语料库构建与数据预处理

文本数据是非结构化的数据，语料库由文本数据构成。文本数据的处理相对于实验数据而言在预处理部分更为复杂。首先我们要讲各种格式的文本数据导入R语言，然后要进行分句、分词，将文本数据进行结构化。

corpus-based

Tognini-Bonelli (2001) makes a useful dis- tinction between corpus-based and corpus-driven inves- tigations. The former uses a corpus as a source of examples to check researcher intuition or to examine the frequency and/or plausibility of the language contained within a smaller data set. The researcher does not question pre-existing traditional descriptive units and categories. A corpus-driven analysis is a more inductive process: the corpus itself is the data and the patterns in it are noted as a way of expressing regularities (and exceptions) in language. A corpus-driven analysis tends to only use minimal theoretical presuppositions about grammatical structure.

对于语料库研究而言，常用的经典研究方法是词表（wordlist）、关键词（keywords）、共现（concordance）。

Wordlist: a list of words occurring in a corpus, possibly with frequency information.

Frequency: The concept of frequency underpins much of the analytical work that is carried out within the remit of corpus linguistics.

Frequencies can be given as **raw data**, e.g. there are 58,860 occurrences of the word man in the British National Corpus (BNC); or (often more usefully) they can be given as **percentages or proportions** – man occurs 602.91 times per million words in the BNC – allowing comparisons between corpora of different sizes to be made.

**Frequency analyses** also allow comparisons to be made between different words in a corpus – for example man (602.91 per million) tends to occur more frequently than woman (225.43 per million), suggesting that man is the marked or ‘prototype’ term. On the other hand, homosexual (8.41 per million) occurs more than heterosexual (3.86 per million), which in this case is due to the term homosexual being marked because homosexuality has been considered problematical and non-standard by society in the past.

Frequency analyses can also be carried out on grammatical forms, for example to ascertain which past-tense verb forms are more common than their corresponding present- or future-tense forms or to compare different genres of language – for example what are the most (and least) common words in written and spoken English, or how has use of modal verbs shifted over time.

In addition, **word lists**, compiled by frequency counts of each word in a corpus can be used in order to derive **keyword lists**. Frequency counts are also used in the calculation of collocational and dispersion data as well as the type/token ratio of a corpus.

However, care must be taken when using frequency counts. Frequencies do not explain themselves: concordance-based analyses are therefore required in order to explain why certain words are more frequent than others. It may also be necessary to take into account the frequencies of related terms, for example chap, fella, bloke, gent etc. as well as man.

Concordance: an alphabetical index of a search pattern in a corpus, showing every contextual occurrence of the search pattern.

KWIC: key-word-in-context concordance.

Keyword: words in a corpus whose frequency is unusually high (positive keywords) or low (negative keywords) in comparison with a reference corpus.

Reference corpus: a balanced representative corpus balanced for general usage; in keyword analysis, a corpus that is used to provide a reference wordlist

keyword A potential issue with a keyword analysis is that a word may appear to be key simply because it occurs extremely frequently in a very small number of texts in a particular corpus. For example, if a corpus consists of 1,000 equal-sized files, and the word iron- monger only appears fifty-eight times in one single file called ‘The history of the ironmonger’, a keyword analysis may show this word to be key. In order to establish whether a word is key and representative of the corpus as a whole, a dispersion plot of the word across the corpus could be examined or a list of key keywords – words that are key in multiple texts in the corpus – could be calculated therefore allowing for the potentially skewing nature of disproportionate representation.

keyword

1. A word which appears in a text or corpus statistically significantly more frequently than would be expected by chance when compared to a corpus which is larger or of equal size. Usually log-likelihood or chi-squared tests are used to compare two word lists in order to derive keywords. Keywords can be calculated automatically using Scott’s WordSmith Tools program. Commonly found keywords include (1) proper nouns; (2) grammatical words that are often indicators of a particular stylistic profile; (3) lexical words that give an indication of the ‘aboutness’ of a text. For example, the following is a list of keywords found when comparing a small corpus of essays written by learners of English with the 1-million- word Frown Corpus of L1 American English: my, I, very, nice, good, big, holiday, dog, hobby, lovely, tenis, beatiful, voleyball. This list reveals to us what is distinctive about the learner essays – for instance, significant spelling errors (tenis, beatiful, voleyball), over-reliance on a small set of simple adjectives (nice, good, lovely, big), the main essay topics (dog, holi- day, hobby) and the fact that the essays are generally written as first person narratives (I, my). (See also key keywords.)

2. Any word that is considered ‘focal’ in a text, but not through statistical measures (see, for example, the discussion by Stubbs (1996: 166)).

3. A word which is made the subject of a concordance (see, for example, Kennedy 1998: 251).

本章将使用quanteda程序包讲解如何将文本数据导入R并转化成为语料库。

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. <https://doi.org/10.21105/joss.00774>

文本导入

文本数据有很多格式，在语言研究中，文件内部的一些结构信息对于数据分析非常重要。比如，莎士比亚的十四行诗，可以将每一行作为一个分析单位导入。而小说中的章节信息非常重要。新闻报道可以以篇为单位导入，但是如果数量很多，需要批量导入。下面我们分别介绍不同文体的导入方式。

Token: an occurrence of any given word form

Type: a word form

Type-token ratio: the ratio between types and tokens, useful when comparing samples of roughly equal length

下面我们展示3个案例分别为：由本地文本文件导入进行文本数据的整理加工，从网络获取的数据导入机加工，以及双语数据的对齐机。

Silge J, Robinson D (2016). “tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” *JOSS*, **1**(3). doi: [10.21105/joss.00037](https://doi.org/10.21105/joss.00037), <http://dx.doi.org/10.21105/joss.00037>.

gutenbergr package

*David Robinson (2017). gutenbergr: Download and Process Public Domain Works from Project Gutenberg. R package version 0.1.3.*

[*https://cran.rstudio.com/package=gutenbergr*](https://cran.rstudio.com/package=gutenbergr)

这个程序包可以帮你从古登堡项目（<https://www.gutenberg.org/>）的公共与当中下载并处理一些版权公开的经典作品。这个程序包中包含了1系列的嗯工具以及关于股东宝项目的元素局元数据你可以使用这些数据来找到一些你感兴趣用来进行文本分析的作品。

我们可以使用这个函数来下载书籍你只需要在函数中输入书籍在股东报项目当中的ID编号就可以了。而获取他的ID编号我们可以用这个数据集同时这个数据集可以帮助我们了解作家对一些信息比如他们的化名出生和死亡年月这个函数则包含了国会图书馆该书的分类以及主题内免。

A function gutenberg\_download() that downloads one or more works from Project Gutenberg by ID: e.g., gutenberg\_download(84) downloads the text of Frankenstein.

Metadata for all Project Gutenberg works as R datasets, so that they can be searched and filtered:

gutenberg\_metadata contains information about each work, pairing Gutenberg ID with title, author, language, etc

gutenberg\_authors contains information about each author, such as aliases and birth/death year

gutenberg\_subjects contains pairings of works with Library of Congress subjects and topics。“lcc” means Library of Congress Classification, while “lcsh” means Library of Congress subject headings

xml