INTRODUÇÃO

O **Processamento de Linguagem Natural (PLN)** tem como objetivo extrair representações e significados mais completos de textos livres escritos em linguagem natural, que é utilizada para fins de comunicações de uso comum por humanos; em idiomas como Português e Inglês, utilizando conceitos linguísticos como classes de palavras (adjetivo, substantivo, verbo, etc.), e estruturas gramaticais, também lida com situações mais complexas, como anáforas e ambiguidades. Isso se dá através de várias representações de conhecimento, como léxicos de palavras e seus significados considerando a estrutura hierárquica da linguagem, analisando a linguagem pelo seu significado.

O **PLN** pode ser utilizados em uma diversa gama de papéis, como corretores gramaticais, na conversão de fala para texto, na tradução automática de texto entre linguagens e na análise de sentimentos/mineração de opiniões.

OBJETIVO

O objetivo central do trabalho é desenvolver uma aplicação, utilizando a biblioteca **NLTK**, para ler em uma base os dados e analisar os resultados dos testes, Observando o comportamento do Stemming, Stopwords, Tokenização, Classificação (Naive Bayes) e Analise dos sentimento do teste aplicado.

HISTÓRIA DO PYTHON E DA BIBLIOTECA NLTK

Python foi desenvolvido em 1989, no centro de Matemática e Tecnologia da Informação (CWI, Centrum Wiskunde e Informática), na Holanda por Guido van Rossum, em 1991 Guido publicou o código (nomeado versão 0.9.0) no grupo de discussão alt.sources (newsgroup), e foi feita a abertura para a programação, baseado para ser uma linguagem de programação de alto nível, interpretada de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica, simples e forte de excelente funcionalidade para o processamento de dados linguísticos, com uma sintaxe clara, precisa e reduzida, que auxiliar a legibilidade do código fonte

O **NLTK - Natural Language Toolkit** foi desenvolvido em 2001, como objetivo de uma disciplina em linguística computacional no Department of Computer and Information Science da University of Pennsylvania, foi criado como uma suíte de aplicativos e módulos de código-fonte aberto denominada como biblioteca, para prover o aprendizado da linguagem natural.

OS PRINCÍPIOS DO NLTK E SUAS FERRAMENTAS

O NLTK foi desenvolvido como um kit ferramentas e pensado com quatro objetivos primários em mente sendo eles **simplicidade, consistência, extensibilidade e modularidade**. E suas ferramentas sendo elas **Stemming, Stopwords, Tokenização, Classificação (Naive Bayes) e Análise.**

Simplicidade - Oferece um framework intuitivo em conjunto a substanciais blocos de construção, provendo aos usuários de um conhecimento prático de NLP, sem dificultar as tarefas associadas ao processamento de dados linguísticos anotados.

Consistência - Oferece um framework unificado com interfaces e estruturas de dados consistentes, e nomes de método facilmente prognosticáveis.

Extensibilidade - Oferece uma estrutura no objetivo em que novos módulos de software possam ser utilizados de forma simples e fácil, adicionando implementações alternativas a abordagens variadas para uma mesma tarefa.

Modularidade - Oferece componentes (módulos) que possam ser utilizados independentemente sem a necessidade de compreender o restante da ferramenta.

FERRAMENTAS DO NLTK

- O **Stemming** é realizada uma técnica de redução de termos a um radical comum, no ponto inicial de uma análise das características gramaticais dos elementos, como partículas localizadas no final das palavras com finalidade de indicar as flexões de gênero, número dos nomes e as flexões de número, gênero, pessoa, tempo e modo dos verbos. Permitindo que você se concentre-se no significado básico de uma palavra ao contrário de todos os detalhes de como ela está sendo utilizada, reduzindo as palavras agrupando-as por meio de seu sufixo.
- O **Classificador** é um algoritmo baseado em **Naive Bayes** que faz uma abordagem probabilística que utiliza a probabilidade baseado na regra do **Teorema de Bayes** assumindo independência nos atributos do objeto.

FERRAMENTAS DO NLTK

- O **Stopwords** é uma lista de palavras que podem ser consideradas desnecessárias para o entendimento do sentido de um texto, ou seja, palavras semanticamente irrelevantes. Essas palavras são geralmente removidas de um texto durante a fase de pré-processamento realizando uma filtragem e limpeza dos textos, tornando as tarefas de mineração de dados mais fácil e ágil.
- A **Tokenização** tem como finalidade a segmentação de palavras, que é a quebra a sequência de caracteres em um texto localizando onde uma palavra ou por frases terminam e outra começa, possibilitando trabalhar com textos menores que ainda são relativamente coerentes e significativos, mesmo estando do contexto do restante do texto.

FERRAMENTAS DO NLTK

A **Análise** tem como seu relacionamento ao processo de análise de informação que realiza uma varredura procurando reduzir o espaço de busca, recuperando apenas as informações que são relativamente importantes para a resolução de problemas determinados, a análise realiza a validação da eficiência ou acurácia do processo como um todo, entre elas estão a análise **morfológica, sintática, semântica e léxica**.

- Na **Análise Morfológica** tem como finalidade ser responsável por definir artigos, substantivos, verbos e adjetivos, armazenados em um tipo de dicionário.
- Na **Análise Sintática** o utiliza o dicionário procurando por mostrar relacionamento entre as palavras e, em seguida, verifica sujeito, predicado, complementos nominais e verbais, adjuntos e apostos.
- Na **Análise Semântica,** há o encontro de termos ambíguos, de sufixos e afixos, entretanto, em questões de significado associados aos morfemas componentes de uma palavra, o sentido real da frase ou palavra,
- Na **Análise Léxica** faz converter uma sequência de caracteres em um seguimento de palavras que foram as palavras candidatas a serem termos do índice e colocando em ordem alfabética de caracteres de palavras e separadores de palavras.
- Na **Análise Semântica** tem como finalidade procurar e identificar a função que determinados termos exercem no texto.

ESTUDO DAS EMOÇÕES

Segundo o psicólogo Paul Ekman, afirma que as principais emoções básicas são: Surpresa, Alegria, Tristeza, Medo, Desgosto ou Nojo e Raiva.

- **A surpresa** é uma das emoções mais rápida e passageira, apresentando em alguns instantes de tempo
- **A alegria** é uma emoção positiva que tem como finalidade a sensação de bem-estar subjetivo e satisfação com a vida.
- **A tristeza** é uma emoção humano que expressa desânimo ou frustração em relação a alguém ou algo.
- **O Medo** é uma emoção humano que expressa o alerta e o perigo.
- **O desgosto** é uma emoção humano que expressa em ocasiões desagradáveis e aversivas.
- **A raiva** é uma emoção humana que expressa sensação de a revolta, a hostilidade, a irritabilidade, o ressentimento, a indignação, o ódio e a violência,

REQUESITOS DO PROJETO ECONSTRUÇÃO DA APLICAÇÃO

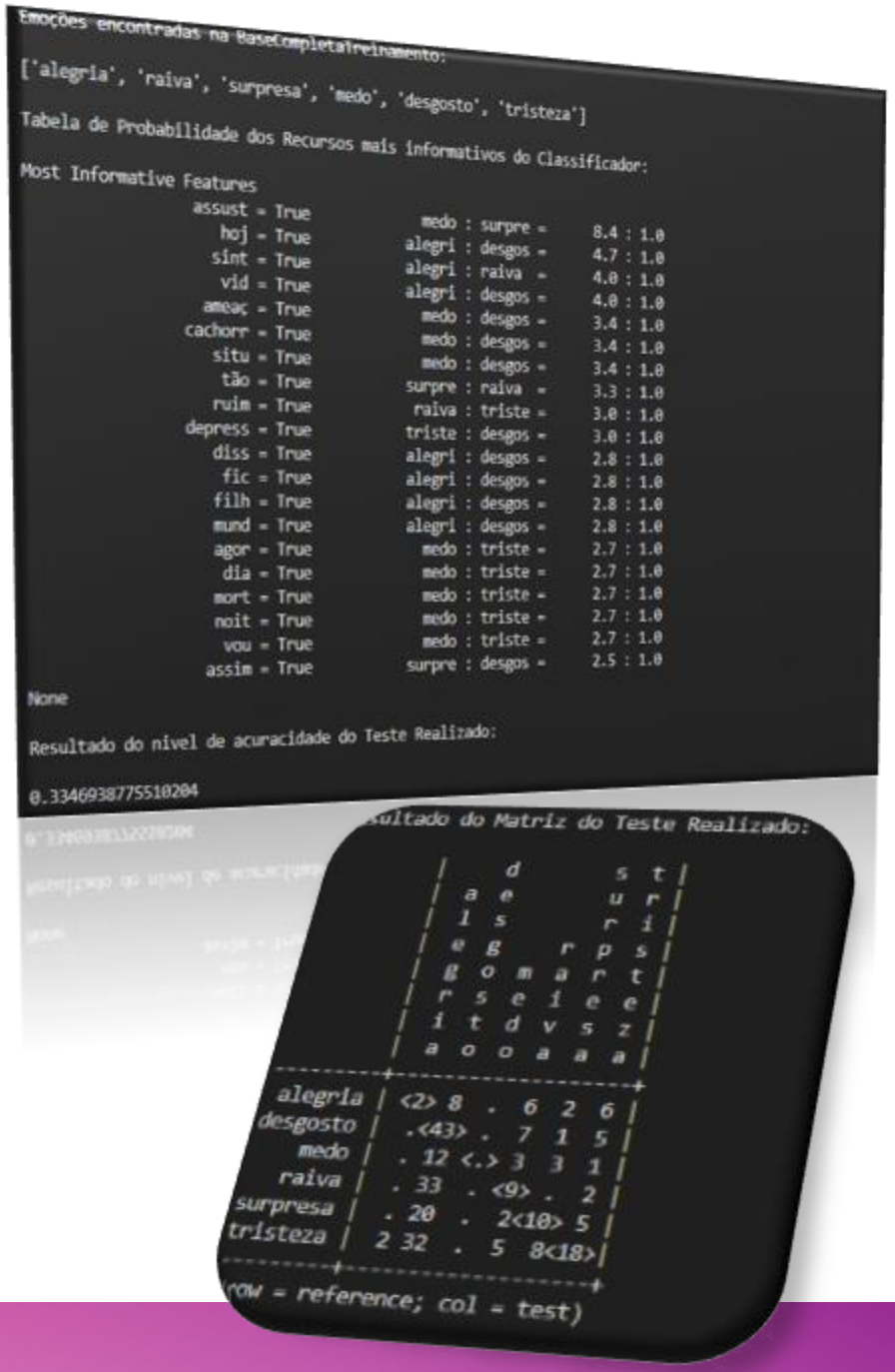
Os requisitos para a realização do projeto é a instalação do Python e do pacote NLTK 3.5 em diante, e a Instalação da IDE e suas dependências (a utilizada foi a VSCODE).

Foi realizado a construção da aplicação para o estudos das emoções em maneiras linguísticas usando o Python com a biblioteca NLTK, foi criando uma base de dados, disponibilizadas pelo autor Luís F. Dias e atualizada pelo autor Washington L. S. Menêzes. Com todas as frases já rotuladas para fazer que o algoritmo aprenda com os dois padrões abordados para frases de alegria, raiva, surpresa, medo, desgosto e tristeza. Que será dividida em duas variáveis sendo elas base de treinamento e base de teste, será necessária a divisão em duas bases para que o algoritmo verifique a assertividades em uma dessas bases para a esta finalidade.

Todos os trechos de código retirados do código-fonte foram criados pelo autor, disponibilizado na plataforma Github :
https://github.com/WashingtonLuiz89/TCC-ENGENHARIA-DA-COMPUTA-O

TESTES REALIZADOS

Foram realizados três testes com as frases, (odeio ir à escola), (gostamos de ir à escola), (não acredito que amanhã vai fazer sol). e recebemos os resultados das emoções encontradas no banco de dados, a Tabela de Probabilidade dos Recursos mais informativos do Classificador, Resultado do nível de acuracidade do Teste Realizado, Resultado do Matriz, Resultado do Classificador Naive Bayes, Porcentagem do classificador de Naive Bayes, e as Sentenças do Tokenize do Teste Realizado.

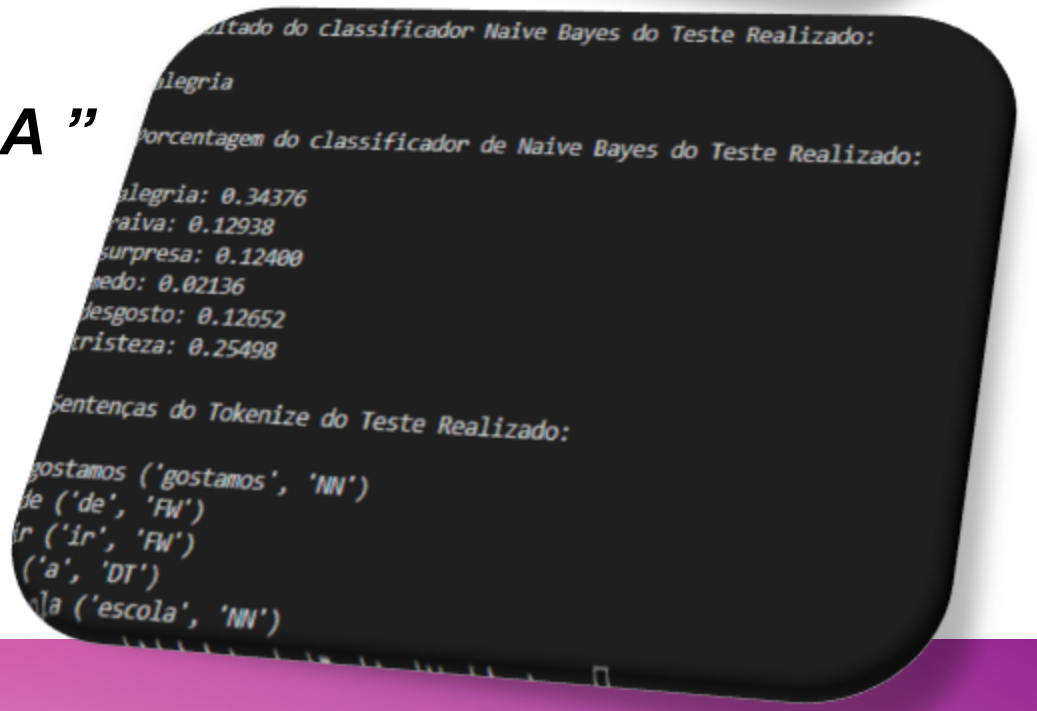


RESULTADO
“ODEIO IR À ESCOLA”

Com o resultado do teste, obtivemos a resposta que a frase ***“odeio ir à escola”,*** significa tristeza, com 38% de assertividade, e que o Tokenizador possui substantivo massivo, frases estrangeiras, determinante. Conforme na figura 26

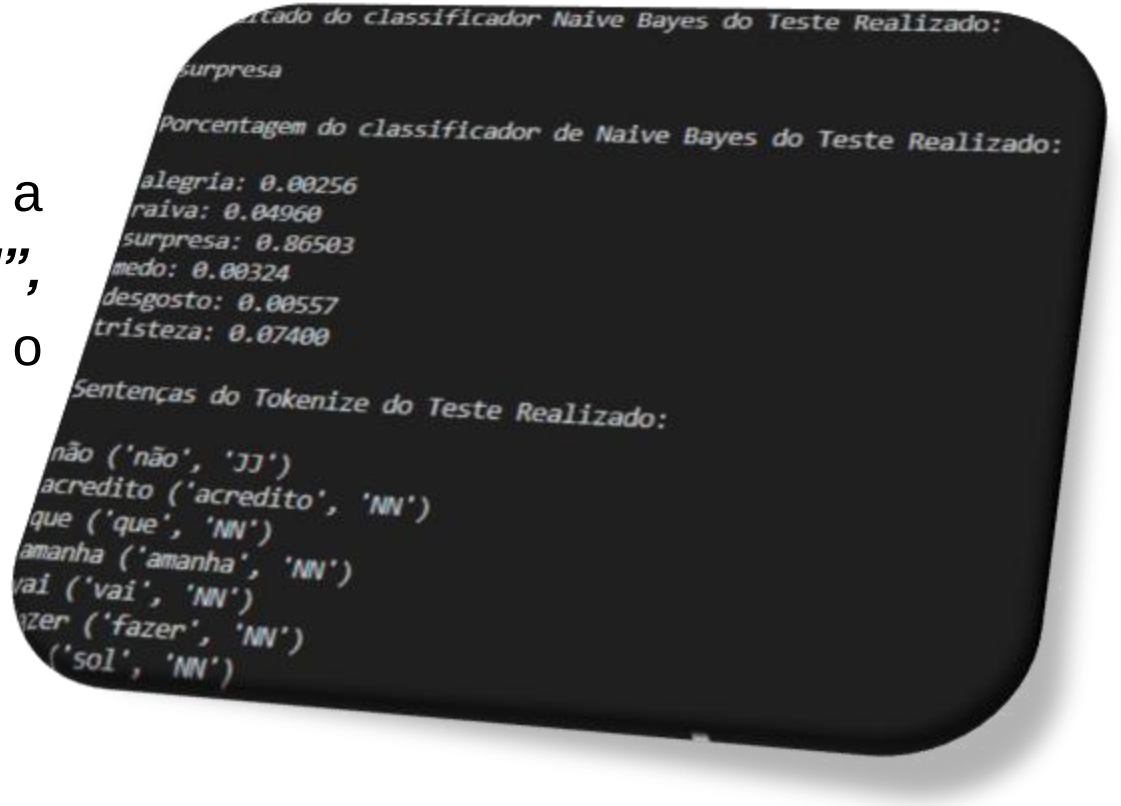
RESULTADO
“GOSTAMOS DE IR À ESCOLA”

Com o resultado do teste, obtivemos a resposta que a frase ***“gostamos de ir à escola”,*** significa alegria, com 34% de assertividade, e que o Tokenizador possui substantivo massivo, frases estrangeiras, determinante.



RESULTADO
“NÃO ACREDITO QUE AMANHÃ VAI FAZER SOL”

Com o resultado do teste, obtivemos a resposta que a frase ***“não acredito que amanhã vai fazer sol”,*** significa Surpresa, com 86% de assertividade, e que o Tokenizador possui substantivo massivo e adjetivo.



CONCLUSÃO

Podemos chegar à conclusão que ao inserir novos de dados na Base.py, e possível aumentar a precisão dos resultados, refinando a assertividade dos sentimentos abordados, assim como a porcentagem do classificador de Naive Bayes, que obtivemos sucesso nos testes realizados, em um dos três resultados obteve uma precisão de 86%.

Identificamos que este método e muito importante para tomadas de decisões, que pode-se descobrir o que o usuário está sentindo, tomando medidas para a solução de um problema ou opinião sobre algo abordado. Por exemplo em um Chatbox com Inteligência Artificial, e possível detectar a melhor abordagem a ser realizada com a resposta do sentimento.

DEMONSTRAÇÃO DA APLICAÇÃO