
fowler.corpora Documentation

Release 0.3

Dmitrijs Milajevs

March 30, 2016

1	Chnagelog	3
1.1	0.3	3
2	Content	5
2.1	Installation	5
2.2	Quick start: Similarity experiments	6
3	Indices and tables	11
	Bibliography	13

fowler.corpora is software to create vector space models for distributional semantics.

It is possible to instantiate a vector space from

- Brown corpus
- British National Corpus
- ukWaC and WaCkypedia

The weighting schemes include:

- PMI
- PPMI
- nITTF

The implemented experiments are:

- Word similarity
 - [SimLex-999](#)
 - [Men](#)
- Sentence similarity
 - [KS14](#)

1.1 0.3

- Documentation update: installation instructions, similarity experiment quick start.
- Correlation and Euclidean similarities are computed.
- PMI variants and parameters.
- Frobenious operators.
- Word2vec space import.

2.1 Installation

It's recommended to use Anaconda and install some packages with it. Refer to [miniconda homepage](#) for links to installers for other platforms.

```
# Install miniconda
wget https://repo.continuum.io/miniconda/Miniconda3-3.7.3-MacOSX-x86_64.sh
sh Miniconda3-3.7.3-MacOSX-x86_64.sh -b

# Conda-install some packages
wget https://bitbucket.org/dimazest/phd-buildout/raw/tip/requirements.txt
~/miniconda3/bin/conda install -c https://conda.anaconda.org/dimazest --file requirements.txt pip
```

You also need NLK data:

```
~/miniconda3/bin/python -c 'import nltk; nltk.download("brown")'
```

2.1.1 The package itself

The package is available on [PyPi](#) and can be installed with `pip`:

```
~/miniconda3/bin/pip install fowler.corpora
```

It's also possible to install a development version right from [GitHub](#):

```
~/miniconda3/bin/pip install https://github.com/dimazest/fowler.corpora/archive/master.zip
```

2.1.2 The final step

Run the package to see whether it works.

```
~/miniconda3/bin/corpora -h
usage: corpora <command> [options]

commands:
  help          Show help for a given help topic or a help overview.
  ...
  ...
  ...
```

2.2 Quick start: Similarity experiments

This tutorial explains how to run similarity experiments. It assumes that a vector space is already built.

2.2.1 SimLex-999

The *[SimLex-999]* data set consists of 999 word pairs judged by humans for similarity. You can download the whole data set from [here](#).

These are some of the records, the similarity score is in the SD (SimLex) column:

word1	word2	POS	Sim-Lex999	conc(w1)	conc(w2)	concQ	As-soc(USF)	SimAs-soc333	SD(SimLex)
old	new	A	1.58	2.72	2.81	2	7.25	1	0.41
smart	intelligent	A	9.2	1.75	2.46	1	7.11	1	0.67
hard	difficult	A	8.77	3.76	2.21	2	5.94	1	1.19
happy	cheerful	A	9.55	2.56	2.34	1	5.85	1	2.18
hard	easy	A	0.95	3.76	2.07	2	5.82	1	0.93
fast	rapid	A	8.75	3.32	3.07	2	5.66	1	1.68
happy	glad	A	9.17	2.56	2.36	1	5.49	1	1.59
short	long	A	1.23	3.61	3.18	2	5.36	1	1.58

Our task is to predict the human judgment given a pair of words from the dataset.

```
# Download the dataset
wget http://www.eecs.qmul.ac.uk/~dm303/static/data/SimLex-999/SimLex-999.txt

# Download the vector space
wget http://www.eecs.qmul.ac.uk/~dm303/t/space_corpus.ukwac_wackypedia-weighting.ppmi.neg.1.base.e-c

# Run an experiment
corpora wsd similarity \
--space space_corpus.ukwac_wackypedia-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.SimLex-999 \
--dataset simlex999://$PWD/SimLex-999.txt?tagset=ukwac \
--output simlex.h5

Similarity |#####| 999/999, elapsed: 0:00:03
Spearman correlation (head), cosine): rho=0.359, p=0.00000, support=999
```

For this space (weighting: PPMI, corpus: ukWaC+Wackypedia, context: 2000 most frequent POS tagged words), the result is **0.359**.

It is possible to access individual similarity results based not only on cosine, but also on correlation and inner product:

```
>>> import pandas as pd

>>> pd.read_hdf('simlex.h5', key='dataset').head()
   unit1                unit2  euclidean    cos  correlation  inner_product  score
0  (old, J, ())          (new, J, ())  0.044349  0.137955  -0.009947    36.827502  1.58
1  (smart, J, ())      (intelligent, J, ())  0.050431  0.504371   0.418022   180.106291  9.20
2  (hard, J, ())       (difficult, J, ())  0.054279  0.483636   0.419472   142.161781  8.77
3  (happy, J, ())     (cheerful, J, ())  0.050953  0.469045   0.403718   149.552762  9.55
4  (hard, J, ())       (easy, J, ())  0.050773  0.436153   0.356595   134.926696  0.95
```

2.2.2 MEN

MEN is a word similarity and relatedness dataset [MEN]:

```
# Download the datasets
wget http://www.eecs.qmul.ac.uk/~dm303/t/MEN_dataset_lemma_form_full
wget http://www.eecs.qmul.ac.uk/~dm303/t/MEN_dataset_lemma_form.dev
wget http://www.eecs.qmul.ac.uk/~dm303/t/MEN_dataset_lemma_form.test

# Download the space

# Run an experiment on the full dataset
corpora wsd similarity \
--space space_corpus.ukwac_wackypedia-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.SimLex-99
--dataset men://$PWD/MEN_dataset_lemma_form_full \
--output men_full.h5

Similarity |#####| 3000/3000, elapsed: 0:00:09
Spearman correlation (head), cosine): rho=0.699, p=0.00000, support=3000

# Dev
corpora wsd similarity \
--space space_corpus.ukwac_wackypedia-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.SimLex-99
--dataset men://$PWD/MEN_dataset_lemma_form.dev \
--output men_dev.h5

Similarity |#####| 2000/2000, elapsed: 0:00:06
Spearman correlation (head), cosine): rho=0.698, p=0.00000, support=2000

# Test
corpora wsd similarity \
--space space_corpus.ukwac_wackypedia-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.SimLex-99
--dataset men://$PWD/MEN_dataset_lemma_form.test \
--output men_test.h5

Similarity |#####| 1000/1000, elapsed: 0:00:03
Spearman correlation (head), cosine): rho=0.701, p=0.00000, support=1000
```

2.2.3 KS14

```
# Download the dataset
wget http://compling.eecs.qmul.ac.uk/wp-content/uploads/2015/07/KS2014.txt

# Download the spaces
wget http://www.eecs.qmul.ac.uk/~dm303/t/space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2

# Addition
corpora wsd similarity \
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2013_turk.reduct
--dataset ks13://$PWD/KS2014.txt \
--composition_operator add \
--output ks14_add.h5

Similarity |#####| 108/108, elapsed: 0:00:01
Spearman correlation (add), cosine): rho=0.780, p=0.00000, support=108

# Head
```

```
corpora wsd similarity \  
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2013_turk.reduct  
--dataset ks13://$PWD/KS2014.txt \  
--composition_operator head \  
--output ks14_head.h5  
  
Similarity |#####| 108/108, elapsed: 0:00:00  
Spearman correlation (head), cosine): rho=0.697, p=0.00000, support=108  
  
# Multiplication  
corpora wsd similarity \  
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2013_turk.reduct  
--dataset ks13://$PWD/KS2014.txt \  
--composition_operator mult \  
--output ks14_mult.h5  
  
Similarity |#####| 108/108, elapsed: 0:00:01  
Spearman correlation (mult), cosine): rho=0.721, p=0.00000, support=108  
  
# Kronecker  
corpora wsd similarity \  
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2013_turk.reduct  
--dataset ks13://$PWD/KS2014.txt \  
--composition_operator kron \  
--output ks14_kron.h5  
  
Similarity |#####| 108/108, elapsed: 0:01:04  
Spearman correlation (kron), cosine): rho=0.805, p=0.00000, support=108  
  
# Relational  
corpora wsd similarity \  
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2013_turk.reduct  
--verb_space out/verb_space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2  
--dataset ks13://$PWD/KS2014.txt \  
--composition_operator relational \  
--output ks14_relational.h5  
  
Similarity |#####| 108/108, elapsed: 0:01:04  
Spearman correlation (relational), cosine): rho=0.522, p=0.00000, support=108  
  
# copy-object  
corpora wsd similarity \  
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2013_turk.reduct  
--verb_space out/verb_space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2  
--dataset ks13://$PWD/KS2014.txt \  
--composition_operator copy-object \  
--output ks14_copy-object.h5  
  
Similarity |#####| 108/108, elapsed: 0:00:38  
Spearman correlation (copy-object), cosine): rho=0.346, p=0.00025, support=108  
  
# copy-subject  
corpora wsd similarity \  
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2013_turk.reduct  
--verb_space out/verb_space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2  
--dataset ks13://$PWD/KS2014.txt \  
--composition_operator copy-subject \  
--output ks14_copy-subject.h5
```

```

Similarity |#####| 108/108, elapsed: 0:00:35
Spearman correlation (copy-subject), cosine): rho=0.446, p=0.00000, support=108

# frobenious-add
corpora wsd similarity \
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2013_turk.reduct
--verb_space out/verb_space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2
--dataset ks13://$PWD/KS2014.txt \
--composition_operator frobenious-add \
--output ks14_frobenious-add.h5

Similarity |#####| 108/108, elapsed: 0:00:39
Spearman correlation (frobenious-add), cosine): rho=0.486, p=0.00000, support=108

# frobenious-mult
corpora wsd similarity \
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2013_turk.reduct
--verb_space out/verb_space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2
--dataset ks13://$PWD/KS2014.txt \
--composition_operator frobenious-mult \
--output ks14_frobenious-mult.h5

Similarity |#####| 108/108, elapsed: 0:00:39
Spearman correlation (frobenious-mult), cosine): rho=0.354, p=0.00017, support=108

# frobenious-outer
corpora wsd similarity \
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2013_turk.reduct
--verb_space out/verb_space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.emnlp2
--dataset ks13://$PWD/KS2014.txt \
--composition_operator frobenious-outer \
--output ks14_frobenious-outer.h5

Similarity |#####| 108/108, elapsed: 0:01:37
Spearman correlation (frobenious-outer), cosine): rho=0.522, p=0.00000, support=108

```

2.2.4 GS11

```

# Download the dataset
wget http://compling.eecs.qmul.ac.uk/wp-content/uploads/2015/07/GS2011data.txt

# Download the sapces
wget http://www.eecs.qmul.ac.uk/~dm303/t/space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2
wget http://www.eecs.qmul.ac.uk/~dm303/t/verb_space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context

# Add
corpora wsd similarity \
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.gs2011.reduction.raw
--dataset gs11://$PWD/GS2011data.txt \
--composition_operator add \
--output gs11-add.h5

Similarity |#####| 199/199, elapsed: 0:00:01
Spearman correlation (add), cosine): rho=0.192, p=0.00670, support=199

```

```
# copy-object
corpora wsd similarity \
--space space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.gs2011.reduction.raw
--verb_space verb_space_corpus.ukwac-weighting.ppmi.neg.1.base.e-context.nvaa2000_dataset.gs2011.redu
--dataset gs11://$PWD/GS2011data.txt \
--composition_operator copy-object \
--output gs11-copy-object.h5

Similarity |#####| 199/199, elapsed: 0:01:15
Spearman correlation (copy-object), cosine): rho=0.024, p=0.73779, support=199
```

2.2.5 References

Indices and tables

- `genindex`
- `modindex`
- `search`

- [SimLex-999] Felix Hill, Roi Reichart and Anna Korhonen. *SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation*. Computational Linguistics. 2015
- [MEN] E. Bruni, N. K. Tran and M. Baroni. *Multimodal Distributional Semantics*. Journal of Artificial Intelligence Research 49: 1-47.