Lenguajes de Data Science

1-Introduccion al curso

-Datos, son centrales para la CD, se mostrara como analizar, y organizar esto

-se debe familiarizar con la programación y los lenguajes mas usados en cd

-estas herramientas ahorran tiempo y esfuerzo

-estas herramientas ayudan a los no programadores a entrar a la cd

-como el uso de la cloud familita y hace mas rápido el manejo de datos

-se tienen laboratorios para facilitar esto y exámenes para reafirmar los conocimientos

2-Lenguajes de la Ciencia de Datos

-variedad de opcione técnicas, demasiadas opciones, dependen de que se quiera utilizar

-recomientan aprender primero Python, R y SqL como primeras opciones a considerar, aunque hay algunos otros populares en el mercado y que tienen diversas aplicaciones Scala, Java, C++ y Julia, también hay otras que tienen sus propios casos de uso como javascript, PHP, Go, Ruby, y Visual Basic

Las lenguas que aprendas dependerán de los problemas que quieras resolver, también dependerá de la compañía en la que trabajes y el rol que tendras

Hay diferentes roles para los que se pueden utilizar, Analista de Negocios, Ingeniero de Base de Datos, Analista de Datos, Ingeniero de datos, Cientifico de Datos, Cientifico de Investigacion, Ingeniero de Software, Gestor de Prodcto Estadistico, Gestor de Proyecto, etc.

-Las 3 principales lenguas de la ciencia de datos: Python, R y SQL

Cada una tiene sus propias lecciones

3-Introduccion a Python

-Caracteristicas del alto nivel del lenguaje de programación de Python, es el mas popular de ciencia de datos

-Encuesta Kaggle de ciencia de datos y aprendizaje automatico, el 75% de los de los 10,000 encuestados reportaron utilizarlo con regularidad

-Glassdor informo que en 2019, 75% de los puesots de ciencia de atos incluianpython en descripciones de datos

-Por que es tan popular?

-Python usa sintax clara, se puede hacer cosas como en otros programas de programación, pero usando menos códigos., es de los mejores para empexzar a aprender,

-80% de los profesionales lo usan

-Python es útil en muchas situaciones: ciencia de datos, inteligencia artificial, aprendizaje automatico, desarrollo web y dispositivos IoT (como Rasberry Pi)

-La utilizan IBM, Wikipedia, Google, Yahoo, CERN, Nasa Facebook, Istagra, reddit, etc.

-Es un lenguaje poderoso de propósito general que puede hacer muchas cosas

-Puede ser usada para diferentes problemas

-Tiene unalibreria estándar que proporciona herramientas para tareas diferentes, incluyendo pero no limitado a las bases de datos, automatización, web scraping, procesamiento de texto, procesamiento de imágenes, machin larning y analitic

Para Ciencia de datos

-se pueden usar las librerías de computaccion científica de Python: Pandas, NumPy, SciPy y Matploylib

Para Inteligencia artificial

TensorFlow, PyTorch, Keras, Scikit-learn

Para Procesamiento del Lenguaje Natural (NLP)

-Utilizando Natural Language Toolkiy (NLTK)

-La comunidad de Python tiene una historia bien documentada

-El lenguaje de Python tiene un código de conducta ejecutado por la Python Software Foundation, que busca la seguridad e inclusión de todos en las comunidades de Python en línea como en persona

Tambien hay comunidades como PyLadies, creando espacios para personas (mujeres) interesadas en Python que aprendan en entornos seguros e inclusivos

4-Introducción a R

-Encuesta Kaggle de la CD 2019, el aprender hasta 3 lenguajes puede aumentar tu salario

-Es gratuito, es un groyecto GNU

-No es de código abierto, es un software libre

¿Codigo abierto o software libre?

-Amplias se refieren al mismo conjunto de licencias, muchos códigos abiertos usan la Licencia Publica general GNU

-Ambas soportan colaboración

-En muchos casos los términos pueden ser usados intercambiablemente

Diferencias

-La Open Source Initiative (OSI) apoya el código abierto, mientras que la Free Software Foundation defiende el software libre

-EL código abierto se enfoca mas en empresas y el software libre se enfoca en un conjunto de valores

-Al usar R puedes usar el lenguaje de lamisma forma que contribuyes al código abierto y permite colaboración publica, al igul que uso privado y comercial

-Quien usa R

.Estadisticos, matemáticos y mineros de datos para desarrollar softwares, estadísticos, graficas y análisis de datos

-La sitaxis orientada a los arrays del lenguaje facilitan la traducción de las matemáticas al código (especialmente para personas sin experiencia de programación)

Encuesta Kaggle de la CD y Aprendizaje automatico: la mayoría de las persons aprende R, aunque ya tengan años en su carrera, y sigue siendo un lenguaje comodo

R es popular en la academia,

Entre las empresas que lo usan están IBM, Google, Facebook, Microsoft, Bank of America, Ford, TechCrunch, Uber y Tulia

\_R es el mayor repositorio de conocimieto estadístico a nivel mundial

-Desde 2018 tiene mas de 15,000 paquetes públicamente lanzados haciendo posible realizar complejos análisis exploratorios de datos

Se integra bien con otros lenguajes como C++, Java, C, .Net y Python

-Operaciones matemáticas convencionales como multiplicación de matricies, la cuales funcionan desde lavase

-R ofrece mas facilidades de programación orientadas a objetos en comparación de la mayoría de los lenguajes de computación estadística

Comunidades a nivel mundia

-UseR¡, WhyR?, SatRdays y R-Ladies

5-Introduccion a SQL

-Structured Query Lenguage

-No es un lenguaje procedimental y su alcance esta limitado a consulya y gestion de datos

-Es usada por que es simple y poderosa

-Es util en el manejo de datos estructurados (datos incorporados relaciones entre entidades y variables

-Fue diseñado para la gestion de datos en bases de datos relacionales

-Bases de datos relacionales: formadas por colecciones de tablas de dos dimencinales

-Fue originalmente diseñado para bases de datos relacionales, porque es generalizado y facil de usar, ha desarrollado interfces SQL para muchos No SQL y grandes repositorios de datos

-El lenguaje de SQL esta subdividido en varios elementos de lenguaje: clusulas, expresiones, pronosticos, consultas y declaraciones

-SQL ayidara a realizar trabajos de Ciencia de datos, incluyendo empresas y analisis de datos,

-Es necesaria en ingenieria de datos y ciencia de datos

-Con operaciones en SQL, se accede a os datos directamente (no necesidad de copiarlos)

-SQL es el interprete entre tu y la base de datos

-SQL esta bajo el Instituto nacional Estadounidense de Estandares (estandares ANSI)

-Bases de Datos SQL

-MySQL, IBM DB2, PostgreSQL, Apache, Open Office Base, SQLite, Oracle, MariaDB, Microsoft SQL Server, etc.

-La sintaxis utilizada en SQL que escribes podria cambiar basado en el sistema de gestion de bases de datos relacionales que uses

-Para prender SQL es mejor centrarse en una base de datos relacional especidica y conectarte a esa comunidad

6-Otros Lenguajes

-Scala, Java, C++, Julia, son los lenguajes tradicionales de la ciencia de datos

-JavaScript, PHP, GO, Rubi y Visual Basic Tambien han Ganado lugar en la comunidad de CD

Java

-Es un lenguaje de programacion orientado a objetos de proposito eneral probado y verdadero

-Adoptado en las empresas, diseñado para ser rapido y escalable

-Sus aplicaciones estan compiladas en codigo Bytes y se ejecutan en la maquina vistual Java (JVM)

-Herramientas para ciencia de datos: Weka (para mineria de datos), Java-ML (librería de aprendizaje automatico), Apache Mllib (hace el aprendizaje automatico escalable) y Deeplearning4j (aprendizaje profundo), Apache Hadoop (gestiona procesamiento y almacenamiento de datos para grandes aplicaciones de datos ejecutadas en sistemas cluster.

Scala

Scala es un lenguaje de programación de uso general que proporciona el soporte a la programación funcional  y a un sistema de tipo estático fuerte.

-Hecha para responder criticas de java

-Interoperable con Java, se ejecuta en JVM

-Java (Scalable Language)

-Diseñado para crecer con las demandas de usuarios

-Apache Spark: programa mas popular con Scala

-Es un sistema de computacion en cluster rapido y de proposito general

-Proporciona API’s que hacen mas facil la escritura de los trabajos paralelos y un motor optimizado que soporta los graficos de computacion general

-Incluye Shark: es un motor de consulta, Mllib para el aprendizaje autmatico, GraphX para el procesamiento grafico y Spark Streaming

-Diseñado para ser mas rapido que Hadoop

C++

-lenguaje de programacin de proposito general

-Extension del lenguage C o C con clases

-Mejora la velocidad de procesamiento, permite la programacion de sistema y un mejor control sobre la aplicación de software

-Otros lenguajes mas avanzados confian en C++ para desarrollar programas que alimentan esos datos a los clientes en tiempo real

-TensorFlow: librería popular de deep learning para el flujo de datos, creada on C++, pero ejecutada en interfaz de Python

-MongoDB (una base de datos NoSQL para la gestion de grandes vlumenes de datos fue contruida con C++)

-Cadde es un repositorio de algoritmos de aprendizaje profundo construido con C++, Python y MatLab blindings

Javascript

-Es un lenguaje de proposito general que se extendio mas alla del navegador con la creacion de Node.js y otros alcances del server side Javascript No esta relainado a Java

-La implementacion mas popular es TensorFlow.js

-TensorFlor.js hace posible el deep learning y el aprendizaje profundo bien sea en Node.js o en el navegador

Fue adoptado por otras librerias de codigo abierto (Brain.js y machinelearn.js)

Proyecto R-js

Ha rescrito las especificaciones de algebra lineal del lenguaje R en Typescript

-Permitira una base para que otros proyectos implementen marcos de trabajo de base de datos mas poderosos como Numpy y SciPy de Python

Typescript

Es un super conjunto de JavaScript

Julia

-creada en MIT para el analisis numerico de alto desempeño y ciencia cumputacional

-Produce un desarrollo rapido como Python o R mientras produce programas que se ejecutan tan rapido como C o Fortran

-Esta compilado (el codigo es ejecutado directamente en el procesador como codigo ejecutable y llama a las liberias de C, Go, Java, MATLAB, R, Fortran y Python e igualmente ha refinado el paralelismo

-Fue escrito en 2012, pero tiene futuro en CD

-JuliaDB es una aplicación util par CD

-Es un paquete para trabajar con conjuntos de datos grandes y persistentes