DiShIn

Apr 22, 2021

Contents:

1 Getting started **3**
- 1.1 Installation . . . 3
- 1.2 Quick start . . . 3

2 Other Examples **5**
- 2.1 Gene Ontology (GO) and UniProt proteins . . . 5
- 2.2 Chemical Entities of Biological Interest (ChEBI) Example . . . 6
- 2.3 Human Phenotype (HP) Example . . . 7
- 2.4 Human Disease Ontology (HDO) Example . . . 7
- 2.5 Medical Subject Headings (MeSH) Example . . . 8
- 2.6 Radiology Lexicon (RadLex) Example . . . 8
- 2.7 WordNet Example . . . 9

3 Data Sources **11**
- 3.1 Gene Ontology (GO) . . . 11
- 3.2 ChEBI . . . 11
- 3.3 Human Phenotype ontology (HPO) . . . 11
- 3.4 Human Disease Ontology (DO) . . . 11
- 3.5 Medical Subject Headings (MeSH) Example . . . 12
- 3.6 RadLex . . . 12
- 3.7 WordNet . . . 12

4 API **13**

5 Reference: **15**
- 5.1 Indices and tables . . . 15

This software package provides the basic functions to start using semantic similarity measures directly from a rdf or owl file.

CHAPTER 1

Getting started

1.1 Installation

Either clone this repository or install from pypi:

```
pip install ssmpy
```

1.2 Quick start

```
import ssmpy
```

1.2.1 Metals Example

To create the semantic base file (*metals.db*) from the metals.owl file:

```
ssmpy.create_semantic_base("metals.owl", "metals.db", "https://raw.githubusercontent.
↪com/lasigeBioTM/ssm/master/metals.owl#", "http://www.w3.org/2000/01/rdf-schema
↪#subClassOf", "metals.txt")
ssmpy.semantic_base("metals.db")
```

The *metals.txt* contains the a list of occurrences. For example, the following contents has one occurrence for each term, except gold and silver with two occurrences.

```
gold
silver
gold
silver
copper
platinum
```

(continues on next page)

(continued from previous page)

```
palladium
metal
coinage
precious
```

Now to calculate the similarity between *copper* and *gold* execute:

```
e1 = ssmpy.get_id("copper")
e2 = ssmpy.get_id("gold")
ssmpy.ssm_resnik (e1,e2)
ssmpy.ssm_resnik(e1,e2)
ssmpy.ssm_lin(e1,e2)
ssmpy.ssm_jiang_conrath(e1,e2)
```

Output:

```
0.22599256187152864
0.1504595366201814
0.281527889373394
```

1.2.2 Options

We can choose to calculate the measures using either the extrinsic or intrinsic Information Content (IC), and using the Most Informative Common Ancestors (MICA) or Disjunctive Common Ancestors (DCA). By default, the measures are calculated using extrinsic IC and DCA.

```
ssmpy.ssm.mica = False # determines if it uses MICA or DCA
ssmpy.ssm.intrinsic = False # determines if it uses extrinsic or intrinsic IC
```

Now calculate the similarity between *copper* and *gold* using intrinsic IC and MICA:

```
ssmpy.ssm.mica = True
ssmpy.ssm.intrinsic = True
e1 = ssmpy.get_id("copper")
e2 = ssmpy.get_id("gold")
ssmpy.ssm_resnik (e1,e2)
ssmpy.ssm_resnik(e1,e2)
ssmpy.ssm_lin(e1,e2)
ssmpy.ssm_jiang_conrath(e1,e2)
```

Output:

```
0.587786664902119
0.39079549108439265
0.35303485982596094
```

CHAPTER 2

Other Examples

The following examples will assume the default options, i.e. the values shown are calculated using extrinsic IC and DCA.

2.1 Gene Ontology (GO) and UniProt proteins

Download the latest version of the database we created:

```
wget http://labs.rd.ciencias.ulisboa.pt/dishin/go202104.db.gz
gunzip -N go202104.db.gz
```

Now to calculate the similarity between *maltose biosynthetic process* and *maltose catabolic process*, first we need to obtain the semantic base IDs of those concepts:

```
ssmpy.semantic_base("go.db")
e1 = ssmpy.get_id("GO_0000023")
e2 = ssmpy.get_id("GO_0000025")
ssmpy.ssm_resnik(e1,e2)
ssmpy.ssm_lin(e1,e2)
ssmpy.ssm_jiang_conrath(e1,e2)
```

Output:

```
4.315813746201754
0.38793452313030363
0.06840605034663635
```

Now to calculate the similarity between proteins Q12345 and Q12346, first we retrieve the GO terms associated with each one:

```
e1 = ssmpy.get_uniprot_annotations("Q12345")
e2 = ssmpy.get_uniprot_annotations("Q12346")
```

Next we use the `ssm_multiple` to calculate the average maximum semantic similarity, using the resnik measure

```
ssmpy.ssm_multiple(ssmpy.ssm_resnik, e1, e2)
ssmpy.ssm_multiple(ssmpy.ssm_lin, e1, e2)
ssmpy.ssm_multiple(ssmpy.ssm_jiang_conrath, e1, e2)
```

Output:

```
0.6015115682274214
0.12201023476842265
0.09317326288224918
```

To create an updated version of the database, download the ontology and annotations:

```
wget http://purl.obolibrary.org/obo/go.owl
wget http://geneontology.org/gene-associations/goa_uniprot_all_noiea.gaf.gz
gunzip goa_uniprot_all_noiea.gaf.gz
```

The annotations will be used to calculate the extrinsic information content.

Next create the semantic base:

```
ssmpy.create_semantic_base("go.owl", "go.db", "http://purl.obolibrary.org/obo/",
↪"http://www.w3.org/2000/01/rdf-schema#subClassOf", "goa_uniprot_all_noiea.gaf)
```

This is stored in the form of a sqlite database on the same directory of your project.

2.2 Chemical Entities of Biological Interest (ChEBI) Example

Download the lastest version of the database we created:

```
wget http://labs.rd.ciencias.ulisboa.pt/dishin/chebi202104.db.gz
gunzip -N chebi202104.db.gz
```

Now to calculate the similarity between *aripiprazole* and *bithionol* execute:

```
ssmpy.semantic_base("chebi.db")
e1 = ssmpy.get_id("CHEBI_31236")
e2 = ssmpy.get_id("CHEBI_3131")
ssmpy.ssm_resnik(e1,e2)
ssmpy.ssm_lin(e1,e2)
ssmpy.ssm_jiang_conrath(e1,e2)
```

Output:

```
1.4393842298350599
0.12935491517581163
0.049077257018319796
```

To create an updated version of the database, download the ontology:

```
wget http://purl.obolibrary.org/obo/chebi/chebi_lite.owl
```

And then create the new database:

```
ssmpy.create_semantic_base("chebi_lite.owl", "chebi.db", "http://purl.obolibrary.org/
↪obo/", "http://www.w3.org/2000/01/rdf-schema#subClassOf", '')
```

2.3 Human Phenotype (HP) Example

Download the lastest version of the database we created:

```
wget http://labs.rd.ciencias.ulisboa.pt/dishin/hp202104.db.gz
gunzip -N hp202104.db.gz
```

Now to calculate the similarity between *Optic nerve coloboma* and *Optic nerve dysplasia* execute:

```
ssmpy.semantic_base("hp.db")
e1 = ssmpy.get_id("HP_0000588")
e2 = ssmpy.get_id("HP_0001093")
ssmpy.ssm_resnik(e1,e2)
ssmpy.ssm_lin(e1,e2)
ssmpy.ssm_jiang_conrath(e1,e2)
```

Output:

```
4.593979372426621
0.5118244533189668
0.10242304162282165
```

To create an updated version of the database, download the ontology:

```
wget http://purl.obolibrary.org/obo/hp.owl
```

And then create the new database:

```
ssmpy.create_semantic_base("hp.owl", "hp.db", "http://purl.obolibrary.org/obo/",
↪"http://www.w3.org/2000/01/rdf-schema#subClassOf", '')
```

2.4 Human Disease Ontology (HDO) Example

Download the lastest version of the database we created:

```
wget http://labs.rd.ciencias.ulisboa.pt/dishin/doid202104.db.gz
gunzip -N doid202104.db.gz
```

Now to calculate the similarity between *Asthma* and *Lung cancer* execute:

```
ssmpy.semantic_base("doid.db")
e1 = ssmpy.get_id("DOID_2841")
e2 = ssmpy.get_id("DOID_1324")
ssmpy.ssm_resnik(e1,e2)
ssmpy.ssm_lin(e1,e2)
ssmpy.ssm_jiang_conrath(e1,e2)
```

Output:

```
2.3627836143597176
0.4328907089097581
0.13906777879867938
```

To create an updated version of the database, download the ontology:

```
wget http://purl.obolibrary.org/obo/doid.owl
```

And then create the new database:

```
ssmpy.create_semantic_base("doid.owl", "doid.db", "http://purl.obolibrary.org/obo/",
↪"http://www.w3.org/2000/01/rdf-schema#subClassOf", '')
```

2.5 Medical Subject Headings (MeSH) Example

Download the lastest version of the database we created:

```
wget http://labs.rd.ciencias.ulisboa.pt/dishin/mesh202104.db.gz
gunzip -N mesh202104.db.gz
```

Now to calculate the similarity between *Malignant Hyperthermia* and *Fever* execute:

```
ssmpy.semantic_base("mesh.db")
e1 = ssmpy.get_id("D008305")
e2 = ssmpy.get_id("D005334")
ssmpy.ssm_resnik(e1,e2)
ssmpy.ssm_lin(e1,e2)
ssmpy.ssm_jiang_conrath(e1,e2)
```

Output:

```
1.2582571367910345
0.17390901691859173
0.07719755683816652
```

To create an updated version of the database, download the _NT_ version from ftp://nlmpubs.nlm.nih.gov/online/mesh/rdf/mesh.nt.gz and unzip it:

```
wget ftp://nlmpubs.nlm.nih.gov/online/mesh/rdf/mesh.nt.gz
gunzip mesh.nt.gz
```

And then create the new database:

```
ssmpy.create_semantic_base("mesh.nt", "mesh.db", "http://id.nlm.nih.gov/mesh/",
↪"http://id.nlm.nih.gov/mesh/vocab#broaderDescriptor", '')
```

2.6 Radiology Lexicon (RadLex) Example

Download the lastest version of the database we created:

```
wget http://labs.rd.ciencias.ulisboa.pt/dishin/radlex202104.db.gz
gunzip -N radlex202104.db.gz
```

Now to calculate the similarity between *nervous system of right upper limb* and *nervous system of left upper limb* execute:

```
ssmpy.semantic_base("radlex.db")
e1 = ssmpy.get_id("RID16139")
e2 = ssmpy.get_id("RID16140")
ssmpy.ssm_resnik(e1,e2)
ssmpy.ssm_lin(e1,e2)
ssmpy.ssm_jiang_conrath(e1,e2)
```

Output:

```
9.366531825151093
0.9310964912333252
0.41905978419640516
```

To create an updated version of the database, download the *RDF/XML* version from http://bioportal.bioontology.org/ontologies/RADLEX and save it as *radlex.rdf*

And then create the new database:

```
ssmpy.create_semantic_base("radlex.rdf", "radlex.db", "http://www.radlex.org/RID/",
↪"http://www.w3.org/2000/01/rdf-schema#subClassOf", '')
```

2.7 WordNet Example

Download the lastest version of the database we created:

```
wget http://labs.rd.ciencias.ulisboa.pt/dishin/wordnet202104.db.gz
gunzip -N wordnet202104.db.gz
```

Now to calculate the similarity between the nouns *ambulance* and *motorcycle* execute:

```
ssmpy.semantic_base("wordnet.db")
e1 = ssmpy.get_id("ambulance-noun-1")
e2 = ssmpy.get_id("motorcycle-noun-1")
ssmpy.ssm_resnik(e1,e2)
ssmpy.ssm_lin(e1,e2)
ssmpy.ssm_jiang_conrath(e1,e2)
```

Output:

```
6.331085809208157
0.6792379292396559
0.14327549414725688
```

To create an updated version of the database, download the ontology:

```
wget http://www.w3.org/2006/03/wn/wn20/rdf/wordnet-hyponym.rdf
```

And then create the new database:

```
ssmpy.create_semantic_base("wordnet-hyponym.rdf", "wordnet.db", "http://www.w3.org/
↪2006/03/wn/wn20/instances/synset-", "http://www.w3.org/2006/03/wn/wn20/schema/
↪hyponymOf", '')
```

CHAPTER 3

Data Sources

3.1 Gene Ontology (GO)

- Ontology: http://purl.obolibrary.org/obo/go.owl
- Annotation files (extrinsic): http://geneontology.org/gene-associations/goa_uniprot_all_noiea.gaf.gz
- SemanticBase: http://labs.rd.ciencias.ulisboa.pt/dishin/go202104.db.gz

3.2 ChEBI

- Ontology: http://purl.obolibrary.org/obo/chebi/chebi_lite.owl
- SemanticBase: http://labs.rd.ciencias.ulisboa.pt/dishin/chebi202104.db.gz

3.3 Human Phenotype ontology (HPO)

- Ontology: http://purl.obolibrary.org/obo/hp.owl
- SemanticBase: http://labs.rd.ciencias.ulisboa.pt/dishin/hp202104.db.gz

3.4 Human Disease Ontology (DO)

- Ontology: http://purl.obolibrary.org/obo/doid.owl
- SemanticBase: http://labs.rd.ciencias.ulisboa.pt/dishin/doid202104.db.gz

3.5 Medical Subject Headings (MeSH) Example

- Ontology: ftp://nlmpubs.nlm.nih.gov/online/mesh/rdf/mesh.nt.gz
- SemanticBase: http://labs.rd.ciencias.ulisboa.pt/dishin/mesh202104.db.gz

3.6 RadLex

- Ontology: http://bioportal.bioontology.org/ontologies/RADLEX
- SemanticBase: http://labs.rd.ciencias.ulisboa.pt/dishin/radlex.db

3.7 WordNet

- Ontology: http://www.w3.org/2006/03/wn/wn20/rdf/wordnet-hyponym.rdf
- SemanticBase: http://labs.rd.ciencias.ulisboa.pt/dishin/wordnet202104.db.gz

CHAPTER 4

API

CHAPTER 5

Reference:

- F. Couto and A. Lamurias, “Semantic similarity definition,” in Encyclopedia of Bioinformatics and Computational Biology (S. Ranganathan, K. Nakai, C. Schönbach, and M. Gribskov, eds.), vol. 1, pp. 870–876, Oxford: Elsevier, 2019 https://doi.org/10.1016/B978-0-12-809633-8.20401-9

5.1 Indices and tables

- genindex
- modindex
- search