CONVÊNIO

CONVÊNIO Nº 07/2019 ENTRE A FUNDAÇÃO DE APOIO À PESQUISA DO DISTRITO FEDERAL – FAPDF, A UNIVERSIDADE DE BRASÍLIA – UNB e a FUNDAÇÃO DE EMPREENDIMENTOS CIENTÍFICOS E TECNOLÓGICOS - FINATEC.

A FUNDAÇÃO DE APOIO À PESQUISA DO DISTRITO FEDERAL – FAPDF, Fundação Pública, instituída pela Lei n. ° 347, de 04/11/92, vinculada à Secretaria de Estado de Ciência, Tecnologia e Inovação do Distrito Federal, com sede na Granja do Torto, lote 04 – Parque Tecnológico BIOTIC. Brasília/ DF - CEP: 70.636-000, inscrita no Cadastro Nacional da Pessoa Jurídica, sob o n.º 74.133.323/0001-90, de um lado, doravante denominada **CONCEDENTE** neste ato representada por seu Diretor-Presidente **ALESSANDRO FRANÇA DANTAS**, brasileiro, portador (a) da RG nº 2.347.805 SSP/DF e do CPF n.º 564.874.011-53, residente e domiciliado (a) em Brasília/DF, publicado no DODF nº 202 em 22 de outubro de 2019, a **UNIVERSIDADE DE BRASÍLIA - UnB**, Instituição Federal de ensino superior, criada pela Lei nº 3.998, de 15.12.1961, instituída pelo Decreto nº 500 de 15/1/1962, inscrita no CNPJ sob o nº 00.038.174/0001-43, sediada no Campus Universitário Darcy Ribeiro, Asa Norte, Brasília/DF, doravante denominada **EXECUTORA**, neste ato representado por sua Presidente, Profa. **MÁRCIA ABRAHÃO MOURA**, nomeada no dia 21 de Novembro de 2016 (publicação DOU de 22/11/2016) com a competência constante do respectivo Estatuto, com a interveniência administrativa e Financeira da **FUNDAÇÃO DE EMPREENDIMENTOS CIENTÍFICOS E TECNOLÓGICOS – FINATEC**, pessoa jurídica de direito privado, sem fins lucrativos, inscrita no CNPJ sob o n.º 37.116.704/0001-34, sediada na Universidade de Brasília, Campus Universitário Darcy Ribeiro, Edifício FINATEC, Asa Norte, Brasília – DF, doravante denominada **CONVENENTE**, representada neste ato por seu Diretor-Presidente, **ARMANDO DE AZEVEDO CALDEIRA PIRES**, brasileiro, engenheiro mecânico e professor universitário, portador da Carteira de Identidade nº 3.324.872 IFP/RJ e inscrito no CPF sob o nº 592.226.547-49, residente em Brasília – DF, **RESOLVEM** celebrar o presente Convênio, em conformidade com o disposto na Lei nº 8.666/1993, na Lei Complementar nº 101/2000, na Lei nº 347/1992, na Lei nº 10.973/2004, recepcionada pela Lei 6.140/2018, no Decreto nº 9.283/2018, no Decreto nº 32.598/2010, no Decreto nº 39.570/2018, na Instrução Normativa nº 01/2005 – CGDF e demais legislações aplicáveis, no que couber, mediante as regras e condições a seguir estabelecidas, as quais, mútua e reciprocamente, estipulam, outorgam, aceitam e se obrigam a cumprir:

Assessoria Jurídica FINATEC

CLÁUSULA PRIMEIRA – DO OBJETO

O presente Convênio tem por objeto estabelecer ações de mútua cooperação técnico científica para a execução do Projeto de Pesquisa intitulado **"KnEDLe – Extração de Informações de Publicações Oficiais usando Inteligência Artificial"**, em conformidade com o PLANO DE TRABALHO – Anexo I.

Parágrafo Único – Para atingir o objeto pactuado, os partícipes obrigam-se a cumprir, fielmente, o PLANO DE TRABALHO elaborado pela CONVENENTE e aprovado pela CONCEDENTE, o qual passa a integrar o presente Termo de Convênio, independentemente de transcrição. O PLANO DE TRABALHO contém os seguintes elementos:

a) Justificativa para a elaboração do instrumento;

b) Descrição completa do objeto a ser executado;

c) Descrição das metas a serem atingidas;

d) Definição das etapas ou fases de execução;

e) Cronograma de execução do objeto e cronograma de desembolso;

f) Plano de aplicação dos recursos a serem desembolsados pelo CONCEDENTE.

CLÁUSULA SEGUNDA – DAS OBRIGAÇÕES DOS PARTÍCIPES

Constituem responsabilidades e obrigações, no âmbito de suas respectivas competências institucionais, além dos outros compromissos assumidos neste Convênio:

I – Comuns aos partícipes:

a) Definir e ajustar diretrizes e procedimentos necessários à realização do objeto descrito na cláusula primeira;

b) Indicar representante legal para acompanhamento da fiel execução deste Convênio;

c) Executar as atividades decorrentes do pactuado no presente Convênio com obediência aos objetivos do PLANO DE TRABALHO;

d) Propor alterações, ajustes e aditivos, visando dar continuidade à execução do objeto do Convênio;

e) Fornecer as informações e as orientações necessárias ao melhor desenvolvimento e ao fiel cumprimento deste Convênio;

f) Observar o direito autoral envolvendo métodos, técnicas, cursos, programas ou qualquer material de divulgação institucional utilizado nas ações previstas neste Convênio, devendo ser informados o crédito da autoria e o respectivo instrumento de cooperação que deu amparo à utilização do material do partícipe;

g) Levar, imediatamente, ao conhecimento do outro partícipe, ato ou ocorrência que interfira no andamento das atividades decorrentes desse Convênio, para adoção das medidas cabíveis;

h) Notificar, por escrito, sobre imperfeições, falhas ou irregularidades verificadas na execução das atividades decorrentes do presente Instrumento.

II –De competência da CONCEDENTE:

Cabe à Concedente as seguintes obrigações:

a) Repassar à CONVENENTE, por meio da conta específica do Convênio, os valores pactuados;

b) Acompanhar, fiscalizar e controlar as atividades de execução do PLANO DE TRABALHO, avaliando os seus resultados;

c) Analisar previamente as propostas de reformulação do PLANO DE TRABALHO, acompanhadas de justificativa e desde que não implique na mudança de objeto;

d) Orientar, supervisionar e cooperar com a implantação das ações inerentes ao objeto deste Convênio;

e) Prorrogar a vigência do Convênio, de ofício, quando ocorrer atraso na liberação dos recursos, limitada a prorrogação ao exato período do atraso verificado.

III – De competência da EXECUTORA:

Cabe à EXECUTORA as seguintes obrigações:

a) Executar todas as etapas do Projeto em conformidade com o Plano de Trabalho;

b) Apresentar relatório semestral das atividades executadas;

c) Permitir o livre acesso de servidores da CONCEDENTE e dos Órgãos de Controle Interno e Externo, a qualquer tempo e lugar, a todos os atos e fatos relacionados direta ou indiretamente com o instrumento pactuado, quando em missão de acompanhamento, avaliação e fiscalização;

d) Cumprir a contrapartida pactuada no presente Convênio, caso haja previsão no PLANO DE TRABALHO;

e) Comunicar formalmente à CONCEDENTE, apresentando justificativas, qualquer fato que implique descontinuidade do PLANO DE TRABALHO, no prazo de até 30 (trinta) dias após seu conhecimento;

f) Propor alterações, ajustes e aditivos visando a dar continuidade à execução do objeto do convênio;

g) Emitir o Relatório Técnico Final, comprovando a conclusão das etapas do Projeto.

IV – De competência da CONVENENTE:

Cabe à CONVENETE (FINATEC) as seguintes obrigações:

Assessoria Jurídica FINATEC

a) Executar as atividades decorrentes do pactuado no presente Convênio, com rigorosa obediência aos objetivos do PLANO DE TRABALHO;

b) Responsabilizar-se pela gestão administrativa e pela gestão financeira dos recursos oriundos deste Convênio, a serem repassados pela CONCEDENTE para a execução das atividades do Projeto;

c) Prestar contas parcial e final dos recursos recebidos, nos termos do Decreto nº 39.570, de 26 de dezembro de 2018 e IN nº 01/2005 da CGDF, no que couber;

d) Recolher, à conta do CONCEDENTE, o valor correspondente ao percentual da contrapartida pactuada que não tenha sido aplicado na consecução do objeto do convênio, atualizado monetariamente, caso haja previsão no PLANO DE TRABALHO;

e) Restituir o valor transferido pela CONCEDENTE de eventual saldo de recursos, inclusive os rendimentos decorrentes de sua aplicação financeira na data de conclusão do seu objeto ou da sua extinção;

f) Movimentar os recursos financeiros liberados pela CONCEDENTE em conta específica do Convênio, aberta no Banco de Brasília – BRB;

g) Permitir o livre acesso de servidores da CONCEDENTE e dos Órgãos de Controle Interno e Externo, a qualquer tempo e lugar, a todos os atos e fatos relacionados direta ou indiretamente com o instrumento pactuado, quando em missão de acompanhamento, avaliação e fiscalização;

h) Restituir o valor transferido pela CONCEDENTE, atualizado monetariamente desde a data do recebimento, acrescido de juros legais, na forma da legislação aplicável aos débitos para com a Fazenda Distrital, nos seguintes casos:

1.quando não executado o objeto da avença;

2.quando não apresentada, no prazo exigido, a prestação de contas; e

3.quando os recursos forem utilizados em finalidade diversa da estabelecida no convênio;

i) Recolher, à conta da CONCEDENTE, o valor correspondente a rendimentos de aplicação no mercado financeiro, referente ao período compreendido entre a liberação do recurso e sua utilização, quando não comprovar o seu emprego na consecução do objeto do convênio, ainda que não tenha feito essa aplicação, admitidas, neste caso, justificativas;

j) Responsabilizar-se integralmente pelos encargos tributários, fiscais, previdenciários e trabalhistas, relativos às obrigações com o pessoal utilizado, além de outros decorrentes da execução do objeto;

k) Comunicar formalmente à CONCEDENTE, apresentando justificativas, qualquer fato que implique descontinuidade do PLANO DE TRABALHO, no prazo de até 30 (trinta) dias após seu conhecimento;

l) Propor alterações, ajustes e aditivos visando a dar continuidade à execução do objeto do convênio;

Assessoria Jurídica FINATEC

m) Responsabilizar-se pela aquisição de bens e serviços necessários à execução do Projeto, incluindo a contratação e o pagamento do pessoal envolvido nas atividades de pesquisa;

n) Submeter-se à fiscalização e ao controle finalístico e de gestão de que trata a Lei n. 8.958/1994, Decreto 7.423/2010 e Decreto nº 39.570, de 26 de dezembro de 2018.

CLAÚSULA TERCEIRA – DOS RECURSOS E DA DOTAÇÃO ORÇAMENTÁRIA

Os recursos financeiros necessários à execução do objeto deste Convênio, no montante de **R$ 3.296.470,59 (três milhões, duzentos e noventa e seis mil, quatrocentos e setenta reais e cinquenta e nove centavos)**, serão repassados pela CONCEDENTE à CONVENENTE, de acordo com o cronograma de desembolso contido no PLANO DE TRABALHO, de acordo com sua disponibilidade orçamentária e financeira.

Parágrafo Primeiro - Os valores repassados à CONVENENTE correrão por conta dos seguintes recursos: Fonte de recursos: 100 Natureza da Despesa: 3.3.50.41
Programa de Trabalho: 19.571.6207.6026.0011 - Fomento ao Desenvolvimento Científico e Tecnológico - Projetos Inovadores em Empresas e Entidades - FAPDF

Parágrafo Segundo - As despesas a serem executadas em exercícios futuros serão objeto de termo aditivo, no qual serão indicadas as dotações orçamentárias e empenhos, ou notas de movimentação de crédito, para sua cobertura, conforme determina o inciso XV do art. 7º da IN nº 01/2005 – CGDF.

Parágrafo Terceiro - Os recursos para atender às despesas de exercícios futuros, no caso de investimento, estão consignados no plano plurianual, ou em prévia lei que o autorize e fixe o montante das dotações que, anualmente, constarão do orçamento, durante o prazo de sua execução, nos termos do inciso XVI do art. 7º da IN nº 01/2005 – CGDF.

CLÁUSULA QUARTA – DA LIBERAÇÃO DOS RECURSOS

A liberação de recursos financeiros, em decorrência das atividades constantes do PLANO DE TRABALHO anexo a este Convênio, deve obedecer ao cronograma de desembolso previsto naquele e guardar consonância com as fases ou etapas de execução do objeto do Ajuste.

Parágrafo Primeiro - Os recursos serão mantidos em conta bancária específica, somente sendo permitidos saques para pagamento de despesas constantes do PLANO DE TRABALHO ou para aplicação no mercado financeiro, nas hipóteses previstas em lei, mediante movimentação exclusiva através de cheque nominativo, ordem bancária, transferência eletrônica disponível, ou outra modalidade de saque autorizada pelo Banco Central do Brasil, em que fiquem identificados sua destinação e, no caso de pagamento, o credor.

Parágrafo Segundo – Os custos operacionais da EXECUTORA/CONVENENTE estão limitados em até 15% do valor total dos recursos financeiros destinados à execução do convênio, conforme detalhado no PLANO DE TRABALHO anexo, segundo estabelece o Art. 74 do Decreto nº 9.283 de 07 de fevereiro de 2018, que regulamenta a Lei nº 10.973 de 02 de dezembro de 2004, recepcionada pela Lei Distrital nº 6.140 de 03 de maio de 2018.

Parágrafo Terceiro - A liberação de recursos financeiros, em decorrência de despesas operacionais necessárias à consecução dos objetivos deste convênio, será autorizada pela CONCEDENTE mediante atesto de relatório detalhado contendo as atividades desenvolvidas e seus respectivos custos.

CLAÚSULA QUINTA – DA FORMA DE EXECUÇÃO

O convênio será fielmente executado pelas partes, de acordo com as cláusulas pactuadas e a legislação pertinente, respondendo cada uma, no que lhe couber, pelas consequências de sua inexecução total ou parcial.

CLÁUSULA SEXTA – DAS VEDAÇÕES

O presente Convênio deverá ser executado em estrita observância às cláusulas avençadas e às normas pertinentes, sendo vedado:

a) Aditamento para alterar o objeto;

b) Utilizar recursos em finalidade diversa da estabelecida no respectivo instrumento, ainda que em caráter de emergência;

c) Realizar despesas em data anterior ou posterior à sua vigência;

d) Atribuir vigência ou efeitos financeiros retroativos;

e) Realizar despesas com taxas bancárias, multas, juros ou atualização monetária, inclusive, referentes a pagamentos ou recolhimentos fora dos prazos e manutenção de contas ativas;

f) Transferir recursos para clubes, associações de servidores ou quaisquer entidades congêneres, excetuadas creches e escolas quando destinados ao atendimento pré-escolar regularmente instituído;

g) Pagamento, a qualquer título, a servidor ou empregado público integrante do quadro de pessoal da Administração Direta ou Indireta do Distrito Federal, da União, dos Estados e dos Municípios, por serviços de consultoria ou assistência técnica;

h) Realizar despesas com publicidade, salvo as de caráter educativo, informativo ou de orientação social, das quais não constem nomes, símbolos ou imagens que caracterizem promoção pessoal de autoridades ou servidores públicos.

CLÁUSULA SÉTIMA – DO DESTINO E DO DIREITO DE PROPRIEDADE DOS BENS REMANESCENTES

Os bens remanescentes na data da conclusão ou extinção do presente Convênio, e que, em razão deste, tenham sido adquiridos, produzidos ou transformados serão de propriedade da CONCEDENTE, podendo ser doados à CONVENENTE e ou INSTITUIÇÃO EXECUTORA para utilização na destinação especificada no PLANO DE TRABALHO.

Assessoria Jurídica FINATEC

CLAÚSULA OITAVA – DA VIGÊNCIA E DA PRORROGAÇÃO

O presente Convênio terá vigência de 36 (trinta e seis) meses, a contar da data de sua assinatura, podendo ser prorrogado, por meio de Termo Aditivo, após análise e aprovação pelo Conselho Diretor da FAPDF, mediante solicitação de prorrogação apresentada com antecedência mínima de 30 (trinta) dias corridos, anterior ao término de sua vigência, fundamentada em razões concretas que justifiquem a prorrogação.

Parágrafo único - A CONCEDENTE fica obrigada a prorrogar a vigência do presente Convênio, de ofício, quando ocorrer atraso na liberação dos recursos, limitada a prorrogação ao exato período do atraso verificado.

CLÁUSULA NONA – DAS ALTERAÇÕES

O Convênio, ou plano de trabalho somente poderão ser alterados mediante proposta da CONVENENTE, devidamente justificada, apresentadas no prazo mínimo de 60 dias antes da data que se pretenda implementar as alterações, dentro da vigência do instrumento e desde que aceitas pela CONCEDENTE.

Parágrafo primeiro - As alterações de que trata esta cláusula serão implementadas por meio de Termo Aditivo e sujeitam-se ao registro, pela CONCEDENTE, no SIGGO.

Parágrafo segundo - Fica vedado o aditamento do presente Convênio com o intuito de alterar o seu objeto, sob pena de nulidade do ato e responsabilidade do agente que o praticou.

CLÁUSULA DÉCIMA – DAS HIPÓTESES DE RESCISÃO

Este Convênio poderá, a qualquer tempo, ser denunciado ou rescindido pelos partícipes, , imputando-lhes as responsabilidades pelas obrigações decorrentes do prazo em que tenham vigido e creditando-lhes, igualmente, os benefícios adquiridos nesse mesmo período. Para tanto, o interessado deverá externar formalmente a sua intenção nesse sentido, com antecedência mínima de 30 (trinta) dias da data em que se pretenda que sejam encerradas as atividades, respeitadas as obrigações assumidas com terceiros e realizada por meio de:

a) Distrato, via mútuo consentimento dos partícipes;

b) Resolução, por inadimplemento unilateral das obrigações, por um dos partícipes;

c) Denúncia, resilição do Ajuste por iniciativa dos participantes em notificação ao outro.

CLÁUSULA DÉCIMA PRIMEIRA – DA PERROGATIVA DE AUTORIDADE NORMATIVA

É prerrogativa da CONCEDENTE conservar a autoridade normativa e exercer controle e fiscalização sobre a execução, bem como de assumir ou transferir a responsabilidade pelo mesmo, no caso de paralisação ou de relevante fato superveniente, de modo a evitar a descontinuidade do serviço.

Assessoria Jurídica FINATEC

CLÁSULA DÉCIMA SEGUNDA - DA PROPRIEDADE DOS RESULTADOS

As partes deverão prever, em instrumento jurídico específico, a titularidade da propriedade intelectual e a participação nos resultados da exploração das criações resultantes da parceria, assegurando aos signatários o direito à exploração, ao licenciamento e à transferência de tecnologia, observado o disposto nos §§ 4º a 7º do art. 6º da Lei 10.973/2004

Parágrafo Único – A propriedade de todos os produtos, resultados, dados, informações, tecnologias, procedimentos e rotinas utilizadas para a execução do Projeto e já existentes anteriormente à celebração deste Convênio, continuarão pertencendo à parte detentora.

CLÁUSULA DÉCIMA TERCEIRA – MONITORAMENTO, AVALIAÇÃO E PRESTAÇÃO DE CONTAS

Para o monitoramento, a avaliação e a prestação de contas, a CONCEDENTE, a EXECUTORA e a CONVENENTE observarão as disposições constante dos art. 3º a 16 do Decreto nº 39.570 de 26 de dezembro de 2018.

Parágrafo Primeiro - A prestação de contas observará as seguintes etapas:

I – Monitoramento e avaliação por meio de formulário de resultado;

II – Prestação de Contas Final por meio da apresentação de relatório.

Parágrafo Segundo - O monitoramento e a avaliação deverão observar os objetivos, o cronograma, o orçamento, as metas e os indicadores previstos no plano de trabalho.

Parágrafo Terceiro - É faculdade da CONCEDENTE, durante o monitoramento e a avaliação do projeto, a realização de visitas, para acompanhamento técnico ou fiscalização financeira, bem como o uso de técnicas estatísticas, tais como amostragem e agrupamento em faixas ou subconjuntos de características similares para a utilização de critérios de análise diferenciados em cada um. A visita será comunicada à CONVENENTE e à EXECUTORA, com antecedência mínima de três dias úteis, admitido o uso de meios eletrônicos para a comunicação.

Parágrafo Quarto - O monitoramento será realizado pela CONCEDENTE, que apontará as ocorrências relacionadas com a consecução do objeto, adotará as medidas para a regularização das falhas observadas e deverá manifestar-se fundamentadamente pela aprovação ou pela rejeição das justificativas.

Parágrafo Quinto - Encerrada a vigência do instrumento, a CONVENENTE encaminhará à concedente a prestação de contas final no prazo de até 60 (sessenta) dias.

Parágrafo Sexto - O prazo a que se refere o caput poderá ser prorrogado por igual período, a pedido, desde que o requerimento seja feito anteriormente ao vencimento do prazo inicial.

Parágrafo Sétimo - Se, durante a análise da prestação de contas, a CONCEDENTE verificar irregularidade ou omissão passível de ser sanada, determinará prazo compatível com o objeto e não superior a 60 (sessenta) dias, para que a CONVENENTE e a EXECUTORA apresentem as razões ou a documentação necessária.

Assessoria Jurídica FINATEC

Parágrafo Oitavo - Transcorrido o prazo de que trata o Parágrafo Sétima desta Cláusula, se não for sanada a irregularidade ou a omissão, a autoridade administrativa competente adotará as providências para a apuração dos fatos, nos termos da legislação vigente.

Parágrafo Nono - A análise da prestação de contas final deverá ser concluída pela CONCEDENTE no prazo de até um ano, prorrogável por igual período, justificadamente, e, quando a complementação de dados se fizer necessária, o prazo poderá ser suspenso.

Parágrafo Dez – A prestação de contas será simplificada, privilegiará os resultados obtidos e compreenderá:

I - relatório de execução do objeto, que deverá conter:

a) a descrição das atividades desenvolvidas para o cumprimento do objeto;

b) a demonstração e o comparativo específico das metas com os resultados alcançados; e

c) o comparativo das metas cumpridas e das metas previstas devidamente justificadas em caso de discrepância, referentes ao período a que se refere a prestação de contas;

II - declaração de que utilizou os recursos exclusivamente para a execução do convênio acompanhada de comprovante da devolução dos recursos não utilizados, se for o caso;

III - relação de bens adquiridos, desenvolvidos ou produzidos, quando houver;

IV - avaliação de resultados; e

V - demonstrativo consolidado das transposições, dos remanejamentos ou das transferências de recursos efetuados, quando houver.

Parágrafo Onze - Quando o relatório de execução do objeto não for aprovado ou quando houver indício de ato irregular, a CONCEDENTE exigirá, em prazo não superior a 30 dias, a apresentação de relatório de execução financeira.

Parágrafo Doze - A documentação gerada até a aprovação da prestação de contas final deverá ser organizada e arquivada pela CONVENENTE pelo prazo de cinco anos, contado da data da aprovação da prestação de contas final.

Parágrafo Treze - Fica facultada à CONCEDENTE a solicitação do envio de cópia da documentação original ou digitalizada.

CLÁUSULA DÉCIMA TERCEIRA – DA PUBLICAÇÃO

A publicação do extrato do presente Convênio no Diário Oficial do Distrito Federal é condição indispensável para sua eficácia, devendo ocorrer no prazo de vinte dias a contar da sua assinatura, nos termos do art. 15 da IN nº 01/2005.

CLÁUSULA DÉCIMA QUINTA– DO FORO

Assessoria Jurídica FINATEC

Fica eleito o foro de Brasília, com renúncia expressa a qualquer outro, por mais privilegiado que seja, para dirimir questões decorrentes do presente Convênio, não resolvida administrativamente.

E, como prova de assim haverem livremente pactuado, firmam os PARTÍCIPES o presente instrumento em 03 (três) vias, de igual teor e forma, para que produza entre si os efeitos legais, na presença de 02 (duas) testemunhas que, igualmente, subscrevem.

Brasília – DF, 31 de Dezembro de 2019.

Pela FAPDF: _____

Pela UNB: _____
Márcia Abrahão Moura
Reitora

Pela FINATEC: _____
Prof. Armando de Azevedo Caldeira Pires
Diretor-Presidente
FINATEC

Testemunhas:

1. _____
Nome: Fabio Henrique da Silva Ramos
CPF: 138666 907-54
RG: 2o 936 975-0

2. _____
Nome: Patrícia Magalhães Borges
Assessoria Jurídica
CPF: 040.193.831-03
RG: 2.507.659 SSP/DF

Assessoria Jurídica FINATEC

Universidade de Brasília

PLANO DE TRABALHO

1 – TIPO PROJETO

(x) Pesquisa ( ) Curso de Pós- Graduação ( ) Atividade de Extensão ( ) Ensino de Graduação

2 – TIPO DE INSTRUMENTO PROCESSUAL

( ) Acordo (x) Convênio ( ) Termo de Execução Descentralizada ( ) Contrato ( ) outro

3 – DADOS CADASTRAIS DA UnB

Órgão/Entidade Proponente **UNIVERSIDADE DE BRASÍLIA**				**C.N.P.J** **00.038.174/0001-43**
Endereço **CAMPUS UNIVERSITÁRIO DARCY RIBEIRO – PRÉDIO DA REITORIA - ASA NORTE**				
Cidade **BRASÍLIA**	**UF** **DF**	**CEP** **70910-900**	**Telefone** **(61) 3107-0246**	**UG / Cód. Gestão** **154040 / 15257**
Banco **Banco do Brasil - 001**	**Agência** **1607-1**		**Conta Corrente** **170.500-8**	**Praça de Pagamento** **Brasília**
Nome do Representante Legal **MÁRCIA ABRAHÃO MOURA**				**CPF** **334.590.531-00**
CI / Órgão Exp. / Emissão **960.490 SSP/DF** **Expedição 01/08/1995**	**Cargo** **Professora**		**Função** **Reitora**	**Matrícula FUB** **145378**
e-mail **unb@unb.br**				
Nome do Coordenador (a) do Projeto Teófilo Emidio de Campos				**CPF** 184.452.518-05
CI / Órgão Exp. / Emissão 27.299.234-3 / SSP / SP	**Cargo** **Professor**		**Função** **Professor**	**Matrícula FUB** 1094092
E-mail teodecampos@unb.br				**Telefone** **(61)98327-1023**
Nome do Gestor do Projeto Gladston Luiz da Silva				**CPF** 334.165.591-34
Unidade/Departamento IE / EST				**Matrícula FUB** 1053531
Endereço Eletrônico (e-mail) gladston@unb.br		**Telefone fixo** **(61)** 31077353		**Telefone celular** **(61)**
Assinatura				

Nome do Gestor Substituto do Projeto Li Weigang		**CPF** 150147578-98
Unidade/Departamento IE / CIC		**Matrícula FUB** 149667
Endereço Eletrônico (e-mail) weigang@unb.br	**Telefone fixo** (61) 31073679	**Telefone celular** (61) 9.
Assinatura		

4 – DADOS CADASTRAIS DO CONCEDENTE / CONTRATANTE

Tipo **(X) Público ( ) Privado**	**2 – Nome / Razão Social** Fundação de Apoio à Pesquisa do Distrito Federal – FAPDF		**3 - CNPJ** 74.133.323/0001-90
Endereço sede (Av., Rua, Nº, Bairro) Granja do Torto, Lote 04, Parque Tecnológico de Brasília- BIOTIC, 3° andar.			
Cidade Brasília	**UF** DF	**CEP** 70.333-900	**(DDD) Telefone** (61) 3462-8831 / 3462-8832 / 3462-8806
Nome do representante legal Alessandro França Dantas			**CPF** 564.874.011-53
CI / Órgão Exp. / Emissão 2.347.805 / SSP-DF	**Cargo** Diretor-Presidente		

5-OUTROS PARTÍCIPES

Tipo **( ) Público (X) Privado**	**2 – Nome / Razão Social** Fundação de Empreendimentos Científicos e Tecnológicos - FINATEC		**3 - CNPJ** 37.116.704/0001-34
Endereço sede (Av., Rua, Nº, Bairro) Campus Universitário Darcy Ribeiro – Av. L3 Norte – Ed. Finatec – Asa Norte			
Cidade Brasília	**UF** DF	**CEP** 70.910-900	**(DDD) Telefone** (61) 3348-0407
Nome do representante legal Armando de Azevedo Caldeira Pires			**CPF** 592.226.547-49
CI / Órgão Exp. / Emissão 3.324.872 / SSP-RJ / Emissão em 19.05.1973	**Cargo** Diretor-Presidente		
Nome do responsável pelo Projeto na Fundação Lavocat Galvão de Almeida Coelho Luiza			**CPF** 031.783.761-35

CI / Órgão Exp. / Emissão	**Cargo**
2461123 / SESP-DF / 30/07/2014	Gerente de Projetos

6 - DESCRIÇÃO DO PROJETO

Título do Projeto	**Período de Execução**
"KnEDLe – Extração de informações de publicações oficiais usando inteligência artificial"	36 (trinta e seis) meses a partir da data da assinatura

Valor Total
R$ 3.296.470,59 (três milhões, duzentos e noventa e seis mil, quatrocentos e setenta reais e cinquenta e nove centavos)

RESUMO

Publicações oficiais tais como o Diário oficial do DF são fontes de informação sobre todos os atos oficiais do governo. Tais publicações são ricas em detalhes e oferecem uma quantidade gigantesca de dados, não somente pelo volume e frequência de publicação, mas também pela longevidade. Porém, as edições são publicadas de uma maneira não estruturada e em linguagem natural, o que cria um cenário desafiador para se extrair informações de forma estruturada. Este projeto propõe usar publicações oficiais como objeto de pesquisa e efetuar extração de conhecimento (*KnEDLe – kNowledge Extraction from Documents of LEgal content*). O objetivo é desenvolver ferramentas inteligentes de extração de informação estruturada a partir de tais publicações, visando facilitar a busca e recuperação de informações, aumentando a transparência do governo e facilitando tarefas de auditoria e detecção de problemas relacionados ao emprego de recursos públicos.

Este projeto foi motivado por problemas enfrentados de maneira recorrente em órgãos da administração do Distrito Federal, tais como o Tribunal de Contas e o Controladoria Geral do DF. Porém, espera-se que os resultados gerados sejam relevantes em contextos que extrapolam as esferas desses órgãos, podendo ser aplicado em diversos casos em que há a necessidade de se processar uma vasta quantidade de dados não estruturados para se extrair informação que pode estar escondida em poucos bytes de altíssima relevância, i.e., procurar e encontrar a "agulha no palheiro" (*needle in a haystack*).

6.1 IDENTIFICAÇÃO DO PROPONENTE

- **Nome Completo e vinculação**: Teófilo Emidio de Campos, CiC, IE, UnB
- **Lattes**: http://lattes.cnpq.br/5052452346402051;

O coordenador desta proposta, Teófilo E. de Campos, é doutor pela University of Oxford (concluído em 2006), onde foi parte do Robotics Research Group, uma das principais referências mundiais na área de Visão Computacional. Em 2001 ele completou seu mestrado em Ciência da Computação pelo Instituto de Matemática e Estatística da Universidade de São Paulo, com bolsa da FAPESP. Seu trabalho recebeu prêmio de melhor dissertação de mestrado do Brasil pela Sociedade Brasileira de Computação [25]. Graduou-se em 1998 na Universidade Estadual Paulista, também em Ciência da Computação, onde teve bolsa de iniciação científica da FAPESP.

Foi pesquisador sênior do **Projeto Victor**, um Termo de Execução Descentralizada entre o Supremo Tribunal Federal (**STF**) e a Universidade de Brasília, cujo objetivo foi a criação de métodos de análise automática de processos jurídicos que chegam ao STF usando técnicas de inteligência artificial. Esse projeto está tendo uma grande repercussão devido a seu potencial de agilizar a análise de processos legais, reduzindo significantemente a morosidade do judiciário.

Além disso, o coordenador possui grande experiência como pesquisador em empresas multinacionais, pois trabalhou nos laboratórios europeus da Sharp (2005-2007) e da Xerox (2008-2009) e no laboratório indiano da Microsoft (2007). Trabalhou também como pesquisador nas universidades de Surrey (2009-2016) e Sheffield (2013-2014), onde se envolveu numa grande quantidade de projetos financiados pelo Engineering and Physical Sciences Research Council e também pela União Europeia. Atualmente, é professor adjunto na Universidade de Brasília, desde julho de 2016.

O coordenador desta proposta teve papéis importantes na organização de vários eventos internacionais, incluindo a British Machine Vision Conference, em 2012. Ele foi coordenador local do projeto europeu

PASCAL2 (http://www.pascal-network.org), que envolveu ao todo 1072 pesquisadores na Europa. Possui um histórico de colaboração com autores que são líderes internacionais em Visão computacional e Aprendizado de Máquina, tais como Neil Lawrence, Josef Kittler, Krystian Mikolajczyk, Adrian Hilton, Florent Perronnin, Gabriela Csurka, Manik Varma, Graham Jones, David Murray, Roberto Cesar-Jr e Isabelle Bloch, dentre outros.

Ele possui proeminência internacional, já foi coordenador de área (*area chair*) das conferências WACV 2016, VISAPP 2012 e EURO 2012. Ele é regularmente convidado a compor o painel de programa das principais conferências internacionais de sua área, tais como ICCV, ECCV, CVPR, BMVC e ACCV (todas de nível A1), dentre outras. Ele também é regularmente convidado para revisar artigos das principais revistas da área tais como IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) e Pattern Recognition, IEEE Transactions on Image Processing, dentre outras (todas A1).

Em 2017, foi nomeado como um dos revisores (pareceristas) excepcionais da principal conferência internacional de visão computacional (ICCV). Isso se deveu ao alto nível de qualidade e de detalhes dos seus pareceres.

É autor de mais de 50 artigos publicados internacionalmente e já recebeu 1478 citações (vide Figura abaixo). Possui 5 patentes internacionais, uma delas [26] foi registrada em países da América do Norte, Europa e Ásia e sua tecnologia foi transferida para o computador de bordo de carros de montadoras europeias e japonesas.

Possui experiência como orientador de doutorado [27] e já orientou projetos de pós-doutorado [28], graduação e também estágios (sanduíche) de estudantes de doutorado. Um dos seus ex-orientandos de doutorado *sanduíche* é o Julian McAuley, que hoje em dia é professor University of California San Diego e já foi citado 6379 vezes, de acordo com o perfil de Julian McAuley no Google Scholar, visualizado em maio de 2019 (https://scholar.google.co.uk/citations?user=icbo4M0AAAAJ).

6.2 IDENTIFICAÇÃO DA PROPOSTA

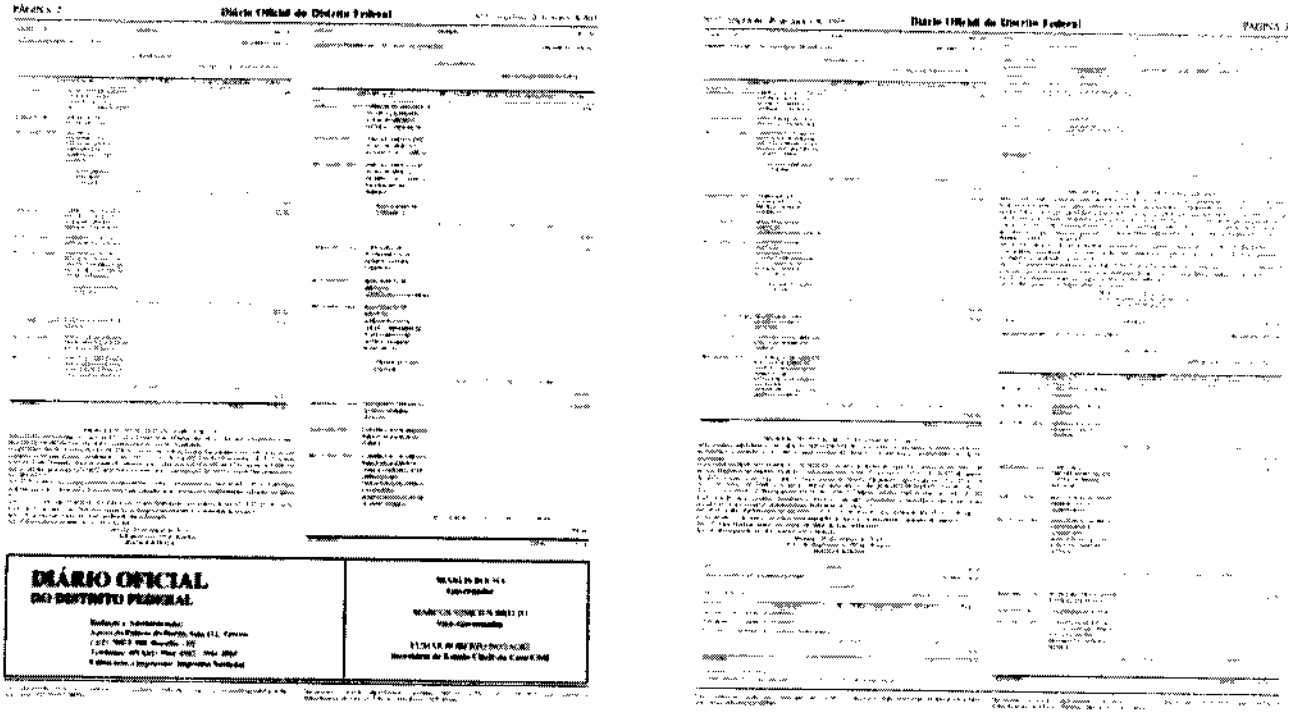
I. INTRODUÇÃO, apresentação e contextualização

As informações contidas em publicações oficiais correspondem a dados relevantes à fiscalização da aplicação do dinheiro público. Esses dados estão disponíveis em forma de texto natural em editais de licitações e de contratos, dentre outros. Além disso, outras informações devem ser extraídas de publicações oficiais e que se mostram relevantes a órgãos do GDF, como cargos e respectivos ocupantes, as datas de nomeação/designação e de exoneração, relação de gestores com os respectivos períodos de ocupação dos cargos, executores, fiscais, gestores de contratos, datas de designação e de dispensa, histórico de fiscais, identificação de acumulações ilegais de cargos, indicação de nepotismo etc.

Diversas publicações oficiais do Distrito Federal não se encontram em formatos estruturados capazes de facilitar a recuperação de informações. O Diário Oficial do DF (DODF) é composto de arquivos em PDF não estruturados contendo informações, na maioria dos casos, em formato de imagem rasterizada. A ilustração na figura abaixo mostra duas páginas de uma publicação recente do DODF que é recorrente no formato atualmente utilizado. Nesse padrão, todas as tabelas foram inseridas no documento como imagens, inclusive a tabela da parte inferior da primeira página, que contém metadados sobre a publicação em si. Especificamente, nessas duas páginas, isso totaliza uma média superior a 80% da área, ou seja, menos de 20% das informações podem ser extraídas direto do arquivo PDF, sendo que esses dados não incluem

nenhuma informação numérica. Todo o restante requer o uso de ferramentas de extração de texto de imagens.

Além disso, o texto do DODF está escrito em linguagem natural do domínio jurídico, o que impossibilita que sejam aplicadas ferramentas genéricas de extração de informação de texto em linguagem natural. Essas deficiências dificultam a distribuição e o uso da informação em curto prazo de forma categorizada para as partes interessadas (entes públicos com funções sociais específicas, auditores especializados em temas etc.).



Páginas 2 e 3 da edição de 26 de março de 2019 do Diário Oficial do Distrito Federal. Todas as tabelas, incluindo o cabeçalho e o bloco na parte inferior da primeira página foram codificados como imagens rasterizadas no arquivo em formato PDF.

- **Descrição do principal problema a ser abordado**

O principal problema a ser abordado é a criação de um sistema automático de extração e estruturação de informação voltado a documentos oficiais. O sistema deverá ser capaz organizar a informação de forma que os itens sejam bem segmentados e classificados. Almeja-se também conectar automaticamente os itens de informação de forma que seja possível estabelecer ligações entre atores, objetos e ações, mesmo que os itens de informação sejam provenientes de publicações diferentes.

- **Objetivos**

Este projeto tem o objetivo de contribuir na tarefa de extração de informações a partir de publicações oficiais, de forma dinâmica, pelas partes interessadas. Para isso, será efetuado um trabalho de pesquisa voltado aos desafios levantados por dados que têm as características do DODF. Esse projeto envolverá diversas áreas relacionadas com a inteligência artificial, incluindo visão computacional, processamento de linguagem natural, aprendizado de máquina, além de mineração de dados, busca e recuperação de informação.

Planeja-se também o desenvolvimento de um protótipo de ferramenta de software que poderá ser aplicado na prática por órgãos do Governo do DF.

II. REVISÃO DA LITERATURA

Esta seção apresenta uma revisão da literatura que engloba diversos aspectos do projeto, organizada de acordo com cada frente de trabalho listada posteriormente na seção de metodologia. Com essa revisão, esperamos também complementar a justificativa, a motivação e a discussão da relevância desta proposta, discutidas anteriormente.

Extração de dados a partir de imagens de documentos

Reconhecimento de caracteres (OCR)

Trata-se de uma área em que já existem muitas ferramentas boas disponíveis para uso, tal como o Tesseract [1], as quais poderiam ser usadas como ponto de partida para o trabalho nessa área, gerando transcrições de texto que podem ser usadas para alimentar as frentes de trabalho de processamento de linguagem natural. Porém, as ferramentas existentes geram transcrições altamente ruidosas, principalmente se aplicadas aos documentos do DODF. Portanto, ainda há um desafio de pesquisa que envolve a combinação de informação de baixo nível (pixels) com alto nível (modelo de linguagem) para melhorar a qualidade do texto extraído.

Detecção de palavras-chave

Os melhores sistemas de OCR levam, em média 1 segundo para processar cada página de um documento usando uma CPU moderna. Considerando o gigantesco volume de dados disponíveis, isso torna o processo de extrair informações legadas em algo excessivamente custoso. Alternativamente, propomos que a análise seja feita somente usando palavras chave, possibilitando o uso de detectores de objetos altamente eficientes voltado a localizar somente as palavras relevantes para extrair informações necessárias. Tais detectores (um exemplo é o Faster RCNN [2]), se implementados usando uma GPU, podem processar dezenas de páginas por segundo. Para tal, primeiro é necessário fazer um levantamento de uma lista de palavras relevantes usando técnicas de análise de texto para criar um conjunto de treinamento desse detector visual.

Classificação visual direta

Para várias tarefas de análise de documentos, é importante identificar a semântica de cada objeto contido nas páginas, por exemplo, para detectar tabelas, elementos gráficos e texto padrão. Técnicas de classificação visual de objetos podem ser empregadas para tal (tal como NASNET [3]). Com isso, é possível extrair informação de partes de documentos mesmo sem usar o texto que elas contêm, como um pré-processamento para se decidir qual técnica de extração de informação deve ser aplicada a cada parte do documento. Tais métodos são ainda mais rápidos do que os detectores de objetos.

Análise morfo-sintática (Part-of-Speech Tagging)

Aqui, o desafio é transformar linguagem humana em informação útil computacionalmente. Diversas técnicas emergem como formas de estruturar os textos, sendo uma delas o Part-of-Speech Tagging (POS Tagging). O POS Tagging se trata de um processo de rotulação de elementos textuais – tipicamente palavras e pontuação – com o fim de evidenciar a estrutura gramatical de um determinado trecho de texto. Existem várias aplicações possíveis de serem construídas usando POS Tagging. Um modelo estado-da-arte é capaz de realizar a tarefa com precisão em torno de 97% em corpora de língua inglesa [4]. A literatura reporta resultados semelhantes para diferentes corpora em português do Brasil [5, 6]. No entanto, ainda existem desafios e campo para avanço, vez que os corpora existentes não são compostos por textos forenses. Neste

projeto, pretende-se contribuir na elaboração e correção de corpora em português com termos específico do domínio jurídico.

Elementos relevantes do texto, tais como substantivos, verbos, advérbios, adjetivos, entre outros, são identificados a partir do POS Tagging. Com isso, é possível criar uma representação estruturada dos dados, tal como uma rede semântica, facilitando a busca e recuperação de informação.

Reconhecimento de entidades nomeadas (Named Entity Recognition - NER)

Na perspectiva de desenvolvimento de um sistema para a estruturação dos dados de publicações oficiais, os desafios estão na escassez de ontologias, conjuntos de regras, dicionários, léxicos, ou mesmos dados textuais devidamente rotulados em língua portuguesa. Assim, existe a necessidade de se investigar técnicas de aprendizado de máquina, destacando os modelos de classificação e detecção de entidade nomeada.

Em uma tarefa de classificação de dados textuais, tem-se o objetivo de atribuir uma categoria adequada para uma sentença ou documento de texto. Já na tarefa de NER, o objetivo é identificar entidades em textos, como pessoas, localizações, organizações, tempo, cargos e respectivos ocupantes. Tal técnica já foi anteriormente empregada em textos forenses para capturar entidades como cortes, título do processo, tipo de documento e juízes [7, 8]. Ainda, há corpus para treinamento de NER em português brasileiro com documentos jurídicos nacionais [9] com marcação de expressões referentes a leis e decisões judiciais. Por outro lado, pode ser interessante a ampliação dos dados existentes para identificação e reconhecimento de outras categorias de entidades que possam ser de interesse.

É importante observar que a tarefa realizada no contexto do NER é frequentemente usada como um passo inicial em sistemas de recuperação e extração de informação, resolução de referências, respostas de questionários, desambiguação de termos e modelos de tópicos. Outra vantagem é que os modelos atuais de NER não demandam conhecimento específico do domínio, como lexicons e ontologias, diminuindo o custo da elaboração e estruturação de recursos. Recentemente, vários sistemas baseados em redes neurais se mostraram eficientes e com resultados que se destacam como o estado-da-arte [4, 10].

Ligação de entidades nomeadas (Named Entity Linking – NEL)

Uma vez extraídas as entidades, deve-se relacioná-las com uma base de conhecimento que possa identificá-las para o uso correto dessas informações. Para isso, tem-se o problema de "Ligação de Entidades Nomeadas" (NEL) [11]. Formalmente, este problema é definido da seguinte forma. Dada uma base de conhecimento com um conjunto de entidades E e uma coleção de texto com menções a entidades M previamente extraídas, o objetivo é criar um mapeamento entre as $m \in M$ para suas correspondentes $e \in E$ na base de conhecimento. Aqui, uma menção a entidade é um conjunto de caracteres m que potencialmente se refere a uma entidade já determinada e descrita na base de conhecimento. É possível que uma entidade extraída não seja especificada na base de conhecimento, sendo necessário determinar a entidade como não mencionável.

O problema de ligação de entidades nomeadas é similar ao problema de desambiguação de palavras. O problema de desambiguação de palavras é determinar o sentido da palavra (sentido de uma palavra em vez de uma entidade nomeada) dado um tesauros ou dicionário. Note que o problema de ligação de entidades não requer um dicionário completo de significados, o que nem sempre é encontrado em uma base de conhecimento.
Após a extração das entidades nomeadas, um sistema para determinar a ligação de entidades deve realizar os seguintes procedimentos:

- **Geração das entidades candidatas:** Para cada menção de entidades $m \in M$ extraídas do texto, o sistema de ligação de entidade deve filtrar as entidades candidatas da base de conhecimento. Aqui, o desafio é, dado o conjunto de entidades definidas na base de conhecimento, deve-se filtrar as possíveis entidades que podem estar relacionadas com aquelas descritas no texto. Para esse objetivo, existem várias técnicas, como aquelas baseadas em dicionários ou técnicas de mecanismos de buscas.
- **Ranque de entidades candidatas:** Em muitos casos, para cada menção de entidades $m \in M$ existe mais de uma relacionada, demandando a criação de um ranque de entidades mais relacionadas. Para isso, abordagens baseadas em aprendizado supervisionado e não supervisionado podem ser modeladas para a criação desse ranque.
- **Detecção de entidades não mencionáveis:** Aqui, a tarefa é identificar entidades que estão presentes no texto, mas não são definidas na base de conhecimento. Para isso, faz-se o uso de técnicas baseadas em aprendizado de máquinas supervisionado.

Apesar das técnicas para o problema de Ligação de Entidades Nomeadas serem bem exploradas na literatura [11], existe uma demanda para a construção e povoamento de dados em uma base de conhecimento de domínio específico – domínio específico como definições jurídicas, descrição de corporações, produtos, etc.

Mineração de Dados (Data Mining)

Mesmo com um sistema adequado para extração de informações e entidades, a grande quantidade de textos publicados em diários oficiais torna a análise manual bastante dispendiosa. É difícil criar relações, ou mesmo identificar potenciais uso de entidades extraídas do texto. Assim, é necessário um processo automático para a descoberta de conhecimento implícito e potencialmente relevante nos dados.

A grosso modo, a tarefa de mineração de dados pode ser beneficiada pelo correto sistema de extração de entidades. Para isso, considere o problema de mineração de correferência entre documentos. Correferência entre documentos ou segmentos de texto ocorrem quando a mesma entidade nomeada é mencionada em mais de uma fonte de texto. Essa tarefa é importante para auxiliar usuários a examinar informações sobre uma entidade particular de múltiplas fontes de texto ao mesmo tempo, identificando as dependências entre as entidades que ocorrem em toda a coleção de textos [12, 13, 14]. Dada a correferência entre documentos, outras aplicações são a elaboração automática de sumários e a fusão de informações [15].

Nas tarefas de mineração de dados, principalmente em dados no formato textual, é necessário estruturar esses dados que estão descrito na forma de linguagem natural. Estruturar automaticamente dados não estruturados está fortemente relacionado com as tarefas de agrupamento e classificação. Técnicas automáticas para agrupar e classificar dados textuais são as tarefas mais importantes na área de aprendizado de máquinas. Modelos de tópicos é uma abordagem que foi aplicada com sucesso nessas tarefas, sendo seu principal objetivo descobrir dimensões latentes de um corpus [16]. Os tópicos são os assuntos tratados em uma coleção de documentos e são extraídos automaticamente, ou seja, tópico é definido como um conjunto de palavras que frequentemente ocorrem em documentos semanticamente relacionados. Esses conjuntos de palavras (que definem os tópicos) são obtidos por um processo de pós-processamento realizado a partir das dimensões latentes descobertas pela aplicação dos métodos de modelo de tópicos [17, 16, 18, 19]. A aplicação dessas técnicas é útil para sistemas de exploração, busca, recomendação e correlação entre uma enorme quantidade de documentos de textos. Aqui, é interessante não apenas aplicar e avaliar essas técnicas, mas também investigar a combinação de entidades extraídas com modelos de organização de coleções de dados textuais [20].

Já em uma perspectiva local da coleção de textos, onde se analisa apenas um documento (ou um trecho do texto de uma publicação oficial), é interessante definir categorias específicas e que auxiliem na classificação de trechos importantes no texto. Por exemplo, suponha que, dado um parágrafo de texto em uma seção do diário oficial, a estratégia é identificar se seu conteúdo se refere a concessão de aposentadoria. Para isso, define-se um problema de classificação supervisionado [21, 22]. Neste tipo de problema, o objetivo é identificar o conjunto de categorias de um documento observado. Esse processo exige uma fase de treinamento, em que um algoritmo de aprendizado de máquinas irá induzir um modelo de classificação a partir de dados previamente rotulados. Note que é possível definir várias categorias correlacionadas ao negócio e ao conteúdo no qual se queira analisar, porém, é necessário ter dados corretamente rotulados.

A tarefa de classificação de texto é normalmente conduzida por um algoritmo de aprendizado de máquina, no qual o objetivo é induzir um modelo para classificar documentos (ou blocos de textos) ainda não conhecidos. Para isso, um considerável número de documentos rotulados é necessário para criar um modelo de classificação o mais acurado possível. Entretanto, um consistente conjunto de textos rotulados para induzir um modelo de classificação não está disponível na maioria das aplicações reais, desde que o processo de categorização manual (rotulação) dos documentos é caro e trabalhoso, exigindo consumo de tempo de pessoal especializado no domínio da aplicação. Assim, uma forma mais prática de realizar a classificação de texto é aplicar métodos que fazem o uso de uma pequena quantidade de documentos rotulados. Assim, propõe-se investigar técnicas semi-supervisionadas indutivas e transdutivas. Nas técnicas de aprendizado semi-supervisionado indutivo, o modelo é induzido a partir de um conjunto de exemplos contendo ambos dados rotulados e não rotulados. A vantagem é que não é necessário utilizar uma grande quantidade de dados rotulados para induzir o modelo. Já nas técnicas de aprendizado semi-supervisionado transdutivos, não existe uma fase de treinamento, fazendo com que o processo de aprendizado seja realizado a partir de instâncias rotuladas para instâncias não rotuladas [23, 24]. Assim, com a aplicação de técnicas semi-supervisionadas, pretende-se diminuir o custo de rotular dados textuais.

A mineração de textos pode ajudar na organização, extração e estruturação de informações de notas oficiais. De forma manual não seria possível realizar esse nível de análise devido ao grande volume de dados e relações existentes, além da falta de disponibilidade de tempo e pessoal treinado para realizar os diversos trabalhos. Por isso, a investigação de métodos apropriados e específicos ao domínio da aplicação são importantes. Como foi anteriormente discutido nesta seção, todo o processo de examinar dados textuais para reunir informações valiosas é complexo, dependendo de técnicas algorítmicas de aprendizado de máquina, estatística e processamento de linguagem natural. Além disso, o desenvolvimento e pesquisa de técnicas de mineração de textos ainda permanece como tópico de pesquisa valioso para comunidade de aprendizado de máquina.

Referências

[1] R. Smith. An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ICDAR '07, pages 629–633, Washington, DC, USA, 2007. IEEE Computer Society.

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[3] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. *CoRR*, abs/1707.07012, 2017.

[4] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics.

[5] Erick R. Fonseca, João Luís G Rosa, and Sandra Maria Aluísio. Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21(1):2, Feb 2015.

[6] Marcelo Rodrigues de Holanda Maia and Geraldo Bonorino Xexéo. Part-of-speech tagging of Portuguese using hidden Markov models with character language model emissions. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, 2011.

[7] Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. Named entity recognition and resolution in legal text. In *Semantic Processing of Legal Texts*, pages 27–43. Springer, 2010.

[8] Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedints of the 16th International Conference on Artificial Intelligence and Law (ICAIL)*, London, United Kingdom, June 2017. Preprint available from https://hal.archives-ouvertes.fr/hal-01541446.X

[9] Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. Lener-br: a dataset for named entity recognition in brazilian legal text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), pages 313–323, Canela, RS, Brazil, September 24-26 2018. Springer.

[10] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[11] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.*, 27(2):443–460, 2015.

[12] Amit Bagga. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL COLing)*, volume 1, pages 79–85, 1998.

[13] Javier Artiles, Julio Gonzalo, and Satoshi Sekine. The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64–69, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[14] Gideon Mann and David Yarowsky. Unsupervised personal name disambiguation. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 33–40, 2003.

[15] Vasudeva Varma, Praveen Bysani, Kranthi Reddy, Vijay Bharath Reddy, Sudheer Kovelamudi, Srikanth Reddy Vaddepally, Radheshyam N, Kiran Kumar N, Santhosh Gsk, and Prasad Pingali. Iiit hyderabad in guided summarization and knowledge base population.

[16] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[17] David M. Blei. Introduction to probabilistic topic models. In *In Communications of the ACM*, 2011.

[18] Thiago de Paulo Faleiros and Alneu de Andrade Lopes. On the equivalence between algorithms for non-negative matrix factorization and latent dirichlet allocation. In *ESANN 2016, 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium, April 26-29, 2016, Proceedings*, 2016.

[19] Thiago de Paulo Faleiros. *Propagation in bipartite graphs for topic extraction in stream of textual data*. PhD thesis, University of São Paulo, Brazil, 2016.

[20] Xianpei Han and Le Sun. An entity-topic model for entity linking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 105–115, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[21] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[22] Rafael Geraldeli Rossi, Alneu de Andrade Lopes, Thiago de Paulo Faleiros, and Solange Oliveira Rezende. Inductive model generation for text classification using a bipartite heterogeneous network. *Journal of Computer Science and Technology*, 29:361–375, 2014.

[23] Thiago de Paulo Faleiros, Rafael Geraldeli Rossi, and Alneu de Andrade Lopes. Optimizing the class information divergence for transductive classification of texts using propagation in bipartite graphs. *Pattern Recognition Letters*, 87:127–138, 2017.

[24] Rafael Geraldeli Rossi, Alneu de Andrade Lopes, and Solange Oliveira Rezende. Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts. *Information Processing & Management*, 52:217–257, 2016.

[25] T. E. de Campos and R. M. Cesar-Jr. Advances on feature selection techniques with applications to face recognition. In *Anais do Concurso de teses e dissertações da Sociedade Brasileira de Computação (CTD-SBC)*, Florianopolis-SC, Brazil, July 2002. Preprint available at http://www.robots.ox.ac.uk/~teo/ctd_sbc/.X

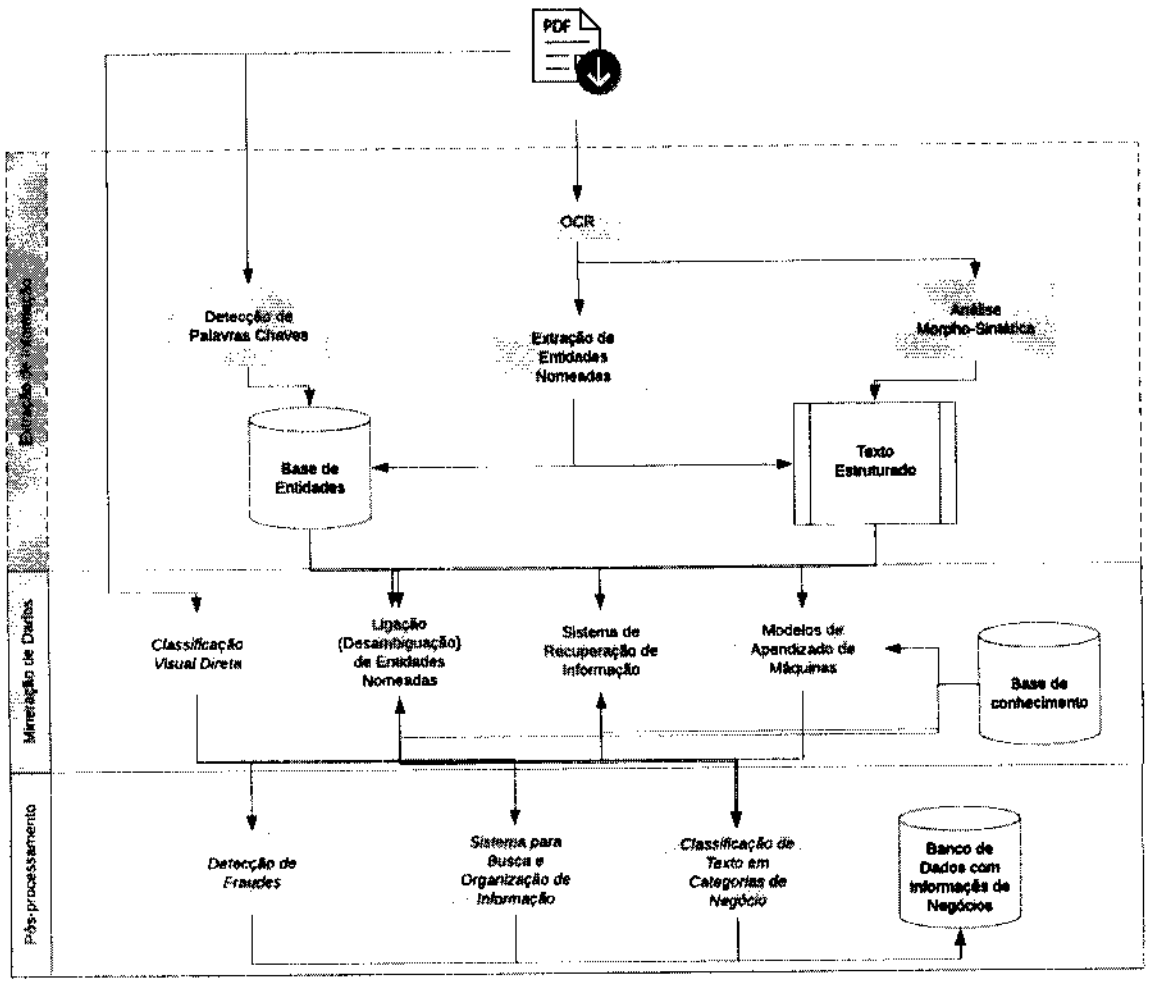
[26] Graham Roger Jones, Benjamin James Hammett, and Teofilo Emidio de Campos. Method of and apparatus for processing image data for display by a multiple-view display device, 2008. US Patent App. 12/523,450.

[27] Nazli FarajiDavar. *Transfer Learning for Computer Vision*. PhD thesis, University of Surrey, Guildford, UK, 2015. Supervised by Teófilo de Campos, available from http://www.cic.unb.br/~teodecampos/TransferLearning/.X

[28] Moacir Ponti, Josef Kittler, Mateus Riva, Teófilo E. de Campos, and Cemre Zor. A decision cognizant Kullback–Leibler divergence. *Pattern Recognition*, 61:470–478, 2017. DOI:10.1016/j.patcog.2016.08.018.X

[29] Frederigo Guth and Teófilo E. de Campos. Skin lesion segmentation and classification – UnB entry at the ISIC challenge. In *ISIC Challenge on Skin Lesion Analysis Towards Melanoma Detection*. International Skin Imaging Collaboration, 2018. https://challenge2018.isic-archive.com/.X

III. MÉTODO



- **Descrição das atividades/etapas a serem desenvolvidas;**

1. Preparação da equipe e das ferramentas:
   a. recrutamento de equipe;
   b. aquisição de equipamentos;
   c. aquisição de dados crus.
2. Processamento de documento para extrair blocos de texto de maneira coerente e identificar elementos de indexação, tais como nomes de seções, listas de itens etc.

   Essa é uma tarefa centrada em desenvolvimento de software. Seu foco é de desenvolver sistemas baseados em expressões regulares e outras heurísticas para se extrair informação especificamente dos objetos de trabalho. Serão implementadas, por exemplo, regras para a extração de dados textuais e de indexação desenvolvidas especificamente para o DODF. Essa tarefa se iniciará logo no início do projeto.
3. Frentes de pesquisa:
   a. Detecção visual de palavras chave
   b. Classificação visual de páginas ou blocos de documentos

c. Análise morfo-sintática e reconhecimento de entidades nomeadas
d. Ligação de entidades nomeadas
e. Busca, recuperação e organização de informação
f. Detecção de fraudes
g. Classificação de textos

Os itens acima envolvem desafios ativos de pesquisa, os quais serão desenvolvidos em paralelo e todos seguirão um cronograma de acordo com os itens abaixo:

A. Estudo da literatura científica e atualização com relação ao estado-da-arte. Essa tarefa deverá ser continuamente ativa, principalmente para os membros com mais senioridade no projeto. Haverá um ciclo regular de reuniões de grupos de leitura, em que a cada reunião, um dos membros do projeto traz para a discussão algum artigo científico recente ou uma técnica que seja do interesse de múltiplos membros do projeto.
B. Desenvolvimento de sistema de base:
    a. Implementação e avaliação de sistemas simples (ou mesmo "ingênuos") para cada frente de trabalho, de forma que possamos tenhamos um ponto de partida para nos familiarizarmos com detalhes dos dados e seus desafios.
    b. Implementação de métodos do estado-da-arte para cada frente de trabalho. Tomando por base os resultados mais recentes publicados cientificamente, a proposta desta tarefa é de implementar, para cada frente de trabalho, o sistema mais recente ou mais promissor. Tais métodos deverão ser avaliados para identificarmos seus pontos fortes e pontos fracos.
C. Proposição de novos métodos:
    Criação de novos métodos de aprendizado de máquina para análise visual de documentos e extração de informação estruturada de texto. Os resultados obtidos na fase anterior nos guiarão a propor novas vias de pesquisa a serem exploradas para tentar obter melhorias em relação a métodos recentes.
D. Experimentos e avaliações: Bateria de experimentos nas bases de dados adquiridas a partir do Diário Oficial do DF e também bases de dados públicas para comparação com outros métodos a nível internacional.
E. Elaboração de artigos para publicação:
    a. Desenvolvimento de experimentos complementares visando a conclusão dos trabalhos e a escrita de artigos para publicações científicas a nível internacional.
    b. Escrita e revisão dos artigos.

4. Desenvolvimento de protótipos de software que demonstram a aplicação das ferramentas desenvolvidas em problemas práticos dos órgãos do DF.
    O foco desta tarefa é o desenvolvimento de protótipos que sejam intuitivos e de fácil adoção por usuários comuns. Ao final de cada ano do projeto haverá a entrega de um protótipo.

5. Análise e gerenciamento de riscos das etapas do projeto, tais como, cronograma de atividades, recursos humanos, custos.

- **Procedimentos e/ou instrumentos a serem utilizados**

Dados

A lei brasileira facilita o acesso a informações oficiais. O diário oficial do DF, da União e de outros estados estão disponíveis pela Internet. Porém, o acesso a grandes quantidades de edições dessas publicações muitas vezes é limitado por uma interface que foi projetada justamente para dificultar o trabalho de "robôs".

Esse será o maior desafio da tarefa (1.c) acima. Entretanto, há um grande potencial de que possamos estabelecer contato direto com o órgão responsável pela publicação do DODF, facilitando esse acesso.

Além disso, planejamos utilizar bases de dados públicas internacionais para avaliar os métodos propostos para diversas tarefas deste projeto. A importância do uso de bases públicas se intensifica quando se deseja publicar artigos científicos que comparem nossos métodos com outros métodos propostos internacionalmente.

Outra fonte de dados relacionada é o projeto Victor, pois seus membros indicaram intenção de tornar toda sua base de dados pública. Vide http://gpam.unb.br/victor/

Recursos computacionais e infra-estrutura
Além dos recursos requeridos nesta proposta, este projeto contará com os recursos descritos na seção de "Equipamentos e materiais".

IV. RECURSOS ENVOLVIDOS

EQUIPE
Além do coordenador, a equipe é composta pelos seguintes pesquisadores:

Thiago de Paulo Faleiros (CV lattes http://lattes.cnpq.br/1193412523364471): Possui graduação em Ciência da Computação pela Universidade Federal de Goiás (2007), mestrado em Ciência da Computação pela Universidade Estadual de Campinas (2010) e doutorado em Ciência da Computação pela Universidade de São Paulo (2016), com período sanduíche na University of Maryland, EUA. É professor/pesquisador do Departamento de Ciência da Computação da UnB. No mestrado trabalhou com problemas de otimização combinatória, em específico, com algoritmos para o problema de particionamento em grafos. Já no doutorado, trabalhou na área de aprendizado de máquina, pesquisando soluções algorítmicas que utilizam representação de texto na forma de grafos. Na tese de doutorado, com o título "Extração de Tópicos em fluxo de documentos textuais", foram desenvolvidos algoritmos para o aprendizado supervisionadas, não supervisionadas e semi-supervisionadas em dados no formato textual. Entre os projetos no qual o pesquisador trabalhou, destacam-se: 1) Participação como bolsista de Treinamento Técnico e apoio a pesquisa nível 3 (TT III) no Projeto PIPE FAPESP LinkDigger-* n.03/07968-9. O LinkDigger é sistema que provê um serviço de criação automática de ligações entre documentos textuais; 2) Participação como bolsista de estágio de pesquisa no exterior no Projeto BEPE FAPESP n. 2013/15353-6. No projeto foram aplicadas metodologias para avaliação do modelo STM (Syntactic Topic Model). O STM é um modelo Bayesiano não paramétrico de documentos analisados sintaticamente; 3) Participação como bolsista de Treinamento Técnico e apoio a pesquisa nível 5 (TT V) no projeto e-SHARE Miner -- gerenciamento de informação apoiado pela descoberta de conhecimento via taxonomia de tópicos (projeto FAPESP n. 2016/19295-9). Logo, dada a breve descrição dos projetos de pesquisa no qual o pesquisador participou, ressalta-se sua experiência na área de Aprendizado de Máquinas, atuando principalmente na investigação de técnicas aplicadas em dados não estruturados em formato textual.

Vinicius Ruela Pereira Borges (CV Lattes: http://lattes.cnpq.br/1841593572448050) é Professor Adjunto no Departamento de Ciência da Computação (CIC) da Universidade de Brasília (UnB), campus Darcy Ribeiro. Obteve o grau de Bacharel (2009) e de Mestrado (2011) em Ciência da Computação pela Faculdade de Computação da Universidade Federal de Uberlândia. Obteve o título de doutor (2016) em Ciências da Computação e Matemática Computacional no Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP), em São Carlos - SP. Realizou doutorado sanduíche (2014-2015) na University of California, Davis (UC Davis), em Davis, Califórnia, Estados Unidos. Foi Professor Substituto