Introducción al Procesamiento de Lenguaje Natural

Diciembre de 2018

Consideraciones generales
i)La prueba es sin material escrito.
ii)Escriba nombre y C.I. en todas las hojas.
iii)Numere todas las hojas.
iv)En la primera hoja, indique el total de hojas.
v)Comience cada ejercicio en una hoja nueva.
vi)Utilice las hojas de un solo lado.
vii)Entregue los ejercicios en orden
viii)El total de puntos es 40

Ejercicio 1 [12 puntos]

a) ¿Qué se conoce por representación one-hot en el contexto de las representaciones vectoriales de palabras?
b) Describa en términos generales en que consisten los métodos basados en predicción para la construcción de representaciones vectoriales de palabras.
c) Indique en qué consiste la medida pointwise mutual information (PMI) y cómo puede utilizarse en los métodos basados en conteo.
d) Considere el siguiente repertorio de oraciones:

El perro come hueso
El gato come pescado
El gato no come hueso
El perro come pescado
El gato no anda en bicicleta
Este perro anda en ómnibus
María come con Juan en el ómnibus
Juan anda en bicicleta
Juan anda en ómnibus
María anda en bicicleta
María anda en ómnibus

(i) Construya la matriz de co-ocurrencias palabra-palabra considerando como contexto de co-ocurrencia la oración completa. Considere para la matriz únicamente las palabras “come”, “anda”, “perro”, “gato”, “Juan”, “María”, “ómnibus” y “bicicleta”.

(ii) Grafique sobre los ejes correspondientes a “come” y “anda”, de acuerdo a las frecuencias de la matriz de la parte a), las palabras: “perro”, “gato”, “Juan”, “María”, “ómnibus” y “bicicleta”

a) La representación one-hot, en el contexto de las representaciones vectoriales de palabras, refiere al índice de la palabra a representar en el vocabulario, estableciendo previamente cierto orden. En términos vectoriales consiste en un vector cuyas componente son cero excepto la de la palabra a representar que es uno.

b) Los métodos basados en predicción son métodos que iterativamente ajustan pesos según una función objetivo. Los pesos ajustados son, o forman parte de, las representaciones vectoriales. Como ejemplo de métodos de este tipo es posible nombre SGNS, CBOW y fasttext.

c) La PMI es una medida de asociación estadística. Su formulación matemática es

$$pmi(x; y) = \log (p(x,y)/(p(x)p(y))$$

Su valor varía entre menos infinito y min(-log(p(x)), -log(p(y))), los valores positivos indican que x e y tienden a co-ocurrir. Esta medida puede ser utilizada para extraer colocaciones de un corpus. En los métodos basados en conteo la pmi ser usada como medida de asociación en la matriz palabra-palabra.

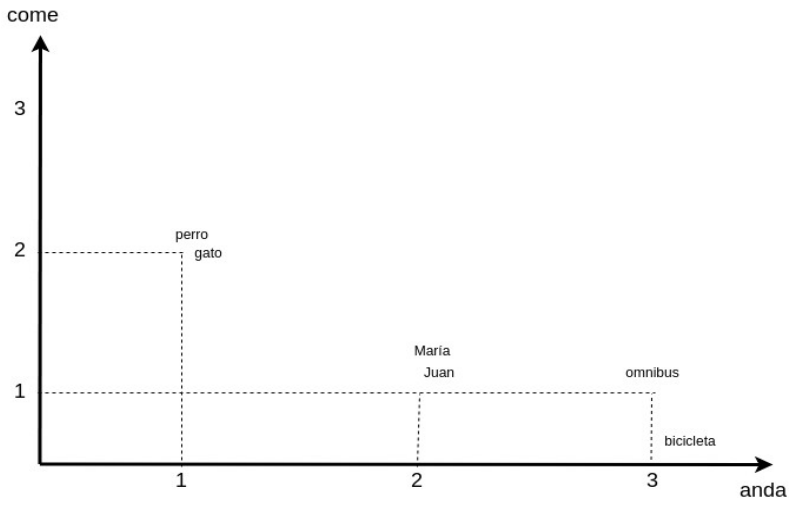
d)

i) Dado que la matriz es simétrica escribimos únicamente la parte triangular superior.

$$\begin{pmatrix} 5 & 0 & 2 & 2 & 1 & 1 & 1 & 0 \\ & 6 & 1 & 1 & 2 & 2 & 3 & 3 \\ & & 3 & 0 & 0 & 0 & 1 & 0 \\ & & & 3 & 0 & 0 & 0 & 1 \\ & & & & 3 & 1 & 2 & 1 \\ & & & & & 3 & 2 & 1 \\ & & & & & & 4 & 0 \\ & & & & & & & 3 \end{pmatrix}$$

Las componentes son: come, anda, perro, gato, Juan, María, ómnibus y bicicleta (en ese orden).

ii)



Ejercicio 2 [6 puntos]

Indique si las siguientes proposiciones son Verdaderas o Falsas. Justifique en cada caso.

a) La reducción β simplifica una fórmula del cálculo lambda aplicando los términos con abstracción funcional a constantes.

Falso: aplica los términos con abstracción funciona a otros términos en general y no únicamente a constantes.

b) El método SVM (Support Vector Machines) es un método secuencial que se utiliza para el etiquetado morfosintáctico.

Falso: SVM no es secuencial.

c) Naïve Bayes es un método de aprendizaje que genera clasificadores haciendo hipótesis de independencia, que no necesariamente se cumplen, entre los atributos de los ejemplos de entrenamiento.

Verdadero: por definición de Naïve Bayes.

d) La medida *tf-idf* asociada a un término y un documento en recuperación de información, se define como la cantidad de ocurrencias de un ese término en ese documento dividido la cantidad de documentos

Falso: tf es la frecuencia de un término en un documento (cantidad de ocurrencias) * la frecuencia inversa de ese término en la colección de documentos; que es el log de cantidad de documentos/cantidad de documentos que contienen al término.

Ejercicio 3 [10 puntos]

Considere una gramática G con las siguientes reglas:

O → GV | GN GV
GN → Det Nom | Nom | GN GP
GV → V | V GN | V GN GP
GP → Prep GN
Det → el | la | los | las
Nom → Carlos | caramelos | sandwiches | manzana | Salta
V → como | come | salta
Prep → con | de | en

a) Aplique el algoritmo CKY para la entrada *"como sandwiches de manzana"* considerando las reglas de la gramática G. ¿Qué salida devuelve el algoritmo? Justifique.

Primero se verifica que la gramática esté en FNC (Forma Normal de Chomsky). Como no está, el primer paso es normalizarla.

Quedan entonces las siguientes reglas de producción:

O → como | come | salta | V GN | Aux1 GP | GN GV
GN → Det Nom | Carlos | caramelos | sandwiches | manzana | Salta | GN GP
GV → como | come | salta | V GN | Aux1 GP
GP → Prep GN
Aux1 → V GN
Det → el | la | los | las
Nom → Carlos | caramelos | sandwiches | manzana | Salta
V → como | come | salta
Prep → con | de | en

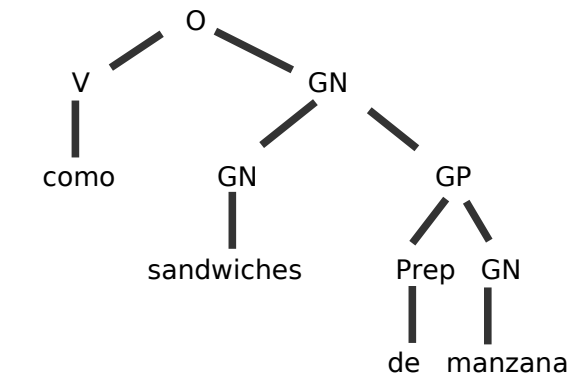
Analizamos utilizando CKY ***como sandwiches de manzana***

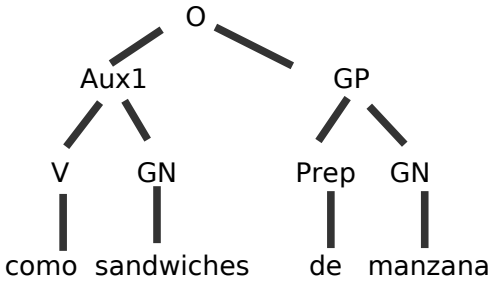
	como (1)	**sandwiches (2)**	**de (3)**	**manzana (4)**
0	O GV V	GV Aux1 O	∅	**O** GV Aux1
1		GN Nom	∅	GN
2			Prep	GP
3				GN Nom

El algoritmo devuelve TRUE, indicando que la secuencia de entrada puede derivarse a partir de la gramática, al estar la variable O en la celda V_{04}

b) Escriba el árbol o los árboles de análisis sintáctico correspondiente(s) a la parte a).

Se considera para esta parte la gramática en la FNC (pero podría haberse utilizado la original)





c) Realice las derivaciones de dos oraciones generadas por la gramática G que no contengan las mismas palabras y que no sean sintácticamente válidas en español. Explique porqué no lo son.

Hay varias posibles oraciones. A modo de ejemplo se realizan 2 derivaciones posibles distintas y con distintas palabras que son agramaticales

(1) O => GN GV => Carlos GV => Carlos como

Tomando ***como*** como verbo, la forma verbal correcta sería come

(2) O => V GN => salta GN => salta Det Nom => salta el Nom => salta el caramelos

Aquí el problema es por ejemplo la concordancia en número del determinante ***el*** y el sustantivo ***caramelos***.

Ejercicio 4 [12 puntos]

Considere los siguientes ejemplos de grupos nominales:

(i) *un gran jardín sombreado*
(ii) *un jarrón rojo*
(iii) *los verdes arbustos*

a) Indicar la categoría gramatical de cada una de las palabras que aparecen en los 3 ejemplos anteriores.

(i) *un gran jardín sombreado*
det adj nom adj

(ii) *un jarrón rojo*
det nom adj

(iii) *los verdes arbustos*
det adj nom

b) Escribir una gramática independiente de contexto que genere al menos los 3 ejemplos anteriores.

gn → det nom
nom → nom adj | adj nom | n
---
det → un | los
n → jardín | jarrón | arbustos |
adj → gran | sombreado | verdes | rojo

c) Escribir las componentes de interpretación semántica asociadas a las reglas de la gramática de la parte 2.

gn → det nom	*gn.sem = det.sem(nom.sem)*
nom → nom adj	*nom.sem = adj.sem (nom.sem)*
nom → adj nom	*nom.sem = adj.sem (nom.sem)*
nom → n	*nomj.sem = n.sem*
n → jardín	*n.sem = λ x . jardín(x)*
n → jarrón	*n.sem = λ x . jarrón(x)*
n → arbustos	*n.sem = λ x . arbustos(x)*
adj → gran	*adj.sem = λ P . λ x . gran (P(x))*
adj → rojo	*adj.sem = λ P . λ x . P(x) ∧ rojo(x)*
adj→ verdes	*adj.sem = λ P . λ x . (P(x) ∧ verdes(x))*
adj → sombreado	*adj.sem = λ P . λ x . (P(x) ∧ sombreado(x))*
det → los	*det.sem = λ P . λ Q . ∀x (P(x) → Q(x))*
det → un	*det.sem = λ P . λ Q . ∃x (P(x) ∧ Q(x))*

d) Realizar una derivación sintáctico-semántica para el ejemplo (i), efectuando las reducciones β que corresponda

gran jardín | *nom → adj nom | nom.sem = adj.sem(nom.sem)*
= λ P . λ x . gran (P(x)) (λ y . jardín(y))
= λ x . (gran (λy .jardín(y) (x)) = λ x .(gran (jardín(x))

gran jardín sombreado | *nom → nom adj | nom.sem = adj.sem(nom.sem)*
= λ P . λ x . (P(x) ^ sombreado(x)) (λ y .(gran (jardín(y)))
= λ x . (λ y .(gran (jardín(y))(x) ^ sombreado(x)))
= λ x . (gran (jardín(x)) ^ sombreado(x))

un gran jardín sombreado | *gn → det nom | gn.sem = det.sem(nom.sem)*
= λ P . λ Q . ∃x (P(x) ^ Q(x))(λ y . (gran (jardín(y)) ^ sombreado(y)))
= λ Q . ∃x ((λ y . (gran (jardín(y)) ^ sombreado(y)) (x) ^ Q(x))
= λ Q . ∃x (gran (jardín(x)) ^ sombreado(x) ^ Q(x))