Master on Artificial Intelligence

Introduction to Human Language Technologies
Lab.6: Word Sense Disambiguation

Natural Language Research Group

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Course 2018/19

Natural Language Research Group

Session requirements

WSD

SensEval Lexical Sample

Outline

1 Session requirements

2 WSD
- Example
- Paraphrases
- Exercise

3 SensEval Lexical Sample
- Optional exercise
- Naïve Bayes

Session requirements

SentiWordnet:

- Both Linux & Windows (via python shell)
  - `import nltk`
  - `nltk.download('sentiwordnet')`

Attached resources:

- `trial.tgz`: trial set of the project
- `line-n.xml`: samples of noun line from SensEval corpora

Outline

1 Session requirements

2 WSD

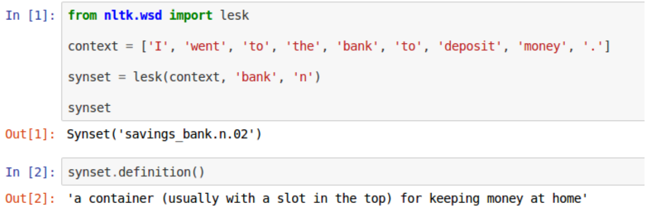
- Example
- Paraphrases
- Exercise

3 SensEval Lexical Sample

- Optional exercise
- Naïve Bayes

Example

Lesk in NLTK:



```
In [1]:  from nltk.wsd import lesk

         context = ['I', 'went', 'to', 'the', 'bank', 'to', 'deposit', 'money', '.']

         synset = lesk(context, 'bank', 'n')

         synset

Out[1]:  Synset('savings_bank.n.02')

In [2]:  synset.definition()

Out[2]:  'a container (usually with a slot in the top) for keeping money at home'
```

Mandatory exercise

Statement:

1. Read all pairs of sentences of the trial set within the evaluation framework of the project.
2. Apply Lesk's algorithm to the words in the sentences.
3. Compute their similarities by considering senses and Jaccard coefficient.
4. Compare the results with those in session 2 (document) and 3 (morphology) in which words and lemmas were considered.
5. Compare the results with gold standard by giving the pearson correlation between them.

Optional exercise

Statement:

1. Implement some of the variants of the Lesk's algorithm.
2. Repeat the previous exercise using this variant.

Note:

- A possible variant is: stopwords, cosine, examples and hypernyms. But the solution is up to you.

Outline

1. Session requirements
2. WSD
   - Example
   - Paraphrases
   - Exercise
3. SensEval Lexical Sample
   - Optional exercise
   - Naïve Bayes

Optional exercise

Statement:

1. Implement a Machine Learning approach to build a WSD classifier for the noun *line*
2. A Naïve Bayes algorithm with bag of words is a possible approach. The next slide shows how to use it in NLTK.

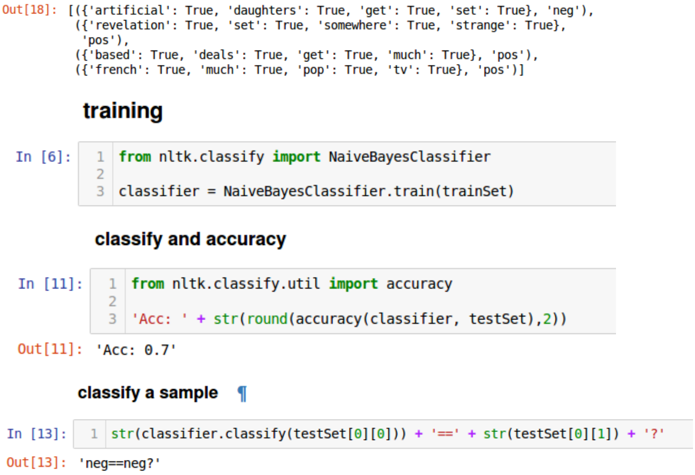
Description of the data:

- 4146 samples
- 6 senses

samples	sense
373	cord
376	division
349	formation
429	phone
2218	product
404	text

NLTK's Naïve Bayes Example

data format:



```
Out[18]: [({'artificial': True, 'daughters': True, 'get': True, 'set': True}, 'neg'),
          ({'revelation': True, 'set': True, 'somewhere': True, 'strange': True},
           'pos'),
          ({'based': True, 'deals': True, 'get': True, 'much': True}, 'pos'),
          ({'french': True, 'much': True, 'pop': True, 'tv': True}, 'pos')]
```

training

```
In [6]:  from nltk.classify import NaiveBayesClassifier

          classifier = NaiveBayesClassifier.train(trainSet)
```

classify and accuracy

```
In [11]:  from nltk.classify.util import accuracy

           'Acc: ' + str(round(accuracy(classifier, testSet),2))

Out[11]:  'Acc: 0.7'
```

classify a sample ¶

```
In [13]:  str(classifier.classify(testSet[0][0])) + '==' + str(testSet[0][1]) + '?'

Out[13]:  'neg==neg?'
```