WP2
Parts-of-Speech-Tagging for sentiment analysis
« I fish a fish! »

Part-of-speech tagging is the process of converting a sentence, in the form of a list of words, into a list of tuples, where each tuple is of the form (word, tag). The tag is a part-of-speech tag, and signifies whether the word is a noun, adjective, verb, and so on.

Most of the taggers are trainable. They use a list of tagged sentences as their training data, such as what you get from the tagged_sents() method of a TaggedCorpusReader class. With these training sentences, the tagger generates an internal model that will tell it how to tag a word. Other taggers use external data sources or match word patterns to choose a tag for a word.
All taggers in NLTK are in the nltk.tag package. Many taggers can also be combined into a backoff chain, so that if one tagger cannot tag a word, the next tagger is used, and so on.

Training a unigram part-of-speech tagger

UnigramTagger can be trained by giving it a list of tagged sentences at initialization.

```
>>> from nltk.tag import UnigramTagger
>>> from nltk.corpus import treebank
>>> train_sents = treebank.tagged_sents()[:3000]
>>> tagger = UnigramTagger(train_sents)
>>> treebank.sents()[0]
['Pierre', 'Vinken', ',', '61', 'years', 'old', ',', 'will', 'join', 'the', 'board', 'as', 'a', 'nonexecutive', 'director', 'Nov.', '29', '.']
>>> tagger.tag(treebank.sents()[0])
[('Pierre', 'NNP'), ('Vinken', 'NNP'), (',', ','), ('61', 'CD'), ('years', 'NNS'), ('old', 'JJ'), (',', ','), ('will', 'MD'), ('join', 'VB'), ('the', 'DT'), ('board', 'NN'), ('as', 'IN'), ('a', 'DT'), ('nonexecutive', 'JJ'), ('director', 'NN'), ('Nov.', 'NNP'), ('29', 'CD'), ('.', '.')]
```

We use the first 3000 tagged sentences of the treebank corpus as the training set to initialize the UnigramTagger class. Then, we see the first sentence as a list of words, and can see how it is transformed by the tag() function into a list of tagged tokens.

You can find a list example of tags used by the program Tree Tagger :
http://courses.washington.edu/hypertxt/csar-v02/penntable.html

---

The main goal of this work package is to identify if reviews of movies are positives or negatives.

Download movie reviews polarity dataset v2.0 at http://www.cs.cornell.edu/people/pabo/movie-review-data

First,to identify positivity or negativity you will pos-tag the reviews and identify adverbes.
Second you will use a lexical resource to classify the adverbes found.

SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity.

Download it at: http://sentiwordnet.isti.cnr.it/

Propose a method based on pos-tagging and SentiWordNet to classify movie reviews.
You can use machine learning on top of your features !

For examples of how use sentiWordNet directly in NLTK please see :

http://www.nltk.org/_modules/nltk/corpus/reader/sentiwordnet.html