QUANDO PALAVRAS SÃO NÚMEROS: A EXPLORAÇÃO DE PALAVRAS PARA AUXILIAR NA TOMADA DE DECISÕES BASEADA EM DADOS NO CONTEXTO DO AGRONEGÓCIO

Luiz Gustavo Teixeira
luiz.teixeira12@fatec.sp.gov.br
Fatec Adamantina

Guilherme Saderio Polino
guilherme.polino@fatec.sp.gov.br
Fatec Adamantina

Paulo Roberto da Silva Ruiz
paulo.ruiz2@fatec.sp.gov.br
Fatec Adamantina

Resumo: Com o objetivo de preenchermos a lacuna entre os algoritmos que nos são expostos e os algoritmos que realmente devemos entender, o nosso projeto tem como objetivo principal realizar o levantamento de palavras relacionadas ao agronegócio, mais frequentes em cada região de desenvolvimento econômico do estado de São Paulo, a fim de auxiliar a empresa parceira estabelecer estratégias mercadológicas e tomada de decisão. Como Fundamentação teórica teremos como base os pressupostos teóricos-metodológicos da linguística de Corpus (BERBER-SARDINHA, 2004; MAHLBERGER, 2007; BIBER, 2011) bem como conceitos básicos do Processamento de Linguagem Natural (INDURKHYA; DAMERAU, 2010). A metodologia consta dos seguintes passos: a partir do uso do Python aplicado à Linguagem Natural, após o levantamento e organização das palavras relacionadas ao Agronegócio, serão desenvolvidas base de dados e páginas web que funcionam com buscadores de termos técnicos presentes no Agronegócio. A aplicação desenvolvida permite que os usuários consultem o banco de dados e encontrem termos relevantes para poderem desenvolver suas tomadas de decisões de maneira mais assertiva e, por sua vez, os alunos possam explorar os termos técnicos, a fim da compreensão de produção e escrita de relevância para a atuação profissional.

Palavras-chave: Colocações; Processamento de Linguagem Natural (PLN); Linguística de Corpus; Agronegócio; Desenvolvimento Regional .

1. Introdução

Talvez por sermos seres extremamente humanos e incapazes de conceber o meio que estamos inseridos, às vezes nos escapa a compreensão de palavras que circundam nosso dia a dia, seja no meio social, do trabalho e até mesmo o do nosso entretenimento. Por um outro lado, números estão em todas as partes e nos assustam com frequência. Diariamente, assistimos a um bombardeio de informações, de propagandas, sugestões de produtos nas redes socias, mil lugares para frequentar e até mesmo dezenas de receitas de bolo de cenoura. Os genericamente chamados de "algoritmos", ou seja, "um conjunto de etapas para executar uma tarefa descrita com precisão suficiente para que um computador possa executá-la" (CORMEN, 2014) invadiram as telas dos nossos dispositivos ditando o nosso próprio ritmo de vida e até mesmo, de consumo.

A fim de preenchermos a lacuna entre os algoritmos que nos são expostos e os algoritmos que realmente devemos entender, o nosso projeto tem como objetivo principal realizar o levantamento de termos técnicos, relacionados ao agronegócio, mais frequentes em cada região de desenvolvimento econômico do estado de São Paulo, a fim de auxiliar a empresa parceira estabelecer estratégias mercadológicas e tomada de decisão. A partir da exploração

estatística de termos técnicos e construção de banco de dados, visamos realizar um projeto de pesquisa que se retroalimenta da necessidade dos arranjos locais de produção, criando um contínuo processo de pesquisa, extensão gerando resultados que contribuam apara a consolidação da política de Pesquisa, Desenvolvimento & Inovação, uma vez que contempla a geração de produtos, processos e serviços inovadores e a transferência e difusão de tecnologia, como forma de consolidar o CPS como agente catalizador do desenvolvimento regional, a partir do incentivo da inovação para aumento da competitividade das empresas paulistas. Sendo assim, dessa maneira, esperamos também interpretar melhor as palavras que cercam a nossa realidade estabelecendo relações com a estatística, buscando termos que realmente são relevantes para a sociedade e para os arranjos produtivos locais, dando um maior sentido a essa vasta relação entre palavras e números.

2. Fundamentação Teórica

Neste capítulo, trataremos dos aspectos teóricos que fundamentam a pesquisa: a Linguística de Corpus. Berber Sardinha (2004) ressalta que a Linguística de Corpus é:

ocupa-se da coleta e exploração de corpora, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador. (BERBER SARDINHA, 2004, p.3)

A Linguística de corpus visa a coleta e exploração de dados, buscando evidências na língua real, observando a recorrência de determinados padrões da língua, partindo do pressuposto de Halliday (1991), o qual vê a língua como probabilidade e não possibilidade. É dentro desse contexto que o corpus ganha importância, devido a sua extensão e facilidade de armazenamento de dados, ou seja, partindo do caráter probabilístico da língua, quanto maior o corpus, maior a frequência de determinados padrões dentro de um contexto.

Tomando o léxico como exemplo, é possível diferenciar as palavras entre aquelas que são mais frequentes e as de menor frequência. Dessa maneira, para que haja probabilidade de palavras de ocorrência rara ocorrerem no corpus, é necessário incorporar uma quantidade grande de palavras. Portanto, quanto maior a quantidade de palavras, maior a probabilidade de ocorrência de palavras de baixa frequência. Com relação à definição de corpus, Beber Sardinha cita que a melhor definição é a de Sanchez (1995, p 8-9):

Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise. (BERBER SARDINHA, 2004, p 18.)

O autor ressalta que o corpus deve ser composto de textos autênticos. Além disso, o corpus também deve ser escolhido criteriosamente, ou seja, o pesquisador tem que delimitar regras para que esse corpus corresponda às suas características. Da mesma forma que a autenticidade do corpus, outros requisitos importantes, no que concerne à compilação de um corpus, são a representatividade e a extensão do corpus, que estão totalmente ligados à necessidade do pesquisador, que delimita os objetivos do corpus a ser compilado. Sendo assim, o pesquisador deve sempre estar ciente do que deseja investigar e compilar um corpus que se enquadre às características almejadas. Dessa forma, mais importante que o tamanho e a representatividade do corpus é, como o próprio autor nos propõe, uma inversão da origem da pesquisa, ou seja, parte-se da pesquisa e não do objeto em si, "coloca-se a questão de pesquisa na frente do objeto. Além de representativo o corpus deve

ser adequado aos interesses do pesquisador, que deve ter uma questão a investigar para a qual necessite de um corpus específico." (BERBER SARDINHA, 2004, p.29).

3. Materiais e Métodos

Este trabalho seguirá uma abordagem aplicada, com foco no desenvolvimento de um sistema web para auxiliar na decisão de tomada de dados, utilizando um banco de dados composto por colocações presentes no Agronegócio extraídas de sites de domínio público e de dados fornecidos pela própria empresa parceira. A metodologia será dividida em duas etapas principais: a coleta das expressões e o desenvolvimento da aplicação. A primeira etapa consiste no levantamento das colocações frequentes utilizadas no Agronegócio a partir de dados públicos e de dados fornecidos pela própria empresa. Essa coleta será realizada totalmente de forma manual, através da navegação por sites especializados no Agronegócio, bem como dados da Agência Paulista de Tecnologia dos Agronegócios (APTA).

Após a coleta e organização das expressões, será desenvolvido um site que funciona como um buscador dessas expressões. A aplicação permite que os usuários consultem o banco de dados e encontrem termos relevantes para compor tomar decisões mais assertivas. O sistema será implementado utilizando Python e bibliotecas como Pandas e Gradio. O banco de dados será estruturado a fim de otimizar a consulta e a categorização das expressões coletadas, garantindo uma experiência eficiente e intuitiva para os usuários envolvidos na pesquisa

4. Resultados e Discussão

Nosso projeto iniciou-se a partir do nosso ingresso, em agosto de 2025, como Pesquisador em Regime de Jornada Integral na Fatec Adamantina. Embora ainda embrionário, já temos alguns direcionamentos relevantes na pesquisa. A partir de constante pesquisas realizadas no site da Agência Paulista de Tecnologia dos Agronegócios (APTA) encontramos um mapa com as culturas pesquisadas de acordo com as Cadeias Produtivas Locais, como podemos ver na figura abaixo:

Figura 1- Cadeias Produtivas do Agronegócio

Mapa das Unidades de Pesquisa da APTA Regional. Fonte: GOMES, D; DUARTE, K.M.R. (2025, p. 2)

De acordo com o mapa das unidades de pesquisa da APTA Regional de Adamantina, conseguimos notar algumas culturas presentes na regional. Organizamos as palavras no quadro abaixo:

Quadro 1- Lista de Palavras pesquisadas

LISTA DE PALAVRAS PESQUISADAS COM BASE NA APTA REGIONAL ADAMANTINA
ACEROLA
ALGODÃO
AMENDOIM
AQUICULTURA
BANANA
BOVINOCULTURA DE CORTE
CAFÉ ARÁBICA E ROBUSTA
HORTALIÇA
MILHO
PESCA
SOJA
SORGO

Quadro 1: Lista de palavras pesquisadas com base na APTA regional de Adamantina

A partir desse primeiro registo de culturas que permeiam a região de Adamantina, o próximo passo da pesquisa é levantar as palavras relacionadas a tais culturas, a fim de criarmos uma base de dados para auxiliar a empresa parceira e por conseguinte, os alunos de ciência de dados durante seu percurso formativo.

Um exemplo pode ser ilustrado a partir da palavra de busca "acerola". Ao pesquisarmos a palavra na Base de Dados da Pesquisa Agropecuária (Embrapa), a palavra tem ocorrência de 1362 registros na base de dados, como ilustrado na figura abaixo.

Figura 2- Ocorrência da palavra acerola na BDP@

Resultado da busca "Acerola". Fonte: BPD@ / Embrapa

No próximo passo dentro de nossa pesquisa, realizaremos o levantamento estatístico de todas as ocorrências das palavras presentes no mapa da APTA dentro da Base de Dados

da Pesquisa Agropecuária (Embrapa), a fim de levantarmos estatísticas sobre essas palavras e, construirmos corpora de textos científicos que se tornará uma base de dados para explorar o comportamento lexical dessas no contexto do Agronegócio, visando auxiliar a nossa empresa parceira na tomada de decisões.

5. Considerações Finais

Nesse Resumo apresentamos o principal objetivo do nosso Projeto de RJI realizado na Fatec Adamantina, o qual tem como objetivo principal realizar o levantamento de termos técnicos, relacionados ao agronegócio, mais frequentes em cada região de desenvolvimento econômico do estado de São Paulo, a fim de auxiliar a empresa parceira a estabelecer estratégias mercadológicas e tomada de decisão Embora a pesquisa ainda esteja na fase embrionária, já delimitamos, as bases de dados que vamos nos pautar, tanto para definirmos os palavras de busca, quanto para checarmos seus comportamentos estatísticos.

Espera-se também realizar a criação de banco de dados para auxiliar os alunos de ciência de dados e, futuramente os alunos de agronegócio, durante seu percurso formativo. A partir do auxílio de ferramentas de IA e de Python aplicadas ao Processamento de Linguagem natural, também trabalharemos na criação de páginas-web para a exploração de base de dados criadas a partir de termos técnicos da área do Agronegócio bem como um protótipo de dashboards para visualização dinâmicas de termos presentes nos bancos de dados criados ao longo do projeto;

6. Referências

BIBER, Douglas. Back to the future. In: The Future of Scientific Studies in Literature. Special Issue of Scientific Study of Literature. Amsterdam, Philadelphia: John Benjamins Publishing Company, v. 4, 2011, p. 15-23

BERBER SARDINHA, T. Linguística de corpus. Barueri, SP: Editora Manole, 2004.

CORMEN, T. H. Desmistificando Algoritmos. Rio de Janeiro: Elsevier, 2014.

INDURKHYA, N; DAMERAU F. J. Handbook of Natural Language Processing. Chapman & Hall/ CRC, 2nd edition, 2010.

MAHLBERG, Michaela. Corpus Stylistics: Bridging the Gap Between Linguistics and Literary Studies. In: HOEY, M.; MAHLBERG, M.; STUBBS M.; TEUBERT, W. (Eds.). Text, Discourse and Corpora: Theory and Analysis. Continuum: London, 2007. p. 219-246

ORENHA, A. A compilação de um glossário bilíngüe de colocações, na área de negócios, baseado em corpus comparável. Dissertação (Mestrado). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2004, 233 f.

RASCHKA, Sebastian; MIRJALILI, Vahid. Python Machine Learning. 2nd ed. Packt Publishing Ltd, 2019. 772 p. ISBN 9781789958294.