MINISTÉRIO DA CIDADANIA

PÁTRIA AMADA BRASIL GOVERNO FEDERAL

Documento de Visão

PROBLEMA DE NEGÓCIO

HISTÓRICO

Data	Versão	Descrição	Autor
09/09/2019	1.0	Documento de definição do projeto contendo a contextualização e fundamentação da situação-problema.	Marcello S. Pinheiro
16/09/2019	1.0	Atualização de informação.	Marcello S. Pinheiro
19/09/2019	1.1	Inclusão de novos requisitos e contextualizações.	Marcello S. Pinheiro
30/09/2019	1.2	Revisão.	Marcello S. Pinheiro
02/10/2019	1.3	Detalhamento textual conforme solicitado pelo gestor do contrato Daniel Brasileiro.	Marcello S. Pinheiro
07/10/2019	1.4	Alterações para adequação ao Catálogo de Serviços.	Marcello S. Pinheiro
10/10/2019	2.0	Ajustes e correções.	Marcello S. Pinheiro

ÍNDICE

INTRODUÇÃO

O presente documento é o primeiro de oito produtos integrantes do projeto de Data Science e Data Engineering a ser desenvolvido no âmbito do Projeto Fala Cidadão do Ministério da Cidadania do Governo Federal. Os produtos são: 1) Documento de Visão; 2) Análise de Viabilidade; 3) Arquitetura Computacional e Modelo de Transcrição de Arquivos de Áudio para Arquivos Texto; 4) Documento de Avaliação da Qualidade do Treinamento e Otimização; 5) Repositório de Armazenamento dos Textos; 6) Análise Exploratória de Dados: estudos preliminares; 7) Documento do Modelo de Aprendizado Automático: proposta preliminar e 8) Documento Final de Projeto (juntada).

Tem por objetivo entender os processos de trabalho, as informações coletadas e como elas são analisadas e tratadas, as dificuldades no atendimento incluindo tecnológicas, enfim, tudo relacionado ao negócio no que se refere ao atendimento ao cidadão do referido Ministério.

Especificamente, consiste em entender as necessidades dos cidadãos quando procuram a Central de Atendimento via contato telefônico, acessar a base de dados que contém as informações pertinentes ao cidadão, mostrar as informações que o cidadão possui e as que ele necessita resolver, se for o caso, para continuar recebendo o auxílio governamental que fazem jus.

Nesse sentido, foram realizadas reuniões *in loco* na Central de Atendimento. Tais reuniões ocorreram nos dias 04, 05, 17 e 20 de setembro de 2019. Nessas reuniões foram identificados os requisitos funcionais e não funcionais da situação-problema em questão, a qual será explicada na sequência.

Sendo assim, nesse documento serão elucidadas as possíveis necessidades e as situações-problema do processo de negócio Fala Cidadão do Ministério da Cidadania, bem como serão propostas as abordagens a serem desenvolvidas, em Machine Learning (ML) e Deep Learning (DL), necessárias e mais apropriadas para a resolução dessas questões. No Anexo A há uma lista de terminologias e as respectivas descrições relacionadas às tecnologias de ML e DL.

1. PROBLEMA DE NEGÓCIO

A Central de Atendimento da Coordenação de Relacionamento da Ouvidoria Geral do Ministério da Cidadania é o órgão governamental responsável pelo atendimento, esclarecimento e resolução da situação do cidadão que faz jus a algum programa de auxílio do Governo Brasileiro.

A Central de Atendimento passou a receber em média 30.000 ligações diárias, depois que começou a atender a telefones celulares. O número de ligações aumentou em 7.000% e, com isso, aumentou o custo dos atendimentos por chamada telefônica para aproximadamente 130.000 reais em média por mês.

O grande aumento da quantidade de atendimentos exigiu, proporcionalmente, a contratação de mais atendentes e, por outro lado, a área de TI da Central de Atendimento buscou aprimorar a qualidade dos atendimentos através da utilização e do aprimoramento da URA, PAA Digital, os scripts, roteiros de atendimento, FAQs e Cartas Resposta.

A URA, Unidade de Resposta Audível, é um subproduto do sistema Tactium CRM [9], o qual interage no primeiro momento com o cidadão em uma ligação. Visa filtrar as informações através de opções de atendimento por telefone.

O Tactium recepciona a ligação e um áudio começa a ser gravado. O tipo de solicitação do cidadão é verificado. Em seguida verifica se as informações são do PBF (Programa Bolsa Família) e, sendo sim, solicita o NIS/CPF. O Tactium registra o percurso das opções selecionadas e gera um CALL_ID do atendimento. Além do CALL_ID o Tactium registra o NIS/CPF, o DDD e o telefone de origem da ligação do cidadão.

Com o NIS/CPF, é feita uma consulta no SGD, Sistema de Gestão de Demandas via web services. Sendo os dados válidos, o atendimento prossegue através da busca das informações pertinentes ao cidadão no SGD e são disponibilizadas as respostas ao cidadão.

No caso do NIS/CPF ser inválido, o SGD é aberto para que o atendente possa registrar o atendimento, sendo esse passo denominado de Atendimento Nível 1 (generalista).

Dando sequência, o atendente consulta as FAQs e/ou outras bases de informações dos cidadãos inscritos nos programas de auxílio. Sendo a resposta encontrada, o atendimento é direcionado para a pesquisa de satisfação e o atendimento é finalizado.

Se a resposta não for encontrada, então, uma INE (Informação não Encontrada) é aberta e a necessidade do cidadão é registrada no SGD, bem como o meio de contato pelo qual a informação será retornada. Esse passo aciona o Atendimento Nível 2 (especialista), que recebe a demanda via SGD e, por outro lado, terá uma SLA de 72h para retornar a informação ao cidadão.

O Analista de Nível 2 consulta as bases de informação e, se encontrada, devolve as informações recuperadas ao Analista de Nível 1, que responde o cidadão via e-mail, telefone ou outros meios. Se a resposta não foi encontrada, o atendimento é direcionado

para a área técnica responsável que o recepciona e o registra no SGD. Esse passo é do Atendimento Nível 3 (técnico), que responde à demanda e passa para o Nível 2 tratar a resposta. Dessa forma, é aberta uma SLA de 24h de tempo de resposta. O Nível 2 devolve para o Nível 1 que responde ao cidadão.

É necessário ficar claro que o presente trabalho, em hipótese alguma, será concorrente da empresa terceirizada que realiza a atual gestão dos atendimentos. O intuito é compreender o fluxo de tarefas e utilizar os dados e informações gerados no desenvolvimento de um modelo de aprendizado automatizado.

2. DEFINIÇÃO DO PROBLEMA

Nessa primeira etapa, o escopo do atual projeto visa, em primeiro lugar, propor e desenvolver um modelo de transcrição dos arquivos dos atendimentos gravados em formato áudio em arquivos no formato texto e, em segundo, propor a primeira versão de um Modelo de Aprendizado Automático que suporte o escopo do Projeto Fala Cidadão.

Para deixar claro, a segundo etapa visa propor um novo projeto de desenvolvimento. Trata-se de um modelo de análise capaz de identificar os problemas e a efetividade das demandas oriundas das ligações da Central de Relacionamento e da Ouvidoria Geral. Essa solução deverá permitir, principalmente, a análise dos diversos Programas Sociais, à luz das relações estabelecidas entre a sociedade e o Ministério da Cidadania.

Essa definição leva a converter os problemas de negócio em problemas de dados. Em projetos de Ciência de Dados essa é uma das partes fundamentais e cruciais, posto que sem os dados não haverá a Ciência. Sendo assim, abaixo está uma lista de problemas *a priori* levantados e relacionados às questões do Projeto Fala Cidadão.

- Como transcrever os áudios em textos?
- Como analisar o discurso nos áudios de atendimento?
- Como utilizar os dados do SGD junto com os textos transcritos dos atendimentos?
    - No campo observação são registradas as mudanças de UF, sendo exceção os cidadãos que moram no DF porque para esses o SGD permite a atualização direto no campo UF;
    - Os processos são registrados no sistema CEBAS;
    - A data do benefício é calculada a partir da tabela do site mds.gov.br/area-de-impressa/noticias/2018/dezembro/calendario-de-pagamento.
- Como melhorar a qualidade dos atendimentos (avaliação do atendimento)?
    - Diminuir o tempo de atendimento;
    - Aumentar a quantidade de respostas solucionadas;
    - Diminuir os erros das situações não sanadas.
- Como usar os dados registrados na URA/Tactium?

- Capturar o NIS/CPF do cidadão;
- Capturar o DDD e o telefone do cidadão. São usados para localizar a origem da ligação do cidadão.

- Como usar a URA/Tactium + SGD + outros sistemas para antecipar as necessidades dos cidadãos?
- Como criar um conjunto de palavras-chave do negócio em questão?
- Como classificar e quais são as classes dos áudios dos atendimentos?
- Como aperfeiçoar o atendimento ao cidadão a partir da análise do fluxo de dados e informações gerados entre a URA/Tactium + SGD + outros sistemas?
- Como utilizar a lista de assuntos do SGD?
- Como usar o SIPASS, sistema de pagamento da CEF, que registra a data, a origem e o valor do pagamento do benefício?
- Quais informações do censo do IBGE (PNAD) serão usadas?

2.1. ATORES E PARTES INTERESSADAS

Ministério da Cidadania

- Ereny – Coordenadora da Gestão de Banco de Dados do Ministério da Cidadania (patrocinadora);

Central de Atendimento

- Gilmar – Interino na Coordenação;
- Danilo – Gerente de Operações;
- Fabiana – Coordenadora de Operações;
- Juliana - Coordenadora de Operações;
- Felipe e Artur – Desenvolvedores de Software;
- Raimundo – Coordenador do 2º Nível de Analistas;
- Rodrigo – Analista de Tráfego.

População

- Beneficiários dos programas sociais;
- Gestores estaduais;
- Gestores municipais;
- Representantes da sociedade civil;
- Cidadãos brasileiros.

2.2. CRITÉRIOS DE SUCESSO

- Melhorar a qualidade dos atendimentos;
- Diminuir o tempo de atendimento;
- Diminuir as situações não sanadas;
- Diminuir os custos com os atendimentos telefônicos sem comprometer a qualidade do atendimento;
- Aumentar a quantidade de respostas sem erros;
- Reduzir a necessidade de contratação de serviços de Call Center.

2.3. RESTRIÇÕES

- A quantidade de áudios está na casa dos 300TB;
- A qualidade dos áudios a serem transcritos;
- Acesso aos dados do SGD;
- Conhecer o modelo de dados do SGD;
- Utilizar soluções pagas para a transcrição dos áudios;
- Obter acesso aos áudios;
- Transcrever todos os áudios;
- Obter acesso às bases externas (ex. CEF/SIPASS);
- Utilizar os dados da PNAD;
- Utilizar a arquitetura Spark/Hadoop.

3. HIPÓTESES

Uma hipótese é uma possibilidade de afirmação de uma informação sobre o problema que está sendo trabalhado. Pode ou não ser verdade. Então, qual a necessidade de construir hipóteses antes de coletar os dados?

A análise de dados que **não** é baseada em hipóteses gera a necessidade de:

- Coletar todos os dados para;
- Analisar todas as variáveis disponíveis e;
- Pode levar a uma visão estreita do negócio.

Por outro lado, a análise baseada em hipóteses permite:

- Identificar uma lista de possibilidades sem depender da obtenção dos dados;
- Mostra *insights* menores e específicos de se trabalhar;

- Permite uma visão sem tendências;
- Envolve todas as pessoas e times relevantes em um processo de *brainstorm*.

Abaixo estão algumas sugestões de hipóteses da situação-problema em questão. Mais hipóteses poderão ser acrescentadas, retiradas ou retificadas ao longo do desenvolvimento desse projeto.

DEMOGRAFIA	SAZONALIDADE	COMPORTAMENTO	BASES EXTERNAS
Há alguma região sem auxílio?	Que período aumenta o número de auxílios?	Cidadãos com dependentes têm mais propensão de continuar recebendo o auxílio?	Existem cidadãos que necessitam completar os dados?
Qual a quantidade de auxílios por região?	Que período diminui o número de auxílios?	Cidadãos com auxílio de esporte possuem maior probabilidade de perderem o auxílio?	Estão os valores dos auxílios desproporcionais em quantos % em relação à renda *per capta* brasileira?
Qual a média de auxílio em R$ por região?	Existe algo que afeta diretamente os auxílios?	Os cidadãos estão perdendo o direito de receberem o auxílio?	Existem cidadãos com mais de um auxílio?
As mulheres são mais propensas a auxílio do que os homens?	Qual o período de diminuição de auxílio?	Os cidadãos com auxílio antigo possuem menos chances de continuar recebendo?	Existe cidadão com auxílio fora do padrão de renda popular?
Os cidadãos de uma região são mais propensos a receber auxílio?	Tem alguma semana específica que aumenta ou diminui a procura por informação sobre os auxílios?	Os atendimentos estão resolvendo as dúvidas dos cidadãos?	
Pessoas casadas são mais propensas aos auxílios?	Há algum período quando ocorrem um maior número de auxílios cancelados?	Algo impacta no recebimento dos auxílios?	
		Houve alguma alteração nos direitos	

DEMOGRAFIA	SAZONALIDADE	COMPORTAMENTO	BASES EXTERNAS
		dos auxílios? Em qual período ocorrem mais ou menos problemas sobre os auxílios? É frequente a diminuição de auxílios? São frequentes novos auxílios?	

4. COLEÇÃO DE DADOS E EXTRAÇÃO

Trata-se da etapa onde se supõe que os dados estarão disponíveis, incluindo as bases de dados externas ao Ministério da Cidadania.

5. TRANSFORMAÇÃO E ANÁLISE EXPLORATÓRIA DE DADOS

Com os repositórios de dados, incluindo os áudios disponibilizados, será feita a análise exploratória de dados, bem como o devido tratamento e transformação dos atributos que serão usados nas análises estatísticas para o formato adequado ao estudo.

6. CONSTRUÇÃO DO MODELO DE APRENDIZADO AUTOMÁTICO

Em um primeiro momento, será desenvolvida ou utilizada uma arquitetura de mercado Deep Learning para a transcrição de arquivos em formato áudio em arquivos texto. Isso inclui selecionar os frameworks mais apropriados para esse desenvolvimento.

Com isso, começará a preparação dos textos e o enriquecimento de informação. No caso, *a priori*, é proposta a utilização das estruturas sintáticas denominadas Sintagmas Nominais (SN) ao invés de palavras como termos simples sem expressão de significado.

É apresentado em PINHEIRO [3] que os descritores denominados Sintagmas Nominais são uma unidade independente da maneira de como foram extraídos ou atribuídos aos documentos através de ferramentas de Recuperação de Informação. Tais descritores devem fazer referência a objetos ou a fatos do mundo real, ou seja, não são símbolos soltos como as palavras. Portanto, os descritores devem constituir-se de unidades extraídas do discurso (frase). Essa unidade deve ser a menor parte do discurso que possa servir de base a uma relação referencial autônoma, a qual se chama Sintagma Nominal (SN). O SN é a menor parte do discurso portadora de informação. Ao contrário das palavras/termos, os SN não são símbolos sem referências, pois possuem uma

estrutura sintática e uma estrutura lógico-semântica dentro do contexto que se apresentam. Exemplos de SN são:

SINTAGMA NOMINAL	ESTRUTURA SINTÁTICA
economia da informação	substantivo + preposição + substantivo
boa informação	adjetivo + substantivo
bolsa família	substantivo + substantivo
recursos próprios	substantivo + adjetivo
atendimento prejudicado pelo tempo	substantivo + adjetivo + substantivo + substantivo

Consequentemente, implicará na fundamentação das métricas mais adequadas porque, necessariamente, deve-se considerar que os SN ocorrem em menor frequência em relação às palavras/termos simples, sem expressividade, que se valem de stemming, dicionários e etc para melhorar a ponderação.

Em PINHEIRO [3] foi usada a métrica de ponderação TF-IDF, mas alterada (invertida) para dar a ponderação apropriada aos SN, que são em menor frequência em relação as palavras/termos simples e sem informação. Os SN, por outro lado, possuem significado e informação em sua estrutura.

TF-IDF é uma abreviação do inglês *Term Frequency-Inverse Document Frequency* e trata-se de uma medida estatística cujo objetivo é o de indicar a importância de uma palavra/termo em um documento em relação a uma coleção de documentos. Ela é frequentemente utilizada como fator de ponderação na recuperação de informações e na mineração de textos. O valor TF–IDF de uma palavra aumenta proporcionalmente à medida que aumenta o número de ocorrências dela em um documento, no entanto, esse valor é equilibrado pela frequência da palavra em relação aos demais documentos. Isso auxilia a distinguir o fato da ocorrência de algumas palavras serem geralmente mais comuns que outras [5], [6], [8] e [10].

No caso dos Sintagmas Nominais (SN) a métrica TF-IDF foi adaptada visando ponderar essas estruturas sintáticas, pois ocorrem em menor frequência em relação às palavras/termos simples. Em outras palavras, nesse caso, quanto maior a ponderação de um termo menor será a sua importância contextual nos documentos (corpus) e, por outro lado, quanto menor a ponderação de um SN maior será a sua importância contextual no corpus.

Para melhorar a equiparação dos SN foi utilizada a métrica Levenshtein Distance [4], que aproxima SN semelhantes no intuito de aumentar a ponderação entre os descritores. Obviamente, será necessário estudar os métodos atuais mais apropriados a fim de fundamentar a utilização dos SN e, sendo o caso, utilizá-los no desenvolvimento da

etapa de pré-processamento seguida pela de enriquecimento (Word Embedding). O algoritmo TF-IDF e a versão para SN estão descritos no Anexo A.

Em PINHEIRO [3] foram criadas ambas BOWs, a tradicional e a com SN. Elas foram aplicadas sobre um dataset “hostil” dos registros de reclamação de uma empresa de fornecimento de energia. Os registros eram textuais, tinham erros sintáticos e foram mal escritos.

Foi constatado que a utilização da BOW com sintagmas nominais apresentou os melhores resultados. Foi melhorada, e muito, a interpretação e caracterização dos clusters (K-Means no caso), atividade essencial em um modelo não-supervisionado.

Ainda, o estudo da diminuição do espaço vetorial através da técnica LSA (*Latent Semantic Analysis*) será considerado tendo em vista a quantidade de áudios que serão transformados em textos e, posteriormente, os descritores em colunas da estrutura BOW – Bag of Words [5] e [11].

A técnica LSA é um método de extrair e representar o significado de uso contextual das palavras/termos por meio de cálculos estatísticos aplicados sobre um conjunto de documentos (corpus). Essa técnica analisa e identifica um padrão de descritores no corpus, bem como o relacionamento entre eles. LSA é um método não-supervisionado de descobrir a sinonímia e a polissemia em uma coleção de documentos (corpus).

Também será necessária a criação de uma lista de palavras (word list) de interesse do negócio do Ministério, a fim de recuperar nos arquivos textos os descritores mais usuais no discurso dos atendimentos entre cidadão e atendente.

Há necessidade, também, de extrair as estruturas NER (Named Entity Recognition) com objetivo de encontrar os descritores mais pertinentes ao negócio em questão, conforme [5] e [6]. Com o resultado dessa etapa espera-se uma estrutura BOW para realizar as tarefas de Classificação e a de Análise de Identificação dos Problemas *versus* Efetividade das demandas nos atendimentos da Central de Atendimento.

Nessa fase do projeto está previsto, somente, o desenvolvimento do Modelo para a Transcrição dos arquivos em formato áudio para arquivos texto. A conclusão do presente documento subsidiará o início do desenvolvimento dos modelos de Classificação e o de Análise de Identificação dos Problemas *versus* Efetividade das demandas nos atendimentos da Central de Atendimento. A grande quantidade de etapas, fases e tarefas para o desenvolvimento dos modelos acima os caracterizam como um novo projeto.

Nas referências [1], [2], [7], [8], [10], [11], [14], [15], [16], [17], [18] e [19] são explicados os frameworks spaCy, Gensim, NLTK, Word2Vec/CBOW, fastText, Scikit-learn, TensorFlow, Keras, Pytorch. As mesmas referências mostram que esses são os frameworks mais utilizados em diversos tipos de modelos de Machine Learning e Deep Learning. Eles são objetos de estudo com o propósito de fundamentar e justificar os recursos de software do próximo Plano de Trabalho. Serão investigados na sequência do desenvolvimento do presente projeto a fim de selecionar os frameworks mais adequados para tal questão. No anexo A há uma tabela de terminologias onde são descritos tais frameworks.

Também, é necessário levar em consideração o ambiente computacional do Ministério da Cidadania, o qual disponibiliza uma Workstation com GPUs e, possivelmente, a utilização de uma arquitetura computacional Spark/Hadoop, em [12] e [13]. Por agora, será utilizada a solução em linguagem de programação Python que roda na GPU, porém, já considerando a arquitetura Spark/Hadoop, pois a quantidade de áudios a serem transcritos é de aproximadamente 300TB, os quais aumentam a cada dia. Haverá uma amostragem estratificada dos áudios, a princípio por tempo e tamanho dos áudios, para o início das primeiras transcrições dos áudios para textos.

CONCLUSÃO

No presente documento foram elucidadas as possíveis necessidades e as situações-problema do processo de negócio Fala Cidadão.

Foi elaborada a contextualização e fundamentação da situação-problema e, em decorrência, ocorrerá o desenvolvimento do modelo Deep Learning para a transcrição dos áudios em textos [7], [8], [10] e [15], pois para realizar uma tarefa de Machine Learning são necessários os dados. No caso, serão utilizados os textos transcritos e devidamente preparados juntos aos dados estruturados provenientes do SGD e da URA/Tactium.

Vale ressaltar que no transcorrer do atual projeto poderão ser acrescidas outras bases de dados externas ao negócio, mas relacionadas às hipóteses já levantadas, assim como novas hipóteses poderão igualmente surgir.

As etapas Coleção de Dados/Extração, Transformação/Análise Exploratória e Construção do Modelo de Aprendizagem Automático serão detalhadas em outra fase, pois, devido ao seu vulto, caracteriza como um novo trabalho. Invariavelmente, haverá a necessidade de continuidade das referidas etapas nesse novo projeto, pois será nele onde serão detalhadas as etapas e fases inerentes às tarefas de Classificação e a de Análise de Identificação dos Problemas *versus* Efetividade das demandas nos atendimentos da Central de Atendimento.

REFERÊNCIAS

[1] CHOUDHURY, Aniruddha. Sentiment Classification with Natural Language Processing on LSTM. medium.com consultado em 21/09/2019.

[2] YERPUDE, Ameya, PHIRKE, Akshay, AGRAWAL, Ayush and DESHMUKH, Atharva. Sentiment Analysis on Product Features Based on Lexicon Approach Using Natural Language Processing. International Jornal on Natural Language Computing (IJNLC), Vol. 8, No. 3, June, 2019.

[3] PINHEIRO, Marcello Sandi. Uma Abordagem Usando Sintagmas Nominais como Descritores no Processo de Mineração de Opiniões. Tese de Doutorado apresentada no Programa de Pós-graduação de Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Sistemas Computacionais, área Text Mining. Orientador D.Sc. Nelson Francisco F. Ebecken. Junho/2009.

[4] Levenshtain distance. http://en.wikipedia.com/wiki/Levenshtein_distance consultado em 11/09/2019.

[5] BENGFORT, Benjamin, BILBRO, Rebecca and OJEDA, Tony. Applied Text Analysis with Python. O'Reilly, 2018.

[6] BEYSOLOW II, Taweh. Applied Natural Language Processing with Python. Apress, 2018.

[7] LYTICA. Speech-to-Text Recognition. medium.com consultado em 10/09/2019.

[8] SRINIVASA-DESIKAN, Bhargav. A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing, 2018.

[9] https://www.softium.com.br/solucoes/tactium-crm/ consultado em 18/09/19.

[10] PERKINS, Jocob. Python 3 Text Processing with NLTK 3 Cookbook. Packt Publishing, ed. 2, 2014.

[11] ZHENG, Alice, CASARI, Amanda. Feature Engineering. O'Reilly, 2018.

[12] NANDI, Amit. Spark for Python Developers. Packt Publishing, 2015.

[13] SHERIF, Ahmed, RAVINDRA, Amrith. Apache Spark Deep Learning. Packt Publishing, 2018.

[14] RAMSUNDAR, Bharath, ZADEH, Reza Bosagh. TensorFlow for Deep Learning: From linear regression to reinforcement learning. O'Reilly, 2018.

[15] DI, Wei, BHARDWAJ, Anurag and WEI, Jianing. Deep Learning Essentials. Packt Publishing, 2018.

[16] ALBON, Chris. Machine Learning with Python Cookbook. O'Reilly, 2018.

[17] DAS, Sibanjan. CAKMAK, Umit Mert. Hands-On Automated Machine Learning. Packt Publiching, 2018.

[18] MANASWI, Navin Kumar. Deep Learning with Applications Using Python. Apress, 2018.

[19] ANKAN, Ankur. PANDA, Abinash. Hands-On Markov Models with Python. Packt Publishing, 2018.

[20] ZACCONE, Giancarlo, KARIM, Md. Rezaul. Deep Learning with TensorFlow, 2ed. Packt Publishing, 2018.

ANEXO A - TERMINOLOGIAS

TERMOS	CONCEITO
BOW	BOW (Bag-Of-Words) é uma representação usada no Processamento de Linguagem Natural (NLP – Natual Language Processing) e Extração de Informações (IE – Information Extraction). Nesse modelo, um texto (como uma sentença ou um documento) é representado como o conjunto de suas palavras, desconsiderando a gramática e até a ordem das palavras, mas mantendo a multiplicidade. BOW é comumente usado em modelos de classificação de documentos em que a ocorrência/frequência de cada palavra é usada como um recurso para o treinamento de um classificador, assim como em análise de agrupamentos [3], [10] e [11].
Deep Learning	Segundo [13], [14] e [15], é explicado que embora as técnicas clássicas de ML permitam identificar grupos de variáveis relacionadas à precisão e à eficácia na fase de treinamento, esses métodos diminuem quando são usados conjuntos de dados grandes e de alta dimensão. A arquitetura Deep Learning (DL), é um dos desenvolvimentos mais importantes em Inteligência Artificial nos últimos anos. DL é um ramo do ML baseado em um conjunto de algoritmos que modelam abstrações de alto nível sobre dados de alta dimensionalidade. O desenvolvimento da DL ocorreu em paralelo com o estudo da inteligência artificial e, principalmente, com o estudo das Redes Neurais Artificiais (RNA). Foi principalmente nos anos 80 que essa área se desenvolveu. Naquela época, a tecnologia de computadores não era suficientemente avançada para permitir uma melhoria real nessa direção. Portanto, foi necessário esperar por uma maior disponibilidade de dados e um poder de computação bastante aprimorado para se obter desenvolvimentos significativos nessa área. Em resumo, os algoritmos de DL são um conjunto de RNA que podem fazer melhores representações de conjuntos de dados em larga escala, a fim de construir modelos que aprendam essas representações extensivamente. Nesse sentido, em ZACCONE e KARIM [20] mostra a seguinte definição de DL: "*Deep learning is a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones.*" Traduzindo: "*O aprendizado profundo é um tipo específico de aprendizado de máquina que alcança grande poder e flexibilidade, aprendendo a representar o mundo como uma hierarquia aninhada de conceitos, com cada conceito definido em relação a conceitos mais simples e representações mais abstratas computadas em termos menos abstratos.*".

TERMOS	CONCEITO
fastText	Trata-se de uma técnica de representação vetorial desenvolvida pelo Facebook AI Research. Como o próprio nome sugere é um método rápido e eficiente de encontrar detalhes morfológicos sobre os textos de um Corpus. O FastText é exclusivo porque pode derivar vetores de palavras para palavras desconhecidas ou fora do vocabulário de termos. Isso porque, levando em consideração as características morfológicas das palavras, pode criar o vetor de palavras para uma palavra desconhecida. Isso se torna particularmente interessante em idiomas onde a estrutura morfológica é importante, por exemplo, espanhol, português, turco, francês e finlandês. Isso também significa que, com um vocabulário limitado, ainda é possível fazer combinações de palavras suficientemente interessantes. Por exemplo, significa que é capaz de entender o que a palavra charmosa ou estranhamente significam. Em outras palavras, de acordo com o FastText, estranho e estranhamente são diferentes de encantador e encantadoramente [10] e [11].
Gensim	Gensim é uma biblioteca de código-fonte aberto para Processamento de Linguagem Natural. É implementado em Python e Cython e foi projetado para lidar com grandes coleções de texto usando fluxo de dados e algoritmos on-line incrementais, que o diferencia da maioria dos outros pacotes de software de aprendizado de máquina que visam apenas o processamento em memória [5], [6], [8], [10] e [16].
Keras	Keras é uma biblioteca de alta abstração que possui uma API de fácil entendimento, a qual é amplamente utilizada em frameworks para Deep Learning como TensorFlow [5], [6], [8], [10] e [16].
LSA	Latent Semantic Analysis (LSA) foi originalmente desenvolvido para prover acurácia e efetividade em técnicas de Recuperação de Informação. O foco principal é a semântica das palavras através de uma série de contextos, em oposição ao uso de comparações de descritores simples. Ao invés de focar na frequência das palavras, LSA fornece uma medida quantitativa de semelhança semântica entre documentos com base no contexto de uma palavra resolvendo, assim, dois problemas do uso de contagem de descritores: a sinonímia e a polissemia. Por exemplo, no caso da sinonímia a LSA aproxima descritores como internacionalização e universalização, e no caso de polissemia distingue banco instituição financeira e banco móvel para sentar. A LSA tenta resolver esses problemas, não com dicionários extensos e mecanismos de processamento de linguagem natural, mas usando padrões matemáticos nos próprios dados para descobrir esses relacionamentos [5], [6], [8], [10] e [16].
Machine Learning	Machine Learning (ML) é uma teoria de que os computadores podem aprender sem serem programados para executar tarefas específicas. O aspecto interativo do ML é essencial, pois necessitam se adaptar sempre a novos dados, ou seja, precisam aprender com os dados históricos, otimizar para melhor assertividade e também serem capazes de generalizar para fornecer resultados adequados. ML é um termo mais amplo, onde vários

TERMOS	CONCEITO
	métodos e algoritmos são usados para aprender a partir dos dados. Como um ramo da inteligência artificial (IA), os algoritmos de ML são frequentemente usados para descobrir padrões ocultos, estabelecer um relacionamento e também para prever algo. ML depende de algumas entradas formatadas e fornece um resultado com base no objetivo pelo qual foi desenvolvido. O formato de entrada é específico ao tipo de técnica de ML utilizado e, também, ao algoritmo considerado. Essa representação específica dos dados de entrada é denominada características ou preditores [13], [14], [15], [16] e [17].
NER	O reconhecimento de entidade nomeada (NER – Named Entity Recognition), também conhecido como extração de entidade, é uma subtarefa de Extração de Informações (Information Extraction) que procura localizar e classificar entidades nomeadas em textos e em categorias predefinidas, por exemplo, indivíduos, empresas, locais, organização, cidades, datas, produto, códigos médicos, expressões de horário, quantidades, valores monetários, porcentagens e etc [6], [10] e [11].
NLTK	O Natural Language Toolkit, ou mais comumente o NLTK, é um conjunto de bibliotecas e programas para processamento simbólico e estatístico da linguagem natural (NLP – Natural Language Processing) para inglês, escrito na linguagem de programação Python. Foi desenvolvido por Steven Bird e Edward Loper no Departamento de Ciência da Computação e Informação da Universidade da Pensilvânia. O NLTK inclui demonstrações gráficas e dados de amostra. É acompanhado por um livro que explica os conceitos subjacentes às tarefas de processamento de idiomas suportadas pelo kit de ferramentas, além de um livro de receitas. O NLTK tem como objetivo apoiar a pesquisa e o ensino na PNL ou em áreas estreitamente relacionadas, incluindo linguística empírica, ciência cognitiva, inteligência artificial, recuperação de informações e aprendizado de máquina. O NLTK foi usado com sucesso como ferramenta de ensino, como ferramenta de estudo individual e como plataforma para prototipagem e construção de sistemas de pesquisa. Existem 32 universidades nos EUA e 25 países usando o NLTK em seus cursos. O NLTK suporta funcionalidades de classificação, tokenização, stemming, marcação, análise e raciocínio semântico [10].
Python	Python é uma linguagem de programação de alto nível, interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte. Foi lançada por Guido van Rossum em 1991. Atualmente possui um modelo de desenvolvimento comunitário, aberto e gerenciado pela organização sem fins lucrativos Python Software Foundation. Apesar de várias partes da linguagem possuírem padrões e especificações formais, a linguagem como um todo não é formalmente especificada [5] e [6].
Pytorch	PyTorch é um framework que permite o desenvolvimento rápido de modelos de ML e DL. Inicialmente desenvolvido na linguagem Lua, a equipe de IA do Facebook (FAIR) desenvolveu APIs em Python que

TERMOS	CONCEITO
	permite a sua utilização em alto nível através de um arquitetura computacional baseada em grafos dinâmicos que, definitivamente, o torna altamente apropriado para trabalhar com outros frameworks como TensorFlow, Scikit-learn dentre outros [5], [6], [8], [10] e [16].
Scikit-learn	É uma biblioteca Python para tarefas e técnicas de Machine Learning. Provê uma variedade significativa de técnicas/algoritmos para tarefas de aprendizado supervisionado e não-supervisionado. Essa biblioteca se concentra mais em aprender com os dados do que manipulá-los, é robusta e provê suporte de sistemas em produção. Tem foco na facilidade de uso, uma boa documentação e bom desempenho [5], [6], [8], [10] e [16].
spaCy	A spaCy é uma biblioteca de software de código aberto para processamento avançado de linguagem natural, escrita nas linguagens de programação Python e Cython. A biblioteca é publicada sob a licença MIT e atualmente oferece modelos estatísticos de redes neurais para inglês, alemão, espanhol, português, francês, italiano, holandês e NER multilíngue, além de tokenização para vários outros idiomas. Ao contrário do NLTK, que é amplamente usado para ensino e pesquisa, o spaCy se concentra no fornecimento de software para uso em produção. A partir da versão 1.0, o spaCy também suporta fluxos de trabalho de aprendizado profundo (Deep Learning) que permitem conectar modelos estatísticos treinados por bibliotecas populares de aprendizado de máquina como TensorFlow, Keras, Scikit-learn ou PyTorch. A biblioteca de aprendizado de máquina da spaCy, Thinc, também está disponível como uma biblioteca Python de código aberto separada. Possui modelos de redes neurais convolucionais para marcação de parte do discurso, análise de dependência e reconhecimento de entidades nomeadas, além de melhorias de API em torno de modelos de treinamento e atualização e construção de pipelines de processamento personalizados [5], [6], [8], [10] e [16].
Speech-to-Text	Trata-se do processo de conversão automática de arquivos em formato de áudio para o de arquivos no formato texto. As arquiteturas Deep Learning Long-Short Term Memory (LSTM) e Convolutional Neural Network (CNN) são muito utilizadas nesse propósito, assim como os modelos estocásticos Hidden Markov Models (HMM) [7], [18] e [19].
TensorFlow	TensorFlow é um framework criado e usado pela equipe de IA Google Brains, o qual permite o desenvolvimento de arquiteturas de Redes Neurais Artificiais e Deep Learning. É diferente do framework que está em produção da Google, mas, por outro lado, é mantido por uma comunidade muito ativa e têm um forte suporte para o processamento em GPU [13], [14], [15] e [20].
TF-IDF	Trata-se de uma técnica comentada em [3], [5], [6], [8] e [10], onde o processo de ponderação (weighting) envolve dar ênfase sobre o aspecto de um fenômeno ou a um conjunto de dados, possuindo um sentido muito

TERMOS	CONCEITO
	maior que dar, simplesmente, um peso no efeito final do resultado. Quer dizer, é análoga a prática de adicionar uma taxa extra ou de beneficiar algum objeto ou termo. Sua relação com a Mineração de Textos é voltada para a identificação de padrões. Já na Recuperação de Informação, seu foco está na criação de índices para os descritores no processo de consulta ou busca pela informação. Mas elas se convergem quando o assunto envolve a melhoria dos resultados. O *tf-idf* é proveniente do produto entre: $$tf\text{-}idf(j) = tf(j) \times idf(j)$$ Onde *tf* é o número de vezes que um termo *j* aparece no documento e *idf* é: $$idf(j) = log(N/df(j)$$ Resultante do logaritmo natural de *N* dividido por *df(j)*, onde *N* é o tamanho do corpus (quantidade de documentos) e *df(j)* é a quantidade de documentos onde o termo *j* está presente [5], [6], [8] e [10]. Assim, *idf* sempre dará valores positivos e no mínimo igual a 1, pois, como é sabido, qualquer função logarítmica passa pelo ponto *1* [5]. No caso dos Sintagmas Nominais (SN) o *tf* foi alterado para a fração *1/tf* [3], gerando a seguinte fórmula: $$(1/tf(j)) \times idf(j)$$ Como consequência, o número de SN é inversamente proporcional ao número de ocorrências, ou seja, sendo menor a ocorrência de um SN em vários documentos maior é a sua importância contextual e vice-versa.
Word Embedding	Trata-se de um tipo de mapeamento que permite que palavras com significado semelhante tenham representação semelhante. É uma das representações mais populares do vocabulário do documento. É capaz de capturar o contexto de uma palavra em um documento, semelhança semântica e sintática, relação com outras palavras, etc. Em termos gerais, são representações vetoriais de uma palavra específica que dependem de como são gerados e, mais importante ainda, como o contexto é capturado. O Word2Vec é uma das técnicas mais populares para Word Embedding e se baseia em uma Rede Neural Artificial Superficial [5], [6], [8], [10] e [11].
Word2Vec	É uma solução eficiente para problemas de Word Embedding, o qual aproveita o contexto das palavras de interesse. Essencialmente, trata-se de usar as palavras ao redor para representar as palavras de interesse com uma Rede Neural Artificial cuja camada oculta codifica a representação de palavras [5], [6], [8], [10] e [11].