- RELATÓRIO

O grande volume de avaliações disponíveis em plataformas como o TripAdvisor dificulta a análise manual dessas opiniões, especialmente quando o usuário deseja encontrar rapidamente experiências semelhantes ou filtrar por aspectos específicos, como limpeza, localização ou atendimento. A diversidade de palavras e expressões utilizadas pelos usuários torna a navegação e a tomada de decisão mais complexas. Este problema é ampliado pelo fato de que muitas avaliações utilizam sinônimos ou variações linguísticas, o que pode dificultar a classificação e comparação direta de avaliações semelhantes.
Diante deste cenário, o presente projeto propõe uma solução automatizada para categorizar, filtrar e comparar avaliações de maneira eficiente, facilitando a navegação do usuário. A proposta visa permitir que o usuário encontre, de forma mais precisa, as avaliações que correspondem aos aspectos específicos que deseja avaliar em um hotel. O objetivo do trabalho é desenvolver uma aplicação interativa capaz de agrupar avaliações de hotéis de acordo com tópicos específicos e recomendar as avaliações mais semelhantes dentro de cada tópico. Para isso, o projeto faz uso de técnicas de PLN, além de métodos de álgebra linear, como o cálculo da similaridade de cosseno entre avaliações.

O dataset utilizado é o arquivo tripadvisor_hotel_reviews.csv, que contém informações sobre avaliações de hotéis feitas por usuários, incluindo o texto da avaliação, a nota atribuída (variando de 1 a 5), o nome do hotel e a data da avaliação. Para otimizar o processamento, foi utilizada uma amostra aleatória de 1000 avaliações. O tratamento do dataset envolveu uma série de etapas para garantir a qualidade dos dados e a adequação ao modelo de análise.
Entre essas etapas estão: Limpeza e Normalização: Consistiu na remoção de caracteres não alfabéticos, conversão de todo o texto para minúsculas e remoção de espaços extras.
Tokenização, Lematização e Stemming: O texto foi segmentado em tokens (palavras), processado para reduzir as palavras às suas formas base (lematização) e também reduzido à raiz (stemming), utilizando ferramentas do NLTK, como o WordNetLemmatizer e o SnowballStemmer.
Remoção de Stopwords: Foram removidas palavras comuns, como preposições e artigos, que não possuem valor semântico significativo para a análise.
Mapeamento de Tópicos: Cada avaliação foi associada a um ou mais tópicos, com base na presença de palavras-chave e seus sinônimos, que foram expandidos automaticamente utilizando o WordNet.

O código do projeto está dividido em dois arquivos principais. O primeiro arquivo, utilis.py, contém funções auxiliares que garantem o download seguro dos recursos do NLTK necessários, realizam o pré-processamento do texto, expandem as palavras-chave por meio de sinônimos e identificam os tópicos presentes em cada avaliação. O segundo arquivo, App.py, é responsável pela leitura e amostragem do dataset, pelo pré-processamento das avaliações, pela identificação dos tópicos, pela vetorização das avaliações utilizando a técnica TF-IDF, pelo cálculo da similaridade de cosseno e pela construção da interface gráfica utilizando a biblioteca Tkinter. Nessa interface, o usuário pode selecionar um tópico, e o sistema irá automaticamente escolher uma avaliação aleatória relacionada

a esse tópico, exibindo as dez avaliações mais semelhantes, com destaque para a nota atribuída, os tópicos relevantes e o grau de similaridade. A aplicação utiliza técnicas de PLN e álgebra linear para realizar as tarefas de processamento e análise das avaliações. A principal técnica utilizada para representar o texto das avaliações de forma quantitativa é a TF-IDF (Term Frequency-Inverse Document Frequency), que transforma o texto em vetores numéricos, ponderando a importância de cada termo dentro do contexto do corpus. A similaridade entre as avaliações é calculada por meio da similaridade de cosseno, que mede o ângulo entre os vetores gerados pelas avaliações, indicando o grau de semelhança entre elas.

O sistema utiliza a expansão semântica dos tópicos, por meio da utilização de sinônimos extraídos do WordNet, para tornar a categorização dos tópicos mais robusta e precisa. Isso permite que o sistema identifique e relacione avaliações que utilizam diferentes variações de uma mesma palavra.
Em termos de usabilidade, a aplicação oferece uma interface gráfica simples e intuitiva, na qual o usuário pode selecionar um tópico de avaliação, como "limpeza" ou "localização". O sistema, então, seleciona uma avaliação aleatória para esse tópico e apresenta as 10 avaliações mais semelhantes a ela, com destaque visual para informações relevantes, como a nota atribuída e o grau de similaridade. Além disso, o sistema também exibe a distribuição dos tópicos nas avaliações, fornecendo uma visão geral do conteúdo do dataset e permitindo que o usuário identifique as áreas mais frequentemente comentadas nas avaliações dos hotéis.

O projeto demonstra como técnicas de PLN e álgebra linear podem ser aplicadas para organizar, filtrar e recomendar avaliações em grandes volumes de dados textuais. A solução desenvolvida permite que o usuário realize uma análise detalhada das avaliações dos hotéis de forma mais eficiente e precisa. O tratamento cuidadoso do dataset, a utilização de técnicas de expansão semântica e a criação de uma interface gráfica interativa contribuem para tornar a solução robusta, eficiente e fácil de usar, facilitando a tomada de decisão com base nas avaliações de outros usuários.