IHLT Laboratory

Jordi Turmo
TALP Research Center
turmo@cs.upc.edu

Session 3

Resources in nltk library

- List of resources: http://www.nltk.org/nltk_data/
- Download non-default resources from nltk

```
import nltk
nltk.download()
```

- **Corpora and lexical resources**: Brown corpus (PoS annotations), sentence_polarity corpus... Lexical resources such as WordNet, SentiWordNet and specialized word lists.
  - http://www.nltk.org/howto/corpus.html
  - corpus reader objects

    ```
    from nltk.corpus import *resource* [as *variable_name*]
    ```

  - corpus reader classes
- **Toy grammars**: grammars for English, Spanish, ...
- **Models**: Named Entity recognizer, taggers for English and Russian, ...

Examples

stopwords reader

Provide the list of stop words of a specific language. Words that do not have individual meaning (pronouns, determiners, auxiliar verbs, ...)

```
from nltk.corpus import stopwords [as *var_name*]
sw=stopwords.words('[language]')
```

Examples

wordnet reader
(`http://www.nltk.org/howto/wordnet.html`)

Provide an interface to access WordNet data, such as:

- synsets of a given lemma+PoS pair,
- lemmas of a given synset,
- hypernyms and hyponyms of a given synset,
- synonyms and antonyms of a given lemma in a synset
- least common subsumers of a pair of synsets
- different measures of synset similarity

...

```
from nltk.corpus import wordnet [as *var_name*]
```

Exercise

Given the following (lemma, category) pairs:

('the','DT'), ('man','NN'), ('swim','VB'), ('with', 'PR'), ('a', 'DT'), ('girl','NN'), ('and', 'CC'), ('a', 'DT'), ('boy', 'NN'), ('whilst', 'PR'), ('the', 'DT'), ('woman', 'NN'), ('walk', 'VB')

For each pair, when possible, print their most frequent WordNet synset, their corresponding least common subsumer (LCS) and their similarity value, using the following functions:

- Path Similarity
- Leacock-Chodorow Similarity
- Wu-Palmer Similarity
- Lin Similarity

Normalize similarity values when necessary. What similarity seems better?