Tópicos selectos de Análisis de Datos
Tarea 2

Para entregar el 19 de septiembre de 2018

1. Este ejercicio es sobre Hidden Markov Models (HMM).

   Vimos en clase que el método más usado para el proceso de POS-tagging es HMM, donde asignamos etiquetas gramaticales POS (variables *latentes u ocultas*) a una secuencia de palabras (variables *observables*).

   Dado un Corpus de entrenamiento y una secuencia de palabras de prueba, HMM calcula la *secuencia de etiquetas POS* más probable mediante la expresión

   $$\hat{t}_1^n = \arg\max_{t_1^n} P(t_1^n|w_1^n) \approx \arg\max_{t_1^n} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}),$$

   o, en palabras,

   $$\hat{t}_1^n \approx \arg\max_{t_1^n} \prod P(\text{emisión})P(\text{transición}).$$

   Estas probabilidades están totalmente definidas por el Corpus, y el proceso de entrenamiento de la red se realiza mediante operaciones de conteo de las palabras y las etiquetas contenidas.

   Vamos a utilizar un *mini* Corpus tomado de la NYU[1], el cual puedes encontrar en el archivo `POSData.zip`, y contiene el corpus de entrenamiento (`training.pos`) y otro corpus de prueba (`development.pos`).

   a) Calcula las probabilidades de emisión y transmisión (verosimilitud y apriori) a partir del corpus de entrenamiento.

   b) Considera el texto de prueba:

   ```
   Your contribution to Goodwill will mean more than you may know.
   ```

   Obten los tokens del texto de prueba. Verifica que, para el corpus de entrenamiento usado, las posibles etiquetas POS para cada token es:

   ```
   $Your
   [1] "PRP$"
   ```

[1] `http://cs.nyu.edu/courses/spring12/CSCI-GA.2590-001/`

```
$contribution
[1] "NN"
$to
[1] "IN" "TO"
$Goodwill
[1] "NNP"
$will
[1] "NN" "MD"
$mean
[1] "JJ"  "VB"  "VBP"
$more
[1] "RB"  "JJR" "RBR"
$than
[1] "IN"
$you
[1] "PRP"
$may
[1] "MD"
$know
[1] "NN"  "VB"  "VBP"
$.
[1] "COMMA" "."
```

c) Verifica que el número de posibles secuencias (*paths*) para nuestro sencillo texto de prueba es de 216. Esto te puede dar una idea de la complejidad computacional de éste tipo de problemas. Puedes verificarlo computacionalmente.

d) Usa el algoritmo Viterbi para estimar la secuencia de etiquetas POS del texto de prueba. Compara tu resultado con el obtenido al usar el Anotador de `coreNLP` de Stanford.

En `R` puedes usar la función `viterbi(hmm,tokens)`, incluida en la librería `HMM`. Esta función recibe como parámetros un modelo HMM y el conjunto de tokens de prueba, y calcula la secuencia más probable de variabes ocultas, en nuestro caso, etiquetas POS.

El modelo HMM debes definirlo con `initHMM()` usando las probabilidades apriori y verosimilitud del Corpus de prueba. Revisa la ayuda de la función.

e) Generalmente, no todas las palabras están incluidas en el Corpus de entrenamiento. Intenta, por ejemplo, asignar POS-tags al texto:

`Coming to Goodwill was the first step toward my becoming totally`.

¿Qué podemos hacer en este caso? Implementa tu idea y verifica su desempeño con textos del corpus de prueba.