	PRUEBA DE CONOCIMIENTOS DATOS NO ESTRUCTURADOS I		
	VICEPRESIDENCIA BANCA PERSONAL Y MERCADEO ANALÍTICA - DEPTO. DATOS NO ESTRUCTURADOS		
	Depto. Datos No Estructurados	Elaborado	Versión:1

Prueba de Conocimiento Datos No Estructurados

Entrega: juan.pena@davivienda.com, javier.herrera@davivienda.com

En el Departamento de Datos No Estructurados buscamos personas con excelentes capacidades técnicas a las que les guste explorar los últimos avances en IA para asumir retos de especial dificultad. Aunque la siguiente prueba busca, en primera instancia, corroborar su idoneidad técnica, para nosotros es muy importante que esta destreza venga acompañada de dos elementos adicionales:

1. Creatividad para encarar desafíos técnicos en procesamiento de datos no estructurados.
2. Habilidad para comunicar el trabajo hecho y sus resultados a un público general o experto.


Por lo tanto, el objetivo de la prueba es tener una visión completa de cómo es su compromiso, capacidad analítica, creatividad técnica y comunicación en la solución de un problema puntual de datos no estructurados, la cual, en la práctica, se verá reflejada en una presentación a los jurados.

Para el desarrollo de esta prueba cuentan con una semana, tiempo en el que deben desarrollar la solución del problema mediante código debidamente documentado, depositado en un cuaderno de Python (Google Colab), así como una presentación gerencial de su trabajo. De parte de los jurados se evaluarán estos dos elementos y se comprometen a respetar la propiedad intelectual de los desarrollos hechos por ustedes donde haya lugar (modelos personalizados, APIs propias, técnicas especiales, etc). Si lo desea, puede hacerlo explícito mediante licencia Creative Commons.

Puede hacer los supuestos que considere necesarios. Esto depende de la forma en que usted aborde el problema. Aunque los cuadernos de Python suministrados sugieren caminos, no hay una solución única. Fundamentalmente queremos ver soluciones ordenadas, creativas (si consideran que esta cualidad es pertinente en la resolución del problema) y bien comunicadas.

Introducción

Según Gartner, en este momento, los datos no estructurados conforman el 80-90% del total de los datos que manejan las compañías. Dentro de ellos, las imágenes de documentos ocupan un lugar importante. Cada año se hace más evidente la necesidad de convertir los datos contenidos en estas imágenes en información útil que pueda ser analizada en bases de datos.

	PRUEBA DE CONOCIMIENTOS DATOS NO ESTRUCTURADOS I		
	VICEPRESIDENCIA BANCA PERSONAL Y MERCADEO ANALÍTICA - DEPTO. DATOS NO ESTRUCTURADOS		
	Depto. Datos No Estructurados	Elaborado	Versión:1

Dentro de los primeros pasos para lograr este objetivo se encuentra, por supuesto, el identificar qué imágenes guardan contenido y cuáles pueden desecharse, lo que se traducirá en menores costos de almacenamiento, procesamiento de archivos y recursos humanos.

El objetivo de este reto es lograr un filtro que discrimine automáticamente un tipo de documento sin información relevante: páginas en blanco. Se busca que este filtro reciba como entrada una carpeta con imágenes de documentos diversos y produzca como salida dos carpetas, una con imágenes de páginas sin contenido y otra con imágenes de páginas con contenido.

Páginas con solo el membrete del documento se consideran páginas sin contenido, así como las que, al momento de ser escaneadas, alcanzan a reflejar contenido ininteligible del reverso de la página.



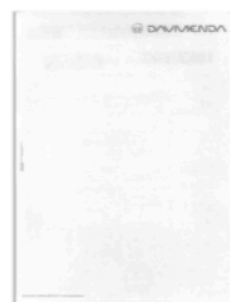
Con contenido



Sin contenido




Sin contenido



Sin contenido

Instrucciones Importantes

- Se suministrará el conjunto de imágenes de documentos a filtrar.
- Debe proporcionar una matriz de confusión que sintetice el desempeño del filtro desarrollado sobre las imágenes suministradas.
- El desarrollo es libre, siempre y cuando quede contenido en el cuaderno de Google Colab suministrado. Siéntase libre de usar la mejor técnica que considere para resolver el reto. Puede utilizar cualquier herramienta, propia o que encuentre en Internet.
- El cuaderno debe tener las líneas de comando que instalen las librerías y versiones utilizadas para correr el desarrollo.
- El desarrollo hecho debe estar documentado, tanto en código como en comentarios del cuaderno suministrado.

	PRUEBA DE CONOCIMIENTOS DATOS NO ESTRUCTURADOS I		
	VICEPRESIDENCIA BANCA PERSONAL Y MERCADEO ANALÍTICA - DEPTO. DATOS NO ESTRUCTURADOS		
	Depto. Datos No Estructurados	Elaborado	Versión:1

- Puede utilizar otros conjuntos de datos en su desarrollo, siempre que la matriz de confusión solo corresponda a imágenes proporcionadas por nosotros.
- En la presentación debe justificar las razones por las que escogió el enfoque desarrollado.
- Si ejecuta algún proceso previo o auxiliar en la solución del reto, por favor justifíquelo en la presentación.
- Si por algún motivo su desarrollo no tiene los resultados esperados por usted, recuerde que esto también es un resultado. Elabore la presentación y justifique su proceso. ¡Es posible que su trabajo sea muy interesante y tenga mucho potencial!

Bono: Adicionalmente, si tiene conocimientos de despliegue, “apificación” o puesta en producción de su desarrollo, puede hacerlos explícitos en la presentación mediante un bosquejo sustentado de lo que usted haría para que su desarrollo sea utilizado por un usuario final. Esto reemplazará el peor puntaje parcial que haya obtenido en la calificación de su presentación. No se tiene que desarrollar.

Objetivo

El propósito de la prueba es medir su capacidad de análisis, programación, elección de técnicas y comunicación de procesos y resultados. Dentro del Departamento de Datos No Estructurados creemos que la unión de estos elementos determina la buena calidad de las soluciones que le aportamos a la organización. Por esta razón, no consideramos esto como una prueba “tipo Kaggle” y sugerimos que no tomen como criterio supremo el desempeño de su desarrollo.

Nos interesa conocer de su propia voz el proceso que hizo para resolver el reto, ver la forma como abordó el problema, el orden en su código y la manera como sintetiza las partes interesantes de su trabajo en una presentación.


Insumos

Para el desarrollo se compartirá una carpeta **‘Prueba_de_conocimientos_davivienda’** con la subcarpeta de este ejercicio **‘Ejercicio_1_imagenes’** con el siguiente contenido:

1. Carpeta de imágenes a clasificar.
2. Cuaderno de Google Colab ‘Prueba_de_conocimientos_I_DNE_Davivienda’.
3. Documento Explicativo del Ejercicio.

Entregables

Se debe entregar en archivo zip una carpeta que contenga los siguientes elementos:

	PRUEBA DE CONOCIMIENTOS DATOS NO ESTRUCTURADOS I		
	VICEPRESIDENCIA BANCA PERSONAL Y MERCADEO ANALÍTICA - DEPTO. DATOS NO ESTRUCTURADOS		
	Depto. Datos No Estructurados	Elaborado	Versión:1

- **Presentación Gerencial:** Una presentación (en formato PDF) donde se encuentre:
 - Exploración de datos.
 - Técnicas usadas o exploradas.
 - Dificultades encontradas.
 - Aspectos importantes.
 - Resultados ejemplo Matriz de confusión.
 - Costos de la ejecución de su desarrollo, si corresponde (Ej: si utilizó una API paga, una VM con hardware especial, etc).
 - Demás detalles que considere relevantes para la presentación de su proceso.
 - **(Opcional)** - Bosquejo de puesta en producción del desarrollo hecho.
- **Códigos:** [Repositorio de Github](#) que incluya Cuaderno de Google Colab suministrado con el desarrollo hecho, debidamente documentado (o su URL). Los resultados de la ejecución de cada celda deben ser visibles sin necesidad de correr las celdas. Tenga en cuenta que no vamos a modificar rutas de ubicación de input/output, archivos auxiliares, etc. por lo que todo el contenido debe ser completo y consistente dentro del archivo zip entregado.
- **Carpetas de resultados de filtro:** El desarrollo hecho debe enviar las imágenes detectadas con contenido a la carpeta “docs_con_contenido” y las páginas en blanco a la carpeta “docs_sin_contenido”.
- **(Opcional)** - Si su desarrollo necesita de archivos auxiliares para funcionar, por favor inclúyalos dentro del archivo zip de manera que sea consistente con la ejecución de los códigos hechos.


Nota: Los resultados deben entregarse por medio de correo electrónico juan.pena@davivienda.com, javier.herrera@davivienda.com

- Con el siguiente asunto: Prueba de Conocimiento Datos no Estructurados [nombre participante]
- Con la siguiente nomenclatura: datos_no_estructurados_01_nombrep participante.zip

Evaluación

Se evaluarán sus capacidades de análisis, manejo de datos, programación y comunicación. Esto se realizará bajo la siguiente metodología:

- Desarrollo de los códigos: 30%
 - Orden (5%)
 - Documentación (5%)

	PRUEBA DE CONOCIMIENTOS DATOS NO ESTRUCTURADOS I		
	VICEPRESIDENCIA BANCA PERSONAL Y MERCADEO ANALÍTICA - DEPTO. DATOS NO ESTRUCTURADOS		
	Depto. Datos No Estructurados	Elaborado	Versión:1

- Técnica (20%)
- Presentación con los resultados: 70%
 - Orden en el relato del proceso (20%)
 - Conocimientos técnicos (20%)
 - Completitud (10%)
 - Claridad (20%)
 - Bono: Si demuestra suficiencia de conocimientos de despliegue de la analítica presentada, obtendrá un buen puntaje en el bono y este reemplazará la peor calificación que tenga en los cuatro criterios anteriores.