ELN 2012
Homework 2

Maximum Entropy Classifier

The following URLs contain revised versions the train and test set for the Evalita 2009 POS task:

http://medialab.di.unipi.it/Project/QA/NamedEntity/evalita09_train.iob
http://medialab.di.unipi.it/ Project/QA/NamedEntity/evalita09_test.iob

The input consists of sentences, one token per line, separated by one empty line.
Each line contains four, tab separated, fields:

FORM POSTAG ART NE

Where NE is in IOB notation denoting the types PER (person), ORG (organization), GPE (geographical physical entity). Hence B-PER marks the first token in a person entity, IPER the following ones and O marks the tokens not belonging to any entity.
For example:

```
il RS O
capitano SS O
della ES O
Gerolsteiner SPN B-ORG
Davide SPN B-PER
Rebellin SPN I-PER
ha VIY O
allungato VSP O
```

Within the file, the start of individual articles is marked by lines such as:

-DOCSTART- adige20041007_id413942 0

which provide the ID of each article.

You will have to train a POS tagger on these data using the nltk module classify.maxent.
This involves defining a set of features to provide to the classifier representing a training event for each token. For example such features may include orthographic features, such as whether the word starts with a capital letter, or contains a number, or any hyphenation, or position of the token the sentence. You can use regular expressions to check patterns in the words. You can also use feature conjunctions, such as whether the word is capitalized AND the value of the previous label.
You can also use global document features, such as the presence earlier in the same document.
In order to train the classifier you will have to create a vector of training examples. Each example consists in a pair:

(features, label)

where features is a dictionary mapping strings to boolean representing the presence or absence of that feature.
Create a classifier on those examples:

classifier = trainer(examples)

Use the same feature extractor for implementing the tagger, and check the accuracy:

acc = accuracy(classifier, test)

MEMM

A maximum entropy Markov model (MEMM) is trained in exactly the same way as a regular maximum entropy model, where each word corresponds to one datum, and the correct label for the previous word can be used in features for the current word.
The primary difference comes at test time, when a Viterbi decoder must be used to find the best possible sequence of labels, instead of greedily finding the best label at each point.

Write such decoder.