Inteligencia Artificial:
Proyecto *etiquetador de partes del habla para una oración*

Luis A. Pineda Cortes
luis at leibniz.iimas.unam.mx
IIMAS, UNAM

Ivan V. Meza Ruiz
ivanvladimir at turing.iimas.unam.mx
IIMAS, UNAM

26 de noviembre de 2009

Objetivo

Crear varios etiquetadores de partes del habla/categorías gramaticales usando Modelos Ocultos de Markov para su comparación.

Condiciones de entrega

Fecha de entrega: 2 de diciembre hasta las 12:00pm.

Archivo a entregar: Un archivo `zip/tgz` con el nombre con el formato *nombre_apellido*.tgz. Este archivo contendrá las salidas del script `entrena` solicitadas y un archivo con el texto de la respuesta a cada una de las preguntas que así lo requieran.

Vía de entrega: email a la cuenta: ivanvladimir at turing.iimas.unam.mx con el subject *PROYECTO IA*.

Descripción

Requerimientos

- Python y Numpy (`http://prdownloads.sourceforge.net/numpy/numpy-1.3.0.tar.gz`)

- Corpus de entrenamiento y de prueba disponible en http://leibniz.iimas.unam.mx/~luis/cursos/IA/index.html#Proyectos
- Natural Language Toolkit modificado para este proyecto disponible en la misma liga anterior.

Consideraciones

- Descomprima el archivo en un directorio de su preferencia. Esto dejará un nuevo directorio `nltk` en su directorio.
- Agregar a la variable de ambiente PYTHONPATH el path al directorio nltk. El siguiente comando en linux hace el truco:
  `export PYTHONPATH=$PYTHONPATH:"$PWD"`
- El script `entretar` dentro del directorio `nlt` entrena un corpus y etiqueta otro, para esto recibe cuatro archivos: el vocabulario, las etiquetas, el corpus de entrenamiento y el corpus a etiquetar.
- En el directorio `corpus_DIME` Existen tres corpus, dos de entrenamiento y uno de prueba. Los de entrenamientos se dividen en con supervisión (con las etiquetas después del símbolo `@@`) y sin supervisión (no contiene etiquetas)
- En el directorio corpus, existe dos archivos que contienen el vocabulario y las etiquetas del modelo. Notar, que vocabulario contiene TODAS las palabras, tanto de entrenamiento como de prueba.

Partes a desarrollar

- Utilizando el programa `entrenar` crear un modelo oculto de Markov utilizando el corpus de entrenamiento con supervisión y etiquetar el corpus de prueba (poner la salida en pantalla en un archivo, un punto).
- Utilizando el programa `entrenar` crear un modelo oculto de Markov utilizando el corpus de entrenamiento sin supervisión y etiquetar el corpus de prueba (poner la salida en pantalla en un archivo, un punto).
- Modificar el corpus con supervisión e eliminar las etiquetaciones de tal forma que sea sin supervisión. Utilizando el programa `entrenar` crear un modelo oculto de Markov utilizando el corpus recién modificado y etiquetar el corpus de prueba (poner la salida en pantalla en un archivo, un punto).
- Crear un nuevo corpus de la unión del corpus original de sin supervisión y el creado en el punto anterior. Utilizando el programa `entrenar` crear un modelo oculto de Markov utilizando el corpus recién creado y etiquetar el corpus de prueba (poner la salida en pantalla en un archivo, un punto).

- Todos los experimentos anteriores utilizan el archivo de vocabulario que contiene todas las palabras de los corpus. Esto no es razonable, modificar los corpus de entrenamiento para sustituir las palabras que ocurren menos de dos veces con el símbolo `unk`. Crear un nuevo archivo de vocabulario que únicamente contenga las palabras de los corpus de entrenamiento y por su puesto el símbolo `unk`. Sustituir las palabras del corpus de prueba que no ocurren en el nuevo archivo de vocabulario con el símbolo `unk`. Finalmente, repetir los cuatro experimentos anteriores (4 puntos)

- Reportar los resultados de *accuracy*/exactitud en una tabla de tal forma que modelos que sea intuitivo la comparación de los modelos creados. Crear una conclusión de los resultados donde se expliquen la naturaleza de los resultados (dos puntos).

- Modificando el script `entrena` crear un modelo HMM utilizando el corpus de entrenamiento con supervisión. Luego utilizar ese modelo como inicial del modelo sin supervisión (poner la salida en pantalla en un archivo e incluir script modificado, dos puntos).

Nombre:

Calificación:

Observaciones: