Solicitação de Propostas: Processamento de Linguagem Natural (PLN/NLP): 2024

Lacuna Fund: Our Voice in Data

27 de junho de 2024

Índice

1 - Introdução

Visão geral e objetivo do Lacuna Fund

O Lacuna Fund apoia a criação, expansão e manutenção de conjuntos de dados que permitem a aplicação sólida e equitativa de ferramentas de aprendizagem automática (ML) de elevado valor social em contextos de baixo e médio rendimento em âmbito mundial.

O Fundo tem por objetivo:

- Atribuir fundos a instituições para criar, expandir e/ou manter conjuntos de dados que preencham as lacunas e reduzam os enviesamentos nos dados rotulados utilizados para a formação e/ou avaliação de modelos de aprendizagem automática.
- Permitir que as populações carenciadas se beneficiem dos avanços proporcionados pela IA.
- Aprofundar os conhecimentos das comunidades de aprendizagem automática e de filantropia sobre o modo de financiar o desenvolvimento e a manutenção de conjuntos de dados rotulados de forma equitativa da forma mais eficaz e eficiente.

Filosofia da concessão de subsídios

O Lacuna Fund valoriza uma abordagem colaborativa e orientada localmente para a criação, expansão e manutenção de dados. Reconhecemos que a utilidade e a manutenção contínuas dos dados abertos derivam de uma comunidade que investe nesses dados. Também vemos a colaboração como uma forma de melhorar a qualidade e o impacto dos conjuntos de dados, bem como de promover uma cultura de cooperação no domínio.

O Lacuna Fund espera financiar conjuntos de dados que contribuam para múltiplas aplicações de elevado valor social, seja através da investigação, da inovação comercial ou da melhoria dos serviços do sector público. **Enquanto a Secção 3: O Objetivo e a Necessidade estabelecem as necessidades identificadas pelo Painel Técnico Consultivo (Technical Advisory Panel, TAP). O Lacuna Fund acolhe ideias novas na área de domínio que tenham um benefício claramente articulado e alinhado com os critérios de seleção abaixo indicados.**

Este convite à apresentação de propostas é apoiado por Google.org.

2 - Visão geral

Elegibilidade da organização

O Lacuna Fund tem como objetivo tornar o seu financiamento acessível ao maior número possível de organizações no espaço da IA para o bem social e cultivar a capacidade e as organizações emergentes neste domínio.

Para serem elegíveis ao financiamento, as organizações devem:

- Ser uma entidade sem fins lucrativos, uma instituição de investigação, uma empresa social com fins lucrativos ou um grupo de tais organizações. Os indivíduos devem candidatar-se

por meio de um patrocinador institucional. As parcerias são grandemente incentivadas como forma de reforçar a colaboração e ampliar os benefícios derivados da utilização dos conjuntos de dados, mas apenas o candidato principal receberá fundos.

- Ter uma missão de apoio ao bem da sociedade, definida em termos gerais.
- Estar sediado no país ou região onde os dados serão coletados. O foco geográfico desta chamada é África e América Latina. As instituições sediadas noutros países ou regiões podem candidatar-se como parceiros da instituição principal. Tal como acima referido, apenas o candidato principal receberá fundos.
- Possuir todas as autorizações nacionais ou outras necessárias para efetuar a investigação proposta. O processo de aprovação pode ser conduzido em paralelo com o pedido de subsídio, se necessário. Os custos de aprovação, caso existam, são da responsabilidade do candidato.
- Possuir a capacidade técnica - ou a capacidade de desenvolver essa capacidade por meio de uma parceria descrita na proposta - para efetuar a rotulagem, criação, agregação, expansão e/ou manutenção de conjuntos de dados, incluindo a capacidade de aplicar as melhores práticas e as normas estabelecidas no domínio específico (por exemplo, Processamento de Linguagem Natural) para permitir que múltiplas entidades conduzam análises de IA/ML de elevada qualidade.

Processo de seleção e critérios de avaliação

O Lacuna Fund busca propostas para criar, expandir, agregar e/ou desbloquear conjuntos de dados para aplicações de aprendizagem automática que permitirão resultados equitativos de Processamento de Linguagem Natural na África e na América Latina. O Lacuna Fund e os seus parceiros farão uma análise inicial da proposta para verificar a sua elegibilidade e viabilidade organizacional. Após o exame inicial, um painel técnico consultivo de peritos no domínio, utilizadores de dados e partes interessadas avaliará as propostas com base nos critérios de seleção a seguir descritos. Os membros do Painel Técnico Consultivo não podem apresentar uma proposta em resposta a um RFP para o qual são revisores (ver Política de Conflito de Interesses do Lacuna Fund).

O Painel Técnico Consultivo do presente convite analisará as propostas apresentadas e escolherá um conjunto de propostas a financiar. As escolhas serão fundamentadas no grau em que as propostas completas satisfazem os seguintes critérios, que se baseiam nos princípios que orientam o funcionamento do Lacuna Fund:

- **Qualidade** - A organização ou o grupo que propõe o projeto inclui peritos qualificados em a) domínio de interesse; b) aprendizagem automática; e c) gestão de dados. São incentivadas as colaborações com agências governamentais e grupos comunitários. O grupo apresenta casos de utilização claros para o conjunto de dados. O proponente situa o conjunto de dados proposto no âmbito dos recursos existentes (ou da falta de recursos) no domínio e propõe a utilização de técnicas e ferramentas eficazes de coleta e rotulagem de dados para acelerar a coleta, a limpeza e a partilha de dados.
- **Impacto transformacional** - O projeto torna o(s) conjunto(s) de dados existente(s) mais representativo(s), inclusivo(s) e/ou sustentável(is) ou cria um conjunto de dados rotulado de elevado valor para uma população carenciada ou problema. Uma proposta pode ser considerada transformacional se tiver potencial para resolver um problema particularmente

importante ou oportuno de equidade no domínio do AM/IA ou se tiver um alcance significativo em termos de número de pessoas carenciadas ou de áreas geográficas atingidas.

- **Equidade** - O grupo indica a questão da equidade a que se propõe abordar e descreve como o conjunto de dados irá preencher as lacunas e tornar o domínio mais representativo e equitativo. Há uma teoria de mudança convincente que demonstra como o conjunto de dados criará um maior acesso aos benefícios do ML/AI para comunidades vulneráveis e carenciadas.
- **Abordagem participativa** - O grupo está sediado na área geográfica onde os dados serão coletados para garantir a manutenção e a utilização sustentadas do conjunto de dados pela comunidade local. Os parceiros nacionais estão envolvidos em elementos estratégicos do projeto (para além da coleta de dados). A proposta descreve a forma como o grupo irá envolver as partes interessadas atingidas, obter o consentimento informado para a coleta e utilização de dados e partilhar os resultados e benefícios do projeto com os fornecedores de dados e/ou a comunidade.
- **Ética** - O projeto tem um plano e é capaz de passar por um exame ético (por exemplo, um conselho de revisão institucional) que investiga: a) preocupações com a privacidade, b) potencial de utilização derivada indevida, c) possíveis vectores de discriminação (por exemplo, género) e d) condições de trabalho justas e equitativas, se estiverem envolvidos no projeto rotuladores pagos. As finalidades e a metodologia propostas para o projeto são imparciais e éticas.
- **Sustentabilidade** e comunicações - O projeto tem um plano para garantir a sustentabilidade e a manutenção futura do conjunto de dados, por exemplo, por meio de um modelo de base, aplicações de ML resultantes, por uma comunidade dedicada ou um conjunto de partes interessadas (com ou sem fins lucrativos), um modelo sólido de governação para o conjunto de dados aberto e possíveis casos de utilização da aprendizagem automática para o conjunto de dados. O plano pode incluir quem irá atualizar e gerir o conjunto de dados; potenciais fontes de financiamento; estratégias de envolvimento propostas para populações atingidas e utilizadores de dados; planos para apresentar o(s) conjunto(s) de dados numa conferência; organização de uma oficina sobre sustentabilidade do conjunto de dados com as partes interessadas; ou criação de um comité de sustentabilidade, bem como medidas para manter os dados abertos e acessíveis.
- **Viabilidade** - O projeto é viável em relação ao orçamento e ao âmbito do trabalho proposto.
- **Acessibilidade** - O conjunto de dados será amplamente acessível ao abrigo de licenças de fonte aberta ou, se tal não for possível, será apresentada uma justificação convincente para uma licença mais restritiva, a fim de proteger a privacidade ou evitar danos. O sub-beneficiário dará prioridade à liberação da propriedade intelectual ao abrigo de uma estrutura de licenciamento de fonte aberta permissiva , como Apache 2.0 para qualquer código ou outras invenções, ou CC-BY 4.0 International, ou CC BY-SA 4.0 para qualquer outra propriedade intelectual (por exemplo, trabalhos criativos que não sejam código ou patenteáveis). A documentação e o acolhimento propostos estão em conformidade com a finalidade do Lacuna Fund Orientações sobre armazenamento de conjuntos de dados e documentação.

Linha do tempo

Abertura do convite para apresentação de propostas	27 de junho de 2024
Webinário para candidatos	9 de julho de 2024
Prazo da pergunta Envie suas perguntas para secretariat@lacunafund.org	12 de julho de 2024
Prazo de mentoria	15 de julho de 2024
Respostas enviadas	29 de julho de 2024
Prazo para apresentação de propostas completas	23 de agosto de 2024

Período de perguntas e respostas: Todas as perguntas relacionadas com o RFP devem ser enviadas para secretariat@lacunafund.org com “NLP 2024 RFP Question (Processamento de Linguagem Natural RFP 2024 Pergunta)" na linha de assunto. As perguntas enviadas até 12 de julho serão desidentificadas e respondidas publicamente até 29 de julho no site do Lacuna Fund, em um documento publicado na página "Candidatura" e partilhado com todos os candidatos por correio eletrónico.

3 - Objetivo e necessidade

Objetivo

O objetivo desta chamada de propostas é apoiar os esforços para desenvolver conjuntos de dados abertos e acessíveis para aplicações de aprendizado de máquina relacionados ao Processamento de Linguagem Natural (PLN) para idiomas e culturas com baixos recursos na África e na América Latina.

A capacidade de se comunicar e de ser compreendido em sua própria variedade de idioma e contexto cultural é fundamental para a inclusão digital e social. As técnicas de processamento de linguagem natural têm o potencial de viabilizar aplicações de IA que facilitam a inclusão digital e melhorias em educação, finanças, saúde, agricultura, comunicação e respostas a riscos naturais, entre outros. Muitos avanços no PLN fundamental e aplicado resultaram de conjuntos de dados licenciados abertamente e disponíveis publicamente.

No entanto, esses conjuntos de dados são escassos ou inexistentes para muitos idiomas africanos e latino-americanos, excluindo essas populações dos benefícios do PLN. Muitos modelos atuais de aprendizado de máquina (ML) são informados por conjuntos de dados anglo-cêntricos e/ou traduzidos, sem nuances culturalmente relevantes e criando modelos tendenciosos ou inutilizáveis para comunidades na África e na América Latina. Quando existem conjuntos de dados relevantes, geralmente se baseiam em textos religiosos ou judiciários do passado, o que leva a uma linguagem desatualizada e tendenciosa. Há uma necessidade de conjuntos de dados de acesso aberto para facilitar as tecnologias de PLN para idiomas com baixos recursos na África e na América Latina e apoiar o desenvolvimento de conjuntos de

dados sólidos de idiomas e culturalmente apropriados que atendam às necessidades específicas de comunidades sub-representadas.

Necessidade

O Lacuna Fund busca propostas de equipes qualificadas e multidisciplinares para desenvolver conjuntos de dados de treinamento e avaliação abertos e acessíveis para aplicações de aprendizado de máquina para PLN em idiomas com baixos recursos e culturas sub-representadas na África e na América Latina.

As propostas podem incluir, mas não se limitam a:

- Coleta e/ou anotação de novos dados;
- Anotação ou liberação de dados existentes;
- Aumentar os conjuntos de dados existentes de diversas fontes para preencher lacunas nos dados locais de verdade, diminuir a parcialidade (como parcialidade geográfica, lacunas de gênero ou outros tipos de parcialidade ou discriminação) ou aumentar a usabilidade dos dados e da tecnologia relacionados ao PLN em contextos de baixa e média renda;
- Vincular e harmonizar conjuntos de dados existentes (tais como entre regiões, hora, variedades linguísticas, bem como conjuntos de dados específicos de domínio, como dados históricos, de saúde e educação).

Embora o foco do Lacuna Fund seja principalmente a criação, a expansão e a manutenção do conjunto de dados, as propostas podem incluir o desenvolvimento de um ou mais modelos básicos para garantir a qualidade do conjunto de dados financiado e/ou para facilitar o uso do conjunto de dados para aplicações socialmente benéficas.

A TAP (Technical Advisory Panel) enxerga a necessidade de conjuntos de dados de treinamento e avaliação que levem em conta a diversidade linguística e as nuances culturais da África e da América Latina. Isso inclui conjuntos de dados sobre gírias regionais, expressões idiomáticas, variedades linguísticas ou dialetos locais e dados culturalmente relevantes. Esses conjuntos de dados são essenciais para o desenvolvimento de ferramentas de processamento de linguagem natural mais inclusivas e eficazes, que possam atender às necessidades únicas de comunidades linguísticas culturalmente diversas.

Buscamos conjuntos de dados identificados por especialistas locais, projetados para atender às necessidades identificadas localmente. Os exemplos a seguir são apenas ilustrativos.

Os conjuntos de dados podem incluir, mas não se limitam ao seguinte:

- **Conjuntos de dados rotulados e não rotulados para tarefas de PLN com baixos recursos,** apoiando o desenvolvimento de modelos de aprendizado de máquina precisos e eficazes. As tarefas derivadas dos conjuntos de dados rotulados podem incluir, entre outras: respostas a perguntas e IA de conversação, conjuntos de dados de análise de sentimentos, detecção de preconceito social, detecção de discurso de ódio e contradiscurso, detecção de má informação e desinformação; resumo automático de texto ou outras tarefas de geração e compreensão de linguagem natural, ou recursos para apoiar o ensino de PLN em

colaboração com as comunidades. Os conjuntos de dados não rotulados incluem corpus de texto que podem ser usados para apoiar o treinamento e a avaliação de modelos de fala.

- **Corpus de fala,** incluindo conjuntos de dados para possibilitar o reconhecimento automático de fala (ASR), que permite que grupos de pessoas analfabetas ou desprivilegiadas acessem informações e/ou serviços em idiomas com baixos recursos.
- **Conjuntos de dados de tarefas de geração de texto,** especialmente outros que não a tradução automática.
- **Conjuntos de dados multimodais e outros inovadores,** como legendas de vídeo ou áudio, respostas visuais a perguntas ou outras interações de imagem-texto.
- **Conjuntos de dados que dão apoio a tarefas com uso intensivo de conhecimento,** tais como garantia de qualidade (QA) e Geração Aumentada de Recuperação (RAG).
- **Conjuntos de dados relacionados a corpus de variação dialetal e texto e fala com alternância de língua,** incluindo a captura de variações linguísticas (gírias regionais, expressões idiomáticas, dados culturalmente relevantes) em idiomas com baixos recursos e ricos em dialetos e em comunidades linguísticas nas quais a alternância de língua é comum.
- **Criação ou aumento de conjuntos de dados de texto e fala específicos de domínio,** tais como saúde, nomes de lugares, agricultura ou educação, que permitem aplicações com impacto social significativo. Exploração de estruturas de aumento de dados generativos para incluir vocabulário, semântica, morfologia e sintaxe especializados do domínio.
- **Conjuntos de dados que apoiam o aprendizado de máquina para linguística,** para a preservação e revitalização de culturas marginalizadas e aspectos de idiomas sub-representados que essas culturas consideram importantes para sua saúde, dignidade, meio ambiente e bem-estar. Esses conjuntos de dados podem incluir anotações fonéticas, morfológicas e sintáticas, além de ferramentas automatizadas para executar essas tarefas, se solicitadas pelo grupo social envolvido.
- **Em todos os conjuntos de dados: capacidade de resposta a questões de gênero e inclusão dos principais grupos vulneráveis,** incluindo mitigação de preconceito para aqueles que vivem em contextos humanitários e de conflito, bem como aqueles que se encontram nas interseções de mais de um grupo socioeconômico (p. ex., deficiência, gênero, idade, minorias). Consulte o parágrafo "Riscos, incluindo Ética e Privacidade" na seção de narrativa da Proposta deste documento e considere cuidadosamente a ética em torno da coleta de dados.

Você poderá analisar os conjuntos de dados dos projetos selecionados no documento 2020 & 2021 NLP solicitações de propostas do Lacuna Fund, para ver o trabalho que está em andamento.

4 - Informações sobre a proposta

Nota: O site Lacuna Fund inclui vários recursos como referências relevantes sobre a qualidade dos dados, documentação e formato para ajudar os candidatos a preparar uma candidatura competitiva.

A apresentação de propostas só será recebida por meio do portal de candidaturas SurveyMonkey Apply, disponível em www.lacunafund.org/apply. As candidaturas podem ser apresentadas em espanhol, francês, inglês e português. A descrição das perguntas relativas à

candidatura está disponível abaixo, a título meramente informativo. São necessárias as seguintes secções:

- Informações sobre o candidato (acessíveis no portal)
- Narrativa da proposta
- Orçamento e calendário

Informações sobre o candidato

Esta secção solicitará ao candidato que forneça:

- Um resumo da proposta com 200-250 palavras;
- Informações sobre a(s) instituição(ões) e/ou o grupo candidato;
- Local de realização dos trabalhos;
- CVs dos principais membros do grupo;
- Informações sobre os processos de análise ética da(s) instituição(ões) afiliada(s);
- Informações sobre a capacidade do grupo para obter aprovações nacionais.

Narrativa da proposta

O texto da proposta deve ter um máximo de 10 páginas, excluindo as referências, com margens de 2,5 cm e um tipo de letra de 11 pontos, no mínimo. Os apêndices ou o material narrativo da proposta com mais de 10 páginas não serão analisados.

Nesta secção, o candidato deve carregar uma narrativa coesa, em formato PDF ou Word, que aborde os seguintes aspectos:

- **Qualificações** - Descreva a(s) organização(ões) ou parceria(s) que se candidatam, de que forma satisfazem os critérios de elegibilidade acima referidos e as suas qualificações específicas para realizar o trabalho proposto.
- **Identificação do problema, solução proposta e conjunto de dados** - Descreva o problema ou a lacuna nos dados de formação ou avaliação e a solução proposta. Resuma o(s) conjunto(s) de dados que pretende criar, aumentar, agregar ou manter. *Indique de que forma o seu projeto aborda uma lacuna e complementa o trabalho existente.*
- **Especificações e produtos para os dados e documentação propostos** - Incluir o seguinte:
    - Quantidade de dados que serão incluídos no conjunto de dados.
    - Tipos e formato dos dados e/ou rótulos, bem como quadro e dimensão da amostra ou um plano para garantir a representatividade, se aplicável.
    - Métricas a utilizar para avaliar os resultados desejados da criação de dados (por exemplo, métricas de equidade no conjunto de dados, QA/QC em relação a uma referência, etc.)
- **Beneficiários Previstos e Casos de Utilização** - Descrever a consulta anterior e/ou a colaboração proposta com os beneficiários previstos e descrever os potenciais casos de utilização existentes e futuros da aprendizagem automática para os conjuntos de dados propostos. Explicar de que forma o conjunto de dados respeita e corresponde à diversidade das comunidades que representa. Indicar de que forma a qualidade proposta, os métodos

de coleta e outros pormenores tornam os dados adequados para utilização nesse contexto específico.

- **Metodologia** - Apresentar uma breve panorâmica das etapas propostas (e dos principais pressupostos) para o desenvolvimento e a execução do projeto. Por favor, inclua:
  - Técnicas propostas de coleta e rotulagem de dados e informações sobre a interoperabilidade. Por favor, inclua a consideração de infraestruturas existentes ou comuns e as mais recentes técnicas e ferramentas para acelerar a coleta, limpeza e partilha de dados.
  - Medidas de controlo de qualidade, tais como a qualidade que todas as amostras de dados devem ter para o conjunto de dados. Inclua a forma como o grupo pretende tratar os valores atípicos que possam afetar a qualidade do conjunto de dados.
  - Um plano para avaliar e atenuar erros e preconceitos (por exemplo, preconceitos de género ou outros preconceitos).
  - Como pretende utilizar os recursos existentes, incluindo métodos ou tecnologias de coleta, ligando conjuntos de dados pré-existentes em todo o domínio, bem como recursos existentes noutros contextos.
  - Autorizações em vigor ou medidas a adotar para obter as autorizações nacionais ou outras necessárias. Considerar quais as jurisdições que exigem aprovações e se a investigação proposta corresponde à definição de investigação nessa jurisdição. Se determinar que *não* são necessárias aprovações locais, nacionais ou regionais, explique a razão.
  - Quaisquer desafios ou incertezas previstas na coleta de dados e contramedidas propostas.
- **Impacto transformacional** - Explicar de que forma a rotulagem ou o conjunto de dados proposto contribuirá para alcançar o impacto desejado. Se for caso disso, descrever a forma como os produtos podem motivar a sua utilização em contextos de investigação ou comerciais. Considere quaisquer limitações práticas que possa enfrentar (por exemplo, penetração da Internet).
- **Gestão de dados e licenciamento** - Descrever:
  - Quaisquer questões previstas relacionadas com os direitos de autor dos dados de base e a colaboração com o detentor dos direitos de autor. Por favor, aborde também quaisquer questões previstas em matéria de direitos de autor e licenciamento de dados secundários.
- Planos de licenciamento para maximizar a utilização derivada responsável. Segundo a Política de propriedade intelectual (PI) do Lacuna Fund, o conjunto de dados e qualquer PI relacionada, como métodos de coleta, fichas de dados, como carregar ou ler conjuntos de dados ou outras informações para garantir a usabilidade, devem ser disponibilizados ao abrigo de uma licença de fonte aberta e de atribuição (CC-BY 4.0 ou CC BY-SA 4.0). Se for proposto um licenciamento mais restritivo, apresentar uma justificação para tal. O orçamento pode incluir recursos para o licenciamento.
- Se pretende utilizar um conjunto de dados existente no seu projeto, indique que o seu grupo recebeu as autorizações necessárias do proprietário do conjunto de dados para que este possa ser divulgado de acordo com o site Política de PI do Lacuna Fund, ou apresente

uma justificação para outra estrutura de licenciamento. *As cartas de apoio dos detentores de dados existentes são opcionais, mas incentivadas.*

- Pretende incluir um ficheiro de metadados e uma folha de dados como documentação para o seu conjunto de dados, de acordo com o Dataset Hosting and Documentation Guidance (Guia de documentação e armazenamento de conjuntos de dados) do Lacuna Fund.
- A plataforma de armazenamento que pretende utilizar. As plataformas de armazenamento devem atribuir um identificador de objeto digital (DOI) ao conjunto de dados, quantificar os downloads do conjunto de dados e coletar informações de contacto para os downloads do conjunto de dados. Para mais orientações sobre as plataformas de armazenamento sugeridas, consulte o Dataset Hosting and Documentation Guidance (Guia de documentação e armazenamento de conjuntos de dados).
- **Riscos, incluindo ética e privacidade** - Identifique problemas e riscos potenciais, incluindo, mas não se limitando a potenciais preocupações éticas e de privacidade, e descreva as medidas que irá tomar para os atenuar. Especificamente:
    - Uma declaração de reflexão sobre as comunidades das quais faz parte e as identidades que detém, e sobre a forma como estas podem afetar o seu trabalho.
    - Indicar a forma como irá garantir o consentimento informado, se for caso disso. (Isto deve incluir a notificação de potenciais casos de utilização futura dos dados)
- Descreva como irá garantir a equidade na mão de obra do projeto, incluindo, mas não se limitando a uma compensação justa pela rotulagem, anotação, fornecimento de dados e outros serviços de IA. Algumas opções para dar prioridade a normas de trabalho justas e práticas de compensação justas incluem:
- Seguir e assinar o Normas de Trabalho Justo para o trabalho de IA quando a anotação e o fornecimento de dados são subcontratados a entidades comerciais.
- Adesão a uma abordagem de fornecimento e anotação de dados orientada para a comunidade, caso exista um modelo desse género.
- Apresentação de uma renúncia/inquérito voluntário aos participantes no projeto para resolver o problema dos encargos indevidos.
- Descreva como a diversidade de género e outras considerações demográficas são incorporadas no grupo de projeto, na coleta de dados de formação e no desenvolvimento de modelos, para que os conjuntos de dados representem com precisão os impactos em diferentes comunidades/grupos.
- Apresentar um plano de anonimização de informações pessoalmente identificáveis (PII) e de conformidade com a legislação em matéria de privacidade, se aplicável. Se não existir um quadro jurídico nacional, a proposta deve indicar ou fazer referência às melhores práticas. Incluir considerações sobre a privacidade, tanto em âmbito individual como comunitário. Consulte o site Dataset Hosting and Documentation Guidance (Guia de documentação e armazenamento de conjuntos de dados) do Lacuna Fund para sugestões.
- Discutir os potenciais impactos adversos na produção e utilização do conjunto de dados e as medidas para os atenuar, incluindo os potenciais riscos para os direitos humanos e o potencial para um elevado consumo de energia na tecnologia de IA, que conduz a um aumento das emissões de dióxido de carbono.
- **Plano de Sustentabilidade e Comunicação**

- Descreva o modo como o conjunto de dados será mantido e/ou expandido para além do financiamento inicial (por exemplo, por meio de um modelo de base, de aplicações de aprendizagem automática resultantes, de uma comunidade dedicada ou de um conjunto de partes interessadas com um modelo de governação sólido para o conjunto de dados aberto) e o modo como um potencial caso de utilização poderá sustentar o projeto.
- Explicar de que forma o conjunto de dados respeitará os princípios dos dados FAIR (https://www.go-fair.org/resources/faq/what-is-fair/). Indique as medidas que irá tomar para garantir que o conjunto de dados seja localizável, **acessível, interoperável e reutilizável**.
- Definir as ações de comunicação para dar a conhecer o(s) conjunto(s) de dados. Estas podem incluir ações de ligação em rede com potenciais utilizadores de dados; apresentação do(s) conjunto(s) de dados numa conferência; organização de uma oficina sobre sustentabilidade do conjunto de dados com as partes interessadas; ou criação de um comité de sustentabilidade.

Cronograma do projeto e Produtos

Esta secção solicitará ao candidato que apresente um quadro com um calendário para a conclusão das principais ações e produtos. O calendário pode incluir, mas não se limita a formação de empregados, coleta de dados, rotulagem, garantia de qualidade, validação/limpeza e publicação dos dados. As prestações podem incluir, mas não se limitam a partes do conjunto de dados para demonstrar a prova de conceito, o conjunto de dados completo e a documentação ou métodos de coleta que o acompanham, que devem ser de fonte aberta.

Todos os calendários devem incluir uma data em que os dados estarão disponíveis ao público com toda a documentação.

Nota: Os projetos propostos devem estar concluídos, os conjuntos de dados publicados e os relatórios finais apresentados até outubro de 2026. Para efeitos de planeamento, espera-se que os acordos estejam concluídos e os trabalhos possam começar até abril de 2025.

Orçamento

Fornecer um orçamento para a conclusão do conjunto de dados proposto apresentado através do portal SurveyMonkey Apply. Este deve ser formatado no modelo de orçamento do Lacuna Fund, disponível no portal do candidato.

O montante total disponível é de aproximadamente 1 milhão de dólares americanos. Gostaríamos de financiar projetos em cada uma das regiões-alvo na África e América Latina, e prevemos apoiar de 6 a 8 projetos reduzidos com orçamentos de até USD 100,000 e de 6 a 8 projetos maiores e mais complexos com orçamentos variando entre USD 100 mil e USD 250 mil. O Painel Técnico Consultivo avaliará a **viabilidade e a adequação do orçamento**, bem como a **relação entre o orçamento e a narrativa do subsídio**, como parte dos critérios de seleção. Os orçamentos podem incluir, mas não se limitam a custos de:

- reforço das capacidades relacionadas com a coleta de dados e a garantia de qualidade/controlo de qualidade;
- coleta de dados (incluindo uma compensação justa pelo fornecimento de dados);
- rotulagem dos dados (incluindo uma compensação justa pela rotulagem dos dados);

- QA/QC ou verificação;
- pós-processamento de dados;
- publicação de dados;
- licenciamento;
- publicação dos resultados em acesso aberto;
- tempo para preparar uma declaração de dados para o conjunto de dados;
- esforços de crowdsourcing, como os "label-a-thons";
- armazenamento de dados;
- poder de computação;
- oficina
- ações de comunicação, incluindo a participação em conferências para no máximo dois eventos de apresentação dos conjuntos de dados

Os fundos não podem ser utilizados para o pagamento direto de quaisquer direitos aduaneiros, de importação ou outros direitos ou impostos cobrados relativamente à importação de bens ou equipamento para qualquer país ou jurisdição. **As taxas indiretas estão estritamente limitadas a 14.5% dos custos diretos de investigação.**

Os parceiros do Lacuna Fund podem disponibilizar armazenamento em nuvem e capacidade de computação em espécie. Se pretender utilizar este recurso, inclua-o no seu orçamento. Os grupos escolhidos receberão instruções sobre a forma de se candidatarem quando receberem o seu prémio.

Ver a folha de instruções do modelo de orçamento para mais informações sobre as orientações orçamentais, incluindo informações sobre os custos de pessoal admissíveis.

Obrigado pelo seu interesse no Lacuna Fund e pelos seus esforços em apoiar uma maior equidade e acessibilidade para aplicações de aprendizagem automática ao apoiar o processamento de linguagem natural. Aguardamos ansiosamente para analisar sua candidatura!

OBSERVAÇÃO: Oportunidade de mentoria e formação de parcerias

O Lacuna Fund tem o prazer de formar uma parceria com a Masakhane Research Foundation (MRF) para oferecer mentoria e oportunidades de formação de parcerias para candidatos ao PLN. A MRF é uma organização de base, cuja missão é fortalecer e estimular a pesquisa de PLN em idiomas africanos, para africanos e por africanos. O objetivo da MRF é que os africanos moldem e se apropriem desses avanços tecnológicos rumo à dignidade humana, ao bem-estar e à equidade, por meio do desenvolvimento de comunidades inclusivas, da pesquisa participativa aberta e da multidisciplinaridade.

Para esta chamada de propostas do Lacuna Fund, a MRF oferecerá aos candidatos da África e da América Latina uma oportunidade especial de se juntarem à comunidade MRF e de encontrarem um mentor para aqueles que estiverem interessados, que analisará o seu projeto de proposta e discutirá as formas de desenvolvê-la. As partes interessadas poderão solicitar

uma sessão com um mentor ao preencher este Formulário do Google com uma breve descrição (resumo de 250 palavras) da proposta para conjunto de dados. Os mentorandos podem solicitar diferentes formas de assistência, tais como ***"discutir lacunas em PLN com baixos recursos", "redigir uma proposta de pesquisa"*** e ***"preparar orçamentos".*** Fornecer uma descrição detalhada ajudará a unir os candidatos a mentores que estejam alinhados com sua área de interesse. O emparelhamento de mentores e mentorandos será administrado pelos nossos parceiros do Matchmaking & Mentorship Associates, que organizarão várias oficinas para reunir candidatos que estejam trabalhando em áreas relacionadas (p. ex., linguistas e pesquisadores de PLN) e coordenarão e organizarão sessões de comunicação entre os grupos relevantes. Por meio desse processo, também ofereceremos oportunidades de emparelhamento, para que os candidatos colaborem uns com os outros para formar equipes de projeto e enviar propostas em conjunto.

Os candidatos interessados são incentivados a se inscreverem para uma sessão de orientação pelo menos seis semanas antes do prazo de apresentação da proposta, ou seja, até **15 de julho de 2024**. Faremos o possível para identificar um mentor para todos que solicitarem um, mas não podemos garantir que haverá mentores disponíveis para todos. Os mentores serão designados por ordem de chegada. Espera-se que todos os candidatos leiam e cumpram o *código de ética e conduta* do programa de mentoria.