**多语种计量指标提取工具技术文档**

广东外语外贸大学数据挖掘实验室

[2023/5/12]

目 录

[1 项目介绍 4](#_Toc18976)

[2 通用指标 4](#_Toc11402)

[2.1 总词数 4](#_Toc9582)

[2.2 词典数 4](#_Toc4487)

[2.3 单现词数 4](#_Toc25441)

[2.4 TTR 4](#_Toc7188)

[2.5 H-point 5](#_Toc21797)

[2.6 熵 5](#_Toc30378)

[2.7 词汇丰富度 6](#_Toc2401)

[2.8 重复率 6](#_Toc4888)

[2.9 相对重复率 7](#_Toc31051)

[2.10 主题集中度 7](#_Toc8668)

[2.11 次级主题集中度 7](#_Toc10012)

[2.12 活动度 7](#_Toc10545)

[2.13 描写度 8](#_Toc356)

[2.14 文本中词的秩序分布的欧氏长度 8](#_Toc24994)

[2.15 文本中词的秩序分布的R指数 8](#_Toc14805)

[2.16 文本的Lambda值 8](#_Toc27683)

[2.17 校正模数 8](#_Toc29558)

[2.18 基尼系数 9](#_Toc22012)

[2.19 基尼系数补数 9](#_Toc2909)

[2.20 作者视野 9](#_Toc16654)

[2.21 动词间距 10](#_Toc6250)

[3 西方语种文本可读性指标 10](#_Toc31545)

[3.1 ARI 10](#_Toc24978)

[3.2 RIX 10](#_Toc26088)

[3.3 FleschReading 11](#_Toc18358)

[3.4 FlsechKincaidGrade 11](#_Toc22287)

[3.5 GunningFog 11](#_Toc18214)

[3.6 Smog 12](#_Toc10025)

[3.7 ColemanLiauIndex 12](#_Toc5689)

[3.8 DaleChallIndex 13](#_Toc31723)

[3.9 LWIndex 13](#_Toc20898)

[4 中文文本可读性指标 14](#_Toc7624)

[5 文本处理 14](#_Toc27750)

# **项目介绍**

本工具主要面向从事计量语言学研究的学者提供计量指标提取服务，内部涵盖了21个基础指标，包括：总词数、单现词数、单现词比例、TTR、h点、熵、词汇丰富度、重复率、相对重复率、主题集中度、次级主题集中度、活动度、描写度、文本中词的秩序分布的欧氏距离、文本中词的秩序分布的R指数、文本的Lambda值、校正模式、基尼系数、基尼系数补数、作者视野。

总体支持多个国家语种文本处理：包括：汉语、英语、阿拉伯语、孟加拉语、波斯语、土耳其语、韩语、日语、印尼语、柬埔寨语、老挝语、菲律宾语、泰语、越南语、葡萄牙语、西班牙语、捷克语、法语、德语、俄语、乌克兰语、意大利语、瑞典语、希腊语、希伯来语、印地语、马来语、波兰语、塞尔维亚语、乌尔都语。

# 通用指标

## 总词数

经过分词后文本中总的单词个数。

## 词典数

经过分词后文本中去重的单词总数。

## 单现词数

经过分词后出现频次为1的单词总数。

## TTR

型例比，设文本中词的例符总数为N，型符总数为 V。例符即文本总词数，型符即文本去重后的总词数。

## H-point

H点是文本中词的秩序分布的一个临界点。词的秩序分布是这样得到的：将文本中每个词（型符）按照其频次降序排序，并按照1到V的顺序依次编号。每个频序r都对应着一个频次值 f(r)。如下表 1是一个小规模的秩序分布：

表 1 秩序分布表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 35 | 16 | 7 | 3 | 3 | 2 | 2 | 1 |

所谓h点，即为词的秩序分布上r=f(r)的那个点。处于h点之前的词均满足r < f(r)，而h点之前的词均满足r > f(r)。在很多秩序分布上，可能找不到恰好满足r=f(r)的词。在这种情况下，h点实际上是处于频序相邻的某两个词之间。设其频序分别为r1和r2（r2>r1），则必有r1<f(r1)且r2>f(r2)。过(r1, f(r1)) 和 (r2, f(r2))两点的直线与直线y=x的交点即为h点。解方程即得：

以上面的秩序分布为例：

## 熵

熵（Entropy），是信息论中最基本的一个概念，在计量语言学中熵主要是基于文本中词的出现概率来计算。设在文本中任意一词的出现概率为 （即其频次与 N 之比），那么文本的熵即为[11]：

其中V为文本中的型符数，即文本中不重复的单词总数。

## 词汇丰富度

总体衡量词汇丰富性可以从三个角度进行分析：词汇多样性、词汇复杂性、词汇密度。词汇丰富度（vocabulary richness）[12]，是基于h点（文本中词的秩序分布中的一个临界点）对文本中的实词（例词）的比例的估计，属于词汇密度上的分析。词汇密度是从信息承载量方面解释词汇的丰富度，实词传达的信息量大，因而可认为实词数量越多，文本就越丰富。实际计算过程中，基于文本中词的秩序分布f(r)得出相应的累积概率分布（cumulative probability distribution）F(r)。F(r)为频序从1到r的词所累积的概率，也就是说这些词的总频次占文本例符总数的比例。如果h点之前的词均为功能词，那么F(h)就是功能词在文本中所占的比例。然而事实上，h点之前也可能存在实词，而h点之后也可能存在功能词，因此需要对值进行校正：

由此可得到词汇丰富度的计算公式：

## 重复率

重复率（RR），是衡量文本中词的丰富程度的一个计量指标。重复率越低，说明文本用词越丰富；反之，文本用词较为单一。重复率基于文本中词的出现概率进行计算，设文本中任意一词的出现概率为 （即其频次与N之比），那么文本的重复率为：

重复率反映的也是文本中词的丰富程度，但其取值与词的丰富程度成反比。

## 相对重复率

其中 RR 为2.5中介绍的重复率。RRmc可以大致视为对RR的标准化处理，其取值区间为[0, 1]，从而使不同文本的计算结果具有可比性。

## 主题集中度

主题集中度（thematic concentration），反映文本专注于某个主题的程度，基于文本中的主题词来计算。根据现有研究得出来的规律，h点之前的词大多为功能词，但仍有一部分实词，而这些实词往往与文本的主题有关，因此称为主题词。具体的计算过程为：设h点之前的任意实词的频序为r’，其频次为f(r)’。该词的主题性 Tr’等于其频次与权重(h-r’)之积除以h点之前所有词的权重之和 与最大频次（即f(1)）之积：

对 h 之前全部主题词（设其总数为T）的主题性求和，即得到文本主题集中度：

TC的值越大，文本就越有可能专注于某个主题。

## 次级主题集中度

次级主题集中度（secondary thematic concentration），是基于频序在h到2h之间的主题词来计算的，计算方法同TC。h点之前如果没有实词，那么文本的主题集中度TC就是0。此时就可以考虑采用次级中体集中度。

## 活动度

活动度（Activity），是文本中的动词总数占动词与形容词总数之和的比例。

## 描写度

描写度（Descriptivity），是文本中形容词总数占动词与形容词总数之和的比例。

## 文本中词的秩序分布的欧氏长度

文本中词的秩序分布的欧氏长度是所有相邻秩序的词之间的欧氏距离的总和：

L是对文本中词的秩频分布特征的一种几何描述。

## 文本中词的秩序分布的R指数

设秩频分布在h点前的部分的欧氏长度为，则R指数即为/L 的一个补数，即：

## 文本的Lambda值

文本中词的秩频分布的欧氏长度L会受到N的影响。而 的提出就是要在L的基础上找到一种描述词频结构的稳定指标。Λ 的计算方式如下：

消除了文本规模的影响，适用于诸如语言类型学和文体学领域的计量研究。

## 校正模数

校正模数（adjusted modules），是一种反映词频结构的指标。词频结构指的是文本中不同单词的出现频率和数量之间的关系，不同的词频结构可以反映出文本不同的特征和风格。而实际在评估词频结构的过程中，由于文本长度的差异可能会导致结果存在较大误差，因为通常较长的文本会包含更多的单词和词汇。而校正模数可以帮助消除文本长度对词频结构的影响。具体来说，校正模数是总单词数与不同单词数之间的比值。如果文本长度较短，校正模数比较小，说明文本中存在一些高频单词。如果文本长度较长，校正模数比较大，说明文本中存在一些低频词。校正模数的计算主要基于以下这个向量：

其中f(1)为词频的最大值，h为h点。向量P的模数为：

校正模数即是在M的基础上考虑N这一因素：

## 基尼系数

它是从经济学中借来用于描述文本中词的秩频分布特征的一个指标，计算方法如下：

如果将文本中词汇的丰富程度看作一个从绝对丰富到绝对不丰富之间的连续区间，那么基尼系数反映的就是文本中词汇的实际丰富程度在这两个极端之间的位置。

## 基尼系数补数

## 作者视野

前文提到，h点是实词与功能词的分界点。因此，可以将h点视作文本作者控制整个文本中实词与功能词之间平衡的立足点。h点与秩频分布起点与终点 组成一个三角形，其中以h点为顶点的角（记作 ）的度数（以弧度记）即为作者视野，该角的余弦值计算公式如下：

## 动词间距

动词间距（verb distance），指的是文本中动词之间的距离总和。计算的方法：找出文本中所有相邻的动词对（期间没有其他动词，可能在同一个句子中，也可能在不同的句子中），统计每对相邻动词之间的例符个数，并求均值。动词间距可能与文本的语言有关，也可能与文本的风格有关。

# 西方语种文本可读性指标

## ARI

ARI (Automated Readability Index) —— 自动可读性指数（ARI）是一种评估文本的可理解性的指标。它从字符数与单词数的比率以及单词数与句子数的比率来衡量文本的可读性。ARI 输出一个近似于 所需年级等级的数字。例如，如果 ARI 输出数字 3，则表示 3 年级（8-9 岁）的学生应该能够理解该文本。公式为：

这里characters指的是字母和数字总数，words指的是文本中单词的数量，而sentences指的是句子的数量

## RIX

RIX (Rate Index) —— RIX 可以适用于任何语言西方语言，RIX 考 虑的是文本中包含的“长”词，定义包含 6 个以上字符的单词为“长”词，RIX 输出是介于 0（非常容易）和 55+（非常困难）之间的分数。公式为：

## FleschReading

FRE (Flesch Reading Ease Formula) —— FRE 是一种评估读者的年级等级的简单方法。该评估方法主要从文本的平均句子长度和单词的平均音节数两个方面来衡量文本的可读性。Flesch 认为，句子越长，阅读起来就越困难，单词的发音越复杂，文本理解起来就越困难。公式为：

ASL指通过将单词总数除以句子总数后得到的平均句子长度，

ASW指通过将音节总数除以单词总数得到的每个单词的平均音节数。

FRE表示可读性得分。Flesch Reading Ease Formula 将输出从 0 到 100 的数字，得分越高表示阅读越容易。90-100 的分数表示“非常容易”， 80-89 的分数表“容易”，70-79 的分数表示“非常容易”，60-69 的 分数表示“标准”，50-59 的分数表示“非常困难”，30-49 的分数表示 “困难”，0-29 表示“非常令人困惑”。

## FlsechKincaidGrade

FKG (Flesch-Kincaid Grade Level) —— FKG 评估方法是 Kincaid 的 FRE 方法的改进。该方法更合理地描述了和在文本评分上的 关系。此方法最初是美国海军开发应用于教育领域。FKG 输出一个 美国学校的年级水平值，表明该年级水平的普通学生可以阅读此文 本。公式为：

ASL指将单词总数除以句子总数后得到的平均句子长度，ASW指将音 节总数除以单词总数得到的每个单词的平均音节数。9.3 分意味着九 年级学生可以阅读该文档。因为 FKG 认为最低的阅读水平是三年级， 所以其输出范围在 3-13 之间。

## GunningFog

GF (Gunning Fog Index) —— GF 与 Flesch 指标类似，因为它采用了音节和句子长度。FRE 和 FKG 使用平均字节数来表示文本中单词的难易程度，而 GF 直接计算文本中困难单词的比例，并定义 “Foggy”的单词。“Foggy”的单词是指包含 3 个或者更多音节的单词。得分为 5 表明文本是可读的，得分为 10 表示难度较低，得分为 15 表示难度中等，得分为 20 表示难度很大。公式为：

ASL指将单词总数除以句子总数后得到的平均句子长度，PHW是文本中“Foggy”词所占的百分数。使用 Fog 指标衡量可读性的理想得分是 7 或 8。任何得分高于 12 的文本对于大多数人来说都很难阅读。我们 将分数分为 14 个等级。其中，第 14 级代表最难.

## Smog

Smog 旨在衡量文本的可理解性，输出完全理解该文本所需的美国教

育年限。Smog 定义三个或三个以上的音节的单词为困难的单词。公式为：

## ColemanLiauIndex

CL (Coleman-Liau Index) —— CL 依赖于字符和句子的长度，而不

是每个单词的音节。它根据每一百个单词的平均字符数和句子数来 衡量文本的可读性。该公式将输出一个等级。例如，10.6 表示您的 文字适合 10-11 年级的高中学生。公式为：

L指的是每 100 个单词的平均字母数，S指的是每 100 个单词的平均句子数。CL 与 ARI 相似，但不同于其他大多数基于单个单词的音节进行可读性评估的指标。

## DaleChallIndex

DCI (Dale-Chall Index) —— DCI 是一种文本可读性评测，它提供了一种数字量度，可以用来衡量读者在阅读文本时遇到的理解难度。它使 用了 3000 个单词的列表，四年级的美国学生可以容易地理解这些单 词，并认为列表中没有的单词很困难。Dale-Chall Index 公式如下：

如果困难单词的百分比高于 5％，则将 3.6365 添加到原始分数以获得调整后的分数。困难单词是所有不在单词列表中的单词，但是必须考虑到单词列表包含动词和名词等基本形式，故必须添加名词的常规复数形式，常规的过去时形式，动词的渐进形式等等。

## LWIndex

LW (Linsear Write Formula) —— Linsear Write Formula 是英语文本的可读性公式，最初是为美国空军开发的，旨在帮助他们计算技术手册的可读性。Linsear Write Formula 是根据句子长度和三个以上音节单词的数量来计算文本样本的等级水平。LW 得分计算方法如表 2所示：

表 2 LW 得分计算方法

|  |
| --- |
| （1）计算容易的单词的数量（定义两个或两个以下的音节的单词 为容易的单词），并将每个简单的单词替换“1”，包括 a，an，the 和其他简单的单词。  （2）计算困难的单词的数量（定义三个或三个以上的音节的单词 为困难的单词），并将每个困难的单词替换“3”。  （3）用容易的单词数乘以“1”。  （4）用困难的单词数乘以“3”。  （5）将前两个数字加在一起。  （6）将（5）得到的总数除以句子数。  （7）如果（6）的答案大于 20，则除以“2”。  （8）如果（6）的答案小于 20 或等于 20，请减去“2”然后除以 “2”。 |

# 中文文本可读性指标

中文文本可读性指标相关可参考NLPCC2017 “Exploring the Impact of Linguistic Features for Chinese Readability Assessment”这篇论文，内部有详细具体特征的计算方式以及可读性回归方程。

# 文本处理

应用中基于Hanlp多任务模型进行多语种文本分词以及词性标注，采用的模型如图 1 所示（包含支持处理的语种列表，具体可参考https://hanlp.hankcs.com/docs/api/hanlp/pretrained/mtl.html）：

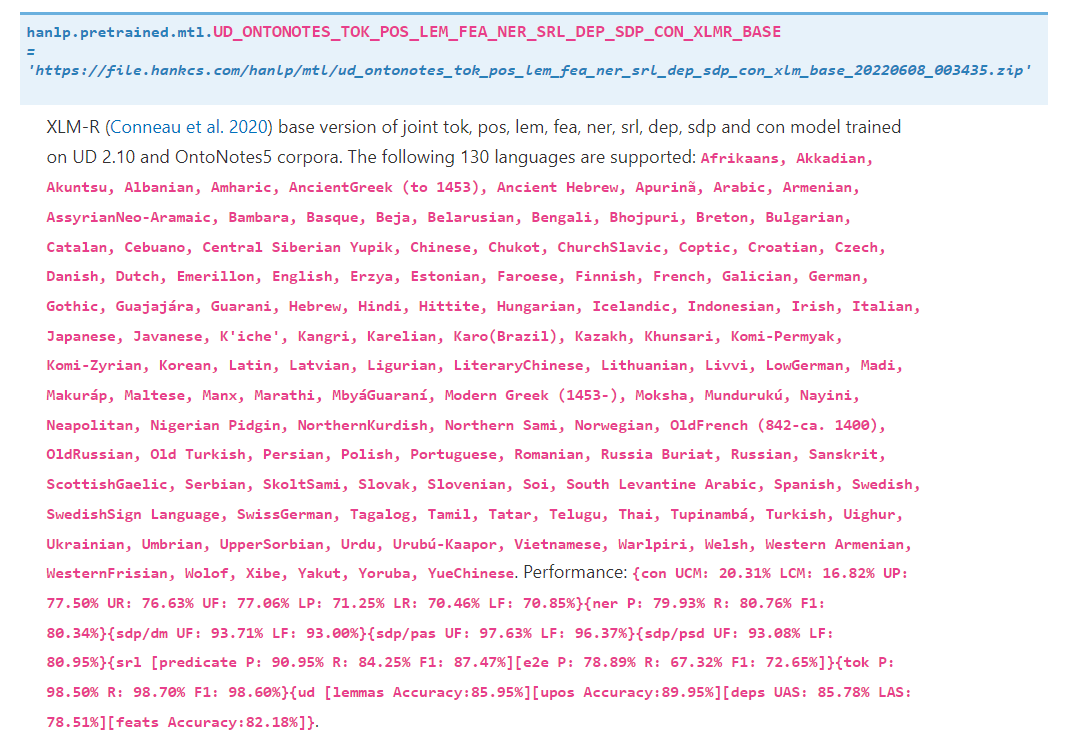


图 1 Hanlp多语种文本处理模型