Versão preliminar (em elaboração)

Plano de trabalho do NIC.br na área de Inteligência Artificial

O Aprendizado de Máquina *(Machine Learning)* é uma categoria de aplicações da área de Inteligência Artificial que permite aos algoritmos aprenderem enquanto processam dados de treinamento. O avanço destas técnicas estabelece uma nova abordagem para a própria computação, na qual deixa de ser necessário programar e codificar manualmente todas as possíveis decisões, uma vez que os algoritmos passam a ser capazes de identificar padrões a partir de exemplos para construir uma lógica própria de decisão.

A evolução das técnicas de *machine learning* está trazendo uma nova perspectiva no desenvolvimento de produtos e serviços - nas mais diferentes áreas - que até então não eram factíveis com a programação tradicional. Carros autônomos e reconhecimento facial são duas aplicações que ganharam notoriedade nos últimos anos justamente pelos avanços das pesquisas de aprendizado de máquina. É muito difícil para um cientista desenvolver um programa para reconhecer o rosto de uma pessoa e suas variações (uso de bonés, barba, maquiagem, entre outras), mas um algoritmo é capaz de fazê-lo, com uma acurácia superior a 95%, utilizando fotos do indivíduo como exemplos para treinamento. Situações parecidas também são encontradas na medicina, como em um projeto do Laboratório de Inteligência Artificial do MIT que utilizou 90 mil mamografias para treinar um modelo de *machine learning* capaz de prever o câncer de mama com cinco anos de antecedência[1].

A inteligência artificial é um assunto em voga, porém amplo e com aplicações nas mais diversas dimensões da sociedade: saúde, transporte, segurança, emprego, entretenimento, emprego, só para citar algumas. Neste plano de trabalho não temos como objetivo contemplar a Inteligência Artificial em toda as sua amplitude e complexidade, mas pretendemos limitá-lo a temas pertinentes no campo da governança da Internet.

Um recorte possível é entender como os sistemas de aprendizado de máquina podem modificar o fluxo informacional e as dinâmicas sociais em rede. Hoje os serviços oferecidos na Internet utilizam essas técnicas para entender melhor o comportamento e preferências de seus usuários para aperfeiçoar a experiência de uso por meio de personalizações, retendo assim a atenção das pessoas e, consequentemente, o tráfego dentro de seus ambientes. A proposta dos grandes *players* parece tentadora aos usuários: use nossos serviços à vontade para que ele aprenda a lhe entregar o que você realmente deseja, sem que tenha que pagar qualquer valor monetário por isso. Exemplos como sistemas de recomendação são implementados por serviços de *streaming*, *e-commerce* e redes sociais, que usam diferentes técnicas para conhecer as nossas preferências e indicar uma música, um vídeo, um produto ou uma notícia que julguem de maior relevância para nós, reduzindo o nosso agenciamento, por exemplo. No mesmo sentido, algoritmos de publicidade online oferecem anúncios tão assertivos que chegam até nos assustar devido seu grau de assertividade.

1 http://news.mit.edu/2019/using-ai-predict-breast-cancer-and-personalize-care-0507

É por meio da coleta de dados que as empresas de tecnologia se fortalecem na economia digital e criam modelos de negócios orientados às decisões automatizadas em diferentes níveis, que passam por um simples classificador de imagem em um buscador, uma publicidade mais segmentada em uma rede social, ou até mesmo um sistema de reincidência criminal. Um algoritmo de aprendizado de máquina é capaz de aprender padrões que mesmo para um especialista seria muito difícil, e com isso passa a tomar decisões que nem mesmo nós podemos entender o porquê. Esta situação em que os sistemas aprendem a tomar decisões a partir de dados está estimulando um debate global sobre governança, ética, transparência, *accountability* e até mesmo o agenciamento humano. Esse assunto vem sendo discutido em foros internacionais, como Unesco, OCDE, ITU, IGF, ACM, IEEE, entre outros. Nesse contexto podemos apontar algumas áreas para projetos e pesquisas no campo da Internet, como a disponibilidade e qualidade de dados para treinamento e a governança dos sistemas inteligentes.

Os dados são os ativos para os projetos de inteligência artificial. Sem eles os algoritmos não aprendem - ou aprendem errado se os dados não tiverem qualidade. Quem os controla está em vantagem. As organizações que lideram o cenário de desenvolvimento de IA possuem seus modelos de negócios orientados à Internet, e exploram esse ambiente como uma fonte para o treinamento de seus algoritmos. Em alguns casos essas organizações coletam dados disponíveis em diversos formatos - imagens, publicações em redes sociais, preferências, *likes* - muitas vezes sem que os seus donos tenham ciência que eles existam. Em outros, elas desenvolvem mecanismos para que os usuários possam produzir dados para abastecer ainda mais o seu acervo, como é o caso da plataforma *Crowdsource* do Google, que por meio de uma proposta gamificada incentiva os usuários da rede a gerar dados nos mais diferentes formatos para treinamento.

Nesse sentido, existem desafios e oportunidades para que sejam investigados os processos de geração, coleta e controle de dados no campo da Internet para treinamento de inteligência artificial. A nossa hipótese inicial é que existe uma assimetria na distribuição de dados, o que causa um desequilíbrio no desenvolvimento tecnológico e no empoderamento social. Percebe-se, atualmente, que as organizações do Norte extraem dados do Sul global em uma dinâmica em que a região se limita apenas ao papel de consumidor de tecnologia. Uma das dificuldades enfrentadas por pesquisadores e empresas brasileiras é o acesso a dados para treinamento de seus sistemas, os quais continuam trancados em outro hemisfério. Nem mesmo os conteúdos coletados pelo Google *Crowdsource* são disponibilizados como dados abertos. Outra questão não menos importante é que por meio dessa lógica as organizações do Norte começam a entender a nossa dinâmica social melhor do que nós mesmos, em atividades online e offline. A Prefeitura de São Paulo, por exemplo, teve que estabelecer uma parceria com o Waze para conseguir entender o trânsito em tempo real da própria cidade.

Inteligência artificial, governança algorítmica e economia dos dados são assuntos presentes no discurso global por apresentar oportunidades e desafios. Por esse motivo, entendemos ser oportuno que a temática seja discutida no campo da governança da Internet por não ser um assunto que se esgote no plano técnico, mas que demanda debates aprofundados sobre privacidade,

ética, transparência e segurança. Este texto de introdução não dialoga e nem esgota todas as intersecções possíveis entre Inteligência Artificial e Internet, mas busca apresentar motivações e justificativas iniciais para que o assunto seja aprofundado. Existe clareza que o tema é amplo e que precisa ser tratado com foco. Nesta perspectiva, este documento apresenta uma proposta inicial de um plano de trabalho que será desenvolvido por uma iniciativa interdepartamental do NIC.br, levando em consideração o escopo de atuação da instituição.

O que já foi feito pelo NIC.br

O tema de Inteligência Artificial e *Data Science* está sendo estudado, pesquisado e aplicado em projetos em andamento há pelo menos dois anos por alguns departamentos do NIC.br. Abaixo há uma relação de atividades, publicações e iniciativas já executadas por cada departamento.

Ações do Ceweb.br

As iniciativas do Ceweb.br relacionadas ao temas Inteligência Artificial têm como foco estudos sobre o impacto de IA no desenvolvimento de tecnologias Web. Desenvolveram-se pesquisas, experimento e construção de debates relacionados a: fundamentos de Inteligência Artificial; ética e não discriminação; Design centrado nas pessoas e dados abertos; processamento de linguagem natural e *machine learning* na Web.

1. Pesquisa e elaboração do capítulo *"Fairness and Non-Discrimination"* do livro *"Responsible AI - A Global Policy Framework*", publicação global do ITechLaw. Mais informações: https://www.itechlaw.org/ResponsibleAI

2. Elaboração do Workshop *"Student-centered Design for AI Projects"* para o simpósio Mobile Learning Week 2019, da Unesco. Mais informações em: https://en.unesco.org/sites/default/files/mlw2019-programme.pdf

3. Proposta do Workshop *"Human-centered Design and Open Data: how to improve AI"* aprovada para o Internet Governance Forum 2019. Mais informações: http://www.intgovforum.org/multilingual/content/igf-2019-ws-179-human-centered-design-and-open-data-how-to-improve-ai

4. Proposta do Workshop *"Design centrado nas pessoas e dados abertos na Web: inclusão e ética na Inteligência Artificial"* aprovada para o Fórum de Governança da Internet 2019. Mais informações em: https://minhaagenda.nic.br/api/submissao/74

5. Elaboração de pesquisas e experimentos sobre ética e vieses em Processamento de Linguagem Natural (em andamento)

6. Co-organização do Workshop "Latin-america Web - LA-WEB 2019", que aconteceu durante a The Web Conference 2019. Mais informações em: https://www2019.thewebconf.org/media/TheWebConf2019-Companion.pdf

7. Participação no grupo de trabalho de "*Machine Learning on the Web*" do W3C. Mais informações em: https://www.w3.org/community/webmachinelearning/

8. Oficina de *"Fundamentos de Inteligência Artificial"* durante o Fórum de Governança da Internet 2018. Mais informações em: https://forumdainternet.cgi.br/2018/programacao/detalhe/1080/

9. Oficina sobre *"Inteligência Artificial, Design e Ética"* para o Centro de Tecnologia de Informação - CTI Renato Archer

Ações do Cetic.br

As iniciativas do CETIC.br relacionadas aos temas *Data Science*, Big Data/Analytics e IA se inserem no contexto de medição e produção de dados estatísticos. Neste sentido, as iniciativas que vem sendo desenvolvidas pelo Departamento visam inovar o processo de produção de dados buscando a redução de custos na produção de estatísticas, aumento da qualidade e uso de fontes alternativas de dados, sejam dados estruturados, como dados administrativos, ou dados desestruturados como dados provenientes de web scrapping.

Adicionalmente, o Cetic.br vem introduzindo este tema nos cursos de capacitação em metodologia de pesquisas, nas publicações voltadas aos usuários de dados e participação em fóruns de definição de indicadores como a UIT, OCDE, UNSD, CEPAL, UNESCO e UNICEF.

a) Produção de dados e estatísticas TIC e Fóruns de Indicadores:

1. Condução de estudos e aplicação de técnicas estatísticas de regressão e modelagem, fundamentais em algoritmos de aprendizado de máquina. Atualmente, os resultados desses estudos estão sendo utilizado na modelagem de dados para construção de indicadores de e-commerce da Pesquisa TIC Empresas, nos estudos de dados do SIMET (publicação Banda Larga no Brasil); dados de tráfego do IX.br; e no projeto de cooperação com o IBGE/CEPAL para uso de dados do Registro.br.
2. Acordo de cooperação com o IBGE no âmbito do projeto de produção de indicadores de comércio eletrônico utilizando fonte de big data privado (DataProvider) e dados do Registro.br e do Cadastro de Empresas do CEMPRE (IBGE) utilizando metodologias de record linkage e regressão logística.

3. Adequação do questionário da pesquisa TIC Empresas alinhado ao novo questionário da Eurostat que trata da medição de novas tecnologias nas empresas, inclui temas como: Robótica, AI, IoT, Computação em Nuvem, Big Data/Analytics. Para todos esses temas foram realizados testes cognitivos relativos aos novos indicadores e a nova versão do questionário encontra-se em campo neste momento.

4. Participação nos fóruns de discussão de Indicadores: UIT (EGH/EGTI), onde o Cetic.br coordenou durante 6 anos o grupo de indicadores e

acompanhou uma das ações ligados ao uso de fontes de big data (mobile data – CDRs) para a construção de indicadores TIC de acesso e uso da Internet; OCDE no grupo de trabalho de indicadores para a medição da economia digital (WP-MADE); e grupos ad hoc da UNESCO e UNICEF que tratam de medição de AI em educação e uso de plataformas digitais por crianças e adolescentes.

5. Pesquisa: estudo exploratório sobre o uso de inteligência artificial entre prefeituras, no âmbito da pesquisa Governo Eletrônico (em andamento).

6. Participação no fórum da União Internacional de Telecomunicações (UIT) “AI for Good - Global Summit”. Mais informações em: https://aiforgood.itu.int/

7. Participação no grupo de trabalho “AI and Child’s Rights” liderado pelo UNICEF, em parceria com o IEEE Standards Association, Berkman Klein Centre for Internet & Society e o World Economic Forum. O Cetic.br participou no workshop “Towards Global Guidelines on Artificial Intelligence and Child Rights” e fez a apresentação: “Kids Online Survey in Brazil: Multistakeholder engagement strategies”. Mais informações em: https://ai4children.splashthat.com/

8. Participação nas conferências da UNSD (United Nations Statistical Division) sobre o uso de big data para a produção de estatísticas oficiais – UN Big Data Conference. Em 2017 o Cetic.br foi responsável pelo keynote speech em Bogotá e apresentou o projeto piloto de produção de indicadores de comércio eletrônico por meio de web scrapping de dados na web.

9. Participação no Congresso Mundial de Estadística (World Statistics Congress – International Statistics Institute) em trilhas sobre data science e big data nos anos de 2015, 2017 e 2019 com apresentação de trabalhos.

b) Publicações e capacitação abordando as temáticas de *Data Science*, Big Data e Inteligência Artificial:

10. Publicação: Panorama Setorial da Internet: edição sobre Inteligência Artificial.
Mais informações em: https://www.cetic.br/publicacao/ano-x-n-2-inteligencia-artificial-e-etica/

11. Publicação: autoria de capítulo de livro sobre IA e juventude:
Adib, Luisa; Senne, Fabio. "Plataformas digitales y aprendizaje: indicadores sobre el acceso, actividades y habilidades digitales de niños y adolescentes en Brasil". In: Brossi, Lionel; Dodds, Tomás. Inteligencia Artificial y Bienestar de las Juventudes en América Latina. LOM Ediciones. 2019.

12. Publicação: tradução para o português da publicação 'UN eGovernment Survey 2018', que contempla um capítulo próprio dedicado ao tema Inteligência Artificial.
Disponível em:
https://publicadministration.un.org/egovkb/Portals/egovkb/Documents/un/2018-Survey/E-Government%20Survey%202018_Portuguese.pdf

13. Capacitação Online:
MOOC "Tech for Good: The Role of ICT in Achieving the SDGs", em parceria com Unesco e SDG Academy, inclui um módulo sobre IA.
Mais informações em: https://www.edx.org/course/tech-for-good-the-role-of-ict-in-achieving-the-sdgs-2

14. Capacitação Presencial:
9o Worskhop NIC.br de Metodologia (2019): “Data for public statistics: Data Science, Big Data & Artificial Intelligence”.
Mais informações em: https://workshop.metodologia.cetic.br/

15. 7o Worskhop NIC.br de Metodologia (2017): “Integration of different data sources for measuring the SDGs”; “State of affairs on SDG monitoring through Big Data”; “Political economy of Big Data for the SDGs”; and “Investing in collabora
Mais informações em: https://cetic.br/semana-metodologias-pesquisas/eventos-anteriores/2017/

16. Realização do curso em parceria com a CEPAL e Data-Pop Alliance (2017): “Global Professional Training Programme on Big Data for Measuring the Digital Economy”.
Mais informações em: https://cetic.br/semana-metodologias-pesquisas/eventos-anteriores/2017/

17. Colaboração no evento “AI LATIN AMERICA SUMMIT”, organizado pelo Massachusetts Institute of Technology, MIT.
Mais informações em: http://ailatinsum.mit.edu/

Ações do Ceptro.br

1. Apresentação no *FORUM ON ARTIFICIAL INTELLIGENCE IN AFRICA* (Marrocos 2018)
   High-Level Plenary Session: “What future for AI in Africa?”
   Título: *DEVELOPING COMPREHENSIVE AI AND INTERNET GOVERNANCE SYSTEM: THE CASE OF BRAZIL*
   Autores: Murilo Vieira Komniski (Governo Brasileiro)
   Paulo Kuester Neto (NIC.br)
   https://en.unesco.org/artificial-intelligence/africa-forum

2. Apresentação no Mobile Learning Week (UNESCO Paris 2018)
   Título: *Mapping a national public policy of digital skills development in Brazil*
   Autores: Daniela Costa (CETIC.br)

Paulo Kuester Neto (CEPTRO.br)

A apresentação foi mencionada como um destaque na publicação Skills for a connected world - Report of the UNESCO Mobile Learning Week 2018
http://unesdoc.unesco.org/images/0026/002658/265893E.pdf

Texto do Destaque: "National policies for ICT in education have to be constantly reviewed and adapted. Governments in developing countries can build partnerships with organizations and institutions to overcome the lack of initial governmental statistics and obtain systematic data to support and assess the implementation of public policies and the effectiveness of the programmes (see for example NIC.br)."

3. Acordo de cooperação entre o Ministério da Educação (MEC), o CEPTRO e o CETIC (NIC.br) no desenvolvimento de uma plataforma utilizando metodologias de Big Data e Clusterização Geográfica. A plataforma (link) foi desenvolvida tendo como objetivo o mapeamento do acesso, da conectividade e do uso das TIC em âmbito nacional, de acordo com as diretrizes da política educacional do Programa Inovação Educação Conectada. O sistema hoje já conta com mais de 9000 escolas públicas mapeadas, com previsão de crescimento de até 20.000 escolas até o final de 2019.

4. Desenvolvimento de parcerias internas no NIC.br para produção de trabalho científico a ser publicado usando técnicas de Text Mining em redes sociais.

5. Desenvolvimento de um novo Mapa de Qualidade de Internet a partir do Sistema de Medição de Tráfego Internet (SIMET), utilizando metodologias de Big Data e técnicas de Clusterização, Sistemas de Armazenamento indexados geograficamente e Correlação Espacial (Regressão Espacial - Índice de Moran). (em andamento)

6. Desenvolvimento de estudos aplicados de sistemas de computação distribuída, análises, usando técnicas de detecção de anomalias em fluxo em tempo real para construção de aplicações no âmbito de projetos do CEPTRO.br, a partir de dados do Sistema de Medição de Tráfego Internet (SIMET).

7. Estudos em andamento para a construção de sistemas de clusterização, usando distância Gaussiana e Euclidiana para mapeamento de uso das TIC em contextos diversos, tendo como fonte o Sistema de Medição de Tráfego Internet (SIMET).

8. Avaliação de modelos mistos (*Ensemble methods)* no âmbito da subárea de Machine Learning para detecção de comportamento no Sistema de Medição de Tráfego Internet (SIMET).

9. Desenvolvimento de trabalhos aplicados na área de Ciência de Dados para avaliação, visualização de grandes volumes de dados usando Spark/R e Python. Estatística Descritiva, Modelagem e Transformação de Dados, Enriquecimento de variáveis de contexto e processamento em sistemas de pipeline.

O que será feito pelo NIC.br

Considerando a reflexão exposta na Introdução deste documento, as atividades propostas para o NIC.br no tema Inteligência Artificial (IA) e *Data Science* têm como objetivos identificar as possíveis relações e impactos entre IA e:

- Infraestrutura da rede de Internet no Brasil;
- Privacidade e proteção de dados pessoais;
- Segurança da informação e do ambiente computacional;
- Aplicações na camada da Web;
- Medição de dados e produção de indicadores.

Esta proposta contempla atividades iniciais e resultados esperados que possam servir de subsídios para projetos de interesse do CGI/NIC.br. Abaixo há uma relação de iniciativas que planejamos para os próximos 18 meses.

Criação do Grupo de Trabalho Interdepartamental do NIC.br

A elaboração deste documento uniu pessoas de diferentes departamentos do NIC.br para a troca de informações e experiências com IA dos diferentes departamentos. Na primeira reunião foi possível perceber a sinergia entre as diferentes equipes, que se complementam com suas especialidades. Neste caminho, propomos a criação de um Grupo de Trabalho Interdepartamental do NIC.br para conduzir ações, pesquisas e iniciativas relacionadas à Inteligência Artificial, especialmente no que se refere às atividades a seguir deste documento.

Mapeamento das iniciativas internacionais e nacionais

Uma das primeiras atividades será mapear as principais iniciativas e organizações que têm como objetivo investigar ou fomentar questões sobre a governança de Inteligência Artificial. Esta ação nos permitirá ter uma visão ampla da discussão que está sendo conduzida no mundo sobre o tema, propor parcerias com organizações nacionais e internacionais, além de adquirirmos subsídios teóricos e técnicos para nossos trabalhos.

Fórum brasileiro de Inteligência Artificial

A organização de evento multissetorial, que agregue especialistas sobre Inteligência Artificial, com o intuito de debater sobre as implicações de IA para a internet brasileira. O primeiro Fórum brasileiro de Inteligência Artificial deverá ser realizado no dia 9 de dezembro, um dia antes do evento “Regional Forum on Artificial Intelligence in Latin America**”**, que reunirá especialistas internacionais e está sendo organizado pelo NIC.br, em parceria com o Governo Federal e a Universidade de São Paulo.
Além de proporcionar a criação de uma rede de especialistas liderada pelo NIC.br, o Fórum terá como resultados as gravações das discussões e uma publicação com análises sobre as questões debatidas durante o evento.

Criação de rede de especialistas

Antes, durante e depois da organização do evento estaremos em contato com especialistas nacionais e internacionais de diferentes especialidades e setores. A nossa proposta é criarmos um grupo com estes especialistas para que possamos manter vivo um debate sobre a temática com reuniões periódicas e produção de pesquisas e estudos.

Publicação de *Web Trends* baseada nas discussões do evento

Baseada nas discussões multissetoriais ocorridas no primeiro Fórum brasileiro de Inteligência Artificial, produziremos uma publicação com análises sobre as oportunidades e os desafios relacionados à Inteligência Artificial para a internet brasileira. Dentre as quais, relacionamos:

- Ética, Fairness, Não-discriminação e Gênero em IA
- Web e Inteligência Artificial (Privacidade, *Open Data, Fair Competition* e Transparência)
- Impacto de IA na infraestrutura da rede

Pesquisa e publicações sobre tecnologias emergentes de IA

Estudos e pesquisas sobre tecnologias emergentes de IA que tenham impacto na Internet e na Web. Inicialmente, a iniciativa estará focada em tecnologias de processamento de linguagem natural por ser um ramo da IA que permite analisar conteúdos publicados na rede (como publicações em redes sociais, páginas web, relatórios) para que se possa entender o livre fluxo de informações na Internet, ou para nossa compreensão dos problemas que surgem nessa área, como privacidade, segurança, censura, desinformação, entre outros.

Departamentos e pontos focais na iniciativa do NIC.br

Ceweb.br

- Vagner Diniz
- Caroline Burle
- Diogo Cortiz

Ceptro.br - medições

- Paulo Kuester Neto

Cetic.br

- Alexandre Barbosa
- Marcelo Pitta

Jurídico

- Kelli Angelini

Anotações das reuniões internas do NIC.br

13 de agosto de 2019

Participantes:

Caroline Burle - Ceweb.br
Diogo Cortiz - Ceweb.br
Kelli Angelini - Departamento Jurídico
Paulo Kuester Neto - Ceptro.br
Tatiana Jereissati - Cetic.br
Stefania Cantoni - Cetic.br

Paulo:
Evento Marrocos - IA na África
Como IA impactará a produtividade.
Internalizar a IA na educação dos países.
Ética.
Preocupação das grandes empresas guiarem as discussões.

Diogo:
Crowdsource
Discussão sobre quem tem o poder de processamento dos dados.
Empresas que lideram o processo: empresas norte-americanas e chineses
Assimetria de informação e dos dados
Desafios a serem superados

Kelli:
Fortalecer um grupo de especialistas no assunto
Trazer o lado positivo de IA, além do negativo
Estimular ética, transparência, privacidade para servir de modelo
Promover debate sobre impactos para o país

Tatiana:
Como medir IA é pauta nas discussões do Cetic.br
Foi feito o Panorama Setorial com essa temática
Mooc em parceria com a Unesco teve módulo de IA, ministrado por Sara Rendtorff-Smith (MIT - Policy Governance)
Houve discussão temática sobre IA na Semana de Metodologia do NIC.br, atrelada a Big Data
Pesquisadores associados ao MIT estão organizando evento regional sobre IA que acontecerá em janeiro/2020. Chegaram via CEPAL - podemos sugerir palestrantes e propor painel. A ideia é levar três representantes de cada país da América Latina e Caribe
Evento UNESCO faz parte de um ciclo de eventos regionais sobre IA

- Será em dezembro, provavelmente no auditório de RI da USP

Paulo:
Medições começaram a serem usadas para cruzar com outras áreas. Ex. Educação
Wifi livre da PMSP - monitoramento com portal dos dados de contexto

IA entra na quantidade de dados e entendimento de padrões dos contextos analisados
Avaliação de modelos de IA para entender contextos das medições e, posteriormente, fazer cruzamentos
Trabalharam com o olhar da LGP: fizeram matriz com as variáveis
15 ou 20 milhões de medições mês