Bachelor Thesis Documentation

Release v1.0

Bachelor Thesis

Jul 27, 2016

CONTENTS

1 Bachelorarbeit **3**

1.1 src package . . . 3

1.1.1 Subpackages . . . 3

src.clustering package . . . 3

Submodules . . . 3

src.clustering.cluster_mappings module . . . 3

Module contents . . . 5

src.eval package . . . 5

Submodules . . . 5

src.eval.analogy module . . . 5

src.eval.eval_vectors module . . . 5

src.eval.word_similarity module . . . 6

Module contents . . . 7

src.mapping package . . . 7

Submodules . . . 7

src.mapping.mapthreading module . . . 7

Module contents . . . 11

src.misc package . . . 11

Submodules . . . 11

src.misc.helpers module . . . 11

Module contents . . . 12

src.prep package . . . 12

Subpackages . . . 12

Module contents . . . 15

src.trans_e package . . . 15

Submodules . . . 15

src.trans_e.add_inverse_relations module . . . 15

src.trans_e.clean_relations module . . . 15

src.trans_e.contains_entities module . . . 15

src.trans_e.convert_relations module . . . 16

src.trans_e.differentiate_datasets module . . . 16

src.trans_e.partition_data module . . . 16

src.trans_e.trans_we module . . . 16

Module contents . . . 17

1.1.2 Module contents . . . 17

2 Indices and tables **19**

Python Module Index **21**

Index **23**

Contents:

CHAPTER

ONE

BACHELORARBEIT

1.1 src package

1.1.1 Subpackages

src.clustering package

Submodules

src.clustering.cluster_mappings module

Script to cluster mapping vectors created with *src.mapping.mapthreading* .

src.clustering.cluster_mappings. **aggregate_cluster** (*points*, *labels*)

Arranges all clusters in a list, where a sublist with all points at index i corresponds with the custer with label i.

Parameters

- **points** (*list*) – List of datapoints
- **labels** (*list*) – List of unique cluster labels

Returns list of lists of datapoints belonging to the i-th cluster

Return type list

src.clustering.cluster_mappings. **cluster_mappings** (*vector_inpath*, *do_pca=False*, *target_dim=100*, *indices_inpath=None*, *epsilon=2.625*, *min_s=20*)

Cluster mapping vectors created with *src.mapping.mapthreading* or rc.mapping.map_vectors.py . Because just reading about the number of clusters and their sizes, there's an option to resolve the indices of the vectors in the cluster to their original word pairs.

Parameters

- **vector_inpath** (*str*) – Path to vector file. File should have the following format (separated by spaces): <index of original vector #1> <index of original vector #2> <Dimension 1> ... <Dimension n>
- **do_pca** (*bool*) – Flag to indicate whether PCA should be executed before clustering to reduce amount of
- **computation.** –

- **target_dim** (*int*) – Number of dimensions vectors should be shrunk to in case PCA is performed.
- **indices_inpath** (*str*) – Path to file with the indices given to words. The file should have the following format: <index of word> <word> (separated by tab)
- **epsilon** (*float*) – Radius of circle DBSCAN uses to look for other data points.
- **min_s** (*int*) – Minimum number of points in radius epsilon DBSCAN needs to declare a point a core object.

src.clustering.cluster_mappings. **get_cluster_size** (*labels*)
Calculate the size of every cluster found by DBSCAN.

Parameters **labels** (*list*) – List of cluster IDs assigned to every data point.

Returns Dictionary of cluster sizes with cluster id as key and cluster size as value.

Return type defaultdict

src.clustering.cluster_mappings. **init_argparser** ( )
Initialize all possible arguments for the argument parser.

Returns ArgumentParser object with command line arguments for this script.

Return type argparse.ArgumentParser

src.clustering.cluster_mappings. **load_indices** (*indices_inpath*)
Load word indices from a file. The file should have the following format: <index of word> <word> (separated by tab)

Parameters **indices_inpath** (*str*) – Path to index file.

src.clustering.cluster_mappings. **load_mappings_from_model** (*mapping_inpath*)
Load mapping vectors from file.

Parameters **mapping_inpath** – Path mapping vector file.

Returns A tuple of a list of word index pairs and a dictionary (defaultdict) with index pair tuple as key and mapping vector (as numpy.array) as value.

Return type tuple

src.clustering.cluster_mappings. **main** ( )
This is the main function. It uses the parsed command line arguments to pick the right function to execute.

src.clustering.cluster_mappings. **resolve_indices** (*points*, *labels*, *indices_inpath*, *model*)

src.clustering.cluster_mappings. **train_clustering_parameters** (*vector_inpath*)
Functions that tries to figure out the optimal clustering parameters in regard to DBSCAN's epsilon, min_samples and p.

Parameters **vector_inpath** (*str*) – Path to vector file. File has to have the following format (separated by spaces): <index of original vector #1> <index of original vector #2> <Dimension 1> ... <Dimension n>

Module contents

src.eval package

Submodules

src.eval.analogy module

Module to evaluate word embeddings by the means of analogies like "W is to X like Y is to Z". Usually, the system uses the word embeddings of word W, X, Y and tries to find the vector of word Z that is most similar to X and Y and most dissimilar to W. Therefore, the CosMul method (Levy et al., 2015) is used.

The whole module is used in `src.eval.eval_vectors.py`.

`src.eval.analogy.` **`analogy_eval`** (*vector_inpath*, *analogy_path*, *per_section=False*)

Perform analogy evaluation. Usually, the system uses the word embeddings of word W, X, Y and tries to find the vector of word Z that is most similar to X and Y and most dissimilar to W for an analogy like "W is to X like Y is to Z." Therefore, the CosMul method (Levy et al., 2015) is used.

Parameters

- **`vector_inpath`** (*str*) – Path to *word2vec* vector file.
- **`analogy_path`** (*str*) – Path to analogy file.
- **`per_section`** (*bool*) – Flag to indicate whether analogies test should be conducted section-wise or just all in one run.

`src.eval.analogy.` **`read_analogies`** (*analogy_path*, *per_section=False*)

Reads a file with analogies.

Parameters

- **`analogy_path`** (*str*) – Path to analogy file.
- **`per_section`** (*bool*) – Flag to indicate whether analogies test should be conducted section-wise or just all in one run. In this function, the section will be put into a data structure accordingly.

Returns Dictionary with section header as key, list of analogy as 4-tuples as value.

Return type dict

src.eval.eval_vectors module

Main module used to evaluate word embeddings. It offers the following options:

1.) Analogy: The system tries to complete an analogy like "W is to X like Y is to...?" The percentage of correct answers is measured.

2.) Word similarity: The system assign word pairs a similarity score based on the cosine similarity of their word embeddings. Then, to correlation between those and human ratings is measured with Pearson's rho.

3.) Nearest neighbors: Find the nearest neighbors for a list of words based on their word embeddings. Good for a first look on the data, but not quantifiable.

4.) Visualize: Plot word embeddings in 2D or 3D. Fancy plots. Yay!

`src.eval.eval_vectors.` **`find_nearest_neighbors`** (*vector_inpath*, *max_n*, *wordlist*)

Find the nearest neighbors for a list of words based on their word embeddings.

Parameters

- **vector_inpath** (*str*) – Path to vector file. File has to have the following format (separated by spaces): <index of original vector #1> <index of original vector #2> <Dimension 1> ... <Dimension n>
- **max_n** (*int*) – Number of nearest neighbors that should be determined.
- **wordlist** (*list*) – List of words nearest neighbors should be found for.

src.eval.eval_vectors. **init_argparser** ( )
Initialize all possible arguments for the argument parser.

Returns ArgumentParser object with command line arguments for this script.

Return type argparse.ArgumentParser

src.eval.eval_vectors. **main** ( )
This is the main function. It uses the parsed command line arguments, especially *–mode*, to pick the right function to execute.

src.eval.eval_vectors. **plot** (*vector_inpath*, *max_n*, *target_dim*, *show_plot=False*, *display_names=False*)
Plot word embeddings in 2D or 3D. As a heuristic, word will only be plotted after the 50th most frequent words to avoid plotting boring stop words.

Parameters

- **vector_inpath** (*str*) – Path to vector file. File has to have the following format (separated by spaces): <index of original vector #1> <index of original vector #2> <Dimension 1> ... <Dimension n>
- **max_n** (*int*) – Maximum number of vectors to be plotted.
- **show_plot** (*bool*) – Flag to indicate whether a window with the (interactive) plot should pop up after executing the script.
- **display_names** (*bool*) – Flag to indicate whether the words should acutally be shown next to the data point in the plot. Can get very messy with higher *max_n*.

src.eval.word_similarity module

Module used to conduct the word similarity evaluation. The system assign word pairs a similarity score based on the cosine similarity of their word embeddings. Then, to correlation between those and human ratings is measured with Pearson's rho.

The whole module is used in src.eval.eval_vectors.py .

src.eval.word_similarity. **evaluate_wordpair_sims** (*x*, *y*, *number_of_pairs*)
Evaluate results of the similarity score assignments, i.e. calculate pearson's rho and its significance.

Parameters

- **x** (*list*) – List of similarity scores assigned by humans.
- **y** (*list*) – List of similarity scores assigned by the system.
- **number_of_pairs** (*int*) – Number of word pairs evaluated.

Returns **rho** – Pearson's correlation coefficient. t (float): Student's t value. z (float): z value.

Return type float

src.eval.word_similarity. **read_wordpairs** (*wordpair_path*, *format='google'*)
Read wordpair file with wordpairs and their similarity scores assigned by humans.

Parameters

- **wordpair_path** (*str*) – Path to word pair file.
- **format** (*str*) – Format of wor pair file {google|semrel}

Returns Tuple of a list of word pairs and a list of similarity scores for those same pair assigned by humans.

Return type tuple

src.eval.word_similarity. **remove_unknowns** (*x*, *y*)

Remove word pairs from the results where one or two word embedding weren't found.

Parameters

- **x** (*list*) – List of similarity scores assigned by humans.
- **y** (*list*) – List of similarity scores assigned by the system.

Returns x – Purged list of similarity scores assigned by humans. y (list): Purged list of similarity scores assigned by the system.

Return type list

src.eval.word_similarity. **word_sim_eval** (*vector_inpath*, *wordpair_path*, *format='google'*)

Function that let's the system assign word pairs a similarity score based on the cosine similarity of their word embeddings. Then, to correlation between those and human ratings is measured with Pearson's rho.

Parameters

- **vector_inpath** (*str*) – Path to vector file. File has to have the following format (separated by spaces): <index of original vector #1> <index of original vector #2> <Dimension 1> ... <Dimension n>
- **wordpair_path** (*str*) – Path to word pair file.
- **format** (*str*) – Format of word pair file {google|semrel}

Module contents

src.mapping package

Submodules

src.mapping.mapthreading module

Module used to map a pair of vectors into a new combined vector space. Those mappings will be created by multiple threads in a master-slave-pattern. To do so, the user can choose between different vector operations as offset, cosine similarity, euclidean distance and many more.

Warning: Because of $\Omega = \frac{n(n-1)}{2}$, it is recommended to use the co-occurrence constraint Λ, which limits the calculations to word embedding pairs which words occurred together in a corpus in at least *n* sentences (but it will still take quite a while).

class src.mapping.mapthreading. **MappingMasterThread** (*n*, *vector_inpath*, *vector_outpath*, *features*, *lambda_*, *ids_inpath*, *indices_inpath*)

Bases: threading.Thread

Master thread class. The master thread loads all necessary data into suitable data structures and distributes them among all worker threads.

prepare ( )

Loads The master thread loads all necessary data into suitable data structures. To be more specific, word embeddings, sentence IDs and word indices are processed.

read_ids_file (*ids_inpath*)

Read the sentence ID file.

Parameters **ids_inpath** (*str*) – Path to sentence IDs file. The file should be in the following *YAML*-format: - <word>:

- <sentence id>
- <sentence id>

...

Returns Dictionary with words as keys and the IDs of the sentences they occur in in a set as value.

Return type defaultdict

start_threads ( )

Starts all the threads (and ends them if they're all finished).

class src.mapping.mapthreading. **MappingWorkerThread** (*worker_id*, *vector_dict*, *vector_queue*, *vector_outpath*, *features*, *occurrences*, *indices*, *lambda_*)

Bases: threading.Thread

Worker thread class. The worker threads do all the dirty work after they receive all necessary data from the master thread an try to calculate every possible combinations of two word embeddings in a dataset.

All the word embeddings will be stores in a dictionary (*VectorDict* as well as a Queue). An idle thread picks a new vector from the queue and then starts to iterate over all the vectors in the *VectorDict* (this way, the queue gets shorter over time while the size of the dictionary stays fixed).

Before it starts calculations, it checks a) if the co-occurrence constraint is satisfied and b) if this combination of word embeddings has already been processed.

cosine_similarity (*v1*, *v2*)

Calculates the cosine similarity ($cos(\vec{v}_1, \vec{v}_2) \in [-1, -1]$) between two vectors.

Parameters

- **v1** (*numpy.array*) – First vector
- **v2** (*numpy.array*) – Second vector

Returns Cosine similarity between the two vectors.

Return type float

distance (*v1*, *v2*)

Return the vector offset of two vectors:

Parameters

- **v1** (*numpy.array*) – First vector
- **v2** (*numpy.array*) – Second vector

Returns numpy.array: Vector offset.

euclidean_distance1 (*v1*, *v2*)

Return the euclidean distance between two vectors.

$$eucl(\vec{a}, \vec{b}) = \sqrt{\sum_{i=1}^{n} (\vec{b}_i - \vec{a}_i)^2}$$

Parameters

- **v1** (*numpy.array*) – First vector
- **v2** (*numpy.array*) – Second vector

Returns Euclidean distance between the two vectors.

Return type float

euclidean_distance2 (*v1*, *v2*)

Returns the squared euclidean distance between two vectors.

$$eucl2(\vec{a}, \vec{b}) = \sum_{i=1}^{n} (\vec{b}_i - \vec{a}_i)^2$$

Parameters

- **v1** (*numpy.array*) – First vector
- **v2** (*numpy.array*) – Second vector

Returns Squared euclidean distance between the two vectors.

Return type float

hash_indices (*i1*, *i2*)

Combines two vector indices (the indices of the words' embeddings used in vector operations) into a hash s.t. threads can do an easy lookup if a mapping vector has already been calculated. To guarantee this, $h(i_1, i_2) = h(i_2, i_1)$ has to be the case.

Parameters

- **i1** (*int*) – Index of first word's embedding
- **i2** (*int*) – Index of second word's embedding

Returns Unique hash for index pair.

Return type int

manhattan_distance (*v1*, *v2*)

Returns the manhattan distance between two vectors.

$$manhattan(\vec{a}, \vec{b}) = \sum_{i=1}^{n} |\vec{b}_i - \vec{a}_i|$$

Parameters

- **v1** (*numpy.array*) – First vector
- **v2** (*numpy.array*) – Second vector

Returns Manhattan distance between the two vectors.

Return type float

run ( )
Starts a worker thread.

soft_cosine_similarity (*v1*, *v2*)
Calculates the soft cosine similarity between two vectors.

$$S = \begin{bmatrix} eucl(\vec{a}_1, \vec{b}_1) & \dots & eucl(\vec{a}_1, \vec{b}_n) \\ \vdots & \ddots & \vdots \\ eucl(\vec{a}_n, \vec{b}_1) & \dots & eucl(\vec{a}_n, \vec{b}_n) \end{bmatrix}$$

$$softcos(\vec{a}, \vec{b}) = \frac{\sum_{i,j}^{N} S_{ij} \vec{a}_i \vec{b}_j}{\sqrt{\sum_{i,j}^{N} S_{ij} \vec{a}_i \vec{a}_j} \sqrt{\sum_{i,j}^{N} S_{ij} \vec{b}_i \vec{b}_j}}$$

(It considers the similarity between pairs of features.)

Parameters

- **v1** (*numpy.array*) – First vector
- **v2** (*numpy.array*) – Second vector

Returns Soft cosine similarity between the two vectors.

Return type float

class src.mapping.mapthreading. **VectorDict**
Bases: object

VectorDict class that serves two functions:

1.) Storing word embeddings so they don't allocate memory for every worker thread
2.) Providing a set, where are processed vector pairs are stored so no redundant computations are made.

Locks are used for synchronization purposes.

add_skippable (*index_hash*)
Add the hash of an index pair to a set of already processed vector pairs.

Parameters **index_hash** (*int*) – Hash value of index pair. Produced with hash_indices() .

add_vector (*index*, *vector*)
Add a new word embedding.

Parameters

- **index** (*int*) – Index of the word the embedding belongs to.
- **vector** (*numpy.array*) – Word embedding corresponding to given index.

get_keys ( )
Get all the keys (word embedding IDs) of this dictionary.

Returns List of word embedding IDs.

Return type list

get_vector (*index*)
Get a word embedding given its word's index.

Parameters **index** (*int*) – Index of the word the embedding belongs to.

Returns Word embedding corresponding to given index.

Return type numpy.array

skippable (*index_hash*)
Checks whether a pair of vectors has already been processed.

Parameters **index_hash** (*int*) – Hash value of index pair. Produced with `hash_indices()`.

Returns Whether a pair of vectors has already been processed.

Return type bool

src.mapping.mapthreading. **alt** (*func*)
Prepends the local time to the output of a function.

Parameters **func** (*function*) – Function the local time should be prepended to.

src.mapping.mapthreading. **init_argparse** ( )
Initialize all possible arguments for the argument parser.

Returns ArgumentParser object with command line arguments for this script.

Return type `argparse.ArgumentParser`

src.mapping.mapthreading. **main** ( )
Main function that initializes the master thread with command line arguments and starts it.

Module contents

src.misc package

Submodules

src.misc.helpers module

src.misc.helpers. **alt** (*func*)
Prepends the local time to the output of a function.

Parameters **func** (*function*) – Function the local time should be prepended to.

src.misc.helpers. **capitalize** (*word*)

src.misc.helpers. **load_vectors_from_model** (*vector_inpath*, *max_n=None*, *logpath=None*, *indices=False*)

Module contents

src.prep package

Subpackages

src.prep.corpus package

Submodules

src.prep.corpus.convert_to_plain module

src.prep.corpus.convert_to_plain. **alt** (*func*)

src.prep.corpus.convert_to_plain. **contains_tag** (*line*)

src.prep.corpus.convert_to_plain. **convert_decow_to_plain** (*decow_dir*, *out_dir*, *log_path*, *merge_nes*, *log_interval*)

src.prep.corpus.convert_to_plain. **convert_part** (*argstuple*)

src.prep.corpus.convert_to_plain. **convert_part_merging** (*argstuple*)

src.prep.corpus.convert_to_plain. **extract_named_entity** (*line*)

src.prep.corpus.convert_to_plain. **extract_sentence_id** (*tag*)

src.prep.corpus.convert_to_plain. **get_file_number** (*filename*)

src.prep.corpus.convert_to_plain. **log_time** (*logpath='log.txt'*, *interval=5*)

src.prep.corpus.convert_to_plain. **log_time_mp** (*logpath='log.txt'*, *interval=5*)

src.prep.corpus.convert_to_plain. **main** ( )

src.prep.corpus.extract_conll module

src.prep.corpus.extract_conll. **extract_conll** (*inpath*, *outpath*, *column*)
Extract information out of CoNLL files.

Parameters **inpath** (*str*) – Path to input file.

src.prep.corpus.extract_conll. **init_argparse** ( )

src.prep.corpus.extract_conll. **main** ( )

src.prep.corpus.mapper module

src.prep.corpus.prepare_corpus module

src.prep.corpus.prepare_corpus. **construct_yaml_str** (*self*, *node*)
Docstring for git testing purposes.

src.prep.corpus.prepare_corpus. **get_file_number** (*filename*)

src.prep.corpus.prepare_corpus. **init_pool** (*args*)

src.prep.corpus.prepare_corpus. **main** ( )

src.prep.corpus.prepare_corpus. **prepare** (*ne_inpath*)

src.prep.corpus.prepare_corpus. **process_corpora** (*nesi*, *corpus_dir*, *out_dir*, *log_dir*, *logging_interval_meta=10*, *logging_interval_processing=10*)

src.prep.corpus.prepare_corpus. **process_corpus** (*argstuple*)

src.prep.corpus.quick_and_dirty module

src.prep.corpus.quick_and_dirty. **clean_file** (*inpath*, *outpath*)

src.prep.corpus.reducer module

Module contents

src.prep.misc package

Submodules

src.prep.misc.decorators module

src.prep.misc.decorators. **alt** (*func*)

src.prep.misc.decorators. **log_time** (*logpath='log.txt'*, *interval=5*)

src.prep.misc.decorators. **log_time_mp** (*logpath='log.txt'*, *interval=5*)

Module contents

src.prep.nes package

Submodules

src.prep.nes.extractNE module

src.prep.nes.extractNE. **contains_tag** (*line*)

src.prep.nes.extractNE. **extract_named_entity** (*line*)

src.prep.nes.extractNE. **extract_sentence_id** (*tag*)

src.prep.nes.extractNE. **main** ( )

src.prep.nes.extractNE. **print_dict_in_file** (*dictionary*, *out_path*)

src.prep.nes.extractNE. **print_ids_in_file** (*dictionary*, *out_path*)

src.prep.nes.extractNE. **print_list_in_file** (*ne_list*, *out_path*)

src.prep.nes.extractNE. **process** (*inpath*, *outpath*, *logpath*)

src.prep.nes.merge module

src.prep.nes.merge. **dump_ids_dict** (*idsdict*, *outpath*)

src.prep.nes.merge. **freqWorker** (*inpath*)

src.prep.nes.merge. **idWorker** (*inpath*)

src.prep.nes.merge. **load_ids_dict** (*inpath*)

src.prep.nes.merge. **main** ( )

src.prep.nes.merge. **mergeDicts** (*dicttuple*)

src.prep.nes.merge. **merge_frequency_files** (*infiles_path*, *outpath*, *logpath*)

src.prep.nes.merge. **merge_id_dicts** (*dicttuple*)

src.prep.nes.merge. **merge_id_files** (*infiles_path*, *outpath*, *logpath*, *yaml=False*)

src.prep.nes.merge. **print_key_lengths** (*dictionary*)

src.prep.nes.merge. **rl** (*infile*)

src.prep.nes.mwe module

src.prep.nes.mwe. **create_mwe_pickle** (*inpath*, *outpath*, *logpath='./mwes.log'*)

src.prep.nes.mwe. **create_mwe_pickle2** (*inpath*, *outpath*, *logpath='./mwes.log'*)

src.prep.nes.mwe. **dump_dict_pickle** (*d*, *outpath*)

src.prep.nes.mwe. **dump_dict_pickle2** (*d*, *outpath*)

src.prep.nes.mwe. **load_dict_pickle** (*inpath*)

src.prep.nes.mwe. **load_dict_pickle2** (*inpath*)

src.prep.nes.mwe. **main** ( )

src.prep.nes.mwe. **replace_mwes** (*mwe_path*, *corpus_path*, *out_path*)

src.prep.nes.statistics module

src.prep.nes.statistics. **calculate_occurrences** (*freqpath*, *relations_path*)

src.prep.nes.statistics. **main** ( )

Module contents

src.prep.relations package

Submodules

src.prep.relations.relations module

exception src.prep.relations.relations. **MissingTranslationException**

Bases: exceptions.Exception

get_id ( )

src.prep.relations.relations. **fetch_name** (*id*, *lang='en'*)

src.prep.relations.relations. **fetch_relation_triples_of_file** (*inpath*, *outpath*, *logpath*, *lang='en'*)

src.prep.relations.relations. **format_fbid** (*id*)

src.prep.relations.relations. **freebase_request** (*query*, *api_key*, *service_url*)

src.prep.relations.relations. **main** ( )

src.prep.relations.relations. **read_credentials** ( )

src.prep.relations.relations. **rl** (*infile*)

src.prep.relations.relations. **translate_name** (*name*, *lang='en'*)

src.prep.relations.relations. **translate_word2vec_question_phrases** (*inpath*, *outpath*, *lang='en'*)

Module contents

Module contents

src.trans_e package

Submodules

src.trans_e.add_inverse_relations module

src.trans_e.add_inverse_relations. **add_inverse_relations** (*relations_inpath*, *relations_outpath*, *inverse_relations*, *known_relations*)

src.trans_e.add_inverse_relations. **init_argparse** ( )

src.trans_e.add_inverse_relations. **main** ( )

src.trans_e.add_inverse_relations. **read_file_with_inverse_relations** (*inverse_inpath*)

src.trans_e.clean_relations module

src.trans_e.contains_entities module

src.trans_e.contains_entities. **contains_entities** (*entities1*, *entities2*)

src.trans_e.contains_entities. **create_new_dataset** (*entities1*, *dataset*, *outpath*)

src.trans_e.contains_entities. **extract_entities_from_relation_dataset** (*dataset_inpath*)

src.trans_e.contains_entities. **extract_entities_from_tql_file** (*tql_path*)

src.trans_e.contains_entities. **format_fbid** (*id*)

src.trans_e.contains_entities. **init_argparse** ( )

src.trans_e.contains_entities. **main** ( )

src.trans_e.convert_relations module

src.trans_e.differentiate_datasets module

src.trans_e.differentiate_datasets. **compare_entities** (*set1*, *set2*)

src.trans_e.differentiate_datasets. **init_argparse** ( )

src.trans_e.differentiate_datasets. **main** ( )

src.trans_e.differentiate_datasets. **read_dataset** (*inpath*)

src.trans_e.partition_data module

src.trans_e.partition_data. **check_data_integrity** (*data_inpath*, *remove_clones*, *outpath*)

Check whether all triplets in the data are unique.

src.trans_e.partition_data. **check_set_integrity** (*indir*)

src.trans_e.partition_data. **get_stats** (*data*)

src.trans_e.partition_data. **init_argparse** ( )

src.trans_e.partition_data. **main** ( )

src.trans_e.partition_data. **partition_data** (*data*, *prts*, *outdir*, *whole=True*)

src.trans_e.partition_data. **partition_relation_wise** (*data*, *prts*)

src.trans_e.partition_data. **partition_whole** (*data*, *prts*)

src.trans_e.partition_data. **partitions_list** (*l*, *prts*)

src.trans_e.partition_data. **read_only_relations_into_set** (*inpath*)

src.trans_e.partition_data. **read_relations** (*inpath*)

src.trans_e.partition_data. **write_data_in_file** (*data*, *outfile*)

src.trans_e.trans_we module

src.trans_e.trans_we. **convert_data** (*sets_path*, *tql_inpath*, *vector_inpath*)

src.trans_e.trans_we. **create_corrupt_triples** (*grouped_pairs*, *entities*)

src.trans_e.trans_we. **dump_relation_vectors** (*relation_vectors*, *outpath*)

src.trans_e.trans_we. **evaluate** (*model*, *grouped_test*, *relation_vectors*, *entities*)

src.trans_e.trans_we. **extract_data_from_uri** (*uri*)

src.trans_e.trans_we. **get_rank** (*target*, *ranks*)

src.trans_e.trans_we. **init_argparser** ( )

Initialize all arguments for an ArgumentParser object and return it.

@returns {ArgumentParser} argument parser object

src.trans_e.trans_we. **load_relation_vectors** (*inpath*)

src.trans_e.trans_we. **load_vectors** (*vector_inpath*)
@param vector_inpath: Path to word2vec model file

src.trans_e.trans_we. **main** ( )

src.trans_e.trans_we. **prepare_training** (*sets_path*, *vector_inpath*)

src.trans_e.trans_we. **rank_entities** (*reference*, *solution*, *model*, *entities*)

src.trans_e.trans_we. **read_freebase_data** (*sets_path*)

src.trans_e.trans_we. **read_freebase_file** (*fb_inpath*)

src.trans_e.trans_we. **read_tql_file** (*tql_inpath*)

src.trans_e.trans_we. **test_coverage** (*triples*, *model*)
Test the coverage of a dataset consisting of freebase triples on word2vec word embeddings. For every triple (h, l, t), the entities h and t are taken and used for look up in the word2vec model.

@param triples: list of 3-tuples (freebase triples) @param model: gensim word2vec model

src.trans_e.trans_we. **train** (*model*, *grouped_train*, *grouped_corrupted*, *lossf*, *relation_types*, *epochs=1000*, *learning_rate=0.01*, *margin=1*)

src.trans_e.trans_we. **transform_triples** (*triples*, *relation_types*, *entities*)

src.trans_e.trans_we. **write_data** (*triples*, *found_entities*, *outpath*)

Module contents

1.1.2 Module contents

CHAPTER

TWO

INDICES AND TABLES

- genindex
- modindex
- search

S

src, 17
src.clustering, 5
src.clustering.cluster_mappings, 3
src.eval, 7
src.eval.analogy, 5
src.eval.eval_vectors, 5
src.eval.word_similarity, 6
src.mapping, 11
src.mapping.mapthreading, 7
src.misc, 12
src.misc.helpers, 11
src.prep, 15
src.prep.corpus, 13
src.prep.corpus.convert_to_plain, 12
src.prep.corpus.extract_conll, 12
src.prep.corpus.mapper, 12
src.prep.corpus.prepare_corpus, 12
src.prep.corpus.quick_and_dirty, 13
src.prep.corpus.reducer, 13
src.prep.misc, 13
src.prep.misc.decorators, 13
src.prep.nes, 14
src.prep.nes.extractNE, 13
src.prep.nes.merge, 14
src.prep.nes.mwe, 14
src.prep.nes.statistics, 14
src.prep.relations, 15
src.prep.relations.relations, 14
src.trans_e, 17
src.trans_e.add_inverse_relations, 15
src.trans_e.clean_relations, 15
src.trans_e.contains_entities, 15
src.trans_e.convert_relations, 16
src.trans_e.differentiate_datasets, 16
src.trans_e.partition_data, 16
src.trans_e.trans_we, 16

A

add_inverse_relations() (in module src.trans_e.add_inverse_relations), 15
add_skippable() (src.mapping.mapthreading.VectorDict method), 10
add_vector() (src.mapping.mapthreading.VectorDict method), 10
aggregate_cluster() (in module src.clustering.cluster_mappings), 3
alt() (in module src.mapping.mapthreading), 11
alt() (in module src.misc.helpers), 11
alt() (in module src.prep.corpus.convert_to_plain), 12
alt() (in module src.prep.misc.decorators), 13
analogy_eval() (in module src.eval.analogy), 5

C

calculate_occurrences() (in module src.prep.nes.statistics), 14
capitalize() (in module src.misc.helpers), 11
check_data_integrity() (in module src.trans_e.partition_data), 16
check_set_integrity() (in module src.trans_e.partition_data), 16
clean_file() (in module src.prep.corpus.quick_and_dirty), 13
cluster_mappings() (in module src.clustering.cluster_mappings), 3
compare_entities() (in module src.trans_e.differentiate_datasets), 16
construct_yaml_str() (in module src.prep.corpus.prepare_corpus), 12
contains_entities() (in module src.trans_e.contains_entities), 15
contains_tag() (in module src.prep.corpus.convert_to_plain), 12
contains_tag() (in module src.prep.nes.extractNE), 13
convert_data() (in module src.trans_e.trans_we), 16
convert_decow_to_plain() (in module src.prep.corpus.convert_to_plain), 12
convert_part() (in module src.prep.corpus.convert_to_plain), 12
convert_part_merging() (in module src.prep.corpus.convert_to_plain), 12
cosine_similarity() (src.mapping.mapthreading.MappingWorkerThread method), 8
create_corrupt_triples() (in module src.trans_e.trans_we), 16
create_mwe_pickle() (in module src.prep.nes.mwe), 14
create_mwe_pickle2() (in module src.prep.nes.mwe), 14
create_new_dataset() (in module src.trans_e.contains_entities), 15

D

distance() (src.mapping.mapthreading.MappingWorkerThread method), 8
dump_dict_pickle() (in module src.prep.nes.mwe), 14
dump_dict_pickle2() (in module src.prep.nes.mwe), 14
dump_ids_dict() (in module src.prep.nes.merge), 14
dump_relation_vectors() (in module src.trans_e.trans_we), 16

E

euclidean_distance1() (src.mapping.mapthreading.MappingWorkerThread method), 9
euclidean_distance2() (src.mapping.mapthreading.MappingWorkerThread method), 9
evaluate() (in module src.trans_e.trans_we), 16
evaluate_wordpair_sims() (in module src.eval.word_similarity), 6
extract_conll() (in module src.prep.corpus.extract_conll), 12
extract_data_from_uri() (in module src.trans_e.trans_we), 16
extract_entities_from_relation_dataset() (in module src.trans_e.contains_entities), 15
extract_entities_from_tql_file() (in module src.trans_e.contains_entities), 15
extract_named_entity() (in module src.prep.corpus.convert_to_plain), 12
extract_named_entity() (in module src.prep.nes.extractNE), 13
extract_sentence_id() (in module src.prep.corpus.convert_to_plain), 12

extract_sentence_id() (in module src.prep.nes.extractNE), 13

F

fetch_name() (in module src.prep.relations.relations), 15
fetch_relation_triples_of_file() (in module src.prep.relations.relations), 15
find_nearest_neighbors() (in module src.eval.eval_vectors), 5
format_fbid() (in module src.prep.relations.relations), 15
format_fbid() (in module src.trans_e.contains_entities), 15
freebase_request() (in module src.prep.relations.relations), 15
freqWorker() (in module src.prep.nes.merge), 14

G

get_cluster_size() (in module src.clustering.cluster_mappings), 4
get_file_number() (in module src.prep.corpus.convert_to_plain), 12
get_file_number() (in module src.prep.corpus.prepare_corpus), 12
get_id() (src.prep.relations.relations.MissingTranslationException method), 14
get_keys() (src.mapping.mapthreading.VectorDict method), 10
get_rank() (in module src.trans_e.trans_we), 16
get_stats() (in module src.trans_e.partition_data), 16
get_vector() (src.mapping.mapthreading.VectorDict method), 10

H

hash_indices() (src.mapping.mapthreading.MappingWorkerThread method), 9

I

idWorker() (in module src.prep.nes.merge), 14
init_argparse() (in module src.mapping.mapthreading), 11
init_argparse() (in module src.prep.corpus.extract_conll), 12
init_argparse() (in module src.trans_e.add_inverse_relations), 15
init_argparse() (in module src.trans_e.contains_entities), 15
init_argparse() (in module src.trans_e.differentiate_datasets), 16
init_argparse() (in module src.trans_e.partition_data), 16
init_argparser() (in module src.clustering.cluster_mappings), 4
init_argparser() (in module src.eval.eval_vectors), 6
init_argparser() (in module src.trans_e.trans_we), 16
init_pool() (in module src.prep.corpus.prepare_corpus), 12

L

load_dict_pickle() (in module src.prep.nes.mwe), 14
load_dict_pickle2() (in module src.prep.nes.mwe), 14
load_ids_dict() (in module src.prep.nes.merge), 14
load_indices() (in module src.clustering.cluster_mappings), 4
load_mappings_from_model() (in module src.clustering.cluster_mappings), 4
load_relation_vectors() (in module src.trans_e.trans_we), 16
load_vectors() (in module src.trans_e.trans_we), 16
load_vectors_from_model() (in module src.misc.helpers), 11
log_time() (in module src.prep.corpus.convert_to_plain), 12
log_time() (in module src.prep.misc.decorators), 13
log_time_mp() (in module src.prep.corpus.convert_to_plain), 12
log_time_mp() (in module src.prep.misc.decorators), 13

M

main() (in module src.clustering.cluster_mappings), 4
main() (in module src.eval.eval_vectors), 6
main() (in module src.mapping.mapthreading), 11
main() (in module src.prep.corpus.convert_to_plain), 12
main() (in module src.prep.corpus.extract_conll), 12
main() (in module src.prep.corpus.prepare_corpus), 12
main() (in module src.prep.nes.extractNE), 13
main() (in module src.prep.nes.merge), 14
main() (in module src.prep.nes.mwe), 14
main() (in module src.prep.nes.statistics), 14
main() (in module src.prep.relations.relations), 15
main() (in module src.trans_e.add_inverse_relations), 15
main() (in module src.trans_e.contains_entities), 16
main() (in module src.trans_e.differentiate_datasets), 16
main() (in module src.trans_e.partition_data), 16
main() (in module src.trans_e.trans_we), 17
manhattan_distance() (src.mapping.mapthreading.MappingWorkerThread method), 9
MappingMasterThread (class in src.mapping.mapthreading), 7
MappingWorkerThread (class in src.mapping.mapthreading), 8
merge_frequency_files() (in module src.prep.nes.merge), 14
merge_id_dicts() (in module src.prep.nes.merge), 14
merge_id_files() (in module src.prep.nes.merge), 14
mergeDicts() (in module src.prep.nes.merge), 14
MissingTranslationException, 14

P

partition_data() (in module src.trans_e.partition_data), 16
partition_relation_wise() (in module src.trans_e.partition_data), 16
partition_whole() (in module src.trans_e.partition_data), 16
partitions_list() (in module src.trans_e.partition_data), 16
plot() (in module src.eval.eval_vectors), 6
prepare() (in module src.prep.corpus.prepare_corpus), 13
prepare() (src.mapping.mapthreading.MappingMasterThread method), 8
prepare_training() (in module src.trans_e.trans_we), 17
print_dict_in_file() (in module src.prep.nes.extractNE), 13
print_ids_in_file() (in module src.prep.nes.extractNE), 13
print_key_lengths() (in module src.prep.nes.merge), 14
print_list_in_file() (in module src.prep.nes.extractNE), 13
process() (in module src.prep.nes.extractNE), 13
process_corpora() (in module src.prep.corpus.prepare_corpus), 13
process_corpus() (in module src.prep.corpus.prepare_corpus), 13

R

rank_entities() (in module src.trans_e.trans_we), 17
read_analogies() (in module src.eval.analogy), 5
read_credentials() (in module src.prep.relations.relations), 15
read_dataset() (in module src.trans_e.differentiate_datasets), 16
read_file_with_inverse_relations() (in module src.trans_e.add_inverse_relations), 15
read_freebase_data() (in module src.trans_e.trans_we), 17
read_freebase_file() (in module src.trans_e.trans_we), 17
read_ids_file() (src.mapping.mapthreading.MappingMasterThread method), 8
read_only_relations_into_set() (in module src.trans_e.partition_data), 16
read_relations() (in module src.trans_e.partition_data), 16
read_tql_file() (in module src.trans_e.trans_we), 17
read_wordpairs() (in module src.eval.word_similarity), 6
remove_unknowns() (in module src.eval.word_similarity), 7
replace_mwes() (in module src.prep.nes.mwe), 14
resolve_indices() (in module src.clustering.cluster_mappings), 4
rl() (in module src.prep.nes.merge), 14
rl() (in module src.prep.relations.relations), 15
run() (src.mapping.mapthreading.MappingWorkerThread method), 10

S

skippable() (src.mapping.mapthreading.VectorDict method), 11
soft_cosine_similarity() (src.mapping.mapthreading.MappingWorkerThread method), 10
src (module), 17
src.clustering (module), 5
src.clustering.cluster_mappings (module), 3
src.eval (module), 7
src.eval.analogy (module), 5
src.eval.eval_vectors (module), 5
src.eval.word_similarity (module), 6
src.mapping (module), 11
src.mapping.mapthreading (module), 7
src.misc (module), 12
src.misc.helpers (module), 11
src.prep (module), 15
src.prep.corpus (module), 13
src.prep.corpus.convert_to_plain (module), 12
src.prep.corpus.extract_conll (module), 12
src.prep.corpus.mapper (module), 12
src.prep.corpus.prepare_corpus (module), 12
src.prep.corpus.quick_and_dirty (module), 13
src.prep.corpus.reducer (module), 13
src.prep.misc (module), 13
src.prep.misc.decorators (module), 13
src.prep.nes (module), 14
src.prep.nes.extractNE (module), 13
src.prep.nes.merge (module), 14
src.prep.nes.mwe (module), 14
src.prep.nes.statistics (module), 14
src.prep.relations (module), 15
src.prep.relations.relations (module), 14
src.trans_e (module), 17
src.trans_e.add_inverse_relations (module), 15
src.trans_e.clean_relations (module), 15
src.trans_e.contains_entities (module), 15
src.trans_e.convert_relations (module), 16
src.trans_e.differentiate_datasets (module), 16
src.trans_e.partition_data (module), 16
src.trans_e.trans_we (module), 16
start_threads() (src.mapping.mapthreading.MappingMasterThread method), 8

T

test_coverage() (in module src.trans_e.trans_we), 17
train() (in module src.trans_e.trans_we), 17
train_clustering_parameters() (in module src.clustering.cluster_mappings), 4
transform_triples() (in module src.trans_e.trans_we), 17
translate_name() (in module src.prep.relations.relations), 15
translate_word2vec_question_phrases() (in module src.prep.relations.relations), 15

V

VectorDict (class in src.mapping.mapthreading), 10

W

word_sim_eval() (in module src.eval.word_similarity), 7
write_data() (in module src.trans_e.trans_we), 17
write_data_in_file() (in module src.trans_e.partition_data), 16