Datatón 2022

Bancolombia Bancoagrícola Banistmo Bam

Recomendador de Noticias

Centro de Excelencia Analítica -IA- GI

Vicepresidencia de negocios corporativos

Ponentes

Juan David Mejia – Líder Corporativo

Juan Camilo Diaz - Líder Analítica

Walter Arboleda Castañeda - Líder DevOps Analítica

1. Bienvenida
2. Generalidades de la Dataton
    1. Descripción
    2. Historia
    3. Fechas
    4. Premiación
3. Reto 2022
    1. Escenario
    2. Reto 2022
    3. Preguntas
4. Entregables
    1. Categorización
    2. Recomendación
    3. Repositorio
    4. Preguntas
5. Evaluación de Resultados
    1. Categorización
    2. Recomendación
    3. Repositorio
    4. Preguntas
6. Prototipo e Ideas
7. Cierre

Generalidades de la Datatón

Descripción

La Datatón es una **competencia** que realiza el Centro de Excelencia en Analítica, Inteligencia Artificial y Gobierno de Información del Grupo Bancolombia, para analíticos, científicos de datos y demás profesionales afines a la analítica predictiva, el aprendizaje automático y la inteligencia artificial, que sean profesionales, aficionados o estudiantes universitarios **externos** a la entidad.

El propósito es resolver un reto analítico complejo y de actualidad, que les permita a los participantes demostrar sus capacidades, disfrutar solucionarlo, y que nos permita desde el banco conocerlos a ustedes para **RECLUTAR** a los mejores.

Historia

1. Primera Versión (2017)
   Cómo mejorar atención en sucursales
2. Segunda Versión (2018)
   Predicción de categorías de gastos por PSE
3. Tercera Versión (2019)
   Uso de traza digital para crédito en clientes independientes
4. Cuarta Versión (2020)
   Estimación de Gastos Personales
5. Quinta Versión (2021)
   Detección señales: Datos alternativos
6. **NLP: Sistema recomendación noticias clientes corporativos**

¡Sexta versión!
151 equipos seleccionados

Cobertura

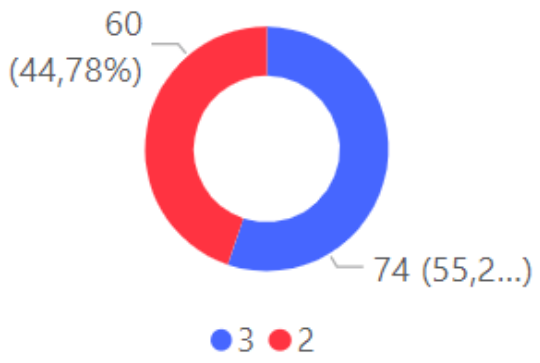
Cantidad equipos

151

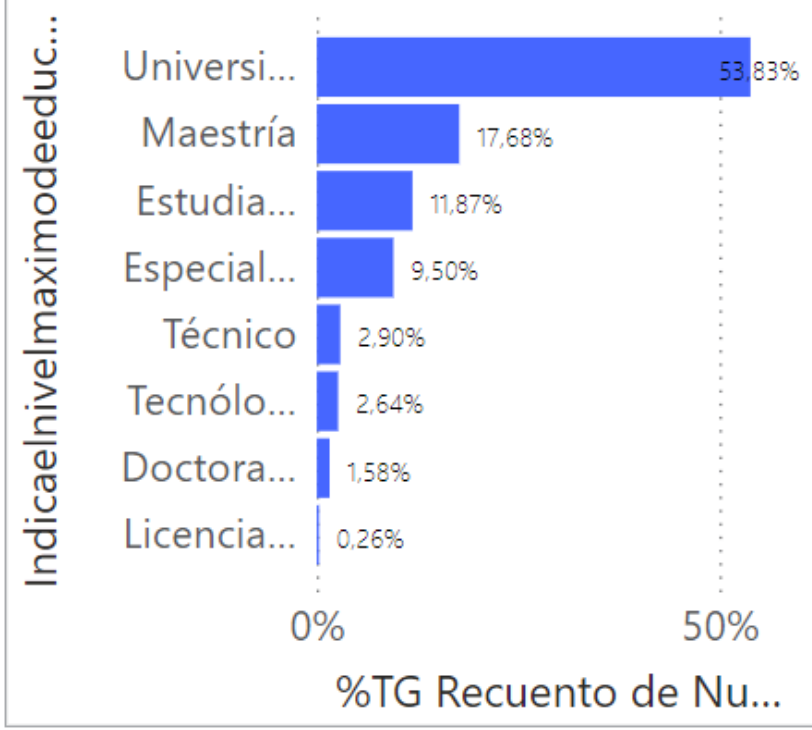
Cantidad Participantes

380

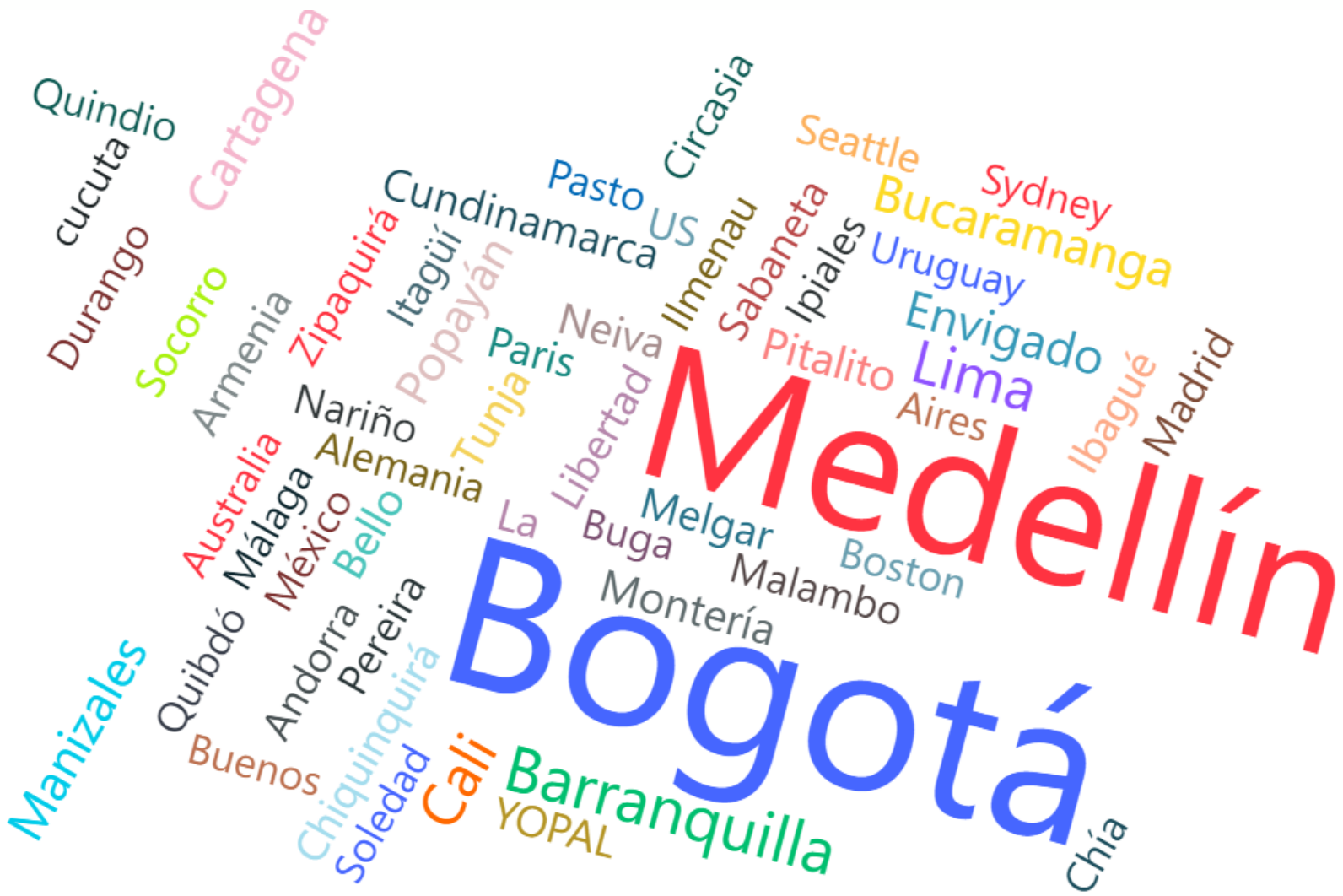
Distribución cnt equipos



Cnt profesiones



Ciudades


Medellín
Bogotá
Barranquilla
Cali
Manizales
Pereira
Bello
Envigado
Lima
Bucaramanga
Cartagena
Sabaneta
Ilmenau
Circasia
Seattle
Sydney
Uruguay
Ipiales
Pasto
US
Cundinamarca
Itagüí
Zipaquirá
Popayán
Neiva
Paris
Tunja
Libertad
Pitalito
Aires
Ibagué
Madrid
Melgar
Buga
La
Boston
Malambo
Montería
Nariño
Alemania
Australia
Málaga
México
Quibdó
Andorra
Chiquinquirá
Soledad
Buenos
YOPAL
Chía
Armenia
Socorro
Durango
cucuta
Quindio

Campos de estudio

Fechas

Debes tener presente las siguientes fechas importantes:

Checkpoint: Espacios personalizados para dudas y bonificación para los 3 equipos mas avanzados.

Entrega final: Entrega del recomendador con todos los requisitos (códigos, documentación, resultados, entre otros).

Premiación

Los cinco finalistas deberán presentarse ante el jurado conformado por integrantes del banco. Las tres mejores soluciones serán premiadas de la siguiente manera, en bonos de Amazon y/o tarjetas débito:

Primer puesto: COP $9’000.000

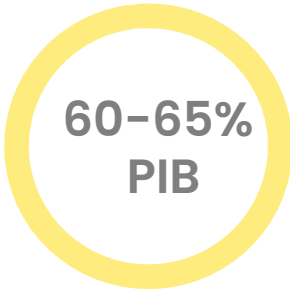
Segundo puesto: COP $6’000.000

Tercer puesto: COP $3’000.000.

Escenario

Característica de nuestro cliente

- **Grandes empresas,** que conforman los **grupos económicos mas grandes del país..**
- **Amplia trayectoria** , y con una antigüedad importante. 10 y 60 años .
- Son referentes y cuentan con **amplio conocimiento de su industria-sector,**
- Tienen personal directivo y administrativo con experiencia y muy bien preparado.
- **Demandan de forma integral todas las soluciones del Banco,** lo que general gran apetito de todos los **Bancos nacionales e internacionales**
- **Estándares de satisfacción muy altos** y siempre quieran tener una **atención premium.**
- Tienen **operación y visibilidad internacional,** continuamente son valorados por las calificadoras (S&P, Moodys, etc) y agentes del mercado de capitales.

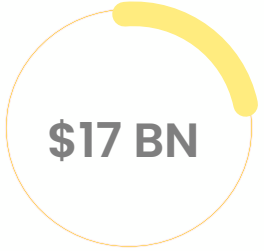


Sus ventas representan

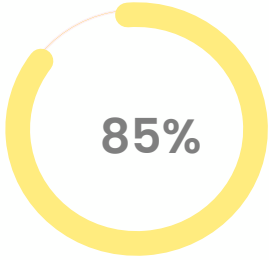
60-65% PIB

Movilizadores de la economía del país.

Pagan la nómina al mes a cerca Aprox 1.150.000 personas***



Pagan impuestos través de Bancolombia cerca de 17 Billones al año**

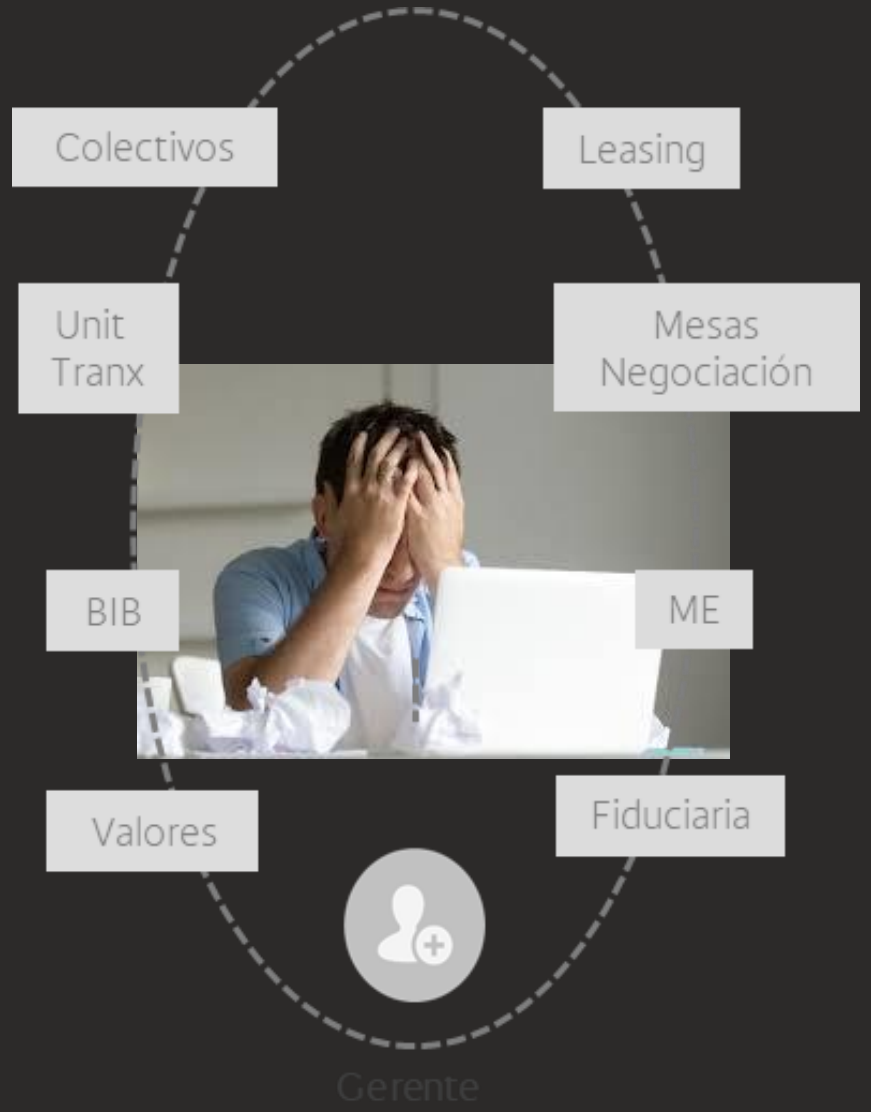


De las operaciones de Comercio Internacional del país corresponde a estos Clientes

¿Qué esperamos ser para nuestros clientes?

Queremos ser banqueros expertos, confiables, cercanos, que se anticipan, proponen futuro y lo hacen posible; que solucionan integralmente las necesidades de nuestros clientes y construyen soluciones de largo plazo a través de una propuesta de valor de clase mundial, relevante y competitiva.

Comercial **Estrategia**



Equipo Comercial

Reto 2022

Recomendador de noticias

Objetivo

Crear un sistema de recomendación de noticias sobre clientes corporativos que le ofrezcan al comercial información relevante, actualizada y confiable.

Impactos

Permitir a los comerciales estar actualizados en la información mas relevante de sus clientes y los sucesos que puedan afectarlos positiva o negativamente en base a su sector:

- Consulta y clasificación automática de noticias.
- Conocimiento y atención oportuna de los clientes.
- Categorización automática por relevancia:
    - Sostenibilidad
    - Alianzas
    - Reputación
    - Desarrollo e innovación
    - Macroeconomía

Categorías

Macroeconomía:

Son noticias sobre variables macroeconómicas que, según el comportamiento actual del mercado, puedan afectar o favorecer la actividad económica del cliente según el sector en que se desempeñen y generar así una alerta u oportunidad de negocio

importancia: 2

Regulaciones:

Noticias en las que el cliente se ve involucrado en nuevas regulaciones, normas, leyes o decretos que debe empezar a cumplir

importancia: 1.6

Otra:

Es la noticia donde está involucrado de cierta forma el cliente mas no cumple o no está dentro de las categorías anteriormente mencionadas

importancia: 0.5

Sostenibilidad:

Son las noticias que detallan al cliente en temas relacionados con innovación en sostenibilidad, compromisos medio ambientales y todo lo relacionado con la naturaleza

importancia: 1.8

Alianzas:

Noticias relacionadas con asociaciones con otras empresas o adquisición de nuevas empresas o compras por parte de empresas más grandes

importancia: 1.4

Descartable:

Es la noticia que no involucra para nada al cliente y no cabe dentro de ninguna de las categorías anteriores

importancia: 0

Innovación:

Son las noticias que dan a conocer los planes y desarrollos futuros del cliente, tales como nuevos proyectos implementación de nuevas tecnologías, herramientas y crecimiento de su negocio

importancia: 1.7

Reputación:

Noticias relacionadas con las malas o buenas experiencias de sus consumidores, escándalos de corrupción, incumplimiento de normativas, lavado de activos, listas de control, etc.

importancia: 1.2

Reto 2022: Categorización

Categorización de Noticias

Crear un sistema automático de categorización de noticias para los clientes corporativos donde se indique si la noticia relacionada a un cliente contiene información directa de este o de su actividad económica (sector); además, se debe indicar si la noticia pertenece a alguna de las clasificaciones dispuestas por el equipo corporativo para determinar su relevancia.

Reto 2022: Recomendador

Recomendación de Noticias

Crear un sistema automático que priorice la relevancia de las noticias relacionadas a un cliente corporativo de acuerdo a la categorización asignada

Recomendador		
Participacion	**Categoría**	**Recomendación**
Cliente	Sostenilibilidad	1
Sector	Macroeconomía	2

Recolección de información

Google News
- Busquedas en Español
- 2 Intervalos `[15/07/2022, 31/07/2022] [1/08/2022, 15/08/2022]`
- Noticias primera página

Iteraciones
- Nombre del cliente + en colombia
- Nombre del cliente + cada categoría
- Descripción división ciiuu
- Subsector

Tratamiento
- Eliminación de acentos
- Eliminación de texto NaN
- Eliminación de mas de un espacio

csv

Datos Clientes

clientes.csv

Clientes	
PK	nit
	nombre
	desc_ciiu_division
	desc_ciuu_grupo
	desc_ciiuu_clase
	subsec

1508 Clientes

- **nit**: Identificador único del cliente
- **nombre**: Nombre corporativo del cliente
- **desc_ciiu_división**: descripción general de la clasificación Industrial uniforme d todas las actividades económicas
- **desc_ciiu_grupo**: descripción por grupo de la clasificación Industrial uniforme d todas las actividades económicas
- **desc_ciiu_clase:** : descripción por clase de la clasificación Industrial uniforme d todas las actividades económicas
- **subsector**: Clasificación de la actividad industrial

Datos Noticias

noticias.csv

noticias	
PK	new_id
	news_url_absloute
	news_init_date
	news_final_date
	news_title
	news_text_content

23377 Noticias

- **new_id**: Identificador único de noticias
- **news_url_absolute**: Url de la noticia encontrada
- **news_init_date**: Fecha mínima del intevalo de tiempo al que pertenece la noticia
- **news_final_date:** : Fecha máxima del del intevalo de tiempo al que pertenece la noticia
- **news_title**: Título relacionado a la noticia
- **news_text_content:** Texto contenido de la noticia

Datos Clientes Noticias

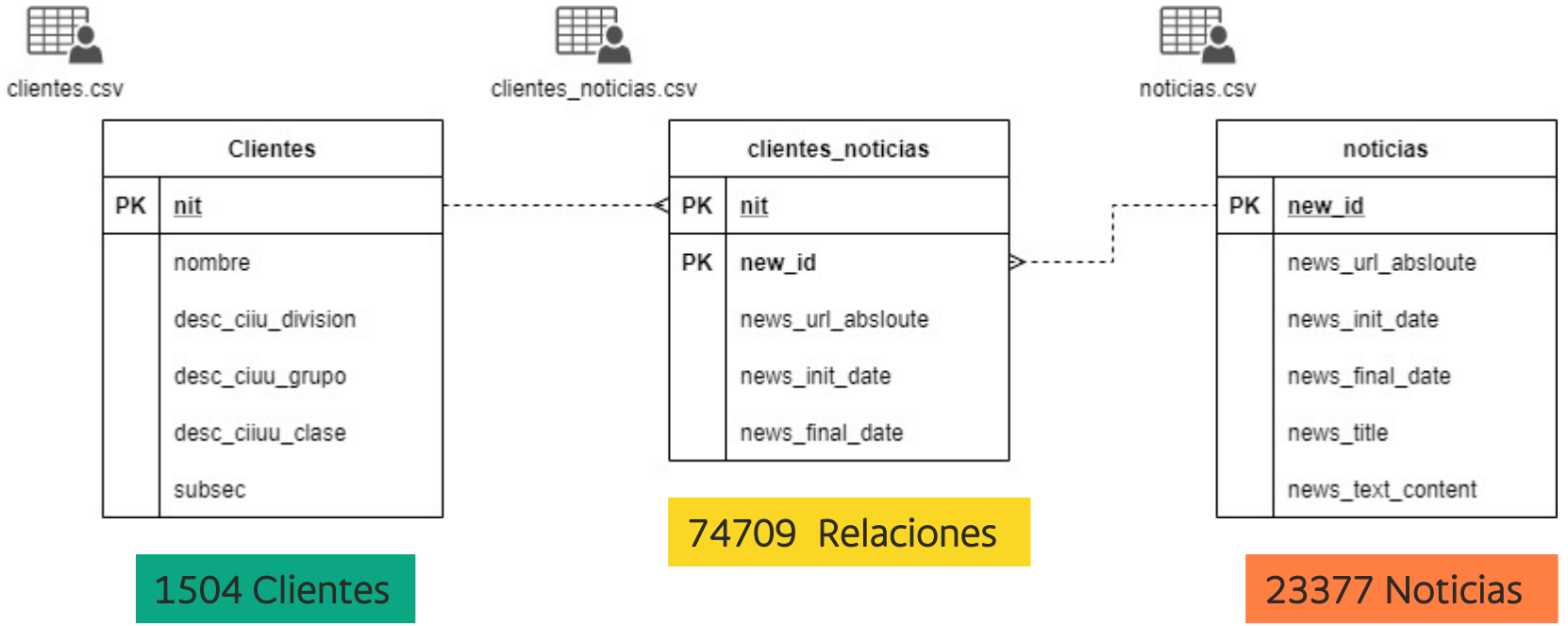
clientes_noticias.csv

clientes_noticias	
PK	nit
PK	new_id
	news_url_absloute
	news_init_date
	news_final_date

74709 Relaciones

- **nit**: Identificador único del cliente
- **new_id**: Identificador único de noticias
- **news_url_absolute**: Url de la noticia encontrada
- **news_init_date**: Fecha mínima del intevalo de tiempo al que pertenece la noticia
- **news_final_date:** Fecha máxima del del intevalo de tiempo al que pertenece la noticia

Datos



- clientes: archivo con el listado de clientes a consultar, la descripción de su actividad económica y el subsector
- clientes_noticias: relación entre cliente y las noticias consultadas mediante el proceso de descarga de información
- noticias: contenido de cada una de las noticias consultadas

Entrega de datos a competidores

clientes.csv

1508 registros
363.81 kB

clientes_noticias.csv

74709 registros
11.98 MB

noticias.csv

23377 registros
104 MB

kaggle

Dataton 2022 | Kaggle

Contenido:

- Datos para el ejercicio
- Descripción de los datos
- Canales de dudas

Preguntas Bloque 1

Entregables

Entregable 1: Firmas de Términos e Inscripción del Repositorio

Fecha Entrega: Lunes 17 de Octubre 2022

Descripción:

- **Términos y condiciones**: Acuerdo que permiten la protección de la información y recursos a los cuales los equipos van a tener acceso y otorgan al grupo Bancolombia los derechos de usabilidad de las soluciones entregadas
- **Repositorio**: Repositorio GitHub donde estará almacenada y versionada la solución

Entrega:

- Registro de Repositorios
  Diligenciar Formulario: Repositorios GitHub entrega Datatón 2022 (office.com)
  Nombre Repositorio: dataton2022-nombre_equipo
  - Este repositorio debe ser privado
  - Otorgar acceso a los usuarios wacGitHub07, jucdiaz
- Firma de acuerdos:
  Documento almacenado en el repositorio en la siguiente ruta:
  dataton2022-nombre_equipo/documentación/terminos_y_condiciones.pdf

Condición:

- Equipo que a la fecha establecida no cuente con este entregable será descalificado de la Dataton 2022

Entregable 2: Categorización

Fecha Entrega: Lunes 14 de Noviembre 2022

categorizacion.csv

nombre_equipo	nit	news_id	participacion	categoria
datatón_owners	8001852951	123	Cliente	sostenibilidad
datatón_owners	8001852952	124	Cliente	macroeconomia
datatón_owners	8001852953	125	Sector	Otra

Participación:

Variable categórica que puede tomar los valores de

- Cliente: el cliente es mencionado relevantemente en la noticia o es el protagonista
- Sector: El cliente no es mencionado relevantemente en la noticia o no es mencionado pero esta asociado con su actividad económica
- No aplica: El cliente no es mencionado y la noticia no esta vinculada con su sector

Categoría:

- Macroeconomía
- Sostenibilidad
- Innovación
- Regulaciones
- Alianzas
- Reputación
- Otra
- Descartable

Entregable 3: Recomendador

Fecha Entrega: Lunes 14 de Noviembre 2022

recomendacion.csv

nombre_equipo	nit	news_id	participacion	categoria	recomendacion
dataton_owners	8001852951	123	Cliente	sostenibilidad	1
dataton_owners	8001852951	124	Cliente	macroeconomia	2
dataton_owners	8001852951	125	Sector	Otra	3

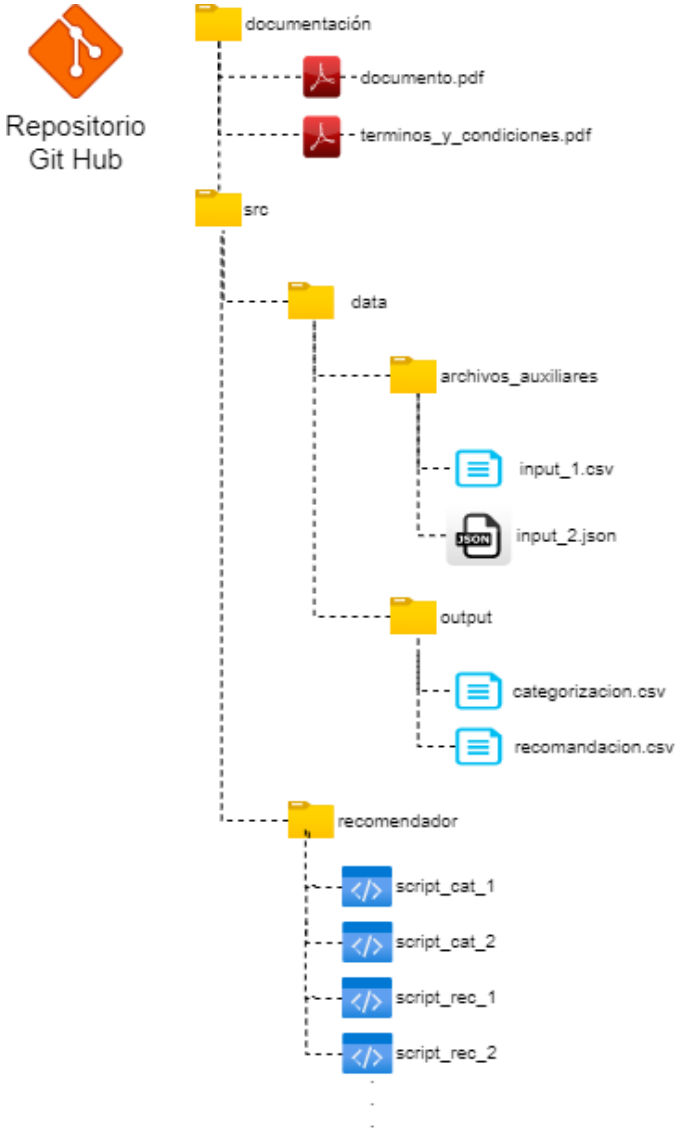
Recomendación:

Variable que ordena por prioridad las noticias relacionadas a cada cliente. Es de esperar una noticia descartable por la categoría sea calificable como no recomendable

La recomendaciones deben responder a las siguientes preguntas:

1. ¿Considera que la lectura le genera un aporte importante al conocimiento del cliente?
2. ¿confiabilidad de la fuente?
3. ¿Qué tan relevante considera que es la lectura?
4. ¿La lectura le ayuda a generar nuevos negocios y/o conversaciones relevantes con el cliente o el sector?
5. ¿La lectura es acorde a la categoría propuesta?
6. ¿la lectura le aportó conocimiento del cliente que usted no sabía?

Estructura Repositorio



Repositorio GitHub:

Nombre: dataton2022-nombre_equipo

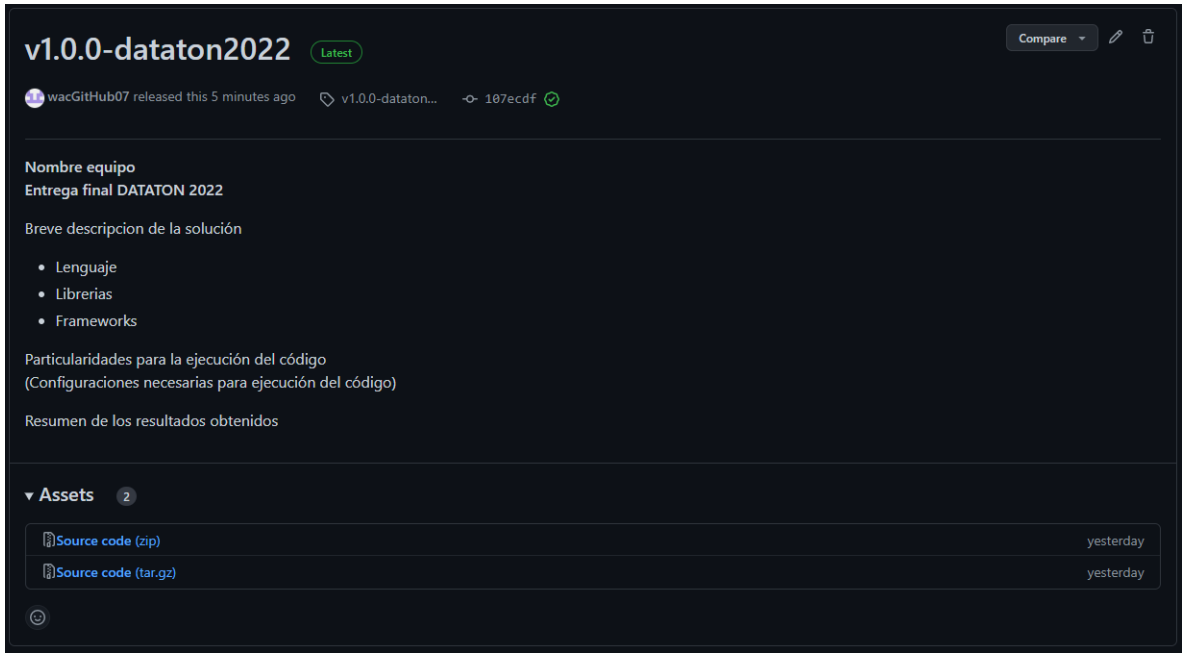
Carpetas:

- **Documentación:**
  - documento .pdf con la descripción completa del proceso realizado
  - terminos_y_condiciones.pdf: confidencialidad y cesión de derechos
- **Src**: carpeta con el código fuente de la solución
- **Data**: archivos utilizados en la solución
- **Archivos auxiliares**: archivos adicionales para proceder con la solución, ejm: parámetros, diccionarios, etc.
- **Output**: archivos resultado de la categorización y la recomendación
- **Recomendador**: scripts para la ejecución de la solución

Entrega Final (Git Hub)

Lunes
14 de Noviembre

Tag con versión final y funcional de la recomendación



Email:

Destinatarios : warboled@bancolombia.com.co; jucadiaz@bancolombia.com.co; canaliti@bancolombia.com.co

Asunto: Dataton 2022 Entrega Final - Nombre Equipo

Contenido:

- Integrantes
- Link Repositorio

Preguntas Bloque 2

Checkpoint

Dudas

- Dudas generales en cuanto a los datos entregables y entendimiento del problema dirigirse en el correspondiente canal del foro de Kaggle.

  Canales:

  Entendimiento del problema | Entregables | Datos entregados

- Dudas en cuanto a la estrategia que está desarrollando, nuevas ideas o temas confidenciales de su solución, escribir un correo a:

  Destinatarios : warboled@bancolombia.com.co; jucadiaz@bancolombia.com.co

  Asunto: Dataton2022 Dudas Nombre Equipo

  Contenido: Duda a escalar

Asesorías

Lunes 31 de Octubre -
Viernes 4 de Noviembre

Solicitud de asesoría personalizada:

- Espacios de **30 min** por Microsoft Teams
- Los espacios serán agendados de **9 a.m. a 12 p.m** y de **2 p.m. a 5 p.m.** por orden de llegada de las solicitudes
- Solicitud de asesoría hasta el miércoles 2 de Noviembre.

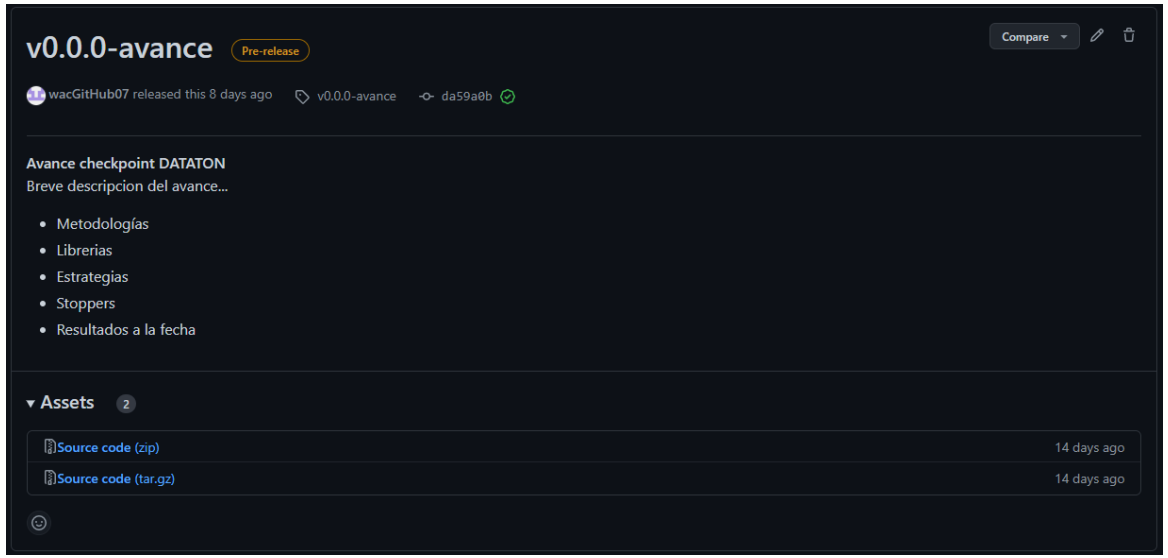
Email:

- **Destinatarios :** warboled@bancolombia.com.co; jucadiaz@bancolombia.com.co
- **Asunto:** Datatón2022 Asesoría Nombre Equipo
- **Contenido:** Tema en el que necesita asesoría

Avance (Opcional)

Domingo 31 de Octubre

Tag con versión funcional de la categorización



Email:

Destinatarios : warboled@bancolombia.com.co;

jucadiaz@bancolombia.com.co

Asunto: Dataton2022 Avance Nombre Equipo.

Bonificación:

Los 3 equipos con los mejores avances reciben un 5% de bonificación sobre la puntuación final

Aquellos que no participen en este avance nos vemos en la entrega final!!

Evaluación de Resultados

Evaluación de Resultados:

Etapa calificación	Criterio	Porcentaje importancia
1	Innovación	10%
1	Metodología	15%
1	Documentación	10%
1	Replicabilidad	10%
2	Calificación Categorías	20%
3	Calificación Recomendaciones	20%
4	Presentación	15%

Evaluación de Resultados: Resumen

Ganadores

Final: Presentación 15%

Semifinal: Calificación Recomendaciones: 20%

Cuartos de final: Calificación Categorías: 20%

Octavos Final: Innovación, Metodología, Replicabilidad, Documentación: 45%

Todos los equipos Con entregables

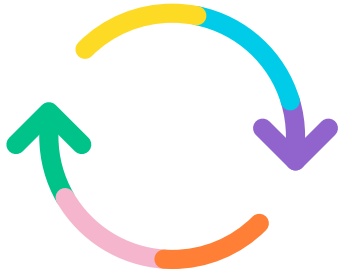
TOP 20

TOP 10

TOP 5

TOP 3

Bancolombia



Innovación:

10%

Innovación

Se evaluará con un **valor de 0 a 5** dado por un conjunto de expertos, dentro del centro de excelencia en analítica e IA, en este criterio se calificará a los equipos que utilicen **herramientas, metodologías, novedosas o recientemente conocidas, originalidad**, también se premiará su **creatividad** al realizar combinaciones de las metodologías o algoritmos clásicos, e incluso un **uso creativo** del manejo de los datos para encontrar un buen desempeño

Metodología:

15%

Metodología:

Se evaluará con un **valor de 0 a 5** dado por un conjunto de expertos, dentro del Centro de Excelencia en Analítica e IA, en este criterio se calificará **el buen uso de las metodologías y herramientas utilizadas**, que cumplan con sus validaciones técnicas y cumplimiento de supuestos si la técnica lo requiere. Se premiará que la metodología tenga un argumento solido para su utilización.

Documentación:

10%

Documentación:

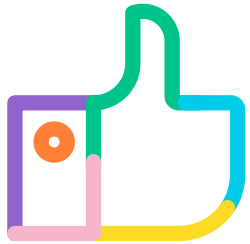
Se evaluará con un valor de 0 a 5 dado por un conjunto de expertos, dentro del Centro de Excelencia en Analítica e IA, en este criterio se calificará por tener un documento claro y conciso máximo 6 paginas, pero a su vez que abarque lo más relevante de la metodología, innovación y resultados obtenidos. También se premiará que el código o funciones propias sean bien documentadas.

Replicabilidad

10%

Replicabilidad:

Se evaluará con un **valor de 0 a 5** dado por un conjunto de expertos, dentro del Centro de Excelencia en Analítica e IA, en este criterio se calificará que la **solución propuesta pueda ser generalizable** para más clientes y a su vez pueda ser **replicable fácilmente por otro equipo analítico.** Adicional se calificará el grado de automatización del código y el hecho de no tener alguna interacción humana.



Evaluación de Resultados:

20%

Calificación Categorías:

Categoría	Cantidad noticas revisar (aleatoriamente)	Peso calificación (w_i)	Porcentaje etiquetas correctas* "participación" (P_{p_i})	Porcentaje etiquetas correctas* "categoría" (P_c_i)
Macroeconómicas	20	2		
Sostenibilidad	18	1.8		
innovación	17	1.7		
regulaciones	16	1.6		
alianzas	14	1.4		
reputación	12	1.2		
Otra	5	0.5		
Descartable	3	0.3		

$$Calif_{eqpo\_etp2} = \sum_{i=1}^{8} W_i * P_{p_i} * P_{C_i}$$

Calificación 20%
Recomendaciones

La etapa 2 será calificada de acuerdo con el **criterio experto de las personas que utilizaran la herramienta**, esta será calificada con la ayuda de **un Chatbot** donde a **cada equipo** se le seleccionaran aleatoriamente

$\mathbf{N_c = 10}$ **clientes** (con características diferentes) con **las primeras 5 noticias recomendadas** y el experto asignado a cada cliente calificará las recomendaciones respondiendo estas 3 preguntas:

$\overline{P_i}$: Promedio de la pregunta i en las 5 noticias del cliente j seleccionado

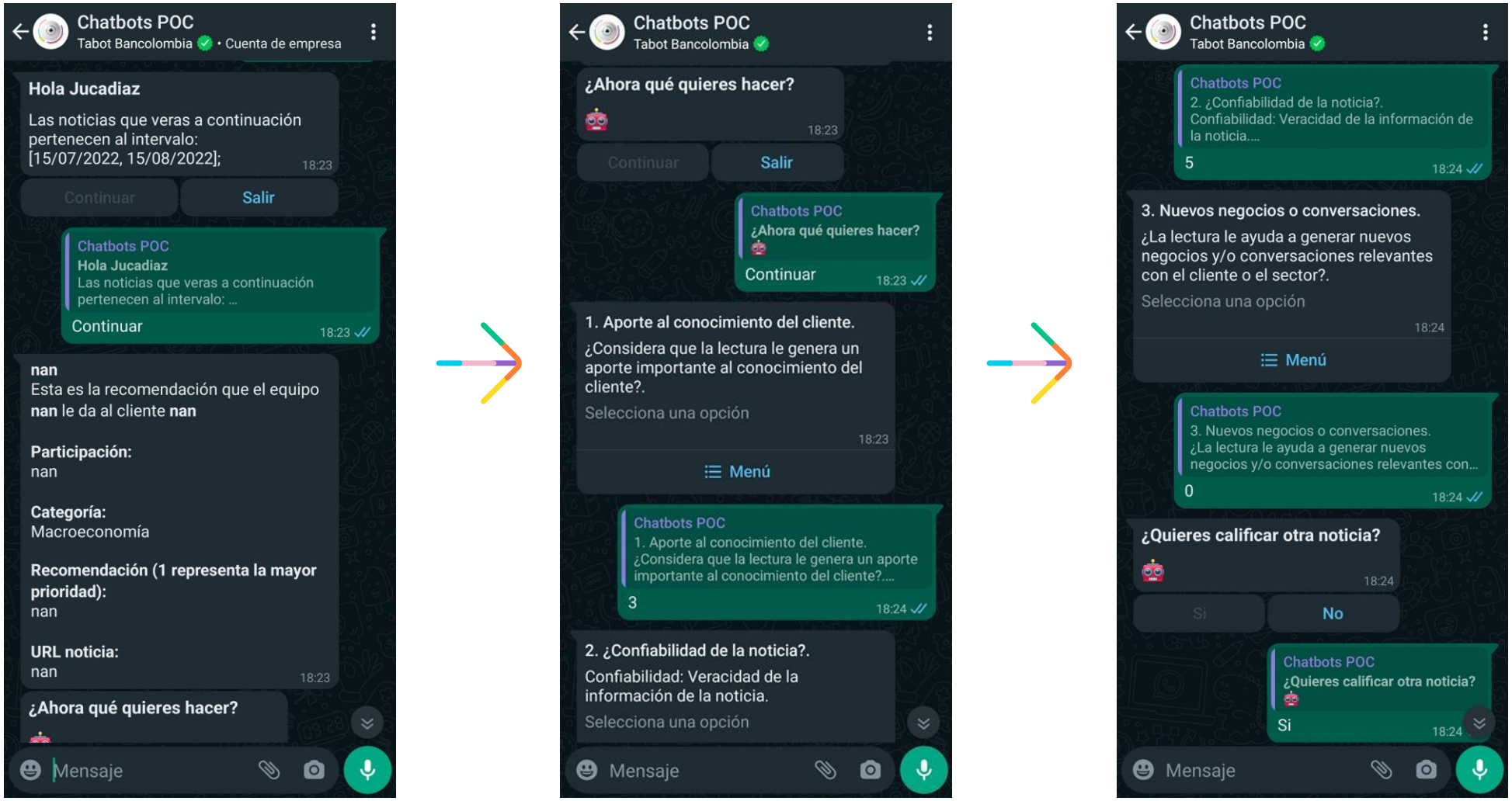
Preguntas finales:

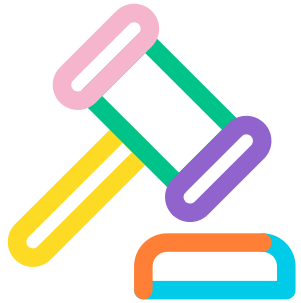
P_1 ¿Considera que la lectura le genera un aporte importante al conocimiento del cliente? De un valor de 0 a 5 donde 0 es ningún aporte importante y 5 un aporte muy impactante

P_2 ¿confiabilidad de la fuente? De un valor de 0 donde 0 es nada confiable y 5 muy confiable

P_3 ¿La lectura le ayuda a generar nuevos negocios y/o conversaciones relevantes con el cliente o el sector? De un valor de 0 a 5 donde 0 es nada relevante y 5 muy relevante

$$Calif_{eqpo\_etp3} = \frac{1}{N_c}\sum_{j=1}^{N_c}\left(\frac{1}{n_p}\sum_{i=1}^{3}\overline{P_i}\right)_j$$


Chatbots POC
Tabot Bancolombia • Cuenta de empresa
Hola Jucadiaz
Las noticias que veras a continuación pertenecen al intervalo: [15/07/2022, 15/08/2022];
18:23
Continuar
Salir
Chatbots POC
Hola Jucadiaz
Las noticias que veras a continuación pertenecen al intervalo: ...
Continuar
18:23
nan
Esta es la recomendación que el equipo nan le da al cliente nan
Participación:
nan
Categoría:
Macroeconomía
Recomendación (1 representa la mayor prioridad):
nan
URL noticia:
nan
18:23
¿Ahora qué quieres hacer?
Mensaje
Chatbots POC
Tabot Bancolombia
¿Ahora qué quieres hacer?
18:23
Continuar
Salir
Chatbots POC
¿Ahora qué quieres hacer?
Continuar
18:23
1. Aporte al conocimiento del cliente.
¿Considera que la lectura le genera un aporte importante al conocimiento del cliente?.
Selecciona una opción
18:23
Menú
Chatbots POC
1. Aporte al conocimiento del cliente.
¿Considera que la lectura le genera un aporte importante al conocimiento del cliente?....
3
18:24
2. ¿Confiabilidad de la noticia?.
Confiabilidad: Veracidad de la información de la noticia.
Selecciona una opción
Mensaje
Chatbots POC
Tabot Bancolombia
Chatbots POC
2. ¿Confiabilidad de la noticia?.
Confiabilidad: Veracidad de la información de la noticia....
5
18:24
3. Nuevos negocios o conversaciones.
¿La lectura le ayuda a generar nuevos negocios y/o conversaciones relevantes con el cliente o el sector?.
Selecciona una opción
18:24
Menú
Chatbots POC
3. Nuevos negocios o conversaciones.
¿La lectura le ayuda a generar nuevos negocios y/o conversaciones relevantes con...
0
18:24
¿Quieres calificar otra noticia?
18:24
Si
No
Chatbots POC
¿Quieres calificar otra noticia?
Si
18:24
Mensaje



Evaluación de Resultados:

15%

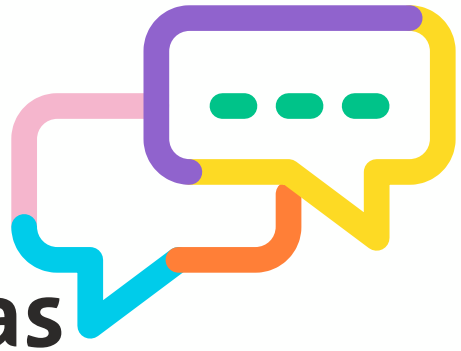
Presentación

Se evaluará con un valor de 0 a 5 dado por un conjunto de expertos, pertenecientes la vicepresidencia de negocios corporativos, en este criterio se calificará los criterios **Storytelling** y calificaran el reto de forma **general e integral**, donde explicaran a los dueños del negocio su solución y porque su solución analítica es la mejor

Numero Jurado	Nombre experto	Cargo
1	Juan Sebastian Ruiz	Líder EVC
2	Leidy Carolina Hernández	Líder Estrategia Comercial
3	Laura Victoria Álvarez	P.O Sector Agro Comercio
4	Rafael Martínez	Gerente de Zona
5	Julian Mora Gómez	V.P de Cumplimiento

Prototipo básico ejemplo

https://github.com/jucdiaz/Dataton_Owners



Ideas Inspiración...

- Inspirarse en The Hugging Face IA – Model.
- Aplicar modelos LDA.
- Descargar noticias de fuentes que tengo la certeza de que categoría son, por ejemplo, una página que solo escriba de macroeconomía, con una red neuronal siamesa para identificar similitudes.
- Implementar técnicas de Named Entity Recognition.
- Implementar análisis de sentimientos para saber la importancia de la noticia dependiendo del grado de positividad o negatividad de la noticia.
- No realizar descarte uno a uno de noticias, existen noticias de clientes relacionadas a otros de su misma o diferente actividad económica.
- No seria lógico recomendar una noticia como descartable en una prioridad alta.
- La aparición de una noticia en ambos intervalos de fechas puede ser indicador de relevancia.

Preguntas Bloque Final

Gracias