FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Distributional Semantics in R with the wordspace Package

Stefan Evert
Corpus Linguistics Group, FAU Erlangen-Nürnberg

Motivation

Research on distributional semantic models (DSMs) is an empirical science, given our limited understanding of DSM parameters and their relation to different aspects of word meaning. Flexible, efficient and easy-to-use software packages are essential for advancing the field and making distributional semantics accessible to a wide range of users and application developers.

HiDEx	C++	*re-implementation of a specific model (HAL)*
Semantic Vectors	Java	*representation based on random indexing*
S-Space	Java	*complex object-oriented framework*
JoBimText	Java	*UIMA / Hadoop framework*
Gensim	Python	*complex framework, focus on scalability*
DISSECT	Python	*user-friendly, focus on compositional semantics*
wordspace	R	*interactive research laboratory, but scalable*

Features

- Minimalist approach: small set of carefully designed functions
  - encapsulate non-trivial procedures in user-friendly manner
  - provide efficient C implementations of key operations
- Harness built-in power of R and 5000+ add-on packages for matrix algebra, statistics, machine learning, data analysis, visualization and shiny GUIs
- Example DSMs and evaluation tasks included or available for download
- Input formats
  - compressed triplet file (*target*, *feature*, *score*)
  - native sparse matrix format exported from UCS toolkit (serves as hub)
  - term-document model from R package `tm` (text mining)
- Feature scaling: many association scores*, transformation, normalization
- Metrics: cosine/angle, Euclidean, Manhattan, Minkowski, Dice*, KL*
- Dimensionality reduction: (randomized) SVD, RI, sparse SVD*, NMF*
- Easy subsetting & merging of DSMs
- Nearest neighbours, pair distances, centroid vectors, . . .
- Evaluation tasks: multiple choice, similarity ratings, clustering

*work in progress

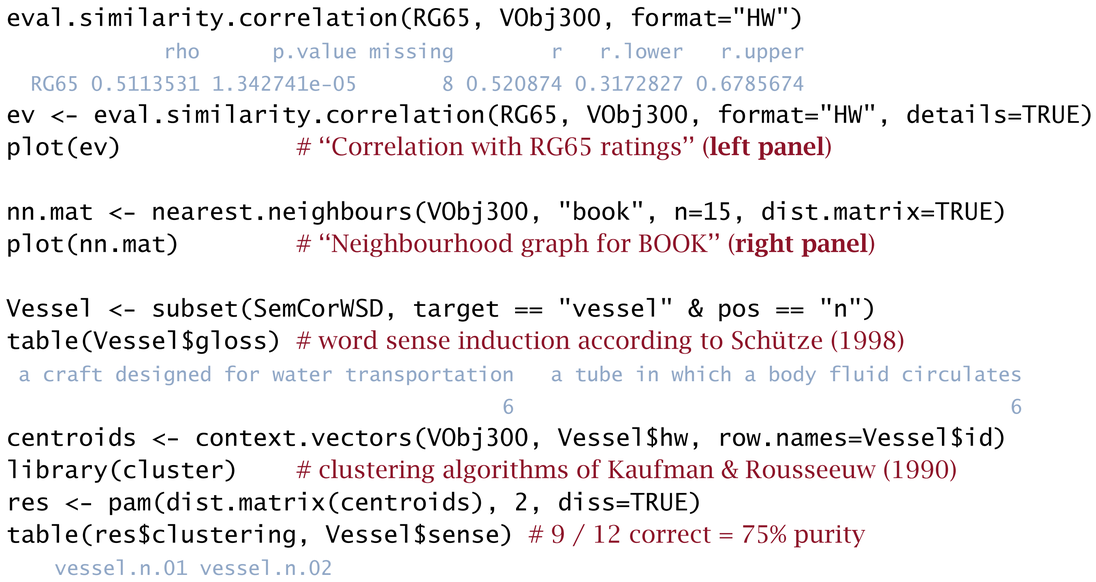
Benchmarks

	wordspace	DISSECT
build model from triples file	186.0 s	503.3 s
save model	57.6 s	1.5 s
file size (`.rda` / `.pkl`)	228.9 MB	725.7 MB
normalize row vectors	0.5 s	1.3 s
SVD projection to 300 latent dimensions	353.6 s	296.6 s
save latent vectors	10.4 s	0.4 s
file size (`.rda` / `.pkl`)	71.5 MB	185.0 MB
20 nearest neighbours (full matrix)	119 ms	1269 ms
20 nearest neighbours (300 dims)	10 ms	92 ms
cosine similarity (full matrix)	4 ms	< 1 ms
cosine similarity (300 dims)	< 1 ms	< 1 ms

An example session

noun	rel	verb	f	mode
dog	subj	bite	3	spoken
dog	subj	bite	12	written
dog	obj	bite	4	written
dog	obj	stroke	3	written
...	...	...	...	...

```
library(wordspace)

Triples <- subset(DSM_VerbNounTriples_BNC, mode == "written")
VObj <- dsm(target=Triples$noun, feature=Triples$verb, score=Triples$f,
            raw.freq=TRUE, sort=TRUE)

VObj <- subset(VObj, nnzero >= 3, nnzero >= 3, recursive=TRUE)
dim(VObj)  # make sure there are ≥ 3 nonzero elements in each row and column
  [1] 12428  3735

VObj <- dsm.score(VObj, score="simple-ll", transform="log", normalize=TRUE)

VObj300 <- dsm.projection(VObj, method="rsvd", n=300, oversampling=4)

pair.distances("book", "paper", VObj300, method="cosine", convert=FALSE)
  book/paper
   0.7322982
nearest.neighbours(VObj300, "book", n=15) # defaults to angular distance
     paper     novel  magazine     works   article textbook     guide      poem
  42.92059 48.03492  49.10742 49.33028  49.54836 49.82660  50.29588 50.37111
     essay  leaflet   edition      text  pamphlet  booklet catalogue
  50.45991 50.53009  50.78630 50.95731  51.12786 51.21351  52.43824

eval.similarity.correlation(RG65, VObj300, format="HW")
            rho      p.value missing        r   r.lower   r.upper
  RG65 0.5113531 1.342741e-05       8 0.520874 0.3172827 0.6785674
ev <- eval.similarity.correlation(RG65, VObj300, format="HW", details=TRUE)
plot(ev)              # "Correlation with RG65 ratings" (left panel)

nn.mat <- nearest.neighbours(VObj300, "book", n=15, dist.matrix=TRUE)
plot(nn.mat)          # "Neighbourhood graph for BOOK" (right panel)

Vessel <- subset(SemCorWSD, target == "vessel" & pos == "n")
table(Vessel$gloss) # word sense induction according to Schütze (1998)
 a craft designed for water transportation   a tube in which a body fluid circulates
                                         6                                          6
centroids <- context.vectors(VObj300, Vessel$hw, row.names=Vessel$id)
library(cluster)     # clustering algorithms of Kaufman & Rousseeuw (1990)
res <- pam(dist.matrix(centroids), 2, diss=TRUE)
table(res$clustering, Vessel$sense) # 9 / 12 correct = 75% purity
   vessel.n.01 vessel.n.02
1            2           5
2            4           1
```



http://wordspace.r-forge.r-project.org/

pre-compiled binaries available from CRAN — GPL v3 license