a-paxton / **Gensim-LSI-Word-Similarities**

Generate word-word similarities from Gensim's latent semantic indexing (Python)

#word2vec #gensim #language #text-analysis #latent-semantic-indexing #latent-semantic-analysis

22 commits | 1 branch | 0 releases | 1 contributor | MIT

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

a-paxton Create license — Latest commit 64c637b on Jan 10, 2017

| | | |
|---|---|---|
| README.md | Readme update | 2 years ago |
| license | Create license | a year ago |
| word2vec_vect_sim_fun.py | Add word2vec similarity function | 2 years ago |
| word_pair_similarity_matrix.py | Comment edits | a year ago |
| wordsim_fun.py | Updated comments | 3 years ago |
| wordvectsim_fun.py | Remove stray square bracket | a year ago |

**README.md**

# Gensim-LSI-Word-Similarities

Two simple little functions to create word-word similarities from Gensim's latent semantic indexing in Python. Both functions produce an inverted cosine similarity score (0 = low, 1 = high) between two words in a Gensim-generated LSA/LSI space across the total number of dimensions specified in the creation of the model (i.e., *num_topics* from *gensim.models.LsiModel*).

Both require Gensim, Pandas, and SciPy.

Includes four functions:

- **wordsim**: Create cosine-derived similarity score (from 0-1) between individual words. Input:
  - *word1* (string or string variable)
  - *word2* (string or string variable)
  - *target_dictionary* (Gensim-created LSI dictionary)
  - *target_lsi_model* (Gensim-created LSI model)

- **wordvectsim**: Same as *wordvect* but created to calculate similarity scores (from 0-1) for word pairs in a 2-dimensional word vector (e.g., using *numpy.apply_along_axis*). Input:
  - *word_vector2d* (2D string vector or 2D string vector variable)
  - *target_dictionary* (Gensim-created LSI dictionary)
  - *target_lsi_model* (Gensim-created LSI model)

- Two additional functions/series of functions added (detailed documentation available in each function and will be added here soon):
  - **word2vec_vect_sim_fun**: similarity score function for gensim's word2vec
  - **word_pair_similarity_matrix**: word-word similarity matrix function for gensim's LSI (LSA) model