Zadanie zaliczeniowe #1, SAD II

Ukryta struktura w... Panu Tadeuszu

Pewien człowiek, który nie zna liter alfabetu łacińskiego zaczął przeglądać polskie teksty. Intuicja podpowiada mu, że symbole nie mają przypadkowego charakteru. Na szczęście ma znajomych statystyków-programistów, którzy chcą spieszyć mu z pomocą.

Nieznajomy spodziewa się, że strukturę tekstów, które przegląda można dobrze modelować Ukrytymi Modelami Markowa - HMM (naturalnie, jest to pierwszy pomysł jaki przychodzi do głowy).

Jego pierwszy pomysł, to sprawdzenie, czy istnieją pewne własności pojedynczych znaków w tekście. W najprostszym ujęciu spodziewa się, że znaki można podzielić na dwie grupy.
W języku Ukrytych Modeli Markowa, oczekuje dwóch ukrytych stanów, które emitują wszystkie możliwe znaki występujące w analizowanym tekście.

Problem 1: Przygotuj HMM, który posiada dwa ukryte stany i wszystkie możliwe znaki występujące w tekście Pana Tadeusza (link w tytule zadania). Prawdopodobieństwa emisji oraz przejść zainicjalizuj w sposób losowy (lekko zaburzone średnie). Używając załączonego tekstu, wytrenuj HMM i sprawdź, czy ukryte stany są związane z jakąkolwiek własnością języka (alfabetu).
Czy wprowadzenie dodatkowych stanów wskazuje dodatkowe własności alfabetu?

Problem 2: Drugi pomysł, to analiza słów. Jest to jednak zadanie dużo bardziej skomplikowane, ponieważ ich liczba i powtarzalność w tekstach pozostawiają wiele do życzenia…
Niemniej, nieznajomy proponuje wykorzystanie tylko i wyłącznie sufiksów słów - unikalnych zakończeń składających się z trzech liter.
Oczekujemy HMM z 5-cioma ukrytymi stanami, które emitują wszystkie możliwe sufiksy (trzy lub mniej literowe) występujące w tekście.
Czy możemy wskazać jakieś własności słów?

Po wytrenowaniu dwóch łańcuchów Markowa przetestuj rozpoznawanie znaków i słów na ulubionych fragmentach poezji.

Rozwiązanie powinno składać się z
- kodu źródłowego
- krótkiego komentarza dotyczącego wyników powyższych eksperymentów

Uwagi techniczne:
- w analizowanym tekście pozostawiamy tylko litery i spacje, pozostałe znaki interpunkcyjne usuwamy.
- sufiksem słowa dwuliterowego jest całe słowo lub słowo z dodanym pustym znakiem na początku (decyzja należy do autora rozwiązania)

Termin nadsyłania rozwiązań - 7.05.2017 (do negocjacji)