R 語言
自然語言處理

王貿

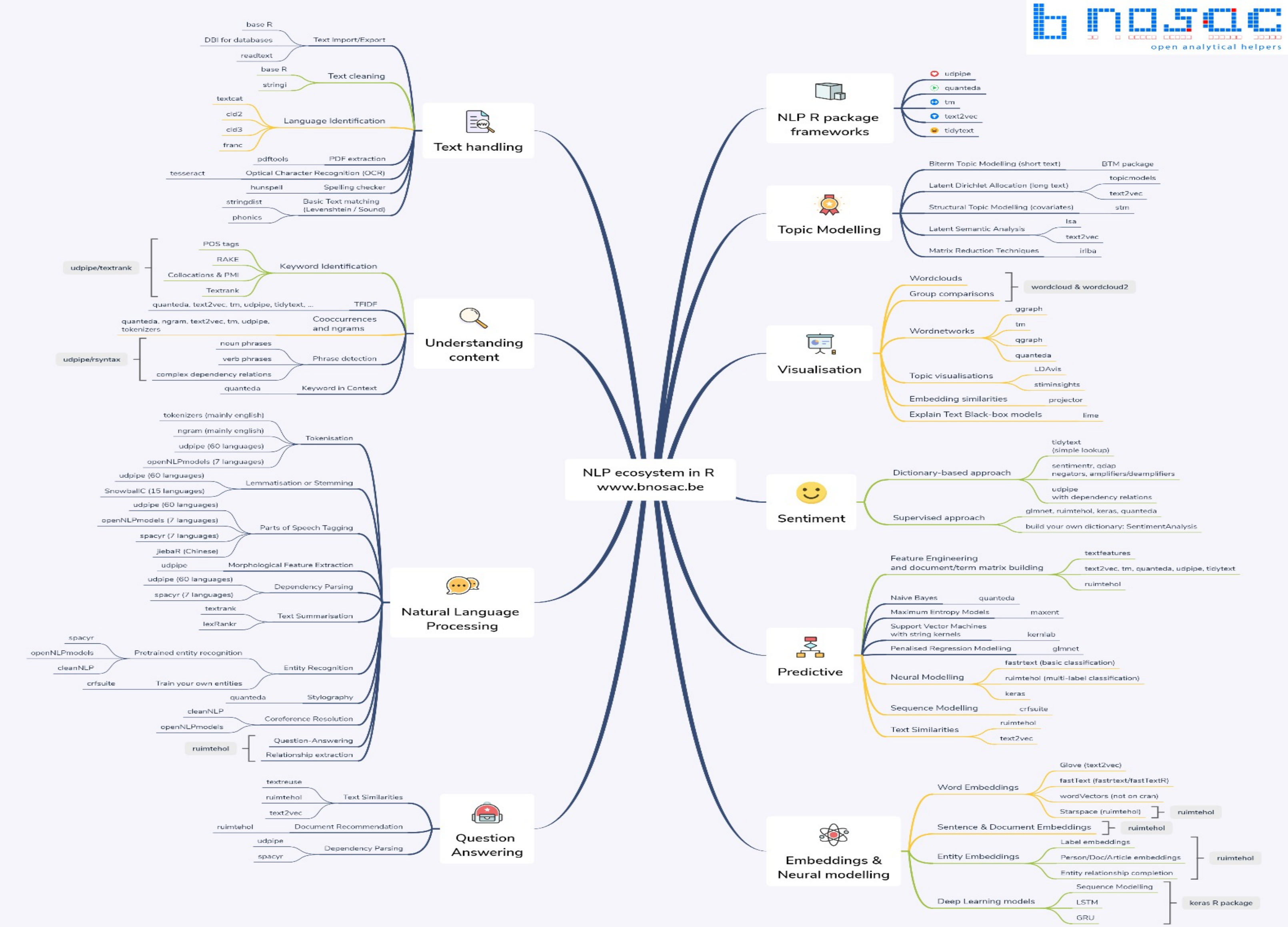
國立臺灣大學行為與資料科學研究中心助理研究員

國立臺灣大學政治學系博士、兼任講師

maowang01@gmail.com

課程主題重點

- 為什麼叫自然語言處理（NLP）？
- 詞幹提取（stemming）
- 詞形還原（lemmatization）
- 詞性標記（part of speech tagging）

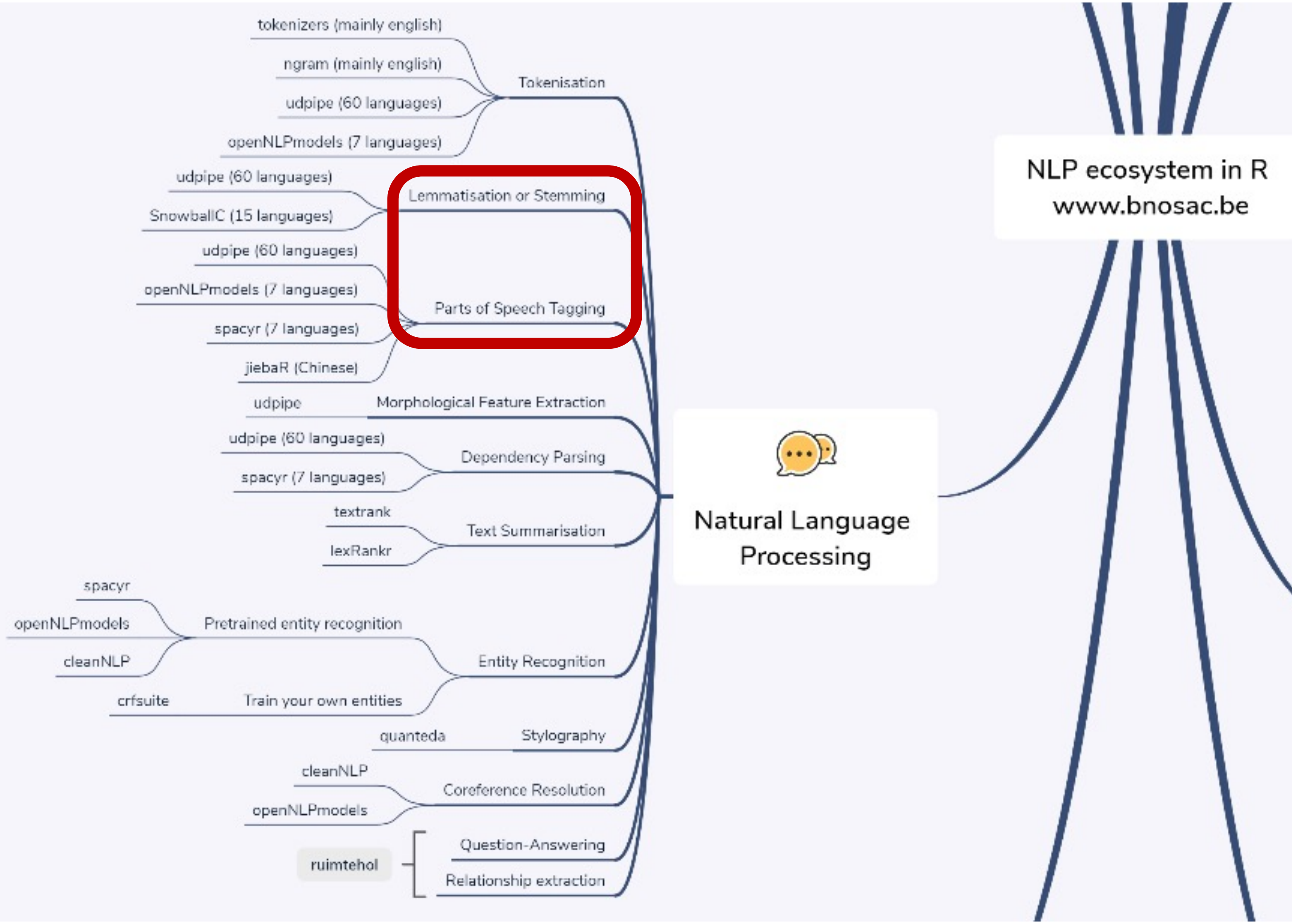


bnosac
open analytical helpers

NLP ecosystem in R
www.bnosac.be

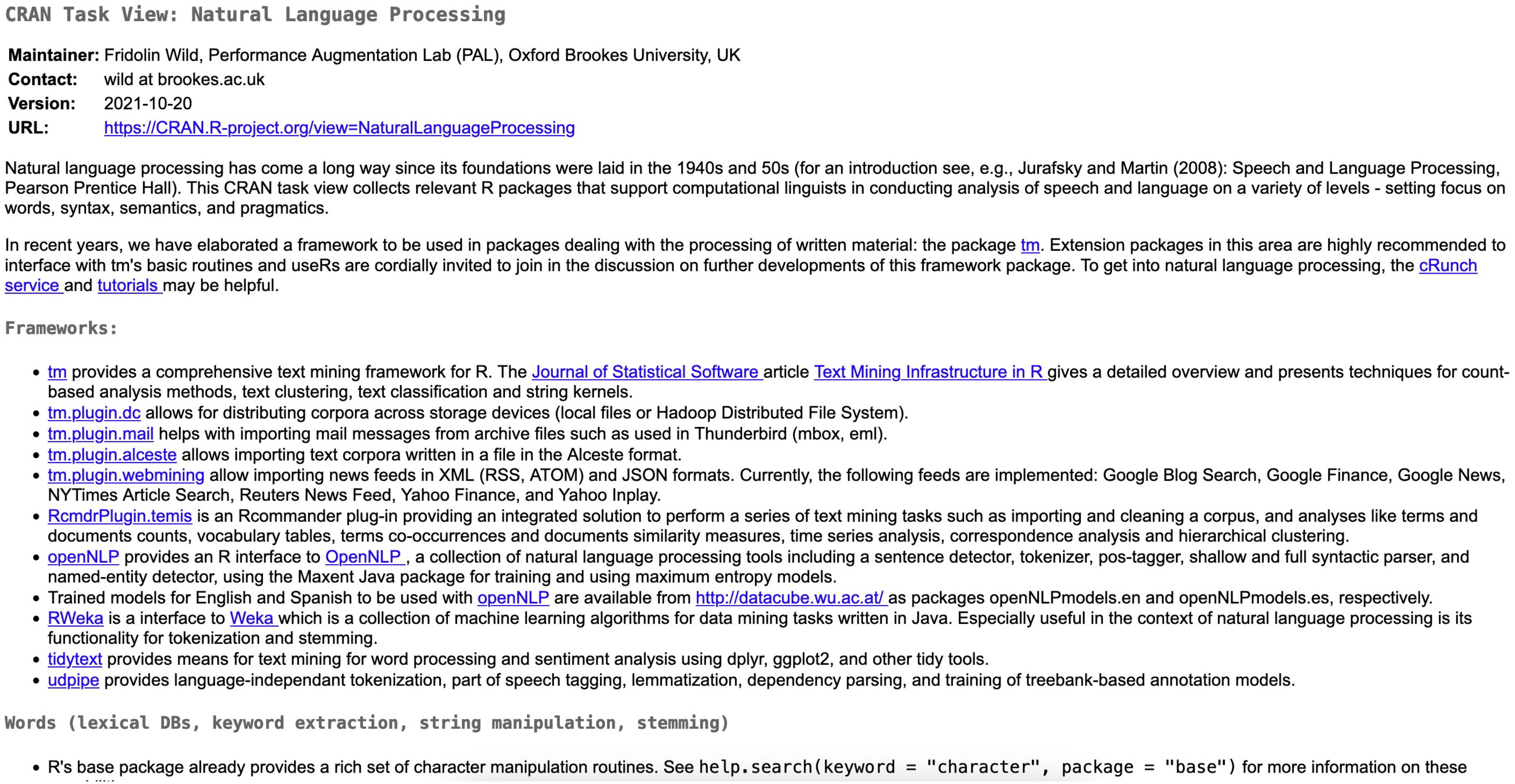
- Text handling
  - Text Import/Export
    - base R
    - DBI for databases
    - readtext
  - Text cleaning
    - base R
    - stringi
  - Language Identification
    - textcat
    - cld2
    - cld3
    - franc
  - PDF extraction
    - pdftools
  - Optical Character Recognition (OCR)
    - tesseract
  - Spelling checker
    - hunspell
  - Basic Text matching (Levenshtein / Sound)
    - stringdist
    - phonics
- Understanding content
  - Keyword Identification
    - POS tags
    - RAKE
    - Collocations & PMI
    - Textrank
    - udpipe/textrank
  - TFIDF
    - quanteda, text2vec, tm, udpipe, tidytext, ...
  - Cooccurrences and ngrams
    - quanteda, ngram, text2vec, tm, udpipe, tokenizers
  - Phrase detection
    - noun phrases
    - verb phrases
    - complex dependency relations
    - udpipe/rsyntax
  - Keyword in Context
    - quanteda
- Natural Language Processing
  - Tokenisation
    - tokenizers (mainly english)
    - ngram (mainly english)
    - udpipe (60 languages)
    - openNLPmodels (7 languages)
  - Lemmatisation or Stemming
    - udpipe (60 languages)
    - SnowballC (15 languages)
  - Parts of Speech Tagging
    - udpipe (60 languages)
    - openNLPmodels (7 languages)
    - spacyr (7 languages)
    - jiebaR (Chinese)
  - Morphological Feature Extraction
    - udpipe
  - Dependency Parsing
    - udpipe (60 languages)
    - spacyr (7 languages)
  - Text Summarisation
    - textrank
    - lexRankr
  - Entity Recognition
    - Pretrained entity recognition
      - spacyr
      - openNLPmodels
      - cleanNLP
    - Train your own entities
      - crfsuite
  - Stylography
    - quanteda
  - Coreference Resolution
    - cleanNLP
    - openNLPmodels
  - Question-Answering
  - Relationship extraction
    - ruimtehol
- Question Answering
  - Text Similarities
    - textreuse
    - ruimtehol
    - text2vec
  - Document Recommendation
    - ruimtehol
  - Dependency Parsing
    - udpipe
    - spacyr
- NLP R package frameworks
  - udpipe
  - quanteda
  - tm
  - text2vec
  - tidytext
- Topic Modelling
  - Biterm Topic Modelling (short text)
    - BTM package
  - Latent Dirichlet Allocation (long text)
    - topicmodels
    - text2vec
  - Structural Topic Modelling (covariates)
    - stm
  - Latent Semantic Analysis
    - lsa
    - text2vec
  - Matrix Reduction Techniques
    - irlba
- Visualisation
  - Wordclouds
  - Group comparisons
    - wordcloud & wordcloud2
  - Wordnetworks
    - ggraph
    - tm
    - qgraph
    - quanteda
  - Topic visualisations
    - LDAvis
    - stminsights
  - Embedding similarities
    - projector
  - Explain Text Black-box models
    - lime
- Sentiment
  - Dictionary-based approach
    - tidytext (simple lookup)
    - sentimentr, qdap negators, amplifiers/deamplifiers
    - udpipe with dependency relations
  - Supervised approach
    - glmnet, ruimtehol, keras, quanteda
    - build your own dictionary: SentimentAnalysis
- Predictive
  - Feature Engineering and document/term matrix building
    - textfeatures
    - text2vec, tm, quanteda, udpipe, tidytext
    - ruimtehol
  - Naive Bayes
    - quanteda
  - Maximum Entropy Models
    - maxent
  - Support Vector Machines with string kernels
    - kernlab
  - Penalised Regression Modelling
    - glmnet
  - Neural Modelling
    - fastrtext (basic classification)
    - ruimtehol (multi-label classification)
    - keras
  - Sequence Modelling
    - crfsuite
  - Text Similarities
    - ruimtehol
    - text2vec
- Embeddings & Neural modelling
  - Word Embeddings
    - Glove (text2vec)
    - fastText (fastrtext/fastTextR)
    - wordVectors (not on cran)
    - Starspace (ruimtehol)
      - ruimtehol
  - Sentence & Document Embeddings
    - ruimtehol
  - Entity Embeddings
    - Label embeddings
    - Person/Doc/Article embeddings
    - Entity relationship completion
    - ruimtehol
  - Deep Learning models
    - Sequence Modelling
    - LSTM
    - GRU
    - keras R package

https://www.bnosac.be/images/bnosac/blog/NLP-R-ecosystem.png



https://www.bnosac.be/images/bnosac/blog/NLP-R-ecosystem.png

視你的需求選擇工具



CRAN Task View: Natural Language Processing

Maintainer: Fridolin Wild, Performance Augmentation Lab (PAL), Oxford Brookes University, UK
Contact: wild at brookes.ac.uk
Version: 2021-10-20
URL: https://CRAN.R-project.org/view=NaturalLanguageProcessing

Natural language processing has come a long way since its foundations were laid in the 1940s and 50s (for an introduction see, e.g., Jurafsky and Martin (2008): Speech and Language Processing, Pearson Prentice Hall). This CRAN task view collects relevant R packages that support computational linguists in conducting analysis of speech and language on a variety of levels - setting focus on words, syntax, semantics, and pragmatics.

In recent years, we have elaborated a framework to be used in packages dealing with the processing of written material: the package tm. Extension packages in this area are highly recommended to interface with tm's basic routines and useRs are cordially invited to join in the discussion on further developments of this framework package. To get into natural language processing, the cRunch service and tutorials may be helpful.

Frameworks:

- tm provides a comprehensive text mining framework for R. The Journal of Statistical Software article Text Mining Infrastructure in R gives a detailed overview and presents techniques for count-based analysis methods, text clustering, text classification and string kernels.
- tm.plugin.dc allows for distributing corpora across storage devices (local files or Hadoop Distributed File System).
- tm.plugin.mail helps with importing mail messages from archive files such as used in Thunderbird (mbox, eml).
- tm.plugin.alceste allows importing text corpora written in a file in the Alceste format.
- tm.plugin.webmining allow importing news feeds in XML (RSS, ATOM) and JSON formats. Currently, the following feeds are implemented: Google Blog Search, Google Finance, Google News, NYTimes Article Search, Reuters News Feed, Yahoo Finance, and Yahoo Inplay.
- RcmdrPlugin.temis is an Rcommander plug-in providing an integrated solution to perform a series of text mining tasks such as importing and cleaning a corpus, and analyses like terms and documents counts, vocabulary tables, terms co-occurrences and documents similarity measures, time series analysis, correspondence analysis and hierarchical clustering.
- openNLP provides an R interface to OpenNLP , a collection of natural language processing tools including a sentence detector, tokenizer, pos-tagger, shallow and full syntactic parser, and named-entity detector, using the Maxent Java package for training and using maximum entropy models.
- Trained models for English and Spanish to be used with openNLP are available from http://datacube.wu.ac.at/ as packages openNLPmodels.en and openNLPmodels.es, respectively.
- RWeka is a interface to Weka which is a collection of machine learning algorithms for data mining tasks written in Java. Especially useful in the context of natural language processing is its functionality for tokenization and stemming.
- tidytext provides means for text mining for word processing and sentiment analysis using dplyr, ggplot2, and other tidy tools.
- udpipe provides language-independant tokenization, part of speech tagging, lemmatization, dependency parsing, and training of treebank-based annotation models.

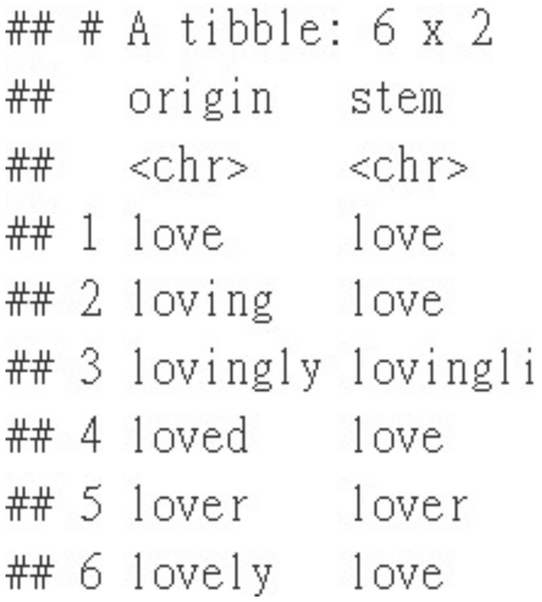
Words (lexical DBs, keyword extraction, string manipulation, stemming)

- R's base package already provides a rich set of character manipulation routines. See `help.search(keyword = "character", package = "base")` for more information on these

https://cran.r-project.org/web/views/NaturalLanguageProcessing.html

詞幹提取（stemming）

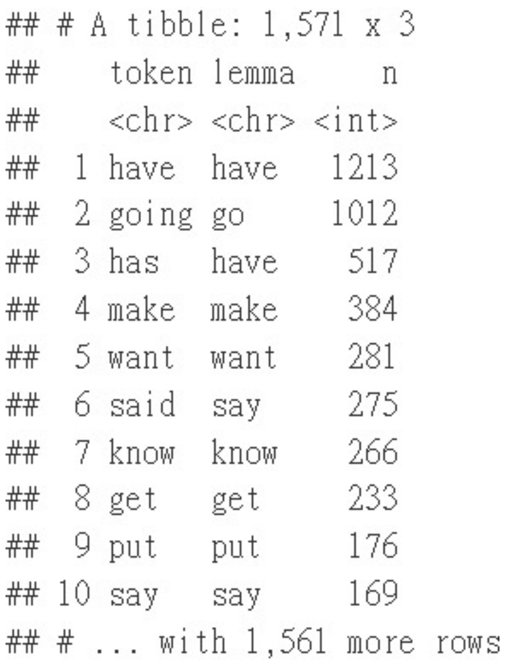
- 將文字的詞幹提取出來，所以詞幹常不會是完整的字。
- 就像是背英文時用的字根、字首、字尾的概念一樣。



```
## # A tibble: 6 x 2
##   origin   stem
##   <chr>    <chr>
## 1 love     love
## 2 loving   love
## 3 lovingly lovingli
## 4 loved    love
## 5 lover    lover
## 6 lovely   love
```

詞形還原（lemmatization）

- 詞形還原會還原語境脈絡下的詞形，如因時態、單複數、變形（比較級、最高級）等因素而改變的字。



```
## # A tibble: 1,571 x 3
##    token lemma      n
##    <chr> <chr> <int>
##  1 have  have   1213
##  2 going go     1012
##  3 has   have    517
##  4 make  make    384
##  5 want  want    281
##  6 said  say     275
##  7 know  know    266
##  8 get   get     233
##  9 put   put     176
## 10 say   say     169
## # ... with 1,561 more rows
```