Certificado Experto en Minería de Texto aplicada

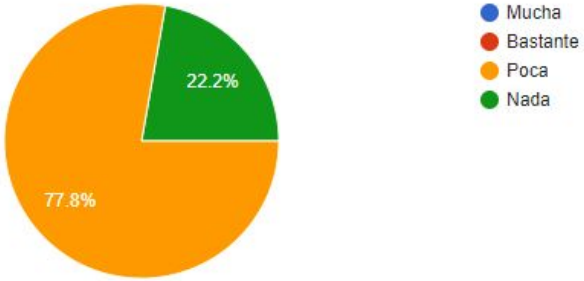
Unidad 1

Introducción a la minería de textos en Python

Perfil de la cursada actual

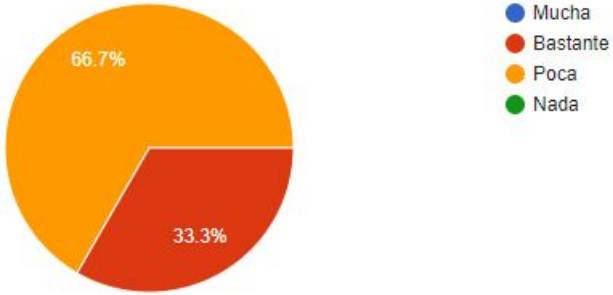
Experiencia en Ciencia de Datos

9 responses



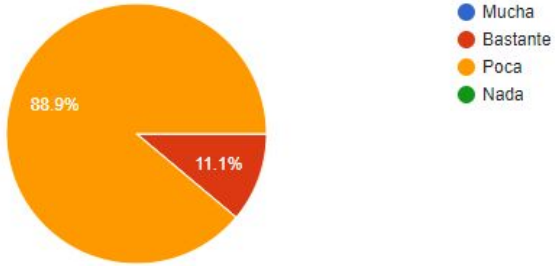
Experiencia en Programación

9 responses



Experiencia en Python

9 responses



Cronograma de clases

Unidad	Tema	Fecha
Introducción	Introducción a Python Conceptos básicos de Minería de Texto	11/8/2021
Creación del corpus	Manejo de archivos locales	11/10/2021
	Web Scraping / APIs	11/15/2021
Expresiones Regulares		11/17/2021
Preprocesamiento		11/22/2021
Clasificación Supervisada		11/24/2021
Identificación de tópicos		11/29/2021
Repaso y Trabajo final		*12/1/2021*

Pasos para la implementación de un modelo de **análisis de texto por computadora**

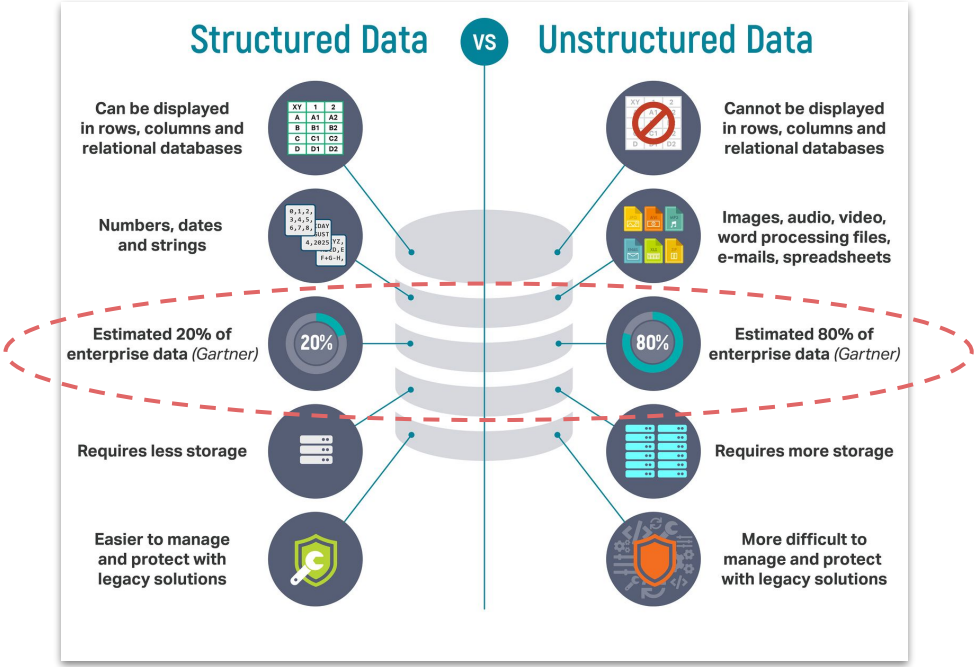
Obtener la información	Preparación de los datos	Etiquetado	Modelado	Análisis
Consolidar un corpus a partir de datos que pueden provenir de diferentes fuentes.	Eliminación de ruido, segmentación, normalización y reducción de léxico del texto.	Señalar la categoría gramatical y los rasgos morfológicos de cada palabra.	Aplicación de algoritmos para obtener inferencias y extraer información de los textos.	Análisis, interpretación y navegación de los resultados

Tres tipos de presentación de los datos:

- **Datos estructurados:** Los datos se encuentran formularios estandarizados o tablas. Son archivos de tipo texto que se suelen mostrar en filas y columnas con títulos. Son datos que pueden ser ordenados y procesados fácilmente por todas las herramientas de minería de datos.

- **Datos no estructurados:** Los datos se encuentran dispersos, no presentan una lógica de ordenamiento en común (Ej., Imágenes, Videos, Sonidos, Comentarios de los usuarios, Textos)

- **Datos semi estructurados:** no presentan una estructura tan clara como la de las tablas, pero tienen cierta organización semántica. Los ejemplos de este tipo de datos son los archivos JSON y los XML.



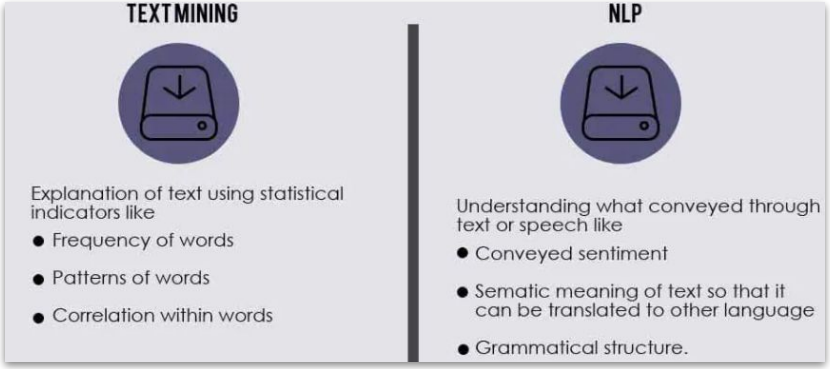
¿Minería de Texto o PLN?

Minería de Texto (Text Mining):

- Extraer información
- Encontrar patrones
- Se trabaja con reglas

Procesamiento de Lenguaje Natural:

- Entender el significado de un mensaje en lenguaje natural a través de la computadora
- Implica comprender la estructura gramatical del texto
- Importa el contexto



Lenguaje Natural vs. Lenguaje Formal

- **Lenguaje natural:** Cualquier lenguaje hablado que evolucionó de forma natural. Español, inglés, etc...

- **Lenguaje formal:** Cualquier lenguaje diseñado por humanos que tiene un propósito específico, como la representación de ideas matemáticas o programas de computadoras; todos los lenguajes de programación son lenguajes formales.

Aprendizaje automático

El aprendizaje automático (Machine Learning) consiste en una serie de técnicas que permiten a los sistemas informáticos predecir, clasificar, ordenar, tomar decisiones y, en general, extraer conocimientos de los datos sin necesidad de definir explícitamente las reglas para realizar esas tareas.

Es un subcampo de la inteligencia artificial que tiene como objetivo lograr que las computadoras realicen tareas basadas en ejemplos sin programación explícita.

Debido al continuo crecimiento de la complejidad de los negocios y de la cantidad de datos a analizar, se pasa de una inteligencia basada en reglas a una basada en datos.

Unidad	Disciplina	Lógica
Introducción	Minería de Textos	Basada en Reglas
Creación del corpus	*Minería de Textos*	*Basada en Reglas*
Expresiones Regulares	Minería de Textos	Basada en Reglas
Preprocesamiento	Minería de Textos	Basada en Reglas
Clasificación Supervisada	Minería de Textos	Basada en Datos
Identificación de tópicos	NLP	Basada en Datos

Introducción a la minería de textos en Python

¿Por qué Python?

Python: Lenguaje de alto nivel.

> ***Lenguaje de alto nivel:*** *Lenguaje diseñado para ser fácil de leer y escribir para los usuarios. La computadora debe traducirlo a un lenguaje de bajo nivel para entenderlo.*
>
> ***Lenguaje de nivel medio:*** *Utilizan estructuras típicas de los lenguajes de alto nivel pero, a su vez, permiten un control a muy bajo nivel. Ej: C.*
>
> ***Lenguaje de bajo nivel:*** *Lenguaje diseñado para ser fácil de ejecutar para una computadora. Ej: Código binario.*

Popular: Actualización constante de las librerías, y comunidad activa de usuarios

Portabilidad: La cualidad de un programa que le permite ser ejecutado en más de un tipo de computadora.

Pandas

Una de las mejores opciones para trabajar con datos tabulares en Python es usar la Python Data Analysis Library (alias Pandas).

Permite:

- Leer y escribir datos en diferentes formatos: CSV, Microsoft Excel, bases SQL, etc.
- Seleccionar y filtrar de manera sencilla tablas de datos en función de posición, valor o etiquetas.
- Fusionar y unir datos (Igual al join en SQL)
- Transformar datos aplicando funciones tanto en global como por ventanas
- Manipulación de series temporales.

```
[ ]  import pandas as pd
```

DataFrame

	A	B	C	D
1		Student	First Name	Score
2	0	60	Olivia	90
3	1	100	John	75
4	2	80	Laura	82
5	3	78	Ben	64
6	4	95	Kevin	45

Un DataFrame es la estructura de datos más usada en Pandas. Permite guardar datos de distintos tipos (como caractéres, enteros, valores de punto flotante, factores y más) en columnas.

Es similar a una hoja de cálculo o una tabla de SQL o el data.frame de R.

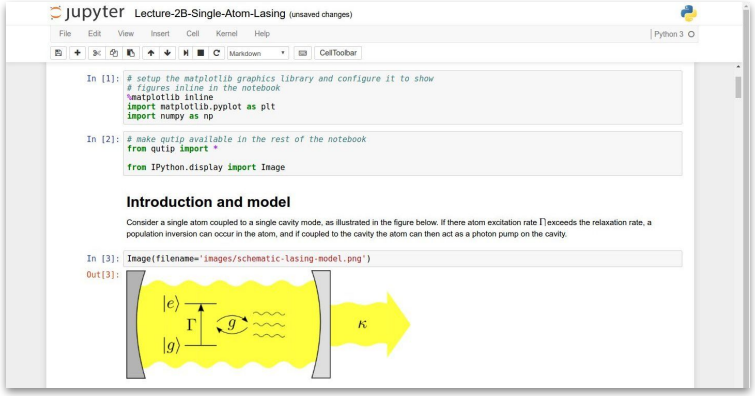
Un DataFrame siempre tiene un índice con inicio en 0. (El índice refiere a la posición de un elemento en la estructura de datos).

Jupyter Notebooks

Es entorno informático que promueve un diálogo entre el código y los textos explicativos.

Crea documentos que siguen una lista ordenada de celdas de entrada y de salida. Estas celdas pueden mostrar código, texto), fórmulas matemáticas, ecuaciones, o contenido multimedia (Rich Media).

Los documentos creados en Jupyter pueden exportarse a HTML, PDF, Markdown o Python.



ANACONDA

INSTITUTO DATA SCIENCE ARGENTINA

Perfil de la cursada actual