Math Modelling Challenge
CoSIAM 2022
Curso Corto

Introducción a la ciencia de datos a través del procesamiento de lenguaje natural

Profesora: M.Sc. Viviana Márquez

¡Hola! Mi nombre es Viviana Márquez. Soy matemática de la Fundación Universitaria Konrad Lorenz y tengo una maestría en Ciencia de Datos de la Universidad de San Francisco, en California, Estados Unidos. Tengo cuatro años de experiencia como científica de datos para diversas empresas en diferentes industrias: TruSTAR Technology (ciberseguridad), Zimmerman Advertising (Publicidad y mercadeo), Royal Caribbean (Industria de viajes y turismo), HBO (streamming) y actualmente trabajo para Dataminr (Inteligencia Artificial) en Manhattan, NY. También tengo experiencia como profesora de inteligencia aritificail para programas de posgrado universitario.

Contacto:

- Página web: http://vivianamarquez.com/
- LinkedIn: https://www.linkedin.com/in/vivianamarquez/

Bienvenidos a este curso corto de CoSIAM llamado “Introducción a la ciencia de datos a través del procesamiento de lenguaje natural”. Donde quiero darles una breve visión general de la ciencia de datos y enseñarles algunas herramientas para que puedan manipular diferentes tipos textos y extraer información de ellos. ¡Comencemos!

Materiales

- Videos: https://www.youtube.com/watch?v=srQTRZzBnoo&list=PLyaMKeUhRqfyr6D4MkAIV6e2ij9V23Ahx

- Jupyter notebooks https://github.com/vivianamarquez/CoSIAM_Challenge_2022
    - resueltos: El código completo de cada modulo
    - por resolver: Los notebooks vacíos por si los estudiantes quieren seguir a la vez que van viendo el curso.

- completados video: Los notebooks exactamente como aparecen en los videos

Contenido

Clase 1

- Introducción: ¿Qué hace un científico de datos?
- Introducción: ¿Qué es el procesamiento de lenguaje natural?

Clase 2

- Crear cuenta de desarrollador en Twitter
- Configurar ambiente de desarrollo e instalar librerías necesarias para el curso

Clase 3

- Repaso de Python.

Clase 4

- Elementos de un modelo de Machine Learning: Input, output, dataset, target, etc.
- Panorama de Machine Learning: Aprendizaje supervisado vs Aprendizaje no supervisado
- Ciclo de vida de un modelo de Machine Learning.

Clase 5

- Adquisición de datos, parte I: Cómo leer datos en diferentes formatos usando Python
- Adquisición de datos, parte II: Datos abiertos

Clase 6

- Adquisición de datos, parte IV: Twitter usando la cuenta de desarrollador
- Adquisición de datos, parte III: Redes sociales usando la librería SNScrape

Clase 7

- Adquisición de datos, parte V: Web Scraping de HTML básico usando Beautiful Soup

Clase 8 (Dividida en dos clases)

Pre-procesamiento de texto

- Feature Engineering
- Representación vectorial de textos
- TF-IDF
- Word2Vec

Clase 9

- Modelos de clasificación y métricas de clasificación
- Modelos de agrupación
- Análisis de sentimiento

Clase 10

- Visualizaciones

Clase 11

- Introducción a las redes neuronales
- Transformers con Hugging Face

Clase 12

- Resumen del curso

Metadata para los videos:

Todos los videos:
CoSIAM Curso Corto 2022
Introducción a la ciencia de datos a través del procesamiento de lenguaje natural
Con: Viviana Márquez - www.vivianamarquez.com
Código disponible en: https://github.com/vivianamarquez/CoSIAM_Challenge_2022

Modulo 1: Introducción
Descripción:

- ¿Qué es la ciencia de datos? ¿Qué es el procesamiento de lenguaje natural?

Tiempo total: 24m27s

Índice:
00:00 Introducción al curso
02:02 ¿Qué hace un científico de datos?
08:34 Artificial intelligence vs Machine Learning vs Deep Learning
12:39 ¿Qué es el procesamiento de lenguaje natural?

Modulo 2: Configuración de ambiente de desarrollo
Descripción:

- Crear cuenta de desarrollador en Twitter
- Configurar ambiente de desarrollo e instalar librerías necesarias para el curso

Tiempo total: 18m50s

Índice:
00:41 Cuenta de desarrollador en Twitter
06:40 Instalar Python
09:56 Ambiente de desarrollo
12:39 ¿Qué es el procesamiento de lenguaje natural?

Modulo 3: Repaso de Python
Descripción:

- Repaso de Python

Tiempo total: 40m:34s

Modulo 4: Panorama general del Aprendizaje Automático (Machine Learning)
Descripción:

- Elementos de un modelo de Machine Learning: Input, output, dataset, target, etc.
- Panorama de Machine Learning: Aprendizaje supervisado vs Aprendizaje no supervisado
- Ciclo de vida de un modelo de Machine Learning

Tiempo total: 23m52s

Índice:
01:26 Elementos de un modelo de ML
05:49 Tipos de modelos en ML
14:50 ML Pipeline

Modulo 5: ¿Cómo adquirir datos?
Descripción:

- Adquisición de datos, parte I: Cómo leer datos en diferentes formatos usando Python
- Adquisición de datos, parte II: Datos abiertos

Tiempo total: 36m14s

Índice:
01:54 Leer archivos en Python
31:52 Datos abiertos

Modulo 6: Cuenta de desarrollador de Twitter
Descripción:

- Adquisición de datos, parte III: Twitter usando la cuenta de desarrollador
- Adquisición de datos, parte IV: Redes sociales usando la librería SNScrape

Tiempo total: 40m22s

Índice:
01:08 Aplicaciones de NLP a las redes sociales
08:26 Crear app de Twitter
16:56 Usar Twitter en Python
35:37 SNScrape

Modulo 7: Web Scraping
Descripción:

- Adquisición de datos, parte V: Web Scraping de HTML básico usando Beautiful Soup

Tiempo total: 18m40s

Índice:
03:01 Repaso de HTML
05:02 BeautifulSoup
08:33 Ejemplo con coinmarketcap.com
35:37 SNScrape

Modulo 8.1: De palabras a vectores - Pre-procesamiento
Descripción:

- Pre-procesamiento de texto

Tiempo total: 36m54s

Índice:
00:52 Limpieza de texto con Python
17:53 Pre-procesamiento de NLP

Modulo 8.2: De palabras a vectores - Vectorización

Descripción:

- Feature Engineering
- Representación vectorial de textos
- TF-IDF
- Word2Vec

Tiempo total: 1h09m41s

Índice:
01:10 Feature Engineering
05:41 Feature Engineering en NLP
12:48 One-Hot Encoding
19:19 Bolsa de palabras (Bag of words)
28:28 TF-IDF
44:10 Medidas de similitud
53:02 Word2Vec

Modulo 9: Modelos de Machine Learning
Descripción:

- Modelos de clasificación y sus métricas
- Modelos de agrupación
- Análisis de sentimiento

Tiempo total: 1h13m12s

Índice:
02:28 Modelos de clasificación
08:10 Medidas de rendimiento
20:11 Modelos de clasificación
41:13 Modelos de agrupación (K-Means)
01:02:15 Análisis de sentimiento

Modulo 10: Visualizaciones
Descripción:

- Visualizaciones

Tiempo total: 30m57s

Índice:
00:00 ¿Por qué son importantes las visualizaciones?
02:19 Herramientas de visualización
06:41 Tips para hacer buenas visualizaciones
12:28 Jupyter Notebook y GitHub
17:38 Nubes de palabras
24:18 La trifecta perfecta: plotly, dash y flask

Modulo 11: Redes neuronales
Descripción:

- Introducción a las redes neuronales
- Transformers con Hugging Face

Tiempo total: 23m39s

Índice: