ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ
ΕΡΓΑΣΤΗΡΙΟ 4

WORDNET: ΕΝΑΣ ΣΗΜΑΣΙΟΛΟΓΙΚΟΣ ΘΗΣΑΥΡΟΣ

Το WordNet είναι ένα σημασιολογικό λεξικό για τα Αγγλικά με πολύ πλούσια δομή (θησαυρός). Το WordNet ομαδοποιεί τις λέξεις σε σετ συνωνύμων (synonym sets ή synsets). Η έκδοση WordNet 3.0 που ενσωματώνεται και στο NLTK περιέχει 117.659 synsets.

ΕΝΝΟΙΕΣ ΚΑΙ ΣΥΝΩΝΥΜΑ

Στην πρόταση

Benz is credited with the invention of the motorcar.

αν αντικαταστήσουμε την λέξη *motorcar* με την λέξη *automobile*, το νόημα της πρότασης παραμένει ουσιαστικά το ίδιο:

Benz is credited with the invention of the automobile.

Μια και το νόημα της πρότασης δεν αλλάζει, μπορούμε να συμπεράνουμε ότι οι λέξεις *motorcar* και *automobile* έχουν την ίδια έννοια, είναι δηλαδή **συνώνυμες**.

Στο WordNet οι λέξεις είναι ομαδοποιημένες σε 4 λεξικά ανάλογα με το μέρος του λόγου στο οποίο ανήκουν: N (ουσιαστικά), V (ρήματα), ADJ (επίθετα), ADV (επιρρήματα).

1. Τρέξτε την Python μέσω του IDLE (Interactive Development Environment).
Start -> All Programs -> Python 2.5 -> IDLE.

2. Φορτώστε το module του WordNet.
>>>from nltk.corpus import wordnet

3. Για να βρείτε την λέξη *motorcar* στο λεξικό των ουσιαστικών πληκτρολογείτε:

```
>>> wordnet.synsets('motorcar')
[Synset('car.n.01')]
```

Τον ορισμό του synset αυτού μπορείτε να τον δείτε πληκτρολογώντας

```
>>> wordnet.synset('car.n.01').definition()
'a motor vehicle with four wheels; usually propelled by an internal combustion engine'
```

Παράδειγμα χρήσης του synset σε μια πρόταση της γλώσσας μπορείτε να δείτε πληκτρολογώντας
>>> wordnet.synset('car.n.01').examples()
['he needs a car to get to work']

Τα υπόλοιπα λήμματα (λέξεις) που ανήκουν στο συγκεκριμένο synset μπορείτε να τα δείτε πληκτρολογώντας
>>> wordnet.synset('car.n.01').lemmas()
[Lemma('car.n.01.car'), Lemma('car.n.01.auto'), Lemma('car.n.01.automobile'), Lemma('car.n.01.machine'), Lemma('car.n.01.motorcar')]

4. Μια λέξη μπορεί μια ή περισσότερες έννοιες. Η λέξη *motorcar* έχει μια μόνο έννοια. Το πόσες έννοιες έχει μια λέξη μπορείτε να το βρείτε χρησιμοποιώντας την συνάρτηση *len()*:
>>> len(wordnet.synsets('motorcar'))
1

Πόσες έννοιες έχει η λέξη *dog*; Ποιές είναι αυτές;

Η συνάρτηση synsets() παίρνει ένα προαιρετικό όρισμα (το pos), το οποίο επιτρέπει, ανάμεσα στις διάφορες έννοιες μιας λέξης, να περιορίσω την αναζήτησή μου ως προς συγκεκριμένο μέρος του λόγου
>>> wordnet.synsets ('dog', pos=wordnet.VERB)
[Synset('chase.v.01')]

Τα υπόλοιπα μέρη του λόγου είναι τα NOUN, ADJ και ADV.

Η ΙΕΡΑΡΧΙΑ ΣΤΟ WORDNET

Τα synsets αντιστοιχούν σε έννοιες. Οι έννοιες αυτές είναι οργανωμένες σε μια ιεραρχία, όπου στην αρχή βρίσκονται πιο γενικές έννοιες και όσο κατεβαίνει κανείς την ιεραρχία οι έννοιες γίνονται όλο και πιο εξειδικευμένες. Για παράδειγμα, στο σχήμα, οι έννοιες *hatchback, compact, gas guzzler* είναι άμεσα υπώνυμα (hyponyms) της έννοιας *motorcar*, ενώ η τελευταία είναι άμεσο υπερώνυμο (hypernym) των πρώτων.

5. Μπορείτε εύκολα να πλοηγηθείτε μέσα στη ιεραρχία. Για να βρείτε τα υπερώνυμα της έννοιας (του synset) 'dog.n.01', πληκτρολογείτε
>>> wordnet.synset('dog.n.01').hypernyms()
[Synset('domestic_animal.n.01'), Synset('canine.n.02')]

Βρείτε τα υπερώνυμα της έννοιας 'car.n.01'

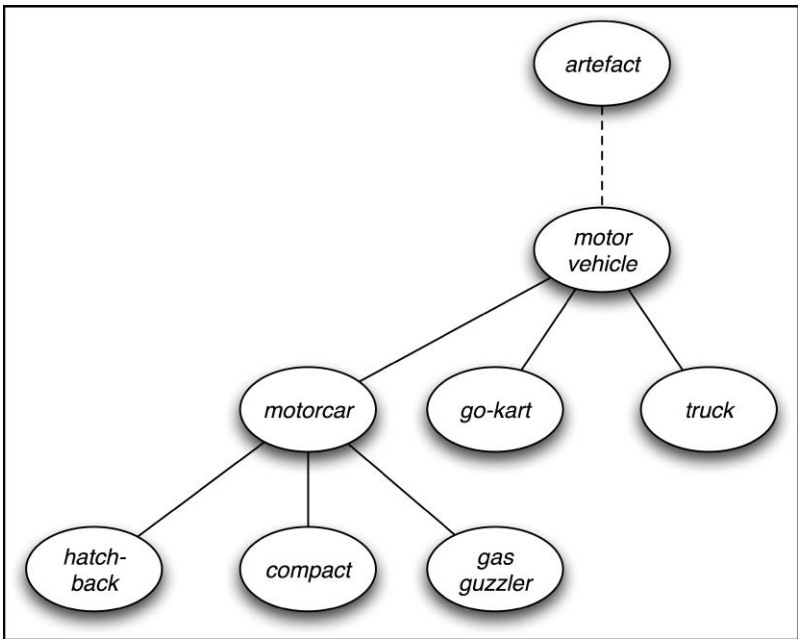
---

6. Για να βρείτε τα υπώνυμα της έννοιας (του synset) ‘dog.n.01’, πληκτρολογείτε

```
>>> wordnet.synset(‘dog.n.01’).hyponyms()
```

[Synset('puppy.n.01'), Synset('great_pyrenees.n.01'), Synset('basenji.n.01'), Synset('newfoundland.n.01'), Synset('lapdog.n.01'), Synset('poodle.n.01'), Synset('leonberg.n.01'), Synset('toy_dog.n.01'), Synset('spitz.n.01'), Synset('pooch.n.01'), Synset('cur.n.01'), Synset('mexican_hairless.n.01'), Synset('hunting_dog.n.01'), Synset('working_dog.n.01'), Synset('dalmatian.n.02'), Synset('pug.n.01'), Synset('corgi.n.01'), Synset('griffon.n.02')]

Βρείτε τα υπώνυμα της έννοιας ‘car.n.01’



7. Ο παρακάτω πίνακας δείχνει όλες τις σημασιολογικές σχέσεις που υποστηρίζει το WordNet.

Hypernym	more general	*animal* is a hypernym of *dog*
Hyponym	more specific	*dog* is a hyponym of *animal*
Meronym	part of	*door* is a meronym of *house*
Holonym	has part	*house* is a holonym of *door*
Synonym	similar meaning	*car* is a synonym of *automobile*
Antonym	opposite meaning	*like* is an antonym of *dislike*
Entailment	necessary action	*step* is an entailment of *walk*

Για να βρείτε τα μερώνυμα της έννοιας (του synset) ‘door.n.01’, πληκτρολογείτε

>>> wordnet.synset(‘door.n.01’).part_meronyms()
[Synset('lock.n.01')]

Για να βρείτε τα ολώνυμα της έννοιας (του synset) ‘door.n.01’, πληκτρολογείτε
>>> wordnet.synset(‘door.n.01’).part_holonyms()
[Synset('doorway.n.01')]

Για να βρείτε τα αντώνυμα του λήμματος 'like', πληκτρολογείτε
>>> wordnet.synset(‘like.v.02’).lemmas()[0].antonyms()
[Lemma('dislike.v.01.dislike')]

ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΟΜΟΙΟΤΗΤΑ (SEMANTIC SIMILARITY)

Η σημασιολογική ομοιότητα δύο εννοιών στο WordNet σχετίζεται με το μήκος του μονοπατιού που τις συνδέει μέσα στην ιεραρχία. Υπάρχουν μια σειρά από μέτρα που υλοποιούν αυτή την σκέψη.

8. Path Similarity

Η συνάρτηση *path_similarity()* επιστρέφει μια τιμή στο διάστημα [0,1], με βάση το μικρότερο μονοπάτι που συνδέει τις έννοιες στην ιεραρχία. Η τιμή -1 επιστρέφεται όταν δεν μπορεί να βρεθεί μονοπάτι. Τιμή 1 σημαίνει ταυτότητα (σύγκριση μιας έννοιας με τον εαυτό της).

path similarity = 1 / αριθμός κόμβων που παρεμβάλλονται (συμπεριλαμβανομένων και των κόμβων των δυο εννοιών των οποίων ψάχνω την ομοιότητα)

```
>>> wordnet.synset(‘dog.n.01’).path_similarity(wordnet.synset(‘cat.n.01’))
0.2
>>> wordnet.synset('poodle.n.01').path_similarity(wordnet.synset('dalmatian.n.01'))
0.1111111111111111
>>> wordnet.synset('poodle.n.01').path_similarity(wordnet.synset('dalmatian.n.02'))
0.33333333333333331
>>> wordnet.synset('run.v.01').path_similarity(wordnet.synset('walk.v.01'))
0.25
```

Βρείτε την σημασιολογική ομοιότητα των παρακάτω ζευγαριών εννοιών:

car-automobile	_____
gem-jewel	_____
journey-voyage	_____
boy-lad	_____
coast-shore	_____
asylum-madhouse	_____
magician-wizard	_____

9. Leacock-Chodorow Similarity

Η συνάρτηση *lch_similarity()* επιστρέφει ένα σκορ που δείχνει πόσο όμοιες είναι δύο έννοιες, δεδομένου του μήκους του κοντινότερου μονοπατιού (p) που τις συνδέει (όπως και η *path_similarity()*), αλλά και του μέγιστου βάθους στην ταξινομία που εμφανίζονται οι έννοιες (d). Η ομοιότητα υπολογίζεται από τον τύπο –log(p/2d).

>>> wordnet.synset(‘dog.n.01’).lch_similarity(wordnet.synset(‘cat.n.01’))
2.0281482472922856

10. Information Content

Το πληροφοριακό περιεχόμενο (Information Content) μιας έννοιας σε ένα σώμα κειμένων δίνεται από την σχέση –logp(έννοια). Το πληροφοριακό περιεχόμενο είναι αθροιστικό. Δηλαδή, για ένα συγκεκριμένο ουσιαστικό σε ένα σώμα κειμένων, η συχνότητα εμφάνισής του (το p(έννοια)) θα αυξάνεται και κάθε φορά που εμφανίζεται και ένα υπερώνυμό του.

>>> from nltk.corpus import wordnet_ic
>>> brown_ic=wordnet_ic.ic(‘ic-brown.dat’)
>>> semcor_ic=wordnet_ic.ic(‘ic-semcor.dat’)

Η συνάρτηση res_similarity() (Resnik Simmilarity) επιστρέφει ένα σκορ που δείχνει πόσο όμοιες είναι δύο έννοιες, βάσει του σημασιολογικού περιεχομένου του πιο εξειδικευμένου κοινού μητρικού κόμβου (Least Common Subsumer)

>>> wordnet.synset('dog.n.01').res_similarity(wordnet.synset('cat.n.01'),brown_ic)
7.9116665090365768
>>> wordnet.synset('dog.n.01').res_similarity(wordnet.synset('cat.n.01'),semcor_ic)
7.2549003421277245

11. Lin Similarity

Η συνάρτηση lin_similarity() επιστρέφει ένα σκορ που δείχνει πόσο όμοιες είναι δύο έννοιες, βάσει του σημασιολογικού περιεχομένου του πιο εξειδικευμένου κοινού μητρικού κόμβου (Least Common Subsumer) καθώς και του σημασιολογικού περιεχομένου των δυο εννοιών εισόδου, σύμφωνα με την σχέση

2 * IC(LCS) / (IC(έννοια1)+IC(έννοια2))

>>> wordnet.synset('dog.n.01').lin_similarity(wordnet.synset('cat.n.01'),brown_ic)
0.87680098437339726
>>> wordnet.synset('dog.n.01').lin_similarity(wordnet.synset('cat.n.01'),semcor_ic)
0.88632886280862277