Primera Tarea
Fecha de Entrega:

Gabriel Infante-Lopez

August 29, 2005

1 Entropia de un texto

En este experimento determinaran la entropia condicional de la distribucion de una palabra en un texto dada otra. Para esto, primero se debera computar $p(i, j)$, que es la probabilidad de encontrar en cualquier posicion del texto, la palabra i seguida de la palabra j y $p(j|i)$ que es la probabilidad de que si la palabra i ocurra en el texto la siga la palabra j. Dada estas probabilities, la entropia condicional de la distribucion de palabras en un texto dada la palabra anterior puede ser computada como sigue:

$$H(J|I) = -\sum_{i \in I, j \in J} p(i, j) \log_2 p(j|i)$$

La perplejidad se computa como

$$PX(p(J|I)) = 2^{H(J|I)}$$

Comput la entropia y la perplejidad para algn texto en espaol de no menos de 150.000 palabras.

Seguidamente, se mezclara el texto y se medira como esto altera la entropia condicional. Para cada *caractr* en el texto, mezclalo con una probabilidad de 0.1. Si un caractr es elegido para ser mezcaldo, mapealo a un caracter elegido aleatoriamente del conjunto de caracteres que ocurren en el texto. Como hay algo de aleatoriedad en la salida del experimento, corr el experimento 10 veces, cada vez midiendo la entropia condicional del texto resultante, y reporta $\min$, $\max$, y la entropia promedio de estos experimentos. Asegurate de resetear la semilla del generador de numeros aleatorios. Adems, de que cada vez ests mezclando el archivo original, y no los archivos que ya habian sido mezclados. Realizar el mismo experimento para probabilidades de mezclado de 0.5, 0.1, 0.01 ,$.001$, y $.0001$.

Ahora hace lo mismo para algun texto en ingles de la misma cantidad de palabras.

Marc, dibuj y explic tus resultados. Adems, intent explicar las differencias entre los dos lenguajes. Para substanciar las explicaciones, intent explicar las diferencias entre los dos textos, evalua, por ejemplo, cantidad de palabras, numero de caracteres (total, y por palabra), la frecuencia de la palabra mas frecuente, la cantidad de palabras con frecuencia 1, etc.

2 Cross-Entropy y modelado de lenguage

Esta tarea mostrara la importancia de smoothing para modelado de lenguaje, y a cierto detalle te dejara sentir su efecto. Primero, tendrs que preparar los datos usados en el experimento anterior.

Prepar 3 dataset de cada archivo: saca las ultimas 20.000 palabras y llamalas *test data*, despues, saca las ultimas 40.000 palabras de lo que queda, y llamal *Heldout Data*, y llam lo restante *Trainning Data*

Aqu viene el codificado: extrae conteo de palabras del material de entrenamiento de manera que puedas calcular probabilidades de unigrams, bigrams y trigrams; calcul ademas la probabilidad uniforme basado en el tamao del vocabulario. Recorda, T es el tamao del texto, y V el tamao del vocabulario, i.e., el numero de tipos diferentes de formas de palabras en el material de entrenamiento.

$$p_0(w_i) = \frac{1}{V} \quad (1)$$

$$p_1(w_i) = c_1(w_i)/T \quad (2)$$

$$p_2(w_i/w_{i-1} = c_2(w_{i-1}w_i/c_1(w_{i-1}) \quad (3)$$

$$p_3(w_i/w_{i-2}, w_{i-1}) = c_3(w_{i-2}, w_{i-1}, w_i/c_2(w_{i-2}, w_{i-1}) \quad (4)$$

Recorda como manejar correctamente los principios y finales del material de entrenamiento con respecto al contero de bigramas y trigramas.

Ahora computa el cuatro parametros de smothing (i.e., coeficientes, pesos, lambdas o como quieras llamarlo) para el trigrama, bigrama, unigrama y la distribucion uniforme utilizando heldout data y el algoritmo EM. Recorda que el model suavizado tiene la siguiente forma:

$$p_s(w_i/w_{i-2}, w_{i-1} = l_0 p_0(w_i) + l_1 p_1(w_{i)} + l_3 p_3(w_i/w_{i-2}, w_{i-1})$$

donde

$$l_0 + l_1 = l_2 + l_3 = 1$$

Finalmente, computa la cross-entropia del material de testeo usando tu modelo de lenguaje suavizado. Ahora afina el suavizado de la siguiente manera. Suma 10%, 20%, 30%, ..., 90%, 95% y 99% de la diferencia entre los parametros de smoothing y 1.0 de su valor, descontando los otros 3 parametros proporcionalmente.

Despues establece los parametros de smooting del trigram a el 90%, 80%, 70%, ... 10%, 0% de su valor, aumentando proporcionalmete los otros tres parametros, de manera que sumen 1 Compute la cross-entropy en el conjuto de testeo para estos 22 casos (original + 11 trigram parametros aumentados + 10 trigram parametros decrementados). Tabula, grafica y explica lo que obtuviste.

Adems, trat de explicar la diferencia entre los lenguajes basandote en estadisticas similares a las usadas en la tarea 1, mas el grafico de cobertura (definido como el porcentaje de palabras en el texto de datos que han sido vistos en el material de entrenamiento). Agreg a tu codigo fuente comentarios para que se pueda entender de una manera simple los experimentos hechos.