RECLA
Red de Educación Continua de Latinoamérica y Europa

PUCP

UNIVERSIDAD RICARDO PALMA

XXIX ENCUENTRO INTERNACIONAL

EL LIDERAZGO ÁGIL Y TRANSFORMADOR DE LA EDUCACIÓN CONTINUA EN LOS NUEVOS ENTORNOS DE CAMBIO

Herramientas de IA para el análisis de mercado en educación continua.

ALEXIS ALVEAR LEYTON
Ingeniero Civil Industrial
Master in Business Administration
Data Scientist
Subdirector Vinculación con el Medio – Facultad de Matemáticas UC
Director Ejecutivo DATA UC

SEBASTIÁN MASSA SLIMMING
Cientista Político
Magíster en Ciencias Sociales
Data Scientist
Encargado Educación Continua – Facultad de Matemáticas UC
Analista Senior DATA UC

DATAUC
Estudios & Servicios Estadísticos

Unidad de la Facultad de Matemáticas que se dedica a la transferencia tecnológica de la investigación matemática y estadística, mediante el desarrollo de aplicaciones de ciencia de datos, algoritmos e inteligencia artificial en problemáticas de negocios, industrias y políticas públicas, mediante asesoría, consultoría, proyectos de innovación y formación de capital humano avanzado.

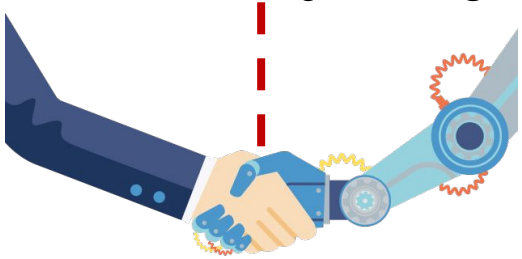
BIG DATA + DATA SCIENCE

DATOS | INFORMACIÓN | CONOCIMIENTO

¿Por qué generamos tantos datos?

¿Dónde guardamos tantos datos?

Ciencia



Tecnología

Es un fenómeno científico – tecnológico, gracias a la integración de estas dos disciplinas, que nos permite transformar la complejidad en simplicidad con ayuda de las tecnologías de la información.

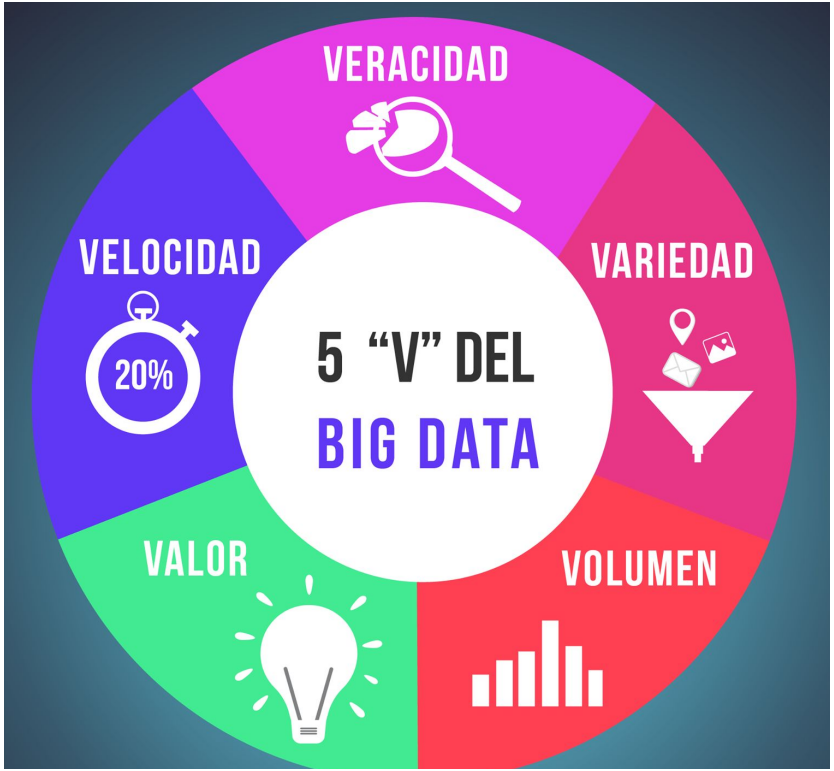
GRANDES VOLÚMENES INFORMACIÓN

SUPERAN NUESTRAS CAPACIDADES

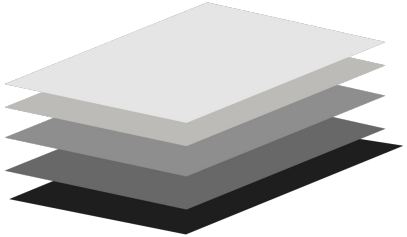
SISTEMAS COMPUTACIONALES

LÓGICAS HUMANAS

Las 5 V del Big Data

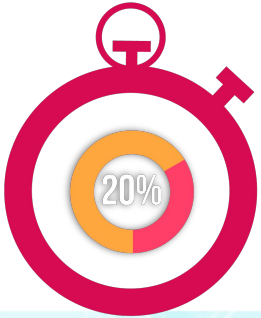


Las 5 V del Big Data



1. Volumen

La primera característica del big data es que implica grandes volúmenes de información. Actualmente, gracias a los dispositivos tecnológicos, es posible capturar, procesar y analizar miles de datos minuto a minuto, por lo que el primer desafío es desarrollar capacidades técnicas para el procesamiento y análisis de datos masivos.



2. Velocidad

El segundo desafío del big data tiene relación con la velocidad de procesamiento. Al trabajar con datos masivos, se hace necesario contar con capacidades robustas que hagan frente a la volatilidad de los datos, ya que muchos de ellos tienen una corta vida útil y es necesario capturarlos y analizarlos en el momento oportuno para que no pierdan valor.

Las 5 V del Big Data

3. Variedad

Los datos que se recopilan pueden provenir de diferentes fuentes y además podemos encontrarlos en diferentes formatos: datos estructurados y no estructurados. Dado el origen diverso de los datos, la configuración de procesos de análisis considerando la variedad de éstos, es la segunda característica del big data.

Estructurados

Datos que tienen un modelo definido o provienen de un campo determinado en un registro.

- Fichas de clientes
- Transacciones comerciales

Semiestructurados

Datos que no tienen formatos fijos, pero contienen atributos o etiquetas.

- Correos electrónicos
- Fichas con imágenes médicas

No estructurados

Datos que no tienen un modelo predefinido o no están organizados de alguna manera.

- Videos
- Fotografías
- Audios

Las 5 V del Big Data

10101100101
10010101011
00110101001
11010110010
11010101111

4. Veracidad

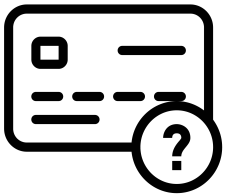
El siguiente desafío es resguardar la calidad de los datos. Al trabajar con grandes volúmenes de información, se pueden presentar problemas como registros incompletos o erróneos, datos faltantes en determinados campos o información que, proveniente de diferentes fuentes, son discrepantes. La veracidad de los datos es el cuarto desafío en big data.

Datos erróneos

Campos o atributos mal consignados por problemas de lectura o tipeo. Ejemplo, RUT´s de 3 dígitos, direcciones inexistentes, etc.

Datos faltantes

Información incompleta de dimensiones considerables que puede impactar los resultados esperados.



Fuentes discrepantes

Información proveniente de más de una fuente de información que presenta antecedentes diversos.

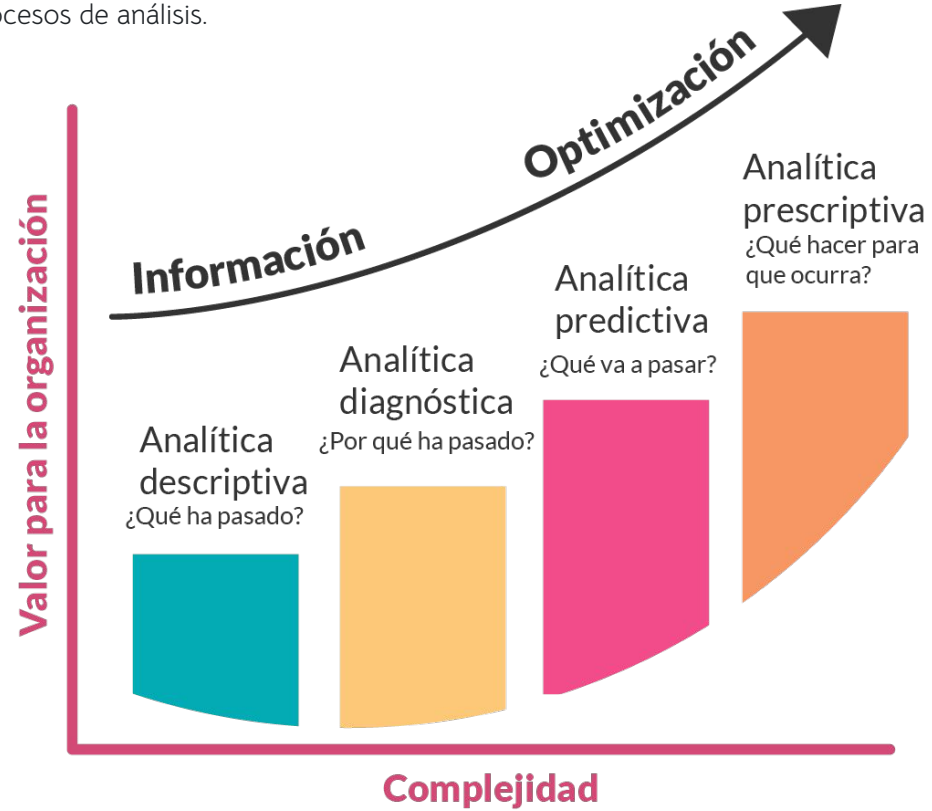
Las 5 V del Big Data

5. VALOR

¿Dónde está el valor del big data?

Capacidad analítica

La complejidad analítica presenta una relación directa entre valor y la dificultad de los procesos de análisis.



BIG DATA + DATA SCIENCE

DATOS

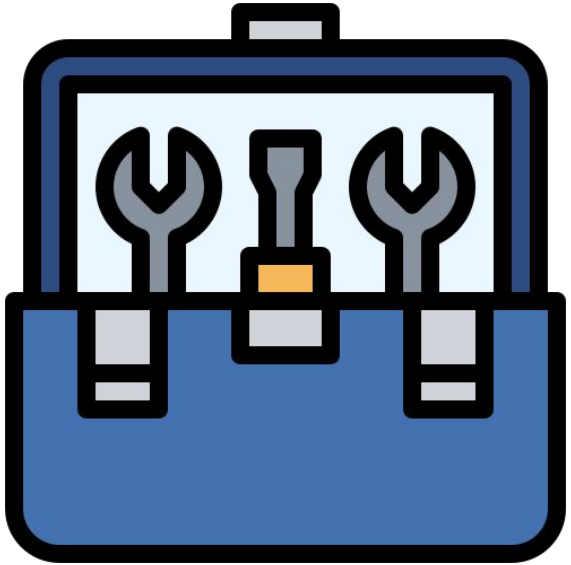
INFORMACIÓN

CONOCIMIENTO

PARTE 2 – APLICACIONES DE DATA SCIENCE EN LA GESTIÓN DE EDUCACIÓN CONTINUA

En casa de herrero, cuchillo de datos…

¿Qué herramientas de la ciencia de datos nos pueden ayudar en educación continua?



- Web scraping para monitorear comportamientos en la demanda.
- Modelos predictivos para pronóstico de matrícula.
- Clusterización para conformación de secciones de cursos.

¿QUÉ DECISIONES TOMAMOS EN FUNCIÓN DE ESTA INFORMACIÓN?

1. Diseñar los programas de cursos, estrategias de aprendizaje y qué tipo de software vamos a utilizar.

2. ¿Qué alianzas relevantes deberíamos tener para impulsar el programa y evitar que caiga en la obsolescencia?

3. ¿Qué capacidades debe tener mi cuerpo académico para desarrollar este programa?

4. ¿Qué tipo de estrategia publicitaria utilizar para comunicar los atributos del diplomado?
    1. Tenemos que explicar muy didácticamente qué es la ciencia de datos a las empresas.
    2. Tenemos que explicar muy didácticamente qué es la ciencia de datos a las personas.
    3. Tenemos que prepararnos para cumplir con las expectativas de nuestros futuros alumnos.

DISEÑO

1. Análisis de ofertas laborales.

- *Los avisos y ofertas laborales contienen datos relativos a necesidades formativas y dominios de lenguajes de programación, software o conocimientos básicos o avanzados que deben tener ciertos perfiles que se desempeñan en el área de la --96*

2. Análisis de sectores estratégicos.

- ***Contribución Significativa al PIB.*** *Los sectores estratégicos tienen un impacto económico sustancial, contribuyendo de manera significativa al Producto Interno Bruto (PIB) del país.*
- ***Generación de Empleo y Efecto Multiplicador.*** *Estos sectores no solo generan empleo a gran escala, sino que también ejercen un efecto multiplicador, impulsando otras industrias relacionadas.*
- ***Influencia en Desarrollo Tecnológico e Innovación.*** *Los sectores estratégicos están vinculados al desarrollo tecnológico y la innovación, impulsando el progreso en estas áreas fundamentales.*
- ***Conexión con Objetivos de Desarrollo Sostenible.*** *La relevancia de estos sectores se manifiesta en su conexión directa con los objetivos de desarrollo sostenible, contribuyendo al avance social y medioambiental.*
- ***Vínculo con Seguridad Energética y Recursos Naturales.*** *Algunos sectores estratégicos están estrechamente vinculados a la seguridad energética, la gestión de recursos naturales o la infraestructura crítica, aspectos cruciales para la estabilidad nacional.*

¿QUÉ BUSCAN LAS EMPRESAS EN UN DATA SCIENTIST?

Buscamos en portales de empleo las características formativas y experiencias requeridas:

¿QUÉ BUSCAN LAS EMPRESAS EN UN DATA SCIENTIST?

Buscamos en portales de empleo las características formativas y experiencias requeridas:

Deloitte.

Data Team Leader of Financial Crime Analytics

Postular

Publicada hoy por **Financial Advisory**

- Completed a bachelor's degree or master's degree preferably in a quantitative discipline such as: Mathematics, Statistics, Engineering or Computer Science
- At least 5 years of leading teams in professional work experience in consulting or analytics, preferably related to financial crime or data analytics.
- At least 3 years of experience querying and analyzing big data using Analytics technologies
- At least 3 years of experience dealing directly with clients and delivering growth proposals or new projects.
- Have proven experience in data preparation and modeling – Spark and SQL; Analytics – Scala or SQL; Cloud – AWS (Amazon Web Services) or GCP (Google Cloud Platform) or Azure; Development Tools – Docker or Kubernetes, Git; Other – Databricks and Azure Data Factory. • Have proven experience in data preparation and modeling – Spark and SQL; Analytics – Scala or SQL; Cloud – AWS (Amazon Web Services) or GCP (Google Cloud Platform) or Azure; Development Tools – Docker or Kubernetes, Git; Other – Databricks and Azure Data Factory.
- Proven experience of building data processing pipelines for use in production batch systems, including either traditional ETL (Extract Transform Load) pipelines and/or analytics pipelines
- Proven experience in manipulating data through cleansing, parsing, standardizing to improve data integrity
- Familiarity with Windows and Linux operating systems
- Knowledge of software engineering best practices across the development lifecycle, including coding

METODOLOGÍA

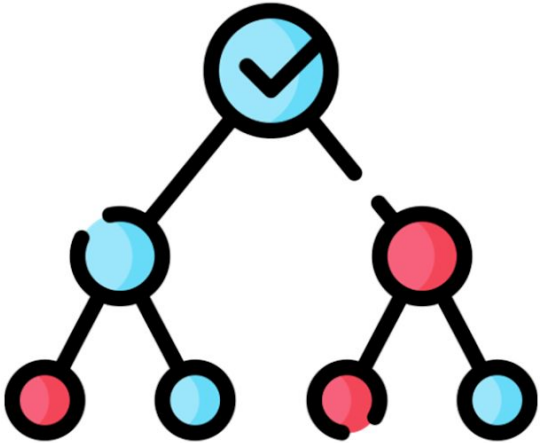
Metodología

Parte N°1: WEB SCRAPING

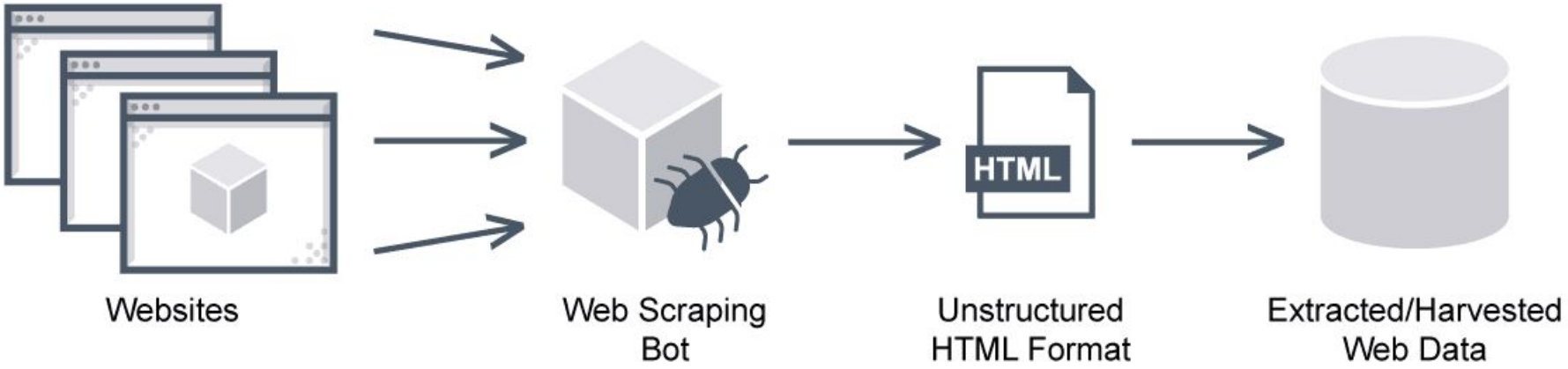
1. Identificación de etiquetas HTML para obtención de información.
2. Utilización de XML, extrayendo los nodos Xpath y construcción de matriz de datos.
3. Creación de sintaxis de código para la ejecución de algoritmo.
4. Aplicación de técnicas de data wrangling y limpieza de patrones a través del dominio de expresiones regulares.
5. Automatización de algoritmo para la extracción de datos estructurados y no estructurados.

Parte N°2: NATURAL LANGUAGE PROCESSING

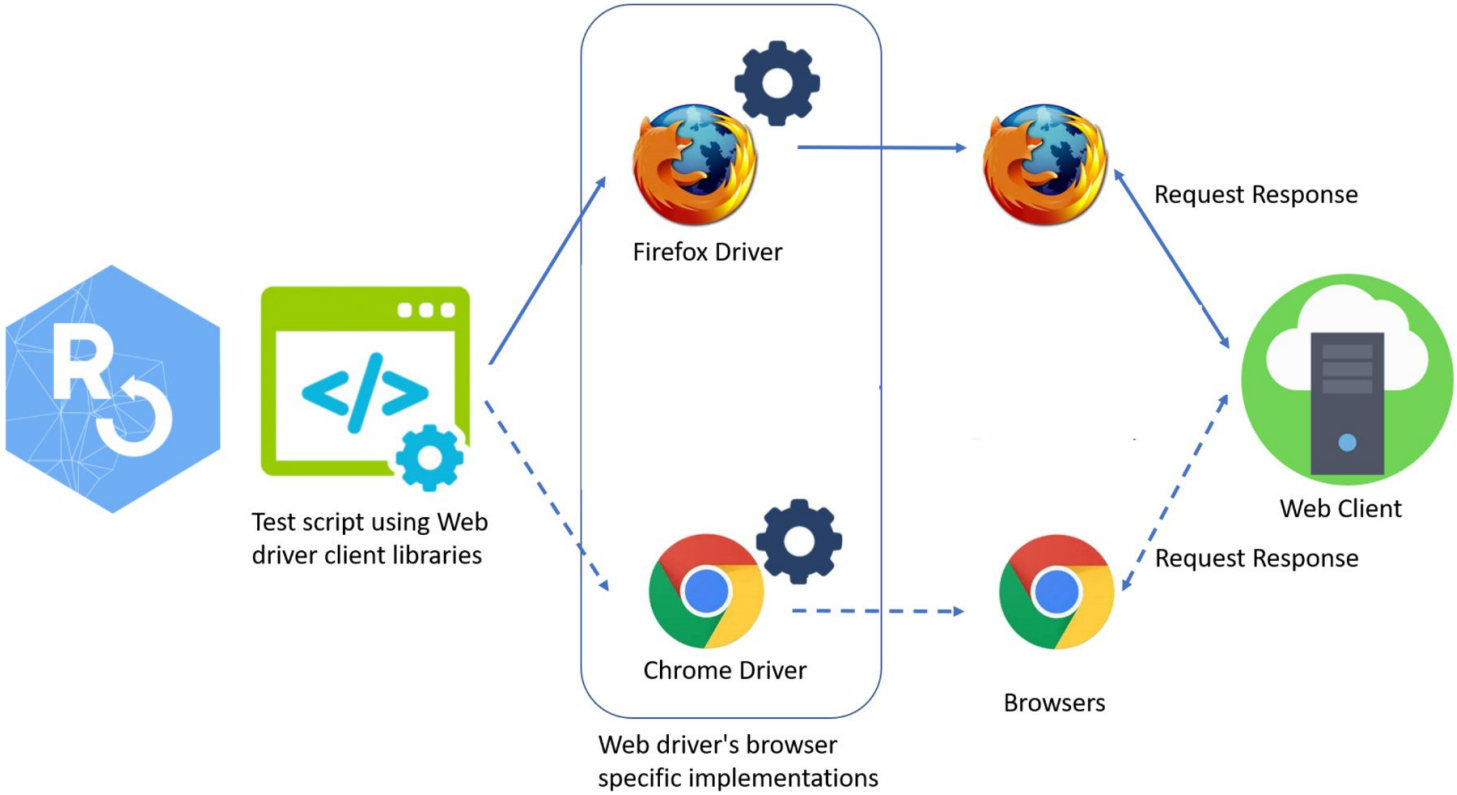
1. Construcción de matriz de frecuencia DFM mediante la transformación de noticias (textos) en vectores que puedan ser representados matemáticamente.
2. Exploración de datasets para hallar términos frecuentes y menciones relativas.
3. Remoción de stopwords (palabras de enlace que no aportan significado léxico).
4. Retrieval Information mediante análisis TF-IDF.
5. Evaluar la factibilidad de clasificador de textos (ya sea por sentiment analysis, topic-modelling o K-means).



METODOLOGÍA



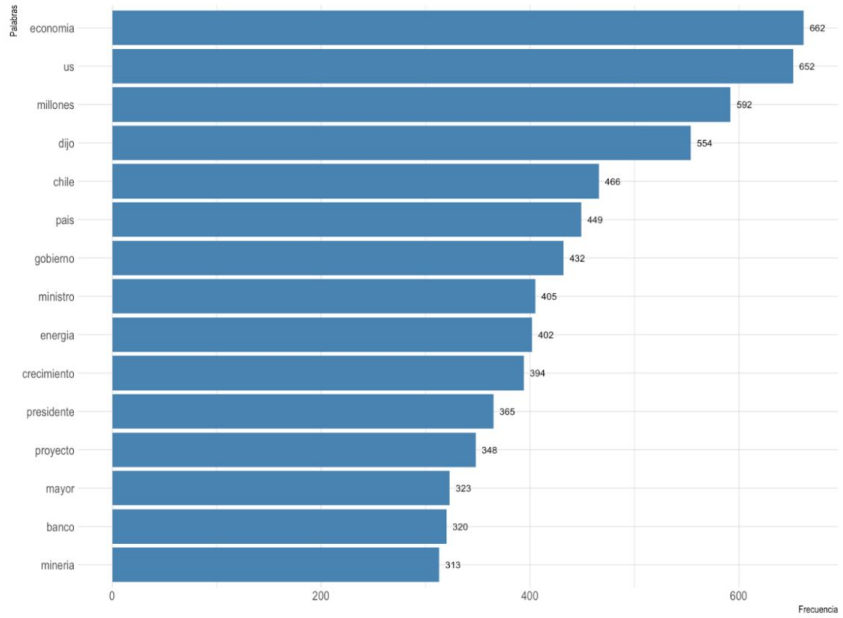
METODOLOGÍA



METODOLOGÍA

METODOLOGÍA

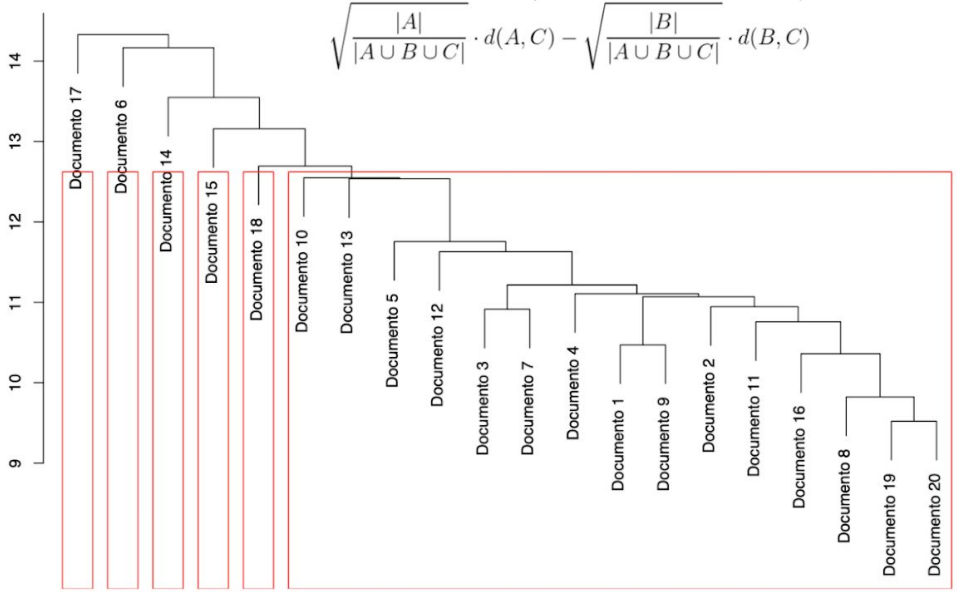
Terms frequency



$$TF(t) = C(t)/N$$

Cluster dendogram

$$d(A \cup B, C) = \sqrt{\frac{|A \cup B|}{|A \cup B \cup C|}} \cdot d(A, C) + \sqrt{\frac{|B \cup A|}{|A \cup B \cup C|}} \cdot d(B, C) - \sqrt{\frac{|A|}{|A \cup B \cup C|}} \cdot d(A, C) - \sqrt{\frac{|B|}{|A \cup B \cup C|}} \cdot d(B, C)$$



NPL model for corpus similarity

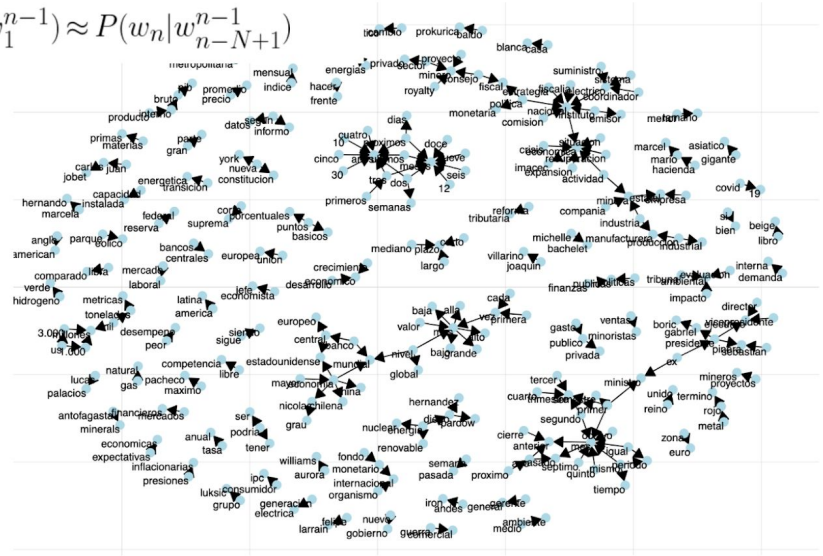
document1	document2	cosine
text1	text2	0.04403080
text1	text3	0.03109661
text1	text4	0.04314325
text2	text4	0.04331827
text1	text6	0.02450680
text3	text6	0.02428393
text5	text6	0.05148814
text1	text7	0.03186468
text3	text7	0.03157490
text5	text7	0.02628647
text6	text7	0.04669942
text1	text9	0.07486055
text7	text10	0.02732874
text5	text11	0.03444307
text13	text14	0.01619272
text1	text15	0.02641548
text3	text15	0.02617526
text6	text15	0.02062835
text7	text15	0.02682177
text1	text16	0.04379785

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

METODOLOGÍA

N-gram analysis

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})$$



METODOLOGÍA

TF-IDF analysis

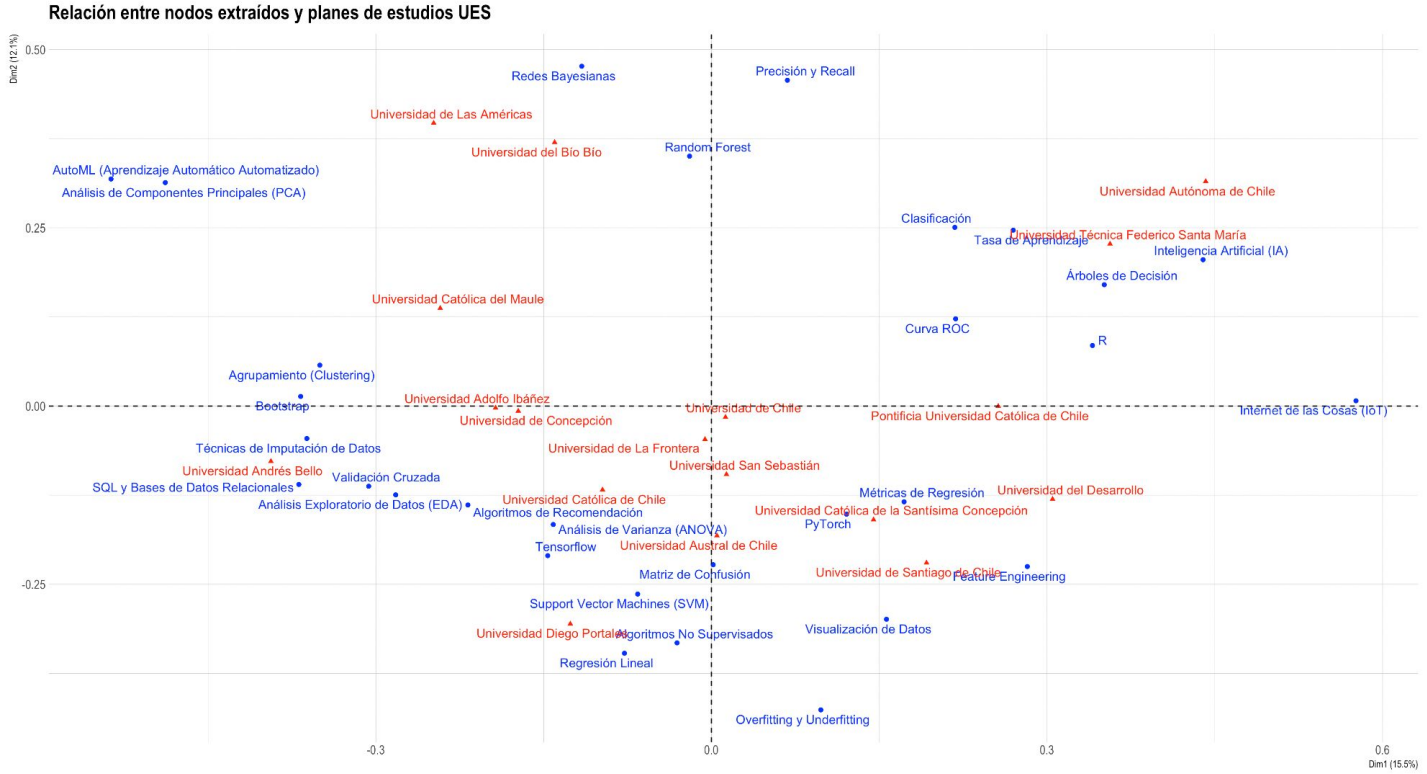
$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i})$$

document	sqm	desmiente	acuerdo	preliminar	codelco	explotacion	litio	salar	atacama	exxon	empezara	producir
Documento 1	2.104816	3.83721	1.75085	3.83721	1.393165	2.93412	1.646878	2.51499	2.305731	0.000000	0.00000	0.00000
Documento 2	0.000000	0.00000	0.00000	0.00000	0.000000	0.00000	1.646878	0.00000	0.000000	3.360088	3.83721	2.83721
Documento 3	0.000000	0.00000	0.00000	0.00000	1.393165	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.00000
Documento 4	0.000000	0.00000	0.00000	0.00000	0.000000	0.00000	1.646878	0.00000	0.000000	0.000000	0.00000	0.00000
Documento 5	0.000000	0.00000	0.00000	0.00000	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.00000
Documento 6	0.000000	0.00000	0.00000	0.00000	1.393165	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.00000
Documento 7	0.000000	0.00000	0.00000	0.00000	1.393165	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.00000
Documento 8	0.000000	0.00000	0.00000	0.00000	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.00000
Documento 9	2.104816	0.00000	0.00000	0.00000	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.00000
Documento 10	0.000000	0.00000	0.00000	0.00000	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000	0.00000	0.00000

METODOLOGÍA

Contenidos	Pontificia Universidad Católica de Chile	Universidad Austral de Chile	Universidad de Chile	Universidad Católica de Chile	Universidad de Santiago de Chile	Universidad Adolfo Ibáñez	Universidad Técnica Federico Santa María	Universidad Diego Portales	Universidad de La Frontera	Universidad Andrés Bello	Universidad Católica de la Santísima Concepción	Universidad de Concepción	Universidad de Las Américas
Big Data	24	21	6	0	26	20	27	29	4	13	2	14	24
Aprendizaje Automático (Machine Learning)	16	14	18	11	17	24	18	1	25	10	17	9	9
Inteligencia Artificial (IA)	13	18	26	9	5	15	21	13	10	16	3	28	3
Visualización de Datos	12	0	18	25	29	7	2	26	10	24	10	17	14
Análisis Predictivo	4	13	3	8	15	7	15	12	24	27	24	20	30
Minería de Datos	8	21	4	24	5	7	13	16	28	16	20	23	29
Internet de las Cosas (IoT)	30	19	26	5	5	6	24	26	9	14	28	12	23
Regresión Lineal	5	10	2	18	8	24	4	29	18	14	7	17	27
Clasificación	14	27	16	21	25	11	29	12	26	3	27	7	4
Agrupamiento (Clustering)	8	26	1	28	11	28	26	0	17	19	14	17	6
Redes Neuronales	29	24	28	26	13	16	13	19	8	29	18	28	27
Algoritmos de Recomendación	26	4	24	30	0	12	0	24	19	29	25	21	29
Feature Engineering	12	18	30	1	12	10	14	10	30	11	22	27	22
Validación Cruzada	12	1	28	11	6	29	9	10	30	16	16	26	0
Overfitting y Underfitting	7	16	22	24	28	16	9	20	30	18	28	13	8
Curva ROC	19	26	4	22	6	1	5	9	14	30	14	24	23
Precisión y Recall	17	10	30	29	13	25	16	27	4	29	14	29	23
R	15	26	7	5	24	17	0	5	15	22	18	16	20
Tensorflow	19	29	1	25	21	18	22	16	29	12	25	27	17

Correspondence Analysis

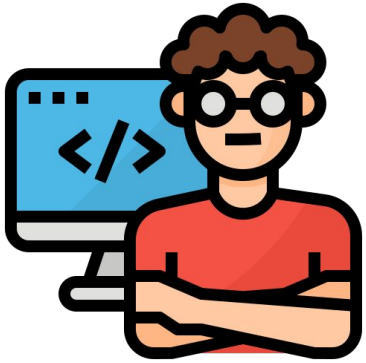


WORKSHOPS Y WEB SERIE “TODO DATO TIENE SU CIENCIA”

WORKSHOPS Y WEB SERIE “TODO DATO TIENE SU CIENCIA”

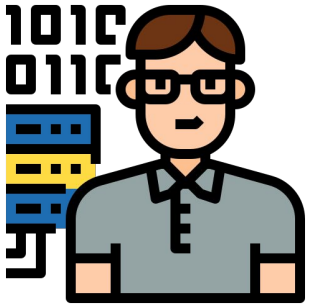
PERFILES DE UN DATA SCIENTIST.
META: CREAR MODELOS QUE NOS PERMITAN PERFILAR A POTENCIALES ESTUDIANTES DE ACUERDO A SUS INTERESES PARA MALLAS FLEXIBLES.

Business analitycs

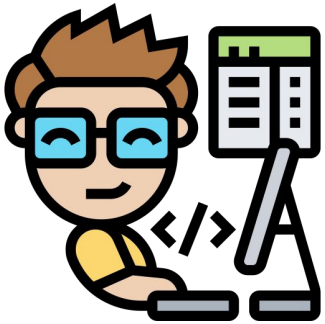


Ciencias sociales

Otras ciencias



Informático Programador



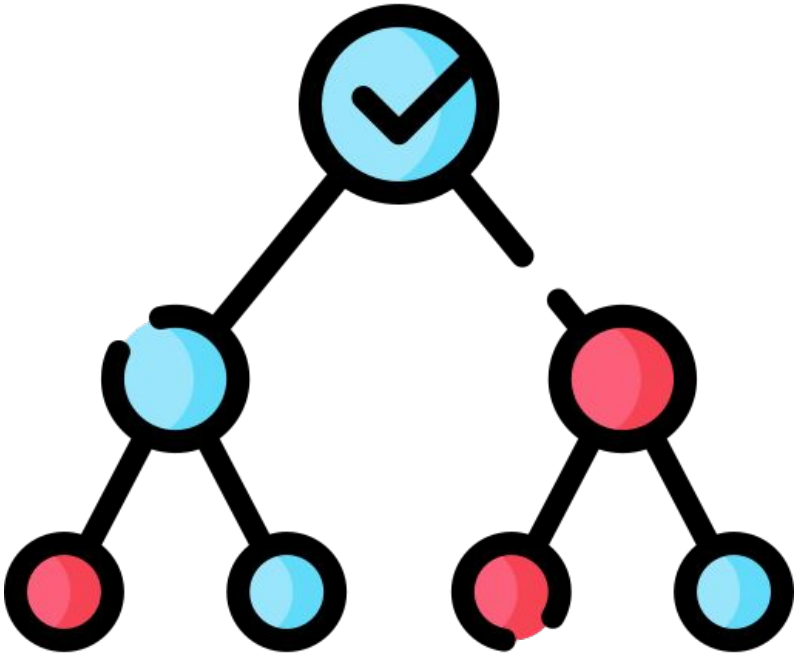
Estoy por moda

PERFILES DE UN DATA SCIENTIST.
META: CREAR MODELOS QUE NOS PERMITAN PERFILAR A POTENCIALES ESTUDIANTES DE ACUERDO A SUS INTERESES PARA MALLAS FLEXIBLES.

¿CÓMO CREARLO Y VALIDARLO?

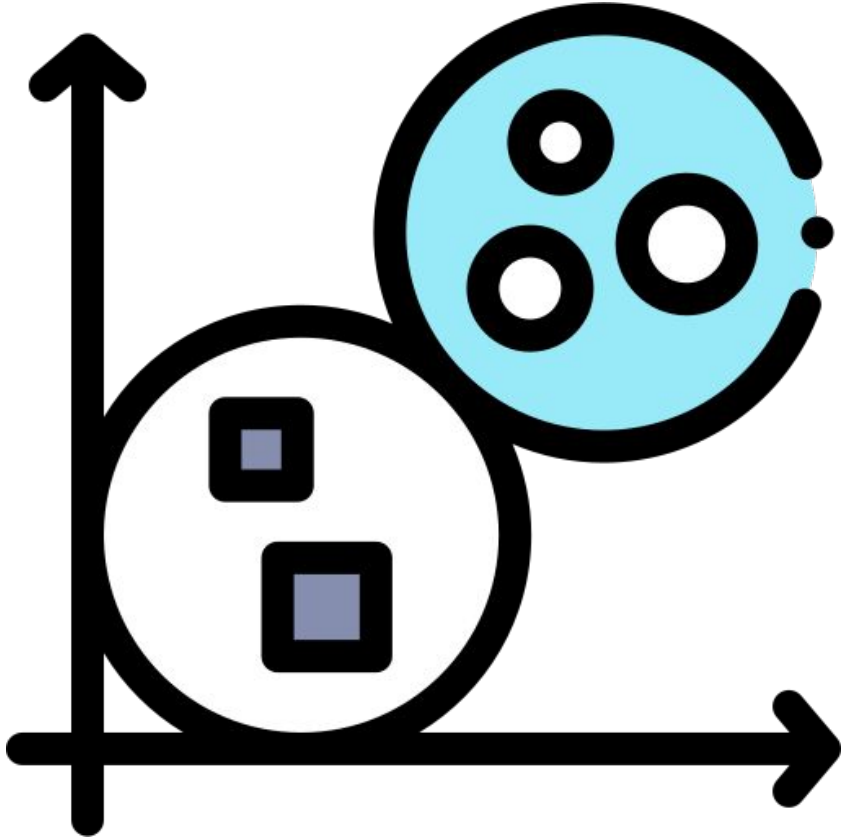
MODELOS PREDICTIVOS

Árboles de decisión para pronosticar probabilidades de matrícula efectiva.



CLUSTERIZACIÓN

Clusterización para la configuración de las secciones de cursos.



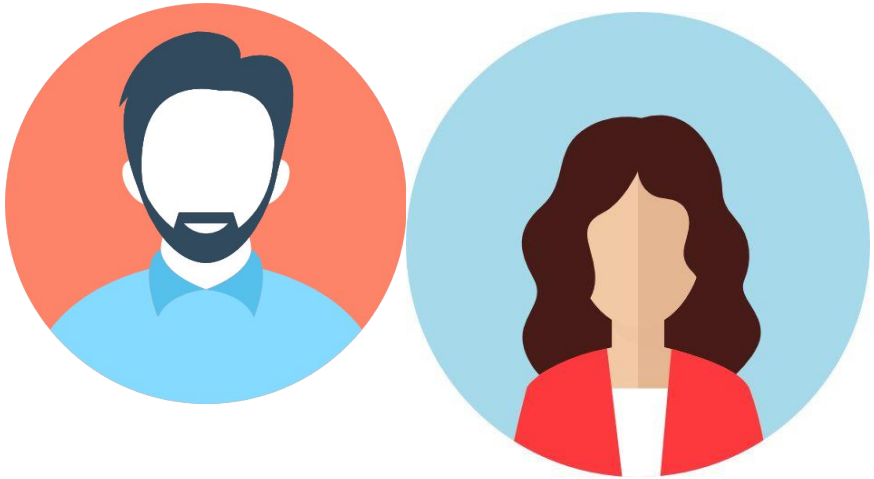
VARIABLES A CONSIDERAR PARA LA ASIGNACIÓN DE CURSO

Edad

Universidad Pregrado

Años de estudio

Años experiencia laboral



Carrera

Años que dejó de estudiar

Estudios y especializaciones

Estado o situación civil

PARTE 3 – ¡TALLER DE INNOVACIÓN!

BIG DATA + DATA SCIENCE

DATOS | INFORMACIÓN | CONOCIMIENTO

¿Qué innovaciones podríamos crear a partir del uso de datos?

¡Manos a la obra!
Taller de web scraping