5. Part-Of-Speech Tagging for Christmas Wishes

Part-Of-Speech Tagging is the process of marking up a word in a text as corresponding to a particular category of words that have similar grammatical properties. For instance, noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection. In this task, we limit our corpus to a set of Christmas wishes given in the course Webpage 1. Moreover, we use relaxed categories consisting of: NOUN, ADJECTIVE, VERB, PRONOUN, and STOP words. The goal is to tag each word of a given Christmas wish using one of the above categories. To achieve this task, you need to use a Hidden Markov Model where:

(a) **States** are the part of speech tags
(b) **Observations** are the words
(c) **Transitions** among states represent their consecutive positions in the text. In other words, there is a transition from state $i$ to state $j$ if a word of category $i$ is followed by a word of category $j$ in the text.

After building the Hidden Markov Model, you will be able to compute the probability of using a given Christmas wish, and also be able to tag each word contained in it. Concretely, the steps to follow are:

1. Compute the transition probabilities using the Christmas wishes dataset 2
2. Implement a recursive algorithm that computes the probability of a given Christmas wish (the probability of a series of observations using $\alpha_t(i)$)

---

1http://www.inf.unibz.it/mkacimi/teaching.shtml
2http://www.inf.unibz.it/mkacimi/teaching.shtml
3. Implement an algorithm that finds POS tags for a given Christmas wish. This means that you need to find the best path that maximizes the probability of the observations to occur. You can do a naive implementation by first finding all paths having the same length of the series of observations (the length of the Christmas wish in words), then you compute the probability of the series of observations given each path, then choose the maximum. Optionally, you can implement the Viterbi Algorithm that would make use of the recursion implemented in the previous question.

The expected outcome of this task is a program that takes as input a Christmas wish (using only the vocabulary provided in the dataset), and provide as output the probability of the wish and its POS tags. Write a report of 1-2 pages about the main functions you have implemented and how they can be used to achieve this task.

**Deadline**: 14/01/2015 at 10:30 am.