1. La minería de texto es un proceso que implica la extracción de información útil y significativa a partir de grandes cantidades de texto no estructurado.

2. Tres diferencias entre minería de texto y minería de datos son:

1.La minería de texto se enfoca en el análisis de texto no estructurado, mientras que la minería de datos se ocupa de datos estructurados.

2. La minería de texto utiliza técnicas específicas para extraer información semántica y significativa del texto, mientras que la minería de datos se centra en descubrir patrones y tendencias en los datos.

3. La minería de texto a menudo implica el procesamiento del lenguaje natural (NLP), mientras que la minería de datos puede utilizar una variedad de técnicas, como aprendizaje automático y estadísticas.

3. La tokenización es el proceso de dividir un texto en unidades más pequeñas, llamadas tokens, que pueden ser palabras, frases o caracteres.

4. Estructura de la tokenización:

- Entrada: Texto sin procesar.

- Proceso: División del texto en tokens.

- Salida: Lista de tokens.

5. Pasos del proceso de tokenización:

1. Recogida de datos: Recopilar datos de texto de diversas fuentes.

2. 2. Preprocesamiento del texto: Normalizar, tokenizar, eliminar stopwords, corregir la ortografía e identificar

partes de la oración.

3. Extracción de características: Representar los datos de texto mediante técnicas como Bag of Words, TF-IDF o

incrustación de palabras.

4. Algoritmos de minería de texto: Aplicar clustering, clasificación, modelado de temas, análisis de

o reconocimiento de entidades con nombre.

5. Evaluación y validación de modelos: Evaluar el rendimiento de los algoritmos utilizando métricas como

exactitud y precisión.

6. Visualización e interpretación: Visualizar los resultados para obtener información y comprender los patrones

en los datos.

7. Despliegue e integración: Implementar soluciones de minería de texto para aplicaciones del mundo real.

6. La tokenización palabra por palabra implica dividir el texto en tokens basados en los espacios en blanco y la puntuación, donde cada token representa una palabra individual.

7. La tokenización letra por letra implica dividir el texto en tokens donde cada token representa una letra individual.

8. Ejemplos donde se utiliza la tokenización palabra por palabra:

a. Análisis de sentimientos en redes sociales.

b. Clasificación de documentos basada en el contenido de palabras clave.

9. Ejemplos donde se utiliza la tokenización letra por letra:

a. Reconocimiento de escritura a mano.

b. Análisis forense de texto en documentos cifrados.

10. NLTK significa Natural Language Toolkit