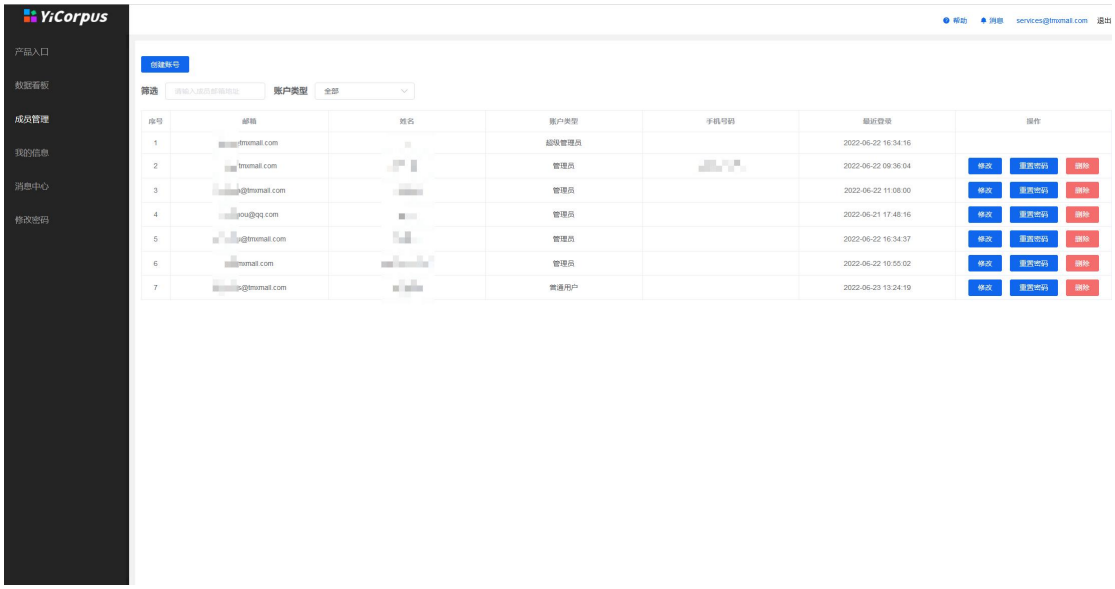
打开网址，登录账号，管理员可在侧边导航栏中看到“数据看板”和“成员管理”功能。



在“数据看板”页面中，可以看到本平台的使用概况、团队成员概况、成员登录情况、检索情况、检索词排行等。



在“成员管理”页面中，可以创建、修改、删除、查询团队成员。本平台共有三类权限角色可选：超级管理员、管理员和普通成员。“超级管理员”只有一名，可创建并管理“管理员”角色，“管理员”可创建并管理“普通成员”角色。其中普通用户无审核权限。



5. 常见问题

5.1 YiCorpus 对形符的定义？

答：除标点以外的单词、数字等。

5.2 可支持标注信息？

答：目前版本只支持词性标注。

5.3 如何检索标注？

答：如需检索标注，请在标注前使用“_”连接符，如在搜索框中输入“troubles_NOUN”。

附录一 YiCorpus 词性赋码集

序号	标注	含义
1	ADV	adverb
2	AUX	auxiliary
3	CCONJ	coordinating conjunction
4	DET	determiner
5	INTJ	interjection
6	NOUN	noun
7	NUM	numeral
8	PART	particle
9	PRON	pronoun
10	PROPN	proper noun
11	PUNCT	punctuation
12	SCONJ	subordinating conjunction
13	SYM	symbol
14	VERB	verb
15	X	other

附录二 语料库常用功能术语对照表

平台名称	近义词	译名（释义）	近义词1	近义词2
索引		Concordance		
KWIC		上下文关键词		
检索词	节点词	search word	node	
搭配		Collocate	collocation	
词频	词表、词频表、词单	word frequency	word list	word
关键词	关键词表、主题词表、关键词单	keyword	keyword list	
词分布	索引定位	plot		
词簇	词串、词丛、多词序列	cluster	lexcical bundles	multiword expressions
词簇词数	词簇长度	cluster size		
N元	N元组	N-gram		
模糊匹配词		Open slots	p-frame	
只显示文本		plain text		
只显示标注		test with POS		
频数	频次	frequency（freq）	occurrence	raw frequency
标准化频数	频率	norm frequency	frequency	
搭配强度		collocability		
共现频数		co-occurrences		
正则表达式		regular expressions	regex	regexp
目标库	观察库	target corpus	observed corpus	
参照库		reference corpus		
分词	词语切分	tokenize	segement	
词形还原		lemmatization		
Lemma表		Lemma list		
Rx/Lx	跨距	检索词（不含）右/左第x个词	span	
停用词表		stop list		
文件数		range		

附录三 语料库常用术语释义

1. Lemma 表（Lemma list）

Lemma 表为词形还原对照表。如 be 有很多种屈折变化（inflections），例如 am、is、are 等，但全部归属于一个 lemma（词元）：be。这样的对应方式可在各语言对应的 lemma 表中查找。

2. 停用词表（Stop list）

语料库词频检索时，结果中往往会出现诸如“的”、“和”、“。”之类无实际意义（以汉语为例），或者其他无关的字词和标点。导入停用词表后，检索结果中将不再囊括停用词表中的字词或标点。

3. 词簇（Cluster）

词簇，又称多词序列，是英语语言中的结构和意义单位的呈现形式。

4. N 元（N-gram）

N 元语法指文本中连续出现的 n 个语词。

举例：I have a blue pen.

如果把上方例句中的单词两两分拆（即：N-Gram Size = 2），有 4 种可能：I have / have a / a blue / blue pen；

以上为“I have a blue pen.”的 2 元拆分方式。

5. 索引（Concordance）

使用索引功能，可以调取所有包含检索词的索引行，同时检索词突出显示。

6. KWIC（Key Word in Context）

使用 KWIC 检索模式，可得到包含节点词的索引行（concordance lines），且视觉上，节点词处于中心位置，邻词在左右两侧依次排开。

7. 分词（Tokenize）

分词指将句子、段落分解为字词单位，方便后续的处理的分析。英文中，单词与单词之间用空格分开，即天然分词符。但中文在词级别没有形式上的分界符，因而中文分词具有一定难度。

8. 正则表达式（Regular expressions，简称 regex 或 regexp）

使用正则表达式可以检索符合特定模式的文本，例如，使用 an+可匹配 and、analysis。使用正则表达式将大大节省文本处理消耗的时间精力。

9. 搭配强度（Collocability）

搭配强度体现两个字词之间的联系，通常采用 MI 值（互信息值）等方式作为表征参数。

10. 词性标注（POS，part of speech）

标注某个字词的词性。（例如：国家_NOUN）

11. 模糊匹配词

在 N 元中，模糊匹配词功能会根据设定的模糊匹配词数 m，在每一条 N 元检索结果（N 元词）中随机抽取 m 个词替换为通配符，执行模糊检索并得到相应结果。

例：在 N 元中检索“经济”，设 N 元词数为 3，模糊匹配词数为 1。则该 N 元词的 3 个单词中的 1 个将替换为通配符“+”，得到结果如：

经济 + 运行

经济 增速 +

+ 经济 布局

……

点击可查看任一结果模糊匹配到的全部结果，如点击查看“经济+运行”的模糊匹配词，包含“经济稳定运行”、“经济平稳运行”等。

附录四 YiCorpus 公式整理

为了计算互信息值相关的数据，需要根据单词出现的相关数据构建如下基本参数表。

观测值	Word 1	Not Word 1	行总和
Word 2	O_{11}	O_{12}	O_{1x}
Not *Word 2*	O_{21}	O_{22}	O_{2x}
列总和	O_{x1}	O_{x2}	O_{xx}

预期值	Word 1	Not Word 1
Word 2	E_{11}	E_{12}
Not *Word 2*	E_{21}	E_{22}

O_{11}：单词 *Word 1* 与 *Word 2* 同时出现的频次

O_{12}：单词 *Word 1* 出现并且单词 *Word 2* 不出现的频次

O_{21}：单词 *Word 1* 不出现并且单词 *Word 2* 出现的频次

O_{22}：单词 *Word 1* 与单词 *Word 2* 都不出现的频次

$E_{11} = (O_{1x}O_{x1})/O_{xx}$

$E_{12} = (O_{1x}O_{x2})/O_{xx}$

$E_{21} = (O_{2x}O_{x1})/O_{xx}$

$E_{22} = (O_{2x}O_{x2})/O_{xx}$

关键词计算

观测值	目标文件	参考文件	行总和
Word w	O_{11}	O_{12}	O_{1x}
Not *Word w*	O_{21}	O_{22}	O_{2x}
列总和	O_{x1}	O_{x2}	O_{xx}

预期值	目标文件	参考文件
Word w	E_{11}	E_{12}
Not *Word w*	E_{21}	E_{22}

O_{11}: *Word w* 在目标文件中出现的频次

O_{12}: *Word w* 在参考文件中出现的频次

O_{21}: 目标文件除 *Word w* 外其他单词的频次

O_{22}: 参考文件除 *Word w* 外其他单词的频次

E_{11}= $(O_{1x}O_{x1})/O_{xx}$

E_{12}= $(O_{1x}O_{x2})/O_{xx}$

E_{21}= $(O_{2x}O_{x1})/O_{xx}$

E_{22}= $(O_{2x}O_{x2})/O_{xx}$

Log-likelihood Ratio[1]

$$LL = 2 \times \sum_{i=1}^{2}\sum_{j=1}^{2}(O_{ij} \times ln\frac{O_{ij}}{E_{ij}})$$

MI

$$MI = log_2 \frac{\frac{O_{11}}{O_{xx}}}{\frac{O_{1x}}{O_{xx}} \times \frac{O_{x1}}{O_{xx}}}$$

1 Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), 61-74.

MI3

$$MI3 = log_2 \frac{\frac{{O_{11}}^3}{O_{xx}}}{\frac{O_{1x}}{O_{xx}} \times \frac{O_{x1}}{O_{xx}}}$$

Z-SCORE[2]

$$z = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$

T-SCORE[3]

$$t = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$$

Dice[4]

$$Dice = \frac{2 \times O_{11}}{O_{1x} + O_{x1}}$$

2 Dennis, S. F. (1964). The construction of a thesaurus automatically from a sample of text. In M. E. Stevens, V. E. Giuliano, & L. B. Heilprin (Eds.), Proceedings of the symposium on statistical association methods for mechanized documentation (pp. 61-148). National Bureau of Standards.

3 Church, K., Gale, W., Hanks P., & Hindle D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), Lexical acquisition: Exploiting on-line resources to build a lexicon (pp. 115-164). Psychology Press.

4 Rychlý, P. (2008). A lexicographyer-friendly association score. In P. Sojka, & A. Horák (Eds.), Proceedings of Second Workshop on Recent Advances in Slavonic Natural Languages Processing. Masaryk University

ChiSquare[5] 卡方

$$x^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

LogRatio[6]

$$Log\ Ratio = log_2 \frac{\frac{O_{11}}{O_{x1}}}{\frac{O_{12}}{O_{x2}}}$$

5 Hofland, K., & Johanson, S. (1982). Word frequencies in British and American English. Norwegian Computing Centre for the Humanities. Oakes, M. P. (1998). Statistics for Corpus Linguistics. Edinburgh University Press.

6 Hardie, A. (2014, April 28). Log ratio: An informal introduction. ESRC Centre for Corpus Approaches to Social Science (CASS).