CONSOLIDADO DE OBSERVACIONES Y RESPUESTAS AL SONDEO DE MERCADO

Objeto: "Indexar la información documental generada por los proyectos de la transición energética y disponerla para búsquedas eficientes, utilizando nuevas tecnologías de la información y ciencia de la computación, mediante la adquisición e implementación de una solución tecnológica de indexación y búsqueda de información."

PREGUNTAS PRESENTADAS POR: Lydis Rangel Cojo – ARCETEC SOLUCIONES TECNOLÓGICAS

Preguntas para la construcción de un modelo ML o LLM que nos permita la clasificación de documentos

1. ¿Cuál es el propósito principal de la clasificación de estos documentos? ¿Es para la búsqueda eficiente de información, para la organización de una base de datos, o para otro propósito?

   Respuesta: El propósito principal de la indexación de la información es la búsqueda, acceso y consulta eficiente de información, esto mediante una herramienta de indexación y búsqueda que integre nuevas tecnologías de la información y ciencia de la computación.

2. ¿Qué tipo de documentos se van a clasificar? ¿Son informes técnicos, artículos científicos, mapas, imágenes, datos tabulares u otros formatos?

   Respuesta: Tal como se especifica en el apartado "ASPECTOS TÉCNICOS Y ACTIVIDADES A EJECUTAR" y en la Actividad 2, la información a indexar corresponde a tipo documental la cual no requiere software especializado o petrotécnico para su visualización y manejo. En particular, esta información tiene contenido técnico y no técnico que encapsula los aspectos más destacados y cruciales de los contratos, convenios y proyectos gestionados por la VT.

   > *"Dicha información puede corresponder a imágenes de campo (afloramientos rocosos, muestras de roca, núcleos, ripios, fotografías de secciones delgadas, puntos de agua, entre otros), información derivada de perforaciones y pozos (curvas de registros, ensayos y pruebas de pozos, geoquímica de rocas, hidrogeoquímica, análisis isotópicos, entre otros), perfiles e imágenes sísmicas, imágenes espectrales e hiperespectrales, imágenes de tomografía, DRX, FRX, SEM, espectroscopía Raman, secciones estructurales, mapas estructurales, columnas estratigráficas, bloques diagrama, batimetrías y geofísica marina, cartografía (geológica, geomorfológica, geofísica, temática, entre otras), datos de sensores remotos e imágenes satelitales, entre otros."*
   >
   > *"Es fundamental que durante la implementación se reconozca tanto la información documental técnica como la no técnica que pueda encapsular los aspectos más destacados ruciales de los contratos, convenios y proyectos gestionados por la VT. Esto abarca desde estudios previos, sondeos de mercado, comunicaciones internas, minutas, actas de inicio, actas de finalización, actas de liquidación, otrosí, actas de entregas, hasta productos en formatos no técnicos como informes finales, informes de actividades y supervisión. No obstante, es crucial señalar que, dada la disposición actual de la información, la cual mayormente carece de estructura, se torna imperativo identificar y ubicar la información*

definitiva y esencial almacenada en nuestro directorio colaborativo. Esto se debe a la posibilidad de que existan versiones preliminares de cada uno de los documentos mencionados anteriormente."

3. ¿Cuántas categorías o temas principales deseas incluir en la clasificación? ¿Se enfocará en áreas específicas de la geología, geofísica y geoespacial, como exploración de minerales, prospección petrolera, análisis de terremotos, cartografía, etc.?

   Respuesta: Tal como se indica en la descripción de la necesidad, se requiere una solución tecnológica que permita búsquedas, acceso y consultas eficientes de información. Ahora bien, la manera en que se lleva a cabo la tipología, clasificación y catalogación documental, lo cual es fundamental en la indexación y búsqueda de información, no obedece a una manera en particular de hacerlo, ni a unas etiquetas y categorías específicas. Sin embargo, la propuesta de los interesados en materia de lo anteriormente mencionado debe responder eficientemente a los aspectos técnicos y la necesidad especificados en el sondeo de mercado.

4. ¿Cómo se van a adquirir y preprocesar los documentos? ¿Se utilizará un motor de búsqueda para recopilar documentos relevantes? ¿Necesitarás técnicas de limpieza y normalización de texto?

   Respuesta: Tal como se describe en la descripción de la necesidad, en el apartado "ASPECTOS TÉCNICOS Y ACTIVIDADES A EJECUTAR" y en la Actividad 2, la información actualmente se encuentra en su mayoría no estructurada. lo anterior conlleva a que se requiera un preprocesamiento para identificar, clasificar y validar la veracidad, integridad, confiabilidad y consistencia de la información. Solo así se podrá recopilar la información correcta para la implementación de la herramienta de indexación y búsqueda. Es fundamental que los interesados propongan maneras eficientes e innovadoras para llevar a cabo este preprocesamiento.

5. ¿Cuáles serán las características relevantes de los documentos para la clasificación? ¿Palabras clave, temas recurrentes, estructura del documento, metadatos, etc.?

   Las preguntas 3 y 5 están interconectadas.

   Respuesta: Tal como se indica en la descripción de la necesidad, se requiere una solución tecnológica que permita búsquedas, acceso y consultas eficientes de información. Ahora bien, la manera en que se lleva a cabo la tipología, clasificación y catalogación documental, lo cual es fundamental en la indexación y búsqueda de información, no obedece a una manera en particular de hacerlo, ni a unas etiquetas y categorías específicas. Sin embargo, la propuesta de los interesados en materia de lo anteriormente mencionado debe responder eficientemente a los aspectos técnicos y la necesidad especificados en el sondeo de mercado.

6. ¿Qué modelo de lenguaje natural (LLM) planean utilizar? ¿Una red neuronal convolucional (CNN), una red neuronal recurrente (RNN), un transformer como BERT, o alguna otra arquitectura?

   Respuesta: El Large Language Model (LLM) corresponde a una propiedad fundamental de la herramienta o solución que se busca adquirir e implementar. Es responsabilidad de los interesados proveer información sobre el LLM utilizado en la solución propuesta. Este mismo

será considerado teniendo en cuenta que es una propiedad fundamental que determina las capacidades que posee la solución y la cobertura de esta frente a las necesidades a las que se busca dar solución.

7. ¿Cómo se entrenará el modelo? ¿Tienes acceso a datos etiquetados para el entrenamiento supervisado, o planeas utilizar técnicas de aprendizaje no supervisado?

   Respuesta: Al igual que el LLM, el algoritmo de aprendizaje de los modelos es una característica propia y crucial de las herramientas o soluciones que se busca adquirir e implementar. La elección de los modelos y su aprendizaje recae en los interesados, quienes deben considerar la necesidad y los requerimientos técnicos específicos de la VT, las actividades a ejecutar que se encuentran especificadas en el sondeo de mercado y los aspectos técnicos de los modelos. No obstante, como se especifica en las actividades 4 y 5, las herramientas y/o soluciones pueden ser probadas y ajustadas, incluyendo allí la posibilidad de variar los algoritmos de aprendizaje con miras a mejoras en el rendimiento en el contexto especifico de la VT

8. ¿Cómo manejarás los desafíos específicos de cada campo (geología, geofísica, geoespacial) en términos de terminología y lenguaje técnico? ¿Será necesario incorporar conocimiento experto en el proceso de clasificación?

   Respuesta: Según lo indicado en el apartado “PERSONAL MINIMO”, es necesario la incorporación de geólogos en el personal en la ejecución del proyecto. Lo anterior debido a la naturaleza de la información, la cual en su mayoría corresponde a información derivada de contratos, convenios y proyectos desarrollados por la VT.

9. ¿Cómo se integrará el modelo en un sistema o aplicación existente? ¿Se proporcionará una interfaz de usuario para interactuar con los resultados de la clasificación?

   Respuesta: La interfaz de usuario (IU) es una propiedad fundamental de las herramientas o soluciones que se busca adquirir e implementar. Es responsabilidad de los interesados proveer información detallada sobre la IU en la propuesta de solución. Este mismo será el componente de interacción directa entre los usuarios y la herramienta que busca dar solución a las necesidades de la organización. Al igual que los modelos empleados, las IU serán probadas con usuarios reales para evaluar la facilidad de uso, diseño, accesibilidad, funcionalidad y eficiencia. Finalmente, se realizarán ajustes a partir de las pruebas de usabilidad, si es necesario. Es importante mencionar que la interfaz de usuario debe alinearse al manual de imagen de la Entidad o a los lineamientos que establezca la dependencia encargada de las comunicaciones internas. En el momento no hay otros lineamientos de experiencia de usuario UX.

10. ¿Cómo se actualizará y mantendrá el modelo con el tiempo? ¿Se necesita un sistema de retroalimentación para mejorar la precisión y relevancia de la clasificación?

    Respuesta: Tal como se indica en el apartado “ASPECTOS TÉCNICOS Y ACTIVIDADES A EJECUTAR”, la solución tecnológica debe incluir servicios de mantenimiento y soporte. Los interesados deben detallar los servicios que se brindarán junto con la herramienta o solución, incluyendo allí los de soporte y actualización de los modelos.

11. ¿Cuáles son los requisitos de seguridad y privacidad de los datos que se clasificarán? ¿Se necesita anonimización o cifrado de información sensible?

Respuesta: Los interesados deben incluir en su propuesta lo relacionado a el control de acceso (roles, responsabilidades, mecanismos de control de acceso, monitoreo de accesos), protección de datos (medidas contra intrusiones, robo, perdida o daño, uso indebido, copias de seguridad y planes de recuperación ante desastres), capacitación frente a políticas de seguridad y privacidad de datos relacionados a la herramienta/solución a implementar, plan de respuesta a incidentes, revisión y actualización de plan de gobernanza de datos. En el caso en que la implementación de la herramienta y/o solución involucre flujo o tránsito de datos e información entre o hacia sistemas externos a la organización, se requiere especificar los detalles de encriptación, anonimización y/o seudonimización que garanticen la gobernanza de los datos.

Preguntas para la indexación documental

1. ¿Cuál es el alcance de la información documental que se va a indexar? ¿Incluye informes técnicos, investigaciones científicas, políticas gubernamentales, datos de energías renovables, etc.?

Respuesta: Tal como se especifica en el apartado “ASPECTOS TÉCNICOS Y ACTIVIDADES A EJECUTAR” y en la Actividad 2, la información a indexar corresponde a tipo documental la cual no requiere software especializado o petrotécnico para su visualización y manejo. En particular, esta información tiene contenido técnico y no técnico que encapsula los aspectos más destacados y cruciales de los contratos, convenios y proyectos gestionados por la VT.

> *“Dicha información puede corresponder a imágenes de campo (afloramientos rocosos, muestras de roca, núcleos, ripios, fotografías de secciones delgadas, puntos de agua, entre otros), información derivada de perforaciones y pozos (curvas de registros, ensayos y pruebas de pozos, geoquímica de rocas, hidrogeoquímica, análisis isotópicos, entre otros), perfiles e imágenes sísmicas, imágenes espectrales e hiperespectrales, imágenes de tomografía, DRX, FRX, SEM, espectroscopía Raman, secciones estructurales, mapas estructurales, columnas estratigráficas, bloques diagrama, batimetrías y geofísica marina, cartografía (geológica, geomorfológica, geofísica, temática, entre otras), datos de sensores remotos e imágenes satelitales, entre otros.”*

> *“Es fundamental que durante la implementación se reconozca tanto la información documental técnica como la no técnica que pueda encapsular los aspectos más destacados ruciales de los contratos, convenios y proyectos gestionados por la VT. Esto abarca desde estudios previos, sondeos de mercado, comunicaciones internas, minutas, actas de inicio, actas de finalización, actas de liquidación, otrosí, actas de entregas, hasta productos en formatos no técnicos como informes finales, informes de actividades y supervisión. No obstante, es crucial señalar que, dada la disposición actual de la información, la cual mayormente carece de estructura, se torna imperativo identificar y ubicar la información definitiva y esencial almacenada en nuestro directorio colaborativo. Esto se debe a la posibilidad de que existan versiones preliminares de cada uno de los documentos mencionados anteriormente.”*

2. ¿Cuáles son los criterios de búsqueda más importantes para los usuarios? ¿Palabras clave, fechas, tipos de documentos, temas específicos relacionados con la transición energética?

   Respuesta: Esta pregunta se encuentra interconectada con las preguntas 3 y 5 del apartado "Preguntas para la construcción de un modelo ML o LLM que nos permita la clasificación de documentos".

   Ahora bien, las principales búsquedas que se desarrollan en la VT corresponden a objetos, alcance, fechas, palabras claves temáticas, áreas de estudio, ubicación geográfica, actas o documentos precontractuales-contractuales-postcontractuales, temáticas abordadas, metodología, conclusiones, productos, entre otros aspectos técnicos, de los contratos, convenios y proyectos desarrollados por la VT. De igual forma, como se especifica en las actividades 4 y 5, las herramientas y/o soluciones pueden ser probadas y ajustadas, incluyendo búsquedas específicas por usuarios funcionales finales, para ajustar o incorporar nuevos criterios de búsqueda, metadatos, características y/o etiquetas en los algoritmos o modelos empleados para mejorar el rendimiento de las herramientas y/o soluciones.

3. ¿Qué tecnologías de indexación y búsqueda se consideran para implementar la solución? ¿Motor de búsqueda convencional, sistemas de recuperación de información basados en inteligencia artificial, tecnologías de procesamiento del lenguaje natural, etc.?

   Respuesta: La elección de las tecnologías de indexación y búsqueda de información recae en los interesados, quienes deben considerar la necesidad y los requerimientos técnicos específicos de la VT, las actividades a ejecutar que se encuentran especificadas en el sondeo de mercado y los aspectos técnicos.

4. ¿Cómo se adquirirán los documentos para su indexación? ¿Se utilizarán rastreadores web, sistemas de gestión de documentos, repositorios digitales, etc.?

   Respuesta: Esta pregunta se encuentra interconectada a la pregunta 4 del apartado "Preguntas para la construcción de un modelo ML o LLM que nos permita la clasificación de documentos"

5. ¿Qué métodos se utilizarán para procesar y normalizar los documentos antes de la indexación? ¿Extracción de texto, eliminación de metadatos irrelevantes, detección de idiomas, etc.?

   Respuesta: Al igual que se indicó en la pregunta 7 del apartado "Preguntas para la construcción de un modelo ML o LLM que nos permita la clasificación de documentos", la elección de los modelos y su aprendizaje, el preprocesamiento de la información y la elección de los metadatos a extraer durante la indexación de la información, recae en los interesados, quienes deben considerar la necesidad y los requerimientos técnicos específicos de la VT, las actividades a ejecutar que se encuentran especificadas en el sondeo de mercado y los aspectos técnicos de los modelos.

6. ¿Cómo se estructurará y organizará la información indexada? ¿Por categorías, etiquetas, fechas, o algún otro criterio?

Respuesta: La información indexada será el insumo principal para las búsquedas de información dentro del proyecto. La forma en la que esta información se gestionará y accederá por el buscador hará parte de la propuesta desarrollada por los interesados.

Sin embargo, como se indica en el sondeo, los metadatos resultantes de la indexación son uno de los productos esperados del proyecto. Por lo tanto, los interesados también deberán proponer cómo serán se facilitarán estos metadatos.

7. ¿Qué nivel de granularidad se aplicará en la indexación? ¿Se indexarán documentos completos o partes específicas dentro de los documentos?

Respuesta: Según la necesidad planteada en la organización, la granularidad en la indexación que se espera corresponde a la más detallada posible, que capture la mayor cantidad de información posible para que las búsquedas, acceso y consultas sean más eficientes.

8. ¿Qué métricas se utilizarán para evaluar la eficiencia de la búsqueda? ¿Tiempo de respuesta, precisión, cobertura, relevancia de los resultados, etc.?

Respuesta: Las métricas para evaluar la eficiencia de la herramienta y/o solución corresponden a objetivos y criterios claros y medibles de éxito. Estos como mínimo deben evaluar la efectividad del indexador y buscador en términos de eficiencia de almacenamiento (tamaño total de los datos almacenados versus espacio físico disponible), rendimiento del acceso a los datos (velocidad y capacidad de respuesta al realizar consultas, recuperación y carga de información asociada al resultado de la búsqueda), tasa de transferencia de datos, velocidad de acceso de datos, capacidad de procesamiento adicional (indexación), capacidad de manejo de cargas de trabajo variables y simultaneas (indexación y búsqueda), capacidad de recepción de nuevos usuarios e información, disponibilidad y fiabilidad de los resultados de consultas, capacidad de mantener la integridad y consistencia de datos, tasa de error de datos, falla de software y hardware, auditoria de seguridad, cumplimiento de regulaciones de privacidad y protección de datos, costo total de propiedad del indexador y buscador (incluyendo adquisición, implementación, mantenimiento y operación del sistema), entre otros. Sin embargo, los interesados pueden proponer nuevas métricas, así como considerar que algunas de las mencionadas anteriormente no son representativas.

9. ¿Cómo se asegurará la seguridad y privacidad de la información indexada? ¿Se aplicarán medidas de cifrado, control de acceso, anonimización, etc.

Respuesta: Esta pregunta se encuentra interconectada a la pregunta 11 del apartado “Preguntas para la construcción de un modelo ML o LLM que nos permita la clasificación de documentos”

10. ¿Cómo se manejará la escalabilidad y mantenimiento del sistema a medida que la cantidad de documentos crezca? ¿Se implementarán técnicas de distribución y paralelización?

Respuesta: Según lo indicado en el apartado “ASPECTOS TÉCNICOS Y ACTIVIDADES A EJECUTAR”, el software/herramienta/aplicación de indexación y búsqueda debe garantizar la posibilidad de personalización y escalabilidad, bien sea a mayores volúmenes de datos o

distintos tipos de información. No obstante, la elección de las técnicas para garantizar estas capacidades corresponde a una responsabilidad de los interesados y debe ser incluida en la propuesta de solución.

11. ¿Se considerará la expansión futura de la solución para incluir más tipos de documentos o fuentes de información? ¿Se planificará una arquitectura flexible y modular?

Respuesta: Según lo indicado en la actividad 3 del apartado “ACTIVIDADES”, el software/herramienta/aplicación debe ser capaz de escalarse o expandirse, la manera en que estas capacidades sean desarrolladas será elección de los interesados. No obstante, deben ser sustentadas por diagramas de desligues y arquitecturas de despliegues, al igual que otros documentos relevantes.

Preguntas para la experiencia

¿La experiencia de integrantes del equipo de trabajo en el sector Oil & Gas es válida para el requisito de la validación de la experiencia?

Respuesta: Se recomienda la participación de al menos un profesional en geología o ingeniería geológica debido a la naturaleza técnica de los datos e información relacionados con el proyecto. Para los demás perfiles del personal mínimo, se recomienda consultar la tabla "PERSONAL MÍNIMO" que se encuentra disponible en el sondeo de mercado. La tabla detalla la experiencia específica y las habilidades requeridas para cada perfil.