List of Publicly Available Data on Pre-Trained Word Embeddings (File Format: `.RData`)

Download Link: Google Drive Cloud Storage

R Package for Processing: PsychWordVec

Author: Han-Wu-Shuang Bao

(Update: March 2022)

Source	Algorithm	Corpus	Language	Vocabulary	Filename (`*.RData`)
GloVe	GloVe	Wikipedia + Gigaword	English	400,000	glove_wiki_(50|100|200|300)d
		Twitter		1,193,514	glove_twitter_(25|50|100|200)d
		Common Crawl [1]		1,560,516	glove_commoncrawl_300d
				1,837,608	glove_commoncrawl_300d_cased
Google	word2vec (SGNS)	Google News	English [2]	878,327	word2vec_googlenews_eng_1word
				1,266,655	word2vec_googlenews_eng_2words
				573,228	word2vec_googlenews_eng_3words
				63,247	word2vec_googlenews_eng_nwords
HistWords	word2vec (SGNS)	Google Books (V2) (in *decades*, not *years*)	English (1800s~1990s)	13,045 … 71,097	sgns_eng_1800 … sgns_eng_1990
			English (fiction) (1800s~1990s)	686 … 24,049	sgns_eng-fiction_1800 … sgns_eng-fiction_1990
			Chinese (1950s~1990s)	2,790 … 14,496	sgns_chi_1950 … sgns_chi_1990
			French (1800s~1990s)	10,878 … 26,539	sgns_fre_1800 … sgns_fre_1990
			German (1800s~1990s)	807 … 19,614	sgns_ger_1800 … sgns_ger_1990
		COHA (Corpus of Historical American English) (in *decades*, not *years*)	American English (1810s~2000s)	1,216 … 15,141	sgns_coha_1810 … sgns_coha_2000
				1,321 … 12,065	sgns_coha-lemma_1810 … sgns_coha-lemma_2000
Chinese-Word-Vectors	word2vec (SGNS)	Baidu Encyclopedia (百度百科)	Chinese [3]	299,065 421,462	sgns_baidubaike_word sgns_baidubaike_bigram-char
		Wikipedia (zh) (中文维基百科)		352,217 352,272	sgns_wiki_word sgns_wiki_bigram-char
		People’s Daily News (人民日报)		355,987 356,053	sgns_renmin_word sgns_renmin_bigram-char
		Sogou News (搜狗新闻)		364,990 365,113	sgns_sogou_word sgns_sogou_bigram-char
		Financial News (金融新闻)		467,370 467,211	sgns_financial_word sgns_financial_bigram-char
		Zhihu QA (知乎问答)		259,922 259,753	sgns_zhihu_word sgns_zhihu_bigram-char
		Sina Weibo (新浪微博)		195,202 195,197	sgns_weibo_word sgns_weibo_bigram-char
		Literature (文学作品)		187,959 187,980	sgns_literature_word sgns_literature_bigram-char
		Si Ku Quan Shu (四库全书) [古文]		19,527	sgns_sikuquanshu_word (*character*)
		Mixed-Large (综合)		566,017 865,918	sgns_merge_word sgns_merge_bigram-char

Note. All raw data files have been transformed into the `.RData` format using the R function `PsychWordVec::data_transform()`.
Filenames in red are datasets that involve case-sensitive words. Unless otherwise noted, all word vectors have 300 dimensions (300d).
Regular expression is used to exclude invalid “words” (e.g., meaningless numbers, punctuation) for overlarge datasets.

[1] Words have been filtered by regular expression `[A-Za-z]` to include only English words (83% of the raw vocabulary).

[2] Words have been filtered by regular expression `[A-Za-z0-9_]` to include English words (raw vocabulary: 3,000,000).
Multiple words (i.e., phrases) are separated and joined by _ in the raw data (e.g., “Hong_Kong”, “Steve_Jobs”).

[3] Word vectors have been trained based on context features of word only (“_word”) or word + ngram + character (“_bigram-char”).
The latter appears to be more reasonable than the former, if we scrutinize the most similar words of some words (e.g., “中国”).

SGNS = Skip-Gram with Negative Sampling (an algorithm of word2vec).