homework03

February 25, 2023

1 Homework 3: Wordnet

WordNet is a project that began at Princeton in 1985 and was originally intended to be a tool to support research that theorized the nature of semantic human memory. It consists of a large database of words and their glosses (definitions), attributes and relationships to other words, and somee example usages. There exist wrappers to access it in most languages, including one built in to `nltk` in Python.

[101]:

```python
import nltk
nltk.download("wordnet", quiet=True)

from nltk.corpus import wordnet as wn
import random

nouns = [ n.name().split(".")[0] for n in wn.all_synsets("n") ]
word = random.choice(nouns)
syns = wn.synsets(word, wn.NOUN)
print(syns)
```

```
[Synset('immunology.n.01')]
```

[102]:

```python
synset = random.choice(syns)
print(synset.definition())
print(synset.examples())
print(synset.lemmas())

top = wn.synset("entity.n.01")
print(list(synset.closure(lambda s: s.hypernyms() if not s == top else [])))
```

```
the branch of medical science that studies the body's immune system
[]
[Lemma('immunology.n.01.immunology')]
[Synset('medicine.n.01'), Synset('medical_science.n.01'),
Synset('life_science.n.01'), Synset('natural_science.n.01'),
Synset('science.n.01'), Synset('discipline.n.01'),
Synset('knowledge_domain.n.01'), Synset('content.n.05'),
Synset('cognition.n.01'), Synset('psychological_feature.n.01'),
Synset('abstraction.n.06'), Synset('entity.n.01')]
```

Nouns generally have a fairly straightforward derivation toward `entity.n.01`, as shown in the generalization of "immunology" into "medicine", medical science", etc. While termination is guaranteed for nouns, this is not necessarily so for verbs.

[103]:

```
print(synset.hypernyms())
print(synset.hyponyms())
print(synset.member_meronyms() + synset.substance_meronyms() + synset.
  ↪part_meronyms())
print(synset.member_holonyms() + synset.substance_holonyms() + synset.
  ↪part_holonyms())
print([ l.antonyms() for l in synset.lemmas() ])
```

```
[Synset('medicine.n.01')]
[Synset('immunochemistry.n.01'), Synset('immunopathology.n.01')]
[]
[]
[[]]
```

[109]:

```
verbs = [ v.name().split(".")[0] for v in wn.all_synsets("v") ]
verb = random.choice(verbs)
syns = wn.synsets(verb, wn.VERB)
print(syns)
```

```
[Synset('puree.v.01')]
```

[110]:

```
synset = random.choice(syns)
print(synset.definition())
print(synset.examples())
print(synset.lemmas())

print(list(synset.closure(lambda s: s.hypernyms())))
```

```
rub through a strainer or process in an electric blender
['puree the vegetables for the baby']
[Lemma('puree.v.01.puree'), Lemma('puree.v.01.strain')]
[Synset('rub.v.01'), Synset('guide.v.05')]
```

The hierarchy for verbs is much more shallow, with this word only having two iterations. Unlike nouns, I suppose it's not as easy to generalize actions without delving into archaic etymology. That also explains why there is no top level verb synset akin to `entity.n.01`.

[121]:

```
for pos in [ wn.NOUN, wn.VERB, wn.ADJ, wn.ADV ]:
    if not wn.morphy(verb, pos) == None:
        print(wn.morphy(verb, pos))
```

```
puree
puree
```

[138]:

```python
from nltk.wsd import lesk

words = ( "find", "locate")
syns = ( wn.synset("find.v.03"), wn.synset("locate.v.01") )

print(wn.wup_similarity(syns[0], syns[1]))
print(lesk(syns[0].examples()[0].split(" "), words[0]))
print(lesk(syns[1].examples()[0].split(" "), words[1]))
```

```
0.8
Synset('recover.v.01')
Synset('situate.v.01')
```

The Wu-Palmer metric of 0.8 indicates that the two words are pretty similar, as I expected. The Lesk result for “find” isn’t too surprising, but “situate”, in the sense it is used, seems somewhat different from “locate”, even considering the context of the example sentence in which it was used (“Can you locate your cousins in the Midwest?”).

1.1 SentiWordNet

SentiWordNet assigns scores of positivity, negativity, and objectivity to approximate how provocative words are and in what sense. This is useful in getting a general idea of how instigative a text might be.

[157]:

```python
nltk.download("sentiwordnet", quiet=True)
from nltk.corpus import sentiwordnet as swn

word = swn.senti_synset("fucking.r.01")
print(word.pos_score())
print(word.neg_score())
print(word.obj_score())

sentence = [ "what", "light", "through", "yonder", "window", "breaks" ]
for w in sentence:
    syns = wn.synsets(w)
    if len(syns) < 1:
        continue
    ss = swn.senti_synset(syns[0].name())
    print(f"{w}: {ss.pos_score()}, {ss.neg_score()}, {ss.obj_score()}")
```

```
0.125
0.0
0.875
light: 0.0, 0.0, 1.0
through: 0.0, 0.0, 1.0
yonder: 0.125, 0.0, 0.875
window: 0.0, 0.0, 1.0
breaks: 0.0, 0.0, 1.0
```

I'm surprised "fucking" is regarded as having a positive connotation over a negative one, but otherwise nothing else is too surprising. The sentence having a vaguely positive tone seems apt, even if the methodology in doing so was rather careless (taking the first sense of the word). Being able to programmatically discern subtleties like this can prove useful in fields like content moderation, though systems may fail to recognize certain nuances.

1.2 Collocations

Collocations are two or more words that are generally used together, where substituting one of the words for a synonym would not convey the same meaning. These can pose issues for language processing since they don't conform to the generic rules of language and may not always be obvious without cultural context.

[182]:

```
nltk.download("gutenberg", quiet=True)
nltk.download("genesis", quiet=True)
nltk.download("inaugural", quiet=True)
nltk.download("nps_chat", quiet=True)
nltk.download("webtext", quiet=True)
nltk.download("treebank", quiet=True)
nltk.download("stopwords", quiet=True)
from nltk.book import text4
import math

print(text4.collocations())

coll = ( "years", "ago" )
bigrams = list(nltk.bigrams(text4))
n_bigrams = len(bigrams)
n_tokens = len(text4)

Pxy = bigrams.count(coll) / n_bigrams
Px = text4.count(coll[0]) / n_tokens
Py = text4.count(coll[1]) / n_tokens
print(math.log(Pxy / (Px * Py), 2))
```

```
United States; fellow citizens; years ago; four years; Federal
Government; General Government; American people; Vice President; God
bless; Chief Justice; one another; fellow Americans; Old World;
Almighty God; Fellow citizens; Chief Magistrate; every citizen; Indian
tribes; public debt; foreign nations
None
9.357832298914353
```

The point-wise mutual information value of 9.3578 indicates that there was some positive association between the two words. To what *degree* there is an association however, I'm not really too sure. Without getting into the nitty-gritty of the specific maximum value possible, 9.3578 seems like a value well above 0 (independent) to the point where I would trust that there is a substantial connection there.