30/01/2024

TF-IDF

El termino de frecuencia/frecuencia inversa del documetno, cuantifica la importancia de una palabra dentro de un texto y de una determinda coleccion de corpus. Usos

* Clasificar en un motor de busqueda
* Hacer un resumen de texto
* Construer chatbots inteligentes

Frecuencia de una palabra en un document, frecuencia de una palabra en el corpus.

La tecina asigna puntuaciones a las palabras:

* Si una palabra aparece de forma frecuente en un documento, podemos determinar que tiene una puntuación alta. (TF)
* Si la palabra que aparece mucho en un documento no lo tiene en el resto de documentos, tiene una puntuación baja (IDF)

Pasos a seguir:

1. Se calculan los valores de frecuencia de un termino
2. Se calcula los valores de frecuencia inversa en el doucmetno de corpus
3. Se combina 1 y 2
4. TF(t, d): Frecuencia del termino entre el numero de términos en el documento
5. IDF(t, D): rareza del termino en el conjunto de documentos D

Log(numero total de documentos/numero de documentos que contiene t + 1)

1. Multiplicar paso 1 y 2

Ejemplo

1. Partimos de un corpus
2. Preprocesar el texto
3. Crear una matriz tf
4. Normalizar valores (ocurrencias entre totales)
5. Crear valores IDF)
6. Calcular TF-IDF asociada

Implementación

La librería scikit-learn proporciona 3 clases para este modelo

* TfidfVectorizer: obtiene la matriz tf-idf

06/02/2024

Repaso

TF-IDF que términos tienen mas importancia dentro de un documento (no es un resumen)

Topic modeling

Tecnica no supervisada del NLP que detecta y extra relaciones ocultas entre documentos. Determina la probabilidad que determinados términos aparezcan juntos

El termino tema (topico)

Ejemplo:

Identificar temas en un conjunto de noticias en los medios.

Proceso

Se crea una matriz de documentos y términos (count vectorizer)

27/02/2024

Los embedings son representaciones vectoriales de una palabra. La principal diferencia con una vectorización tipo One-hot, es que esta vectorización tiene en cuenta las ‘distancia’ entre palabras.

Estos vectores permiten realizar operaciones con otros vectores

Word2vec

Algoritmo de aprendizaje estadisco para hacer embeding, la cual busca proximidades entre palabras

LLMs

Grandes modelos de lenguaje, modelos de aprendizaje profundo para procesar y generar texto en lenguaje natural. Se consideran ‘grandes’ por el gran volumen de datos que utilizan

Que hacen los LLMs?

Permiten generar texto, hacer resumenses, y procesos Q&A. Estos modelos están basados en Transformers.

¿Qué características definene un transformer?

Obtiene patrones del lenguaje para buscar relaciones. Utiliza un mecanismo clave llamado ‘atencion’. Es el peso relativo de la palabra en el texto. Las redes neuronales tiene 2 componentes:

* Encoder: genera estados ocultos, información sobre el contexto. Genera los embedings. Usa la auto-atencion para generar estados ocultos.
* Decoder: considera una secuencia y usa la salida del encoder para generar prediciones
  + Se ingesta la secuencia de palabras, incorporando los estados ocultos (estado de contexto de la palabra)
  + Idk
  + Idk

Tipos de LLMs

* Auto-regresivos: orientados a predicción (GPT). Cual es la siguiente palabra dentro de un texto
* Modelos autocodificado: modelos destinados a llenar huecos en textos
* Modelos híbridos: combinan ambos modelos

LangChain

Framework para aplicaciones que generan textos, reponden preguntas. ‘Lang’ viene de lenguaje y ‘Chain’ se refiere a cadena. Agrega funcionalidad donde un LLM no llega de por si. Hace como de intermediario entre los documentos con los grandes LLMs. Componentes

* Modelos: existen varios modelos donde se puede utilizar (GPT, o modelos de huggingface)

Chain

Son cadenas de diferentes tareas para ejecutar un prompt

Memory

Tiene en cuenta el histórico, las acciones consecutivas tendrán en cuenta lo que ha ocurrido anteriormente

Indices

Parte fundamental, permite a la librería inetractura con fuentes externas. Permite conectar con documentos externos.

* Cargadores de documentos: Documentos en nube, contenido web, etc..
* Base de datos vectoriales: vectoriza la base de datos externa
* Divisores de texto: hacer fragmentos de los textos
* Recuperadores: recuperar datos que ha almacenado

Agentes

Herramienta para crear chatbots