Ferramenta de Apoio aos Estudantes da Agricultura para Identificação de Invasores na Cultura da Soja

Carolinne Roque e Faria - Programa de Pós-Graduação da Universidade Estadual de Londrina (UEL) - carolinne.rf@outlook.com

Cinthyan Renata Sachs Camerlengo de Barbosa - Programa de Pós-Graduação da Universidade Estadual de Londrina (UEL) – cinthyan@uel.br

Resumo. *Este artigo aborda o desenvolvimento de uma ferramenta chamada Carolina (Conversação Agronômica RObotizada em LInguagem NAtural) de apoio aos estudantes da agricultura para identificação de invasores na cultura da soja que possibilita consultas em Linguagem Natural para a construção de diálogos na obtenção de diagnósticos precisos sobre ameaças a essa cultura, simplificando o trabalho dos agricultores e das partes interessadas agrícolas que lidam com muitas informações. O tema proposto contou com a utilização de Aprendizado de Máquina em uma estrutura de dados com 101 pragas e suas informações na cultivar da soja e por meio do spaCy, uma biblioteca para análise sintática, foi possível pré-processar os textos, reconhecer as entidades nomeadas e suportar os requisitos para o desenvolvimento da ferramenta.*

Palavras-chave: Sistemas Inteligentes e Aprendizagem. Agricultura Digital. Processamento de Linguagem Natural.

Agriculture Students' Tool to support the Identification of Invaders in Soybean Crop

Abstract. *This article addresses the development of a tool called Carolina (Robotized Agronomic Conversation in Natural Language) to support agricultural students in identifying invaders in the soybean crop, which allows consultations in Natural Language to build dialogues on obtaining accurate diagnoses of threats to this crop, simplifying the work of farmers and agricultural stakeholders who deal with many pieces of information. The proposed theme included the use of Machine Learning in a data structure with 101 pests and their information in soybean cultivation and through spaCy, which is a library for syntactic analysis. Through this library it was possible to pre-process the texts, recognize the named entities and support the requirements for the development of the tool.*

Keywords: Intelligent Systems and Learning. Digital Agriculture. Natural Language Processing.

1. Introdução

Atualmente, existem inúmeros invasores que prejudicam a cultura da soja e que estão distribuídos em regiões produtoras do Brasil. O fator que limita a produção de soja são os problemas fitossanitários, como a presença de organismos que podem causar algum dano ao meio ambiente (Campos et al., 2018).

A cultura da soja desempenha importante papel no mercado brasileiro que passou a liderar o *ranking* de maior produtor mundial do grão. No país, a produção da *commoditie* pretende alcançar 133,67 milhões de toneladas e a área plantada será em torno de 37,88 milhões de hectares na safra 2020/2021 (CONAB, 2020).

A quantidade de informações disponíveis vai além da capacidade cognitiva de processamento do ser humano. Dessa forma, o gerenciamento de dados agrícolas depende de informações adquiridas de diversas tecnologias e sistemas que sejam capazes de auxiliar na tomada de decisão. Portanto, é fundamental mapear os dados (principais características de identificação de pragas e doenças na cultura da soja) e padronizá-los para que o estudante/agricultor tenha facilidade ao utilizar e que tenha um bom desempenho para a exibição da resposta.

Com o enorme fluxo de informações, essas tecnologias necessitam elaborar soluções para entender a linguagem e uma alternativa é o uso de Processamento de Linguagem Natural (PLN) aplicado para grandes volumes de dados. Essas técnicas pretendem analisar e extrair as informações de uma determinada cultura para proporcionar um diagnóstico com a finalidade de potencializar a produção dos agricultores a partir do aprendizado dos estudantes sobre as pragas da cultura da soja.

Mais precisamente, PLN é um conjunto de técnicas computacionais para a análise de textos em um ou mais níveis linguísticos, com o propósito de simular o processamento humano da língua (Barbosa, 2004) e pode ser utilizado em diferentes abordagens para tratar das infindas aplicações pertinentes para o mundo. Assim, essas tecnologias necessitam elaborar soluções para entender a linguagem humana e gerar diálogos semelhantes aos que são reproduzidos naturalmente por pessoas.

Visto a complexidade das linguagens naturais, este trabalho tem o intuito de desenvolver um aplicativo significativo para os estudantes de Ciências Agrárias por meio do PLN para identificar as ameaças que afetam a cultura da soja analisada e auxiliá-los na tomada de decisão para melhorar a produtividade do agricultor. Problemas linguísticos dessas perguntas em LN foram levantados e estão sendo estudadas soluções para a implementação.

Este trabalho possui a seguinte estrutura: a seção 2 é destinada aos Materiais e Métodos; a seção 3 aborda os Resultados e Discussões sobre o sistema conversacional para Identificação das Pragas e Doenças da Soja. Por fim, na seção 4 tem as conclusões.

2. Materiais e Métodos

Para Miura (2019) é imprescindível o desenvolvimento de aplicações automatizadas para analisar e interpretar textos em linguagem natural para que ações sejam executadas como resposta, fornecendo informações.

De acordo com o trabalho de Muller et al. (2011), o uso de recursos de tecnologia da informação e comunicação na mediação pedagógica reflete em maior motivação dos estudantes do Centro de Ciências Rurais, pois promove um despertar para a transição do ensino baseado na transferência de conteúdos para um ensino mais autônomo, não linear e proativo. Os autores ainda ressaltam que além da verbalização, permite uma interação e criação de conhecimento. Assim, o sistema permite uma nova dinâmica nos processos de construção do saber baseados na existência de relações, diálogos e interações.

Devido ao grande volume de dados, propõe-se treinar os modelos de classificação para avaliar o desempenho de pragas e doenças da soja por meio de Aprendizado de Máquina (AM), o que é definido por Mitchell (1997) como a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência. As técnicas de AM e PLN são essenciais para calcular o desempenho dos dados para o futuro e para isso é sugerido, segundo Piles et al. (2019), aplicar técnicas de AM a fim de detectar pragas automaticamente a partir de textos. Com o grande

volume de dados, propõe-se treinar os modelos de classificação para avaliar o desempenho de doenças da soja.

Dale (2010) destaca a aplicação de Análise de Textos para extrair automaticamente informações estruturadas de documentos não estruturados. O autor ainda ressalta que essa área visa desenvolver soluções para grandes problemas, como: Recuperação de Informação, Categorização e Agrupamento de Textos, Reconhecimento de Entidades, Correferência Nominal, Sumarização de Textos, Extração de Informação, Análise de Sentimentos (Polaridade) e Sistemas de Perguntas e respostas.

Segundo Bird et al. (2009) a fase de pré-processamento de dados pode ser considerada a mais importante para a aplicação das tarefas de Aprendizado de Máquina, pois nessa fase são feitas:

- remoção de *stopwords*, que são palavras como artigos, verbos de ligação, que aparecem nos textos várias vezes, mas praticamente não influenciam a classificação;

- *stemização* ou lematização, que é a redução de palavras a seus radicais, removendo flexões de tempo verbal, gênero, número; e

- tokenização que é o processo de criação de um vetor de termos de um documento, onde cada termo ocupa um índice do vetor.

Assim, é possível que o presente trabalho aplique técnicas para extração e classificação de textos na identificação de pragas e doenças a partir das características da praga na planta por meio de AM, a fim de analisar em menor tempo e avaliar o grau de severidade do dano na lavoura.

Os resultados obtidos a partir da predição para representar o comportamento real da identificação é predizer o desempenho do modelo para o futuro e a principal fonte é calculada pela matriz de confusão (Olson e Delen, 2008).

Segundo Piles et al. (2019), AM é frequentemente utilizado para analisar dados de sequenciamento para encontrar padrões de generalização em dados de alta dimensão por meio de uma quantidade pequena de amostras.

Para validar o modelo de Aprendizado de Máquina foi escolhida a técnica de Árvore de Decisão (AD), cuja fase de treinamento é realizada por meio de um sistema de indução de árvores baseada na divisão recursiva de Quinlan (1993) e Steinberg e Colla (1995). Como é afirmado por Silva e Vieira (2007), após a construção da árvore, pode-se classificar novos exemplos a partir dessa. A classificação é feita percorrendo a árvore até chegar à folha que determina a classe a que o exemplo pertence ou sua probabilidade de pertencer àquela classe.

Conforme o exposto acima, faz-se necessária uma arquitetura de um sistema computacional, como é proposta em Silva et al. (2007), que executa a língua natural e pode variar de acordo com as especificidades da aplicação. O tradutor automático é um possível exemplo em um sistema mais completo que deverá ser capaz de:

a) identificar, ou seja, extrair cada uma das palavras da sentença;

b) analisar sintaticamente a sentença, isto é, relacionar a cada palavra suas propriedades e funções sintáticas;

c) construir uma nova sentença para retomar o sentido das informações levantadas anteriormente, ou seja, extrair um significado absoluto da mesma, a partir dos significados das palavras e das relações entre elas;

d) assopciar o significado extraído em uma representação adequada. Essa pode ser independente da língua destino (interlíngua);

e) transformar a representação anterior em uma sentença na língua destino, isto é, traduzir ou associar as palavras da língua origem para a língua destino, desde que sejam equivalentes.

A partir disso, é possível determinar as inúmeras maneiras de escrever ou expressar certo material informacional, fazendo com que o sistema controle automaticamente a geração da tarefa e que o processo dessa seja mais simples do que a de interpretação, como sistemas que têm a função específica de transmitir informações constantes em uma base de dados em PLN. A Geração de Linguagem Natural (GLN), de acordo com Araujo (2020), é o processo de produzir frases, sentenças e parágrafos que são significativos a partir de uma representação interna.

Assim, foi possível desenvolver a ferramenta CAROLINA (acrônimo para **C**onversação **A**gronômica **RO**botizada em **LI**nguagem **NA**tural) e foi adotado o modelo em cascata por meio das fases de análises de requisitos, projeto, implementação, testes (validação), integração e manutenção de *software*. Cada etapa necessita da finalização da anterior para passar para próxima fase.

A seguir, um exemplo de como o sistema analisa/processa as perguntas, extrai as informações armazenadas no banco de dados e fornece a resposta.

I) Pergunta: Qual a localização da Lagarta-da-maçã-do-algodoeiro na soja? Análise: Pronome + artigo + substantivo + preposição + substantivo composto + preposição + substantivo

II) Resposta: "As Lagartas-da-maçã-do-algodoeiro atacam as vagens." Para reconhecer as referências dos diagnósticos, são estabelecidas algumas regras necessárias. Baseado no trabalho de Barbosa (1998; 2004) foi possível trabalhar o domínio proposto para os tipos de construção:

a) Grupos gramaticais, como:

- **Sentenças;**

- **Sintagma nominal**, em que o trecho da oração é que define completamente uma entidade ou conjunto de entidades do mundo do falante (Goddard e Schalley, 2010);

- **Sintagma verbal** determina uma atividade ou estado no tempo, como os verbos. Se o verbo for intransitivo, forma-se um sintagma verbal. Se for transitivo, deve haver um sintagma nominal que é o objeto da atividade para completar o sintagma verbal;

- **Sintagma adjetival** que por meio de um verbo de ligação atribui qualidades a um sintagma nominal ou qualifica um verbo intransitivo ou sintagma nominal; - Sintagma preposicional que composto por uma preposição seguida de um sintagma nominal;

- **Sintagma adverbial** que é formado por um ou mais advérbios seguidos ocasionalmente por uma preposição;

- **Oração subordinada adjetiva restritiva** que são as que limitam a extensão do nome a que se referem. Esse tipo de oração inicia-se por um pronome relativo (que, quem, o(a) qual, os(as) quais etc.);

b) Sentenças Sim/Não: são as que procuram o valor verdade de uma fórmula, seja ela "True/False". Por exemplo: "*A haste está podre?*";

c) Sentenças –wh: procuram valor de instanciação de funções exclamativas, como em "*A folha da soja está manchada!*";

d) Sentenças alternativas na forma normal (não-clivada): sentenças alternativas feitas no predicado. Exemplo: "*Apresenta manchas amareladas ou púrpuras?*";

e) Sentenças de solicitação de explicação: "*Por que?*";

f) S existencial: "*Alguma parte da planta foi analisada?*";

g) S na voz ativa: "*A manifestação compromete alguma parte da planta?*";

h) S na voz passiva: "*A planta está comprometida pela infestação?*";

i) S clivada: extraposição do verbo ser em "É a mancha que confirmou o diagnóstico?".

A representação da linguagem verbal interpretável/gerável é feita pela gramática. Essa contém regras de estruturação sintática e de morfossintaxe (gênero e número) (Barbosa, 2004). Contém valores possíveis das categorias sintáticas de nível mais baixo, como de complementos "tempo", "direção", "lugar" e outros (García, 1995).

3 Resultados e Discussões

A partir dos estudos realizados acerca do presente tema, foi criada uma modelagem para BDs NoSQL do tipo grafo (Neo4j), que é possível fazer consultas e extrair informações de um conjunto de vértices e arestas, como mostrada na Figura 1.

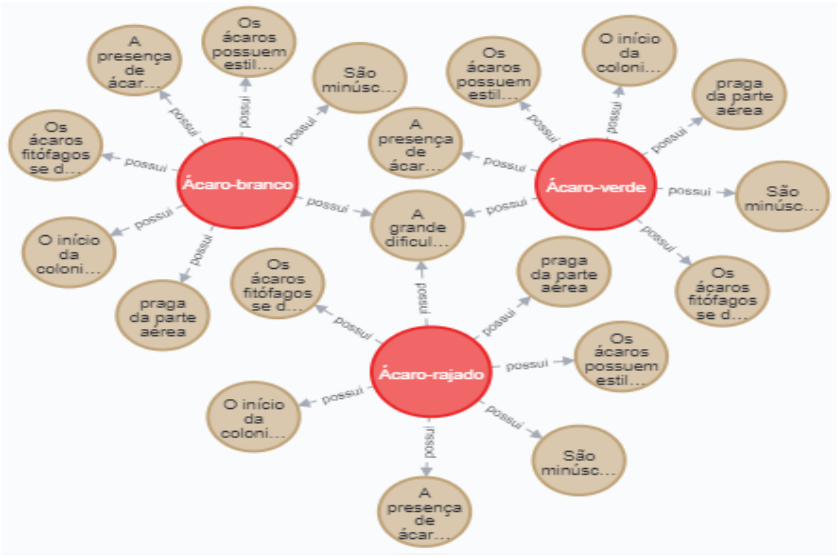

Figura 1 - Dados dos Ácaros Fitófagos no Banco de Dados NoSQL Neo4j

O processo descrito permite a extração das informações que são efetuadas com a finalidade de identificar a qual rótulo específico uma determinada praga ou relacionamento está atrelado. Este projeto tem o objetivo não só de colaborar com o estudante da área agronômica, mas também ganhar vivência no campo, aprender sobre as pragas e conhecer os tipos de danos diretos e indiretos e aplicar na prática os devidos métodos de controles. Visando isso, foi possível realizar perguntas para realizar o diálogo e possibilitar consultas, como:

"Tem um verme provocando uma lesão na raiz da soja."; "A minha semente está com uma mancha-púrpura."; "A soja armazenada está infestada de insetos"; "O broto não está se desenvolvendo."; "Tem uma praga esverdeada na plantação."; "Como posso controlar os nematoides?"; "Tem uma praga atacando a haste da soja."; "O que é um bichinho dourado na soja?"; "O que posso fazer para controlar a Antracnose?"; "Os

percevejos atacam que parte da planta?"; "Qual o dano causado pela Formiga-cortadeira?"; "Os ácaros atacam as folhas da planta?"; "Como previnir a presença de pragas na minha lavoura?"; "Como controlar a Antracnose?".

As perguntas vão surgindo conforme o diálogo entre o usuário e o sistema acontece (pergunta-resposta) e espera-se que o usuário obtenha informações relevantes para tomada de decisão. As palavras para essas perguntas foram catalogadas separadamente de acordo com a sua categoria morfológica para formar o dicionário utilizado na análise léxico-morfológica das palavras. A partir disso, está sendo desenvolvido um sistema para dialogar com os profissionais sobre as principais características das principais pragas e doenças na cultura da soja. Esse sistema recebe o texto e analisa as palavras de uma sentença isoladamente, como são visualizadas na Figura 2, para poder interpretá-lo como um todo.

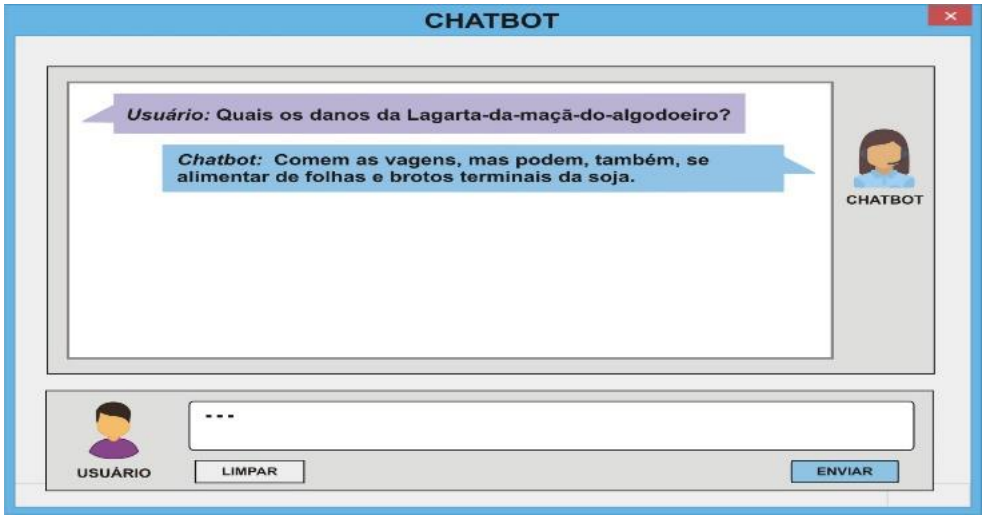

Figura 2 - Protótipo de Interface de Consulta à Base de Dados das Pragas da Soja

Identificar as partes da frase é essencial porque ajuda a entender as frases de entrada e constrói com mais exatidão as frases de saída (resposta). Assim, a primeira etapa é tokenizar os textos e extrair as etiquetas morfológicas para classificação de cada palavra pelo framework spaCy[1].

Posteriormente, esta ferramenta rotula todos os tokens a partir da análise para prever qual tag provavelmente se aplica nesse contexto. Para isso, conta-se com as seguintes tarefas: **Text** (texto puro), **Lemma** (reduz as palavras em seu formato base/raiz), **POS** (tags simples que determinam as categorias gramaticais de um token), **Dep** (Dependência sintática é a relação entre os tokens presentes dentro de uma sentença para entender o seu significado), **Shape** (classificação da palavra em maiúscula ou minúscula), **Alpha** (especificação das palavras em alfanuméricas ou não), **Stop** (indica se as palavras são consideradas *stopwords*), como é detalhada na Figura 3.

[1] https://spacy.io/

```
for token in doc:
    print(token.text, token.lemma_, token.pos_, token.tag_, token.dep_,
            token.shape_, token.is_alpha, token.is_stop)
```

```
As As DET <artd>|ART|F|P|@>N det Xx True True
folhas folhar NOUN <np-def>|N|F|P|@SUBJ> nsubj xxxx True False
atacadas atacar VERB <mv>|V|PCP|F|P|@ICL-N< acl xxxx True False
ficam ficar VERB <mv>|V|PR|3P|IND|@FS-STA ROOT xxxx True False
com com ADP PRP|@<ADVL case xxx True True
grandes grande ADJ ADJ|F|P|@>N amod xxxx True True
áreas área NOUN <np-idf>|N|F|P|@P< obl xxxx True False
recortadas recortar VERB <mv>|V|PCP|F|P|@ICL-N< acl xxxx True False
ou ou CCONJ <co-fcl>|<co-fmc>|<co-vfin>|KC|@CO cc xx True True
são ser VERB <cjt>|<mv>|V|PR|3P|IND|@FS-STA aux:pass xxx True True
completamente completamente ADV ADV|@ADVL> advmod xxxx True False
consumidas consumir VERB <pass>|<mv>|V|PCP|F|P|@ICL-AUX< conj xxxx True False
. . PUNCT PU|@PU punct . False False
```

Figura 3 - Árvore sintática para *"As folhas atacadas ficam com grandes áreas recortadas ou são completamente consumidas?"*

O spaCy possibilita descrever a relação sintática das palavras que se conectam na formação da árvore. Isso permite percorrer toda a árvore e retornar uma sequência ordenada de tokens e verificar os atributos e domínios das palavras. Nessa fase, além do que foi descrito anteriormente, conta-se com o **Head Text** (relação entre as palavras nos tokens), **Head Pos** (rotula as palavras em categorias) e **Children** (dependentes sintáticos do token) e são apresentados na Figura 4.

```
for token in doc:
    print(token.text, token.dep_, token.head.text, token.head.pos_,
            [child for child in token.children])
```

```
As det folhas NOUN []
folhas nsubj ficam VERB [As, atacadas]
atacadas acl folhas NOUN []
ficam ROOT ficam VERB [folhas, áreas, consumidas, .]
com case áreas NOUN []
grandes amod áreas NOUN []
áreas obl ficam VERB [com, grandes, recortadas]
recortadas acl áreas NOUN []
ou cc consumidas VERB []
são aux:pass consumidas VERB []
completamente advmod consumidas VERB []
consumidas conj ficam VERB [ou, são, completamente]
. punct ficam VERB []
```

Figura 4 - Navegação pela Árvore de Análise sintática

Além de ser capaz de processar grandes volumes de textos e extrair informações para realizar tarefas relacionadas ao PLN, o elemento principal do trabalho é a elaboração de um sistema para identificação das pragas da soja utilizando essas técnicas de PLN, uma vez que há um aumento de interesse em utilizar os sistemas de computador como auxílio no aprendizado do estudante da área agronômica.

A principal dificuldade no desenvolvimento é dialogar com o usuário e reconhecer as suas intenções a partir de uma frase e respondê-lo automaticamente. Visto isso, o sistema inteligente de pré-atendimento aos discentes tem o intuito de ser um canal alternativo de comunicação, para facilitar o acesso às informações e auxiliar no ensino para identificar o patógeno.

Perante os problemas levantados, a elaboração deste trabalho permite contribuir com a área agrícola, principalmente no que diz respeito à viabilidade de um assistente virtual, utilizando técnicas de PLN na identificação de pragas e doenças nas plantas da soja.

Os principais aspectos que identificam as ameaças da sojicultura incluem os nomes comum e científico das pragas, descrição da categoria morfológica, ciclo e também das características biológicas, comportamento e danos causados, localização/ ocorrência na planta, distribuição geográfica, métodos de controle e categoria (pragas que atacam plântulas, raízes, caules, folhas ou vagens; pragas subterrâneas; pragas da parte aérea; pragas de solo; pragas mastigadoras; pragas sugadoras; outros).

O dicionário utilizado é uma estrutura de dados, nas quais as palavras são armazenadas e associadas a elas algumas de suas informações (Barbosa, 2004). O repertório das palavras registradas no dicionário conta com 101 pragas na cultivar da soja (Santos, 1995), (Ferreira et. al, 2014), (Moreira e Aragão, 2009) e (Ávila e Grigolli (2014).

Para utilizar técnicas de processamento de linguagem natural para pré-processamento foram aplicados algoritmos de Aprendizado de Máquina à base de dados para a construção de classificadores que pudessem predizer a causa dos sintomas das plantas da soja a partir dos resultados (bom desempenho para auxiliar na tomada de decisão), com o uso de técnicas de PLN para otimizar as atividades guiadas à agricultura.

Posteriormente, para treinar as métricas, foi escolhido o algoritmo *Random Forest (RF)* (Breiman, 2001), termo para aplicar métodos de ensemble utilizando classificadores do tipo árvore. Assim, é possível construir uma grande quantidade de árvores de decisão que compõe a RF para auxiliar e predizer as pragas que atacam a soja, a partir das características descritivas. Os resultados do treino de classificação são apresentados na Tabela 1, composta pela média das métricas de precisão, revocação e f1-*score* para cada praga que danifica a planta. As métricas apresentadas indicam que o modelo obteve uma performance alta, conseguindo uma média de F1-*score* de 84% e acurácia de 81%.

Tabela 1 - Resultados das métricas precisão, revocação e F1-score em relação aos sintomas das plantas

Algoritmo	***Precision***	***Recall***	**F1-*score***	**Acurácia**
Random Forest Classifier	0.83	0.84	0.84	0.81

Por meio do resultado experimental foi possível analisar o desempenho do classificador RF na identificação de 101 pragas que prejudicam a lavoura da soja, o qual se mostrou robusto atingindo uma acurácia de 81%.

4. Conclusões

Para alcançar o objetivo proposto, foi avaliado um conjunto de dados reais na identificação das pragas e doenças na cultura da soja, feitos a partir da extração das características da base de dados e perguntas feitas por meio de consultas com profissionais e estudantes da área de agronomia. Foi feita uma validação utilizando a aplicação da métrica *Random Forest Classifier*, que resultou em 81% de acurácia, considerada uma alta taxa de acertos.

Destaca-se que para que esse modelo seja realmente efetivo para auxiliar na identificação de pragas na cultura da soja, é necessário: analisar outras bases de dados, pois pode causar perda de eficiência treinar somente com uma base de dados estática e também para comprovar a eficácia do modelo com mais textos; melhorias quanto ao método de seleção de palavras de um documento, para uma maior assertividade e

analisar o desempenho de mais classificadores, considerando o processamento e tempo de execução.

Pretende-se ainda descrever mais consultas na base de dados da agricultura, mas para isso faz-se necessário mais entrevistas com os alunos para verificar se há outros dialetos nesse domínio para que sejam abarcadas no dicionário. É preciso validar se as regras gramaticais desse público são as mesmas da norma culta ou não para saber se há a necessidade de acrescentar regras à gramática.

Referências bibliográficas

ARAUJO, E. F. S. **Solução Chatbot no Ambiente acadêmico da UFRJ, Departamento de Eletrônica e de Computação da Universidade Federal do Rio de Janeiro**. Trabalho de Conclusão de Curso. Rio de Janeiro, 77 f. 2020.

ÁVILA, C. J e GRIGOLLI, J. F. **Pragas da soja e seu controle**. Embrapa Agropecuária Oeste, pp. 109-168. 2014.

BARBOSA, C. R. S. C. **Gramática para Consultas Radiológicas em Língua Portuguesa.** Centro de Pós-Graduação em Ciência da Computação da Universidade Federal do Rio Grande do Sul, Dissertação de Mestrado. Porto Alegre, 143f. 1998.

BARBOSA, C. R. S. C. **Técnicas de Parsing para Gramática Livre de Contexto Lexicalizada da Língua Portuguesa**. Departamento de Engenharia Eletrônica e Computação do Instituto Tecnológico da Aeronáutica, Tese de Doutorado. São José dos Campos, 171 f. 2004.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python.** 1ª. ed. Cambridge: O'Reilly Media Inc. 2009.

BREIMAN, L. Random forests. **Machine Learning.** v.45, n.1, p.5–32, out. 2001.

CAMPOS, G. M. J.; ALCANTRA, E.; REZENDE, R. M. Levantamento de Insetos-Praga na Cultura da Soja. **Revista da Universidade Vale do Rio Verde.** v.16, n.3, p.1-8, 2008.

CONAB – COMPANHIA NACIONAL DE ABASTECIMENTO. **Observatório agrícola:** acompanhamento da safra brasileira, Décimo Levantamento, v.8, n.1, 2020.

DALE, R. Classical approaches to natural language processing, In: N. Indurkhya & F. J. Damerau (Eds.). **Handbook of Natural Language Processing.** 2ª Edição. Champman & Hall: CRC machine learning & pattern recognition series. Boca Raton, Florida: CRC Press, 2010. p.3-7.

FERREIRA, B. S. C.; CAMPO, C. B. H.; SOSA-GÓMEZ, D. R.; CORSO, I.; OLIVEIRA, L. J.; MOSCARDI, F. **Manual de identificação de insetos e outros invertebrados da cultura da soja**, Embrapa Soja-Documentos (Infoteca-E). 2014.

GARCÍA, L. S. **LINX: Um Ambiente Integrado de Interface para Sistemas de Informação Baseado em Conhecimento**. CPG Informática da PUCRJ. Tese de Doutorado. Rio de Janeiro, 191f. 1995.

GODDARD, C.; SCHALLEY, C. A. **Semantic Analysis: a Practical Introduction.** 2ª ed. Oxford: Chapman and Hall/CRC. 2010.

MITCHELL, T. **Machine Learning.** 37ª ed. Burr Ridge, IL: McGraw Hill, v.45, 1997. p.870–877.

MIURA, N. K. **Geração incremental de parsers dependentes de contexto para o português brasileiro**. Departamento de Engenharia de Computação e Sistemas Digitais. Tese de Doutorado. São Paulo, 132 f. 2019.

MOREIRA, J. C. e ARAGÃO, F. D. **Manual de Pragas da Soja.** Campinas: FMC Agricultural Products, 144f. 2009.

MÜLLER, L.; BANDEIRA, A. H.; ALVES, B. M.; BARIN, C. S.; MALLMANN, E. M. Recursos das tecnologias de informação e comunicação mediando o ensino aprendizagem e configurando ecologias cognitivas de estudantes do Centro de Ciências Rurais. **RENOTE - Revista Novas Tecnologias na Educação**, v.9, n.2, p.1-10, dez. 2011.

OLSON, D. L.; DELEN, D. **Advanced data mining techniques.** Springer Science & Business Media. 2008.

PILES, M.; LOZANO, C. F.; GALILEA, M. V.; RODRÍGUEZ, O. G.; SÁNCHEZ, J. P.; TORRALLARDONA, D.; BALLESTER, M.; QUINTANILLA, R. Machine learning applied to transcriptomic data to identify genes associated with feed efficiency in pigs. **Genetics Selection Evolution**, v.51, n.1, mar. 2019.

QUINLAN, J. **Programs for Machine Learning.** 1ª ed. San Mateo: Morgan Kaufmann, 1993.

SANTOS, O. S. **A Cultura da Soja 1,** Rio Grande do Sul-Santa Catarina-Paraná. 2ª ed. São Paulo: Globo. 1995.

SILVA, B. C. D.; MONTILHA, G.; RINO, L. H. M.; SPECIA, L.; NUNES, M. G. V.; OLIVEIRA JUNIOR, O. N.; MARTINS, R. T.; PARDO, T. A. S. **Introdução ao Processamento das Línguas Naturais e suas Aplicações**. Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional da Universidade de São Paulo. São Carlos, 121f. 2007.

SILVA, C. F.; VIEIRA, R. Categorização de Textos da Língua Portuguesa com Árvores de Decisão, SVM e Informações Linguísticas. In: **WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA**, 5., 2007, Rio de Janeiro. Anais. Rio de Janeiro: Sociedade Brasileira de Computação, 2007, p.1650-1658.

STEINBERG, D.; COLLA, P. **CART: Tree-Strutured NonParametric Data Analysis.** San Diego, CA : Salford Systems, 1995.