Examen de procesamiento de lenguaje natural

Alumno: Miguel Angel Soto Hernández

1- ¿De las palabras "con, mesas, vio, parece, una, Juan, ella", cuales son stop words (palabras auxiliares)?

con, una, ella

2- ¿Cuáles son los temas de la palabra "trabajo"?

trabajar

3- Explique el modelo de espacio vectorial (5-6 líneas). Sin fórmulas.

Consta de comparar objetos de manera formal, es decir, si queremos representar dos o más objetos, seleccionaremos características y los valores de estas características para cada uno de estos objetos. La selección de estas características es en un principio subjetiva, ya que dependerá de lo que se quiera modelar, y posteriormente en la comparación será objetiva

4- ¿Cómo se formula el problema principal de aprendizaje automático?

Definir el problema que queremos resolver

5.- ¿Qué es el 10 fold cross validation? Explique.

Es una técnica donde K en este caso 10 dividirá todos los datos en 10 volúmenes. Una vez que ya estén divididos, utilizará el volumen 1 para evaluación de resultados y los volumenes del 2 al 10 para entrenamiento, una vez que se haya realizado esta evaluación se tomará el volumen 2 para evaluación y los volumenes restantes para el entrenamiento, y así sucesivamente hasta que los 10 volumenes hayan sido ocupados como evaluadores. Cuando ya se tengan los resultados de las 10 evaluaciones, se toma un promedio y este será el valor final. Con esta técnica podemos neutralizar la influencia de fluctuaciones si es que las hay en los datos que estamos trabajando.

6.- Describa el diseño de experimento en la lingüística computacional moderna (con el uso de aprendizaje automático supervisado).

Para diseñar un experimento se requiere seguir una serie de pasos, explicados a continuación:

1.- Definir la tarea a realizar (por ejemplo, detección de spam)
2- Seleccionar nuestros datos o textos que utilizaremos para el experimento, en este paso hay que tener en consideración que es mejor tomar textos o corpus existentes, esto con el fin de interpretarlos de manera mas objetiva.
3- Preparar el estandar de oro, para este paso será necesario etiquetar nuestros textos.
4- Construir el modelo de espacio vectorial, aquí seleccionaremos las características y sus valores que ocuparemos (podemos utilizar unigramas o bolsa de palabras, n-gramas, n-gramas sintácticos, etc)
5- Definir uno o varios métodos de línea base y uno o varios metodos del estado del arte.
6- Seleccionar uno o varios métodos de aprendizaje automático supervisado

7. Convertir nuestros datos textuales a un formato que pueda aceptar nuestro método de aprendizaje

8- Llevar a cabo los experimentos de clasificación automática

9.- Calcular nuestros valores de precisión, especificidad y F1

7- Escoja un problema de su preferencia (ej. clasificación temática de textos, WSD, NER, etc), y describa el diseño de experimentos para el

1. Detección de spam
2- Seleccionar nuestro conjunto de datos
3- Etiquetar el conjunto de datos si es que no esta etiquetado
4- Seleccionar las características que tomaremos en consideración para la detección de spam, en este caso pueden ser palabras o tokens específicos, así como n-gramas
5- Definir nuestra línea base, en este caso puede ser que detectemos todo como spam o todo como no spam
6- Seleccionamos nuestro metodo de aprendizaje automatico supervisado, en este caso y por simplicidad puede ser Naive Bayes
7- Ahora convertiremos nuestros datos de nuestro modelo de espacio vectorial a algun formato aceptado por nuestro método de aprendizaje automatico, por ejemplo, si estamos ocupando pyton puede ser transformado a un DataFrame de la librería pandas o si estamos ocupando WEKA los datos tienen que ser convertidos a formato ARFF
8- Ejecutaremos nuestros métodos de aprendizaje automático
9.- Calculamos los valores de presición, especificidad y F1 con los resultados que arroje nuestro método de aprendizaje automático.

8- Inventar 3 textos de 5 palabras cada uno (que se repitan 3 o 4 palabras). Construir el índice invertido de esos textos. Construir la matriz termino-documento, llenarlo con tf-idf cada palabra en esta colección. Calcular la similitud entre parejas utilizando coseno.

Texto 1: El va a correr solo

Texto 2: Ella va a correr acompañada

Texto 3: El come solo con tenedor

Vocabulario sin palabras auxiliares

Palabra	Frecuencia
Va	2
correr	2
solo	2
acompañada	1
come	1
tenedor	1

Termino	Documento
va	1, 2
correr	1, 2
solo	1, 3
acompañada	2
come	3
tenedor	3

TF

	Texto1	Texto2	Texto3
va	1/2	1/2	0
correr	1/2	1/2	0
solo	1/2	0	1/2
acompañada	0	1	0
come	0	0	1
tenedor	0	0	1

palabra	IDF
va	1.28
correr	1.28
solo	1.28
acompañada	1.69
come	1.69
tenedor	1.69

Similitudes

Texto 1 y 2: .99

Texto 2 y 3: .99

Texto 1 y 3: .99

	Va	correr	solo	acompañada	come	tenedor	
Texto 1	0.64	0.64	0.64	0	0	0	1.10
Texto 2	0.64	0.64	0	1.69	0	0	1.98
Texto 3	0	0	0.64	0	1.69	1.69	2.47

9.- ¿Cuándo se utiliza el índice invertido?

Cuando queremos estructurar la información que queremos utilizar, de esta manera podemos llevar una búsqueda en los textos más completa ya que guarda los valores de frecuencia

10.- El modelo de espacio vectorial se puede utilizar los valores no numéricos ordenados como valores de una de las características Sin embargo, si los usamos así, el modelo interpretará que los valores ~~[illegible]~~ más lejanos corresponden a los objetos mas distintos, por ejemplo, rojo y azul son mas lejanos, entonces hay mayor diferencia ¿Cómo evitar ese comportamiento y poder usar este tipo de características?

Normalizando, con la norma euclidiana

11.- Si de los 100 objetos, 20 son los que buscamos y el sistema encontró correctamente 10 y nos dió 20 más incorrectos ¿Cuáles son los valores de presición, recall y F1?

$$P = \frac{10}{10+20} = \frac{1}{3} = 0.3\overline{3}$$

$$R = \frac{10}{10+10} = \frac{1}{2} = 0.5$$

$$F1 = \frac{2\left(\frac{1}{3}\right)\left(\frac{1}{2}\right)}{\frac{1}{3}+\frac{1}{2}} = 0.39$$

12: Qué es cero significativo en la lingüística. De un ejemplo
Por qué es importante este concepto?
es un elemento neutro, es decir, cuando no se dice nada
y a pesar de eso se entiende

Ejemplo: mesa (singular)