Curso de IA

Capítulo 7. Cuestionario

Para estudiantes

Samsung Innovation Campus

1. ¿Cuál de las siguientes declaraciones sobre la minería de textos es incorrecta?

   1 Se puede conocer la reacción a un grupo específico.

   2 Se puede obtener información sobre el liderazgo de ventas.

   3 Se pueden establecer estrategias competitivas monitoreando otras marcas.

   4 Se puede aplicar el mismo método a varios idiomas.

2. ¿Cuál de las siguientes afirmaciones no es cierta sobre el uso de las redes sociales?

   ① Es posible conocer cuántos grupos componen la red.

   ② Es posible conocer clientes influyentes.

   ③ Es posible saber ver el cambio a lo largo del tiempo.

   ④ Es posible saber si los clientes se irán la próxima vez.

3. ¿Qué término significa separar una raíz de una palabra cuya forma ha sido modificada para extraer la palabra que es objeto del análisis morfémico?

Es el proceso de reducir una palabra a su forma base o lema, así mismo se considera su significado y contexto gramatical, bajo este término se podría decir que este proceso utiliza diccionarios y reglas lingüísticas para realizar u na reducción

4. ¿Qué término designa un conjunto de materiales que pueden mostrar los aspectos esenciales de la lengua como m aterial de investigación requerido en cada campo de la investigación lingüística?

Corpus es el termino pues este significa : Conjunto de datos de texto sujetos a análisis donde se pueden incluir nume rosos datos como:

Voz Humana, Datos de frecuencia entre otros.

5. Explique qué es la TF (Term Frequency) y escriba cómo calcular la TF (fórmula).

El Term frequency (TF) indica la importancia relativa de cada palabra en este caso también definido como termino de ntro de un documento, si una palabra aparece con un alto nivel de frecuencia en un documento corto TF tendrá un v alor mayor
Para calcular un TF primero se cuenta el numero de veces que aparece un término T en un documento en código se h aría de la siguiente forma:
Convertimos el texto a minúsculas para evitar la diferenciación por mayúsculas con la función Split()
Contamos las palabras con la función Counter() para contar cuantas veces aparece una palabra
Aplicamos la formula donde TF(t, d) Numero de veces que aparece la palabra / Numero total de palabras en el docue mtno
Imprimimos resultados

6. Encuentre la declaración más similar utilizando TF-IDF y similitud coseno con referencia al siguiente código.

```
importar nltk
importar numpy como np
desde nltk.corpus importar stopwords
desde sklearn.feature_extraction.text importar TfidfVectorizer
desde sklearn.metrics importar pairwise_distances

doc = [
   "I es un campo en rápido avance que implica el desarrollo de máquinas inteligentes",
   "El aprendizaje automático es un subconjunto de la IA que se centra en la capacitación de máquinas para apr
ender de los datos.",
   " El aprendizaje profundo es un subcampo del aprendizaje automático que utiliza redes neuronales con múlti
ples capas",
   " La IA se aplica en diversos sectores, como la sanidad, el transporte y el ocio.
",
   " Las consideraciones éticas desempeñan un papel importante en el desarrollo de la IA",
   " La inteligencia artificial puede revolucionar muchos aspectos de la sociedad",
   " Los sistemas de IA, como los chatbots y los asistentes virtuales, son cada vez más comunes",
   "El procesamiento del lenguaje natural es una rama de la IA que permite a las máquinas comprender el lengu
aje humano",
   " El objetivo de la visión por computadora es que las máquinas comprendan la información visual",
]

# preprocesamiento.
doc = [x.lower() for x in doc]

# parámetros
max_features = 18
min_df = 1
max_df = 3
stop_words = stopwords.words('english')

vectorizer = TfidfVectorizer(max_features=max_features,
                     min_df=min_df,
                     max_df=max_df,
                     stop_words=stop_words)
X = vectorizer.fit_transform(doc).toarray()
```