Solicitud de propuestas: Procesamiento del Lenguaje Natural (PLN/NLP) 2024

Lacuna Fund: Our Voice in Data

27 de junio de 2024

Índice

1: INTRODUCCIÓN 3

INFORMACIÓN GENERAL Y PROPÓSITO DE LACUNA FUND 3
FILOSOFÍA DE OTORGAMIENTO DE SUBVENCIONES 3

2: INFORMACIÓN GENERAL 3

ELEGIBILIDAD DE LA ORGANIZACIÓN 3
PROCESO DE SELECCIÓN Y CRITERIOS DE EVALUACIÓN 4
CRONOGRAMA 6

3: PROPÓSITO Y NECESIDAD 6

4: INFORMACIÓN DE LA PROPUESTA 9

INFORMACIÓN DE LOS SOLICITANTES 9
DESCRIPCIÓN DE LA PROPUESTA 9
CRONOGRAMA Y ENTREGAS DEL PROYECTO 12
PRESUPUESTO 13

1: Introducción

Información general y propósito de Lacuna Fund

Lacuna Fund apoya la creación, expansión y mantenimiento de conjuntos de datos que permiten la sólida aplicación de herramientas de aprendizaje automático (ML, por sus siglas en inglés) de gran valor social en contextos de ingresos bajos y medios a nivel mundial.

El objetivo del Fondo es:

- Desembolsar fondos a instituciones para crear, expandir o mantener conjuntos de datos que llenen brechas y reduzcan los sesgos en los datos etiquetados utilizados para la capacitación o evaluación de modelos de aprendizaje automático.
- Hacer posible que las poblaciones desatendidas aprovechen los avances que ofrece la IA.
- Profundizar la comprensión por parte de las comunidades filantrópicas y de aprendizaje automático sobre cómo financiar el desarrollo y el mantenimiento de conjuntos de datos etiquetados equitativamente de la manera más efectiva y eficiente.

Filosofía de otorgamiento de subvenciones

Lacuna Fund valora el enfoque colaborativo e impulsado localmente para la creación, expansión y mantenimiento de datos. Reconocemos que la utilidad y el mantenimiento continuos de los datos abiertos derivan de una comunidad que invierte en ellos. También vemos la colaboración como una forma de mejorar la calidad y el impacto de los conjuntos de datos, así como de promover una cultura de cooperación en este campo.

Lacuna Fund espera financiar conjuntos de datos que contribuyan a múltiples aplicaciones de gran valor social, ya sea a través de la investigación, la innovación comercial o la mejora de los servicios del sector público. **En tanto la sección 3: Propósito y Necesidad establece las necesidades identificadas por el Panel Asesor Técnico (TAP, por sus siglas en inglés), Lacuna Fund da la bienvenida a las ideas innovadoras dentro del área de dominio que tienen un beneficio claramente articulado y alineado con los criterios de selección que se describen a continuación.**

Esta convocatoria de propuestas cuenta con el apoyo de Google.org.

2: Información general

Elegibilidad de la organización

Lacuna Fund tiene como objetivo hacer que su financiación sea accesible para tantas organizaciones como sea posible en el espacio AI para el bien social y cultivar la capacidad y las organizaciones emergentes en este ámbito.

Para ser elegible para la financiación, las organizaciones deben:

- Ser una entidad sin fines de lucro, una institución de investigación, una empresa social con fines de lucro o un equipo perteneciente a estas organizaciones. Las personas deben presentar su solicitud a través de un patrocinador institucional. Se recomiendan las asociaciones como forma de reforzar la colaboración y maximizar los beneficios derivados del uso de los conjuntos de datos, pero solo el solicitante principal recibirá los fondos.
- Tener una misión que apoye el bien social, ampliamente definida.
- Tener su sede en el país o región donde se recopilarán los datos. El ámbito geográfico de esta convocatoria es África y América Latina. Las instituciones con sede en otros países o regiones pueden presentar su candidatura como socios de la institución principal. Como ya se ha indicado, solo el solicitante principal recibirá fondos.
- Tener todas las aprobaciones nacionales o de otro tipo que sean necesarias para realizar la investigación propuesta. El proceso de aprobación puede llevarse a cabo en paralelo con la solicitud de subvención, si es necesario. Los costos de aprobación, si los hubiere, son responsabilidad del solicitante.
- Tener la capacidad técnica, o la capacidad de desarrollar esta capacidad a través de la asociación descripta en la propuesta, para llevar a cabo el etiquetado, la creación, la agregación, la expansión o el mantenimiento de conjuntos de datos, incluida la capacidad de aplicar las mejores prácticas y los estándares establecidos en el ámbito específico (por ejemplo, el procesamiento del lenguaje natural) para permitir que múltiples entidades realicen análisis de IA/ML de gran calidad.

Proceso de selección y criterios de evaluación

Lacuna Fund busca propuestas para crear, ampliar, agregar y/o desbloquear conjuntos de datos para aplicaciones de aprendizaje automático que permitirán resultados equitativos en materia de procesamiento del lenguaje natural (PLN) en África y América Latina. Lacuna Fund y sus socios realizarán una evaluación inicial de la propuesta sobre la elegibilidad y viabilidad de la organización. Después de la evaluación inicial, el Panel Asesor Técnico de expertos en el ámbito, usuarios de datos y partes interesadas evaluarán las propuestas en función de los criterios de selección que se describen a continuación. Los miembros del Panel Asesor Técnico no pueden presentar una propuesta en respuesta a una RFP para la cual son revisores (consulte la Política de Conflicto de Intereses de Lacuna Fund).

El Panel Asesor Técnico para esta convocatoria revisará las presentaciones y seleccionará un conjunto de propuestas para ser financiadas. Las selecciones se basarán en el grado en que las propuestas completas cumplan con los siguientes criterios, basados en los principios que guían las operaciones de Lacuna Fund:

- **Calidad:** La organización o el equipo que propone el proyecto incluye expertos calificados en: a) el área de interés; b) aprendizaje automático; y c) gestión de datos. Se fomenta la colaboración con organismos gubernamentales y grupos comunitarios. El equipo presenta casos de uso claros para el conjunto de datos. El proponente sitúa el conjunto de datos propuesto dentro de los recursos existentes (o la falta de recursos) en el dominio y propone utilizar técnicas y herramientas eficaces de recopilación y etiquetado de datos para acelerar la recopilación, limpieza y puesta en común de los datos.

- **Impacto transformacional:** el proyecto hace que los conjuntos de datos existentes sean más representativos, inclusivos y/o sostenibles o crea un nuevo conjunto de datos etiquetados de gran valor para una población o problema desatendidos. Una propuesta puede ser considerada transformacional si tiene el potencial de abordar un problema particularmente importante u oportuno con equidad en ML/AI o tiene un alcance significativo en términos de número de personas desatendidas o geografías afectadas.
- **Equidad:** el equipo establece el problema de equidad que propone abordar y describe cómo el conjunto de datos llenará las brechas y hará que el dominio sea más representativo y equitativo. Adición de NLP: El conjunto de datos propuesto representa, o conduce a aplicaciones que representan, poblaciones desatendidas por tecnologías lingüísticas.
- **Enfoque participativo:** el equipo tiene su sede en el área geográfica donde se recopilarán los datos o tiene una asociación sustancial con una institución con sede en la región para garantizar el mantenimiento y uso sostenidos del conjunto de datos por parte de la comunidad local. Los socios del país participan en los elementos estratégicos del proyecto (más allá de la recopilación de datos). La propuesta describe cómo el equipo involucrará a las partes interesadas afectadas, buscará el consentimiento informado para la recopilación y el uso de datos y compartirá los resultados del proyecto con los proveedores de datos o con la comunidad.
- **Ética:** el proyecto tiene un plan para abordar y puede pasar una evaluación ética (por ejemplo, una junta de revisión institucional) que investiga: a) inquietudes sobre la privacidad, b) potencial de mal uso posterior, c) posibles vectores de discriminación (por ejemplo, género), y d) condiciones de trabajo justas y equitativas, si en el proyecto participan etiquetadores pagos. Los objetivos y la metodología propuestos para el proyecto son imparciales y éticos.
- **Sostenibilidad y comunicación:** el proyecto tiene un plan para garantizar la sostenibilidad y el mantenimiento futuro del conjunto de datos, por ejemplo, a través de un modelo de referencia, aplicaciones de ML resultantes, por una comunidad dedicada a ello o un grupo de partes interesadas (con o sin fines de lucro), un modelo de administración sólido para el conjunto de datos abierto y posibles casos de uso del aprendizaje automático para el conjunto de datos. El plan puede incluir quién actualizará y gestionará el conjunto de datos; posibles fuentes de financiación; estrategias de compromiso propuestas para las poblaciones afectadas y los usuarios de los datos; planes para presentar el conjunto o conjuntos de datos en una conferencia o conferencias; organización de un taller sobre sostenibilidad del conjunto de datos con las partes interesadas; o creación de un comité de sostenibilidad, así como medidas para mantener los datos abiertos y accesibles.
- **Viabilidad:** el proyecto es viable en relación con el presupuesto y el alcance del trabajo propuesto.
- **Accesibilidad:** el conjunto de datos será accesible bajo licencias de código abierto, o si esto no es posible, se justifica de manera convincente el uso de licencias más restrictivas para proteger la privacidad o evitar daños. El subcesionario dará prioridad a la divulgación de la propiedad intelectual bajo una estructura de licencias de código abierto permisiva, como Apache 2.0 para cualquier código u otras invenciones, o CC-BY 4.0 Internacional, o CC BY-SA 4.0 para cualquier otra propiedad intelectual (por ejemplo, obras creativas que no sean código, o patentables). La documentación y el alojamiento propuestos se ajustan a la Guía para el alojamiento y la documentación de conjuntos de datos de Lacuna Fund.

Cronograma

Se abre solicitud de propuestas	27 de junio de 2024
Seminario web para solicitantes	9 de julio de 2024
Plazo para preguntas y respuestas Envíe las preguntas a secretariat@lacunafund.org	12 de julio de 2024
Fecha límite para solicitar la tutoría	15 de julio de 2024
Respuestas publicadas	29 de julio de 2024
Límite para propuestas completas	23 de agosto de 2024

Período de preguntas y respuestas: Todas las preguntas relacionadas con la RFP deben enviarse a secretariat@lacunafund.org con el asunto "NLP 2024 RFP Question (Pregunta RFP 2024 sobre PLN/NLP)". Las preguntas enviadas antes del 12 de julio serán anonimizadas y respondidas públicamente el 29 de julio en el sitio web de Lacuna Fund, en un documento publicado en la sección "Postularse" y compartido con todos los solicitantes por correo electrónico.

3: Propósito y necesidad

Propósito

El objetivo de esta convocatoria de propuestas es apoyar los esfuerzos para desarrollar conjuntos de datos abiertos y accesibles para aplicaciones de aprendizaje automático relacionadas con el Procesamiento del Lenguaje Natural (Natural Language Processing, NLP) para idiomas y culturas de bajos recursos en África y América Latina.

La capacidad de comunicarse y hacerse entender en la propia variedad lingüística y en el contexto cultural es fundamental para la inclusión digital y social. Las técnicas de Procesamiento del Lenguaje Natural tienen el potencial de integrar las aplicaciones de IA para facilitar la inclusión digital y las mejoras en la educación, las finanzas, la sanidad, la agricultura, la comunicación y las respuestas a los peligros naturales, entre otras cuestiones. Muchos avances en el NLP, tanto fundamental como aplicado, derivan de conjuntos de datos con licencia abierta y de dominio público.

Sin embargo, estos conjuntos de datos son escasos o inexistentes para muchas lenguas africanas y latinoamericanas, lo que excluye a estas poblaciones de los beneficios de dicho sistema. Muchos modelos actuales de aprendizaje automático (AA) reciben información mediante conjuntos de datos anglocéntricos o traducidos, por lo que carecen de matices culturalmente relevantes y crean modelos sesgados o inutilizables para las comunidades de África y América Latina. Cuando se encuentran conjuntos de datos pertinentes, a menudo se basan en textos religiosos o judiciales del pasado, lo que da lugar a un lenguaje anticuado e igualmente sesgado. Se necesitan conjuntos de datos de dominio público que faciliten el uso de las tecnologías de NLP para los idiomas de bajos recursos de África y América Latina y que apoyen el desarrollo de conjuntos de datos lingüísticos sólidos y culturalmente apropiados que satisfagan las necesidades específicas de las comunidades subrepresentadas.

Necesidad

Lacuna Fund busca propuestas de equipos calificados y multidisciplinares para desarrollar conjuntos de datos abiertos y accesibles de entrenamiento y evaluación para aplicaciones de aprendizaje automático para el NLP en idiomas de bajos recursos y culturas subrepresentadas en África y América Latina.

Las propuestas pueden incluir, entre otras cosas, lo siguiente:

- Recopilar o anotar datos nuevos.
- Anotar o publicar datos existentes.
- Aumentar los conjuntos de datos existentes procedentes de diversas fuentes para suplir las brechas en los datos fidedignos de referencia, reducir los sesgos (como los geográficos, los de género u otros tipos de sesgo o discriminación) o aumentar el uso de los datos y la tecnología relacionados con el NLP en contextos de ingresos bajos y medios.
- Vincular y armonizar los conjuntos de datos existentes (por ejemplo, entre regiones, tiempo, variedades lingüísticas, así como los conjuntos de datos específicos de un ámbito, como datos históricos, sanitarios y educativos).

Si bien Lacuna Fund se centra principalmente en la creación, la expansión y el mantenimiento de conjuntos de datos, las propuestas pueden incluir el desarrollo de modelos de referencia para garantizar la calidad del conjunto de datos financiado o facilitar el uso del conjunto de datos para aplicaciones socialmente beneficiosas.

El Panel Asesor Técnico (Technical Advisory Panel, TAP) ve la necesidad de disponer de conjuntos de datos de entrenamiento y evaluación que tengan en cuenta la diversidad lingüística y los matices culturales de África y América Latina. Esto incluye conjuntos de datos sobre jerga regional, expresiones idiomáticas, variedades lingüísticas locales o dialectos y datos culturalmente relevantes. Estos conjuntos de datos son cruciales para desarrollar herramientas de Procesamiento del Lenguaje Natural más inclusivas y eficaces que puedan satisfacer las necesidades únicas de las comunidades lingüísticas culturalmente diversas.

Buscamos conjuntos de datos identificados por expertos locales y diseñados para tratar las necesidades identificadas localmente. Los siguientes ejemplos son solo a modo ilustrativo. **Los conjuntos de datos pueden incluir, entre otras cosas, lo siguiente:**

- **Conjuntos de datos etiquetados y no etiquetados para tareas del NLP de bajos recursos,** que apoyan el desarrollo de modelos de aprendizaje automático precisos y eficaces. Las tareas derivadas de los conjuntos de datos etiquetados podrían incluir, entre otras cosas, la respuesta a preguntas e IA conversacional, los conjuntos de datos de análisis de sentimientos, la detección de prejuicios sociales, la detección y contrarrespuesta a los discursos de odio, la detección de información errónea y desinformación; el resumen automático de textos u otras tareas de comprensión y generación de lenguaje natural, o recursos para apoyar la educación en el NLP en colaboración con las comunidades. Los conjuntos de datos no etiquetados incluyen un corpus de texto que puede utilizarse para el entrenamiento y la evaluación de los modelos del habla.
- **Corpus de habla,** incluidos los conjuntos de datos que permiten el reconocimiento automático del habla (Automatic Speech Recognition, ASR) para que grupos de personas analfabetas o desfavorecidas accedan a información o servicios en idiomas de bajos recursos.
- **Conjuntos de datos para tareas de generación de texto,** en especial aquellas distintas de la traducción automática.
- **Conjuntos de datos multimodales y otros conjuntos innovadores,** como subtítulos de video o audio, respuestas visuales a preguntas u otras interacciones imagen-texto.
- **Conjuntos de datos para tareas basadas en conocimiento,** como la garantía de calidad (Quality Assurance, QA) y la generación aumentada de recuperación (Retrieval Augmented Generation, RAG).
- **Conjuntos de datos relacionados con el corpus de variación dialectal y el cambio de código de texto y habla,** incluida la captura de variaciones lingüísticas (jerga regional, expresiones idiomáticas, datos culturalmente relevantes) en idiomas de bajos recursos, pero ricos en dialectos y en comunidades lingüísticas en las que es habitual el cambio de código.
- **Creación o expansión de conjuntos de datos de texto y habla específicos de un dominio,** como sanidad, toponimia, agricultura o educación, que permitan aplicaciones con un impacto social significativo. Exploración de marco teórico para el aumento generativo de datos con el fin de incluir vocabulario, semántica, morfología y sintaxis especializados en el ámbito.
- **Conjuntos de datos para el aprendizaje automático aplicado a la lingüística,** para la preservación y revitalización de las culturas marginadas y los aspectos de los idiomas subrepresentados que estas culturas consideran importantes para su salud, dignidad, medio ambiente y bienestar. Estos conjuntos de datos pueden incluir anotaciones fonéticas, morfológicas y sintácticas, así como herramientas automatizadas para realizar estas tareas si así lo solicita el grupo social implicado.
- **Sensibilidad a las cuestiones de género e inclusión de los principales grupos vulnerables para todos los conjuntos de datos**, incluida la mitigación de sesgos para quienes viven en contextos humanitarios y de conflicto, así como quienes se encuentran en las intersecciones de más de un grupo socioeconómico (por ejemplo, discapacidad, género, edad, minorías). Consulte el apartado “Riesgos, incluidos los éticos y de privacidad” de la sección Descripción de la propuesta de este documento y tenga en cuenta los aspectos éticos de la recopilación de datos.

Puede consultar los conjuntos de datos de los proyectos seleccionados en la Solicitud de Propuestas para el NLP de Lacuna Fund de 2020 y 2021 para ver qué trabajos se están llevando a cabo actualmente.

4: Información de la propuesta

Nota: El sitio web de Lacuna Fund incluye varios recursos, como referencias relevantes sobre la calidad de los datos, la documentación y el formato, para ayudar a los solicitantes a preparar una solicitud competitiva.

Las presentaciones de propuestas solo se aceptarán a través del portal de solicitud de SurveyMonkey Apply disponible en https://lacunafund.org/es/postularse/. Las solicitudes pueden presentarse en inglés, español, francés y portugués. Una descripción de las preguntas de la solicitud se encuentra disponible a continuación solo a modo de información. Las siguientes secciones son obligatorias:

- Información del solicitante (disponible en el portal)
- Descripción de la propuesta
- Cronograma y presupuesto

Información de los solicitantes

En esta sección, el candidato debe proporcionar:

- Un resumen de propuesta de 200-250 palabras;
- Información sobre la(s) institución(es) o equipo participante;
- Dónde se llevará a cabo el trabajo;
- CV de los miembros clave del equipo.
- Los procesos de revisión ética de las instituciones afiliadas;
- La capacidad del equipo para obtener aprobaciones nacionales.

Descripción de la propuesta

Limite la descripción de su propuesta a 10 páginas sin incluir las referencias, con márgenes de 2,5 cm y una fuente de tamaño mínimo 11. No se revisarán los anexos o el material descriptivo de las propuestas de más de 10 páginas.

En esta sección, el solicitante deberá cargar una narrativa cohesiva, en formato PDF o Word, que aborde lo siguiente:

- **Capacidad:** describa la(s) organización(es) o asociación(es) que participa(n), cómo satisfacen los criterios de elegibilidad expuestos anteriormente y sus capacidades únicas para realizar el trabajo propuesto.

- **Identificación del problema y solución propuesta y conjunto de datos:** describa el problema o la brecha en los datos de capacitación o evaluación y la solución propuesta. Resuma los conjuntos de datos que pretende crear, aumentar, agregar o mantener. *Explique cómo su proyecto aborda una brecha y complementa el trabajo existente.*
- **Especificaciones que deben entregarse para los datos y la documentación propuestos:**
  - Cantidad de datos que se incluirán en el conjunto de datos.
  - Tipos y formato de datos o etiquetas, así como marco y tamaño de la muestra o un plan para asegurar la representación, si corresponde.
  - Métricas que se utilizarán para evaluar los resultados deseados de la creación de datos (p. ej., métricas de imparcialidad en el conjunto de datos, controles de calidad con un parámetro de referencia, etc.)**.**
- **Potenciales beneficiarios y casos de uso:** describa la consulta revisa o la colaboración propuesta con los potenciales beneficiarios y describa los posibles casos de uso actuales y futuros para los conjuntos de datos propuestos. Explique cómo el conjunto de datos respeta y refleja la diversidad de las comunidades que representa. Indique cómo la calidad propuesta, los métodos de recopilación y otra información hacen que los datos sean adecuados para su uso en ese contexto operativo en particular.
- **Metodología:** proporcione una breve descripción general de los pasos propuestos (y los supuestos clave) para desarrollar e implementar el proyecto. Incluya:
  - Propuesta de técnicas de recopilación y etiquetado de datos e información sobre interoperabilidad. Considere la infraestructura existente o común y las últimas técnicas y herramientas para acelerar la recopilación, limpieza y uso compartido de datos.
  - Medidas de control de calidad, como la calidad que deben cumplir todas las muestras de datos para el conjunto de datos. Indique cómo prevé el equipo abordar los valores atípicos que puedan afectar a la calidad del conjunto de datos.
  - Un plan para evaluar y mitigar errores y sesgos (p. ej. prejuicios de género u otros).
  - Cómo piensa aprovechar los recursos existentes, incluidos los métodos o tecnologías de recopilación, la vinculación de conjuntos de datos preexistentes en todo el ámbito, así como los recursos existentes en otros contextos.
  - Permisos vigentes o pasos que llevará a cabo para obtener las aprobaciones nacionales u otras que sean necesarias. Considere qué jurisdicciones requieren aprobaciones y si la investigación propuesta cumple con la definición de investigación en esa jurisdicción. Si determina que no se requieren aprobaciones locales, nacionales o regionales, explique por qué no.
  - También cualquier desafío o incertidumbre prevista en la recopilación de datos y medidas compensatorias propuestas.
- **Impacto transformacional:** explique cómo el etiquetado o el conjunto de datos propuestos contribuirán a lograr el impacto deseado. Si corresponde, describa cómo los productos podrían crear caminos múltiples y duraderos en la investigación o aplicación comercial. Tenga en cuenta las limitaciones prácticas que pueda enfrentar (por ejemplo, el acceso a Internet).
- **Administración de datos y licencias**. Describa:

- Cualquier problema previsto relacionado con los derechos de autor de los datos de origen y la colaboración con el titular de los derechos de autor. Cualquier problema previsto relacionado con los derechos de autor y la concesión de licencias de datos secundarios.
- Planes de concesión de licencias para maximizar el posterior uso responsable. Según la Política de propiedad intelectual (IP) de Lacuna Fund, el conjunto de datos y cualquier propiedad intelectual relacionada, como métodos de recopilación, hojas de datos, cómo cargar o leer conjuntos de datos u otra información para garantizar el uso, debe estar disponible bajo una licencia de código abierto por atribución (CC-BY 4.0 o CC BY-SA 4.0). Si se proponen licencias más restrictivas, proporcione una justificación. El presupuesto puede incluir recursos para la licencia.
- Si tiene la intención de utilizar un conjunto de datos existente para su proyecto, indique que su equipo ha recibido los permisos necesarios del propietario del conjunto de datos para que el conjunto de datos se pueda publicar de acuerdo con la Política de propiedad intelectual (IP) de Lacuna Fund, o proporcione justificación para otra forma de licencia. Las cartas de respaldo por parte de los titulares de datos existentes son opcionales, pero recomendadas.
- Planes para incluir un archivo de metadatos y una hoja de datos como documentación para su conjunto de datos, de acuerdo con la "Dataset Hosting and Documentation Guidance" (Guía para el alojamiento y la documentación de conjuntos de datos) de Lacuna Fund.
- La plataforma de alojamiento que piensa utilizar. Las plataformas de alojamiento deben asignar un identificador de objeto digital (DOI) al conjunto de datos, cuantificar las descargas del conjunto de datos y recopilar información de contacto para las descargas del conjunto de datos. Para más información sobre las plataformas de alojamiento sugeridas, consulte la "Dataset Hosting and Documentation Guidance" (Guía para el alojamiento y la documentación de conjuntos de datos).

- **Riesgos, incluidos la ética y la privacidad:** identifique los riesgos potenciales, como los posibles problemas de privacidad y éticos, y describa los pasos que tomará para mitigarlos. Específicamente:
  - Una declaración reflexiva sobre las comunidades de las que procedes y las identidades que posees, y cómo pueden influir en su trabajo.
  - Indique cómo garantizará el consentimiento informado, si corresponde. (Esto debe incluir la notificación de posibles casos de uso futuros para los datos).
  - Describa cómo garantizará la equidad en el proyecto, lo que incluye un pago justo por el etiquetado anotación, suministro de datos y otros servicios de AI. Algunas opciones para dar prioridad a las normas de trabajo justas y a las prácticas de remuneración justas son:
    - Cumplir y firmar el compromiso de normas de trabajo justo para el trabajo de AI cuando la anotación y el suministro de datos se subcontratan a entidades comerciales.
    - Adherirse a un enfoque de anotación y suministro de datos orientado a la comunidad, si existe tal modelo.
    - Presentar una renuncia/encuesta voluntaria a los participantes en el proyecto para abordar las cargas indebidas.
  - Describa cómo se incorporan la diversidad de género y otras consideraciones demográficas en el equipo del proyecto, la recopilación de datos de formación y el desarrollo de modelos,

con el fin de que los conjuntos de datos representen con exactitud los impactos en las diferentes comunidades/grupos.

- Presente un plan para preservar el anonimato de la información de identificación personal (PII) y el cumplimiento de las leyes de privacidad, si corresponde. Si no se dispone de un marco legal nacional, la propuesta debe describir o hacer referencia a las prácticas recomendadas. Incorpore consideraciones de privacidad tanto a nivel individual como comunitario. Consulte la "Data Anonymization Guide" (Guía de anonimización de datos) de Lacuna Fund para obtener sugerencias.
- Analice los posibles impactos adversos en la producción y el uso del conjunto de datos y los pasos para mitigarlos, incluidos los posibles riesgos para los derechos humanos y el potencial de alto consumo de energía en la tecnología de IA, lo que lleva a un aumento de las emisiones de dióxido de carbono.

- **Plan de sostenibilidad y comunicación**
  - Describa cómo se mantendrá o ampliará el conjunto de datos etiquetado más allá de la financiación inicial (p. ej., a través de un modelo de referencia, de las aplicaciones de ML resultantes, por una comunidad dedicada a ello o un grupo de partes interesadas con un sólido modelo de administración para el conjunto de datos abierto) y cómo un caso de uso potencial podría sostener el proyecto.
  - Explique cómo el conjunto de datos seguirá los principios de datos FAIR (https://www.go-fair.org/resources/faq/what-is-fair/). Indique los pasos que seguirá para garantizar que el conjunto de datos sea localizable, accesible, interoperable y reutilizable.
  - Describa las actividades de comunicación para dar a conocer el conjunto o conjuntos de datos. Puede tratarse de actividades de creación de redes con posibles usuarios de datos, la presentación de los conjuntos de datos en conferencias, la organización de un taller sobre sostenibilidad de los conjuntos de datos con las partes interesadas o la creación de un comité de sostenibilidad.

Cronograma y entregas del proyecto

En esta sección, se le pedirá al solicitante que presente una tabla con un cronograma para la finalización de las actividades principales y las entregas. El cronograma puede incluir la capacitación del personal, recopilación de datos, etiquetado, control de calidad, validación/limpieza y publicación de datos. Las entregas pueden incluir, entre otros, partes del conjunto de datos para la demostración del concepto, el conjunto de datos completo y la documentación que lo acompaña o los métodos de recopilación para que sean de código abierto.

Todos los plazos deben incluir la fecha en la que los datos estarán disponibles públicamente con toda la documentación.

Nota: Los proyectos propuestos deben completarse, los conjuntos de datos deben publicarse y los informes finales deben presentarse para octubre de 2026. A efectos de planificación, los acuerdos deben estar completos y el trabajo podría comenzar para abril de 2025.

Presupuesto

Proporcione un presupuesto para la realización del conjunto de datos propuesto enviado a través del portal SurveyMonkey Apply. Esto debe seguir el formato de la plantilla de presupuesto de Lacuna Fund, disponible en el portal del solicitante.

El fondo total disponible es de aproximadamente 1 millón de dólares estadounidenses. Nos gustaría financiar proyectos en cada una de las regiones objetivo África y América Latina y anticipamos financiar 6-8 proyectos con presupuestos inferiores a 100 mil dólares, y 2 o 3 proyectos más grandes y complejos con presupuestos oscilen entre 100 mil y 250 mil dólares. El Panel Asesor Técnico evaluará **la viabilidad y la idoneidad del presupuesto**, así como **la relación entre el presupuesto y la narrativa de la subvención** como parte de los criterios de selección. Los presupuestos pueden incluir, pero no limitarse a, costos para lo siguiente:

- creación de capacidad relacionada con la recopilación de datos y el control de calidad;
- recopilación de datos (incluida una compensación justa por el suministro de datos);
- etiquetado de datos (incluida una compensación justa por el etiquetado de los datos);
- control de calidad o verificación;
- posprocesamiento de datos;
- publicación de datos;
- licencia;
- publicación de resultados en acceso abierto;
- tiempo para preparar una declaración de datos para el conjunto de datos;
- iniciativas de participación colectiva, como label-a-thons;
- almacenamiento de datos;
- potencia de cálculo;
- taller;
- actividades de comunicación, incluida la asistencia a un máximo de dos conferencias para presentar los conjuntos de datos.

Los fondos no se pueden utilizar para el pago directo de ningún impuesto de aduana, importación u otros aranceles o impuestos recaudados con respecto a la importación de bienes o equipos a cualquier país o jurisdicción. **Las tasas indirectas están estrictamente limitadas al 14.5 % de los costos directos de investigación.**

Los socios de Lacuna Fund pueden ofrecer almacenamiento en la nube y potencia informática en especie. Si desea utilizar este recurso, inclúyalo en su presupuesto. Los equipos seleccionados recibirán instrucciones sobre cómo solicitarlo cuando reciban su premio.

Consulte la hoja de instrucciones en la plantilla de presupuesto para obtener más información sobre las pautas presupuestarias, incluida información sobre los costos de personal permitidos.

Gracias por su interés en Lacuna Fund y por sus iniciativas para lograr mayor equidad y accesibilidad para las aplicaciones de aprendizaje automático para apoyar el procesamiento del lenguaje natural. ¡Esperamos revisar su presentación!

NOTA: Oportunidad de tutoría y emparejamiento

Lacuna Fund se complace en asociarse con la Fundación de Investigación Masakhane (Masakhane Research Foundation, MRF) para ofrecer oportunidades de tutoría y emparejamiento a los solicitantes del NLP. La MRF es una organización comunitaria cuya misión es reforzar e impulsar la investigación en el NLP en idiomas africanos, destinada a africanos y realizada por africanos. El objetivo de la MRF es que los africanos definan y se apropien de estos avances tecnológicos en pos de la dignidad humana, el bienestar y la equidad, a través de la construcción de comunidades inclusivas, la investigación participativa abierta y la multidisciplinariedad.

Para esta convocatoria de propuestas de Lacuna Fund, la MRF ofrecerá a los solicitantes de África y América Latina una oportunidad especial para unirse a la comunidad de la MRF y poner en contacto a las personas interesadas con un mentor que revisará su borrador de propuesta y analizará las opciones para fortalecerla. Los interesados pueden solicitar una sesión con un mentor completando este formulario de Google con una breve descripción (resumen de 250 palabras) de la propuesta de conjunto de datos. Los alumnos pueden solicitar distintas formas de ayuda, como ***"examinar las deficiencias en el NLP de bajos recursos", "redactar una propuesta de investigación"*** y ***"preparar presupuestos".*** Proporcionar una descripción detallada ayudará a emparejar a los solicitantes con mentores afines a su área de interés. El emparejamiento de los mentores con los alumnos será gestionado por nuestros asociados de emparejamiento y tutoría, que organizarán varios talleres para reunir a los solicitantes que trabajen en campos afines (por ejemplo, lingüistas e investigadores del NLP) y coordinarán y organizarán sesiones de comunicación entre los grupos pertinentes. A través de este proceso, también ofreceremos oportunidades de emparejamiento para que los solicitantes colaboren entre sí para formar equipos de proyecto y presentar propuestas juntos.

Se anima a los candidatos interesados a solicitar una sesión de tutoría al menos 6 semanas antes de la fecha límite de presentación de propuestas, el **15 de julio de 2024**. Tomaremos todas las medidas posibles para encontrar un mentor para todos los que lo soliciten, pero no podemos garantizar que haya mentores disponibles para todos. Los mentores se asignarán por orden de llegada. Todos los solicitantes deben leer y respetar el *código ético y de conducta* del programa de tutoría.