Profa. Maria Kelly Venezuela e Prof. Márcio F. Stabile Jr.

PROJETO 1

CLASSIFICADOR NAIVE-BAYES COM TEXTO NATURAL

CLASSIFICADOR AUTOMÁTICO DE TEXTOS

Com a popularização dos marketplaces digitais, como a Amazon, milhares de usuários compartilham suas opiniões sobre comidas em forma de textos corridos. A principal vantagem desse tipo de avaliação é que o cliente pode escrever livremente sua experiência, descrevendo sabor, textura, embalagem, entrega ou outros aspectos. No entanto, essa liberdade também traz um desafio: as avaliações não são estruturadas, o que dificulta extrair informações objetivas diretamente dos textos.

Um problema interessante é tentar prever a nota (de 1 a 5 estrelas) que o usuário atribuiu ao produto apenas com base no conteúdo textual de sua avaliação. Por exemplo, um comentário como "O sabor é maravilhoso e a entrega foi super rápida" provavelmente corresponde a uma nota alta (4 ou 5 estrelas), enquanto "O pacote chegou rasgado e a comida tinha gosto estranho" tende a indicar uma nota baixa (1 ou 2 estrelas).

Para lidar com esse problema, a equipe de projeto pode utilizar técnicas de Ciência dos Dados, em especial modelos de classificação de texto. Um dos métodos mais clássicos é o classificador Naive-Bayes, largamente aplicado em tarefas como filtros de spam. Esse classificador permite estimar a probabilidade de uma avaliação pertencer a cada uma das categorias de nota, considerando as palavras que aparecem no texto.

Na prova de conceito (POC), o objetivo é implementar uma versão do classificador que "aprenda", a partir de uma base de treinamento, a relação entre os textos das avaliações e as notas atribuídas. Em seguida, será necessário validar a performance do modelo em uma base de testes, avaliando quão bem ele consegue prever as notas reais. Após validado, o protótipo poderia até mesmo ser expandido para capturar automaticamente novas avaliações da Amazon e prever suas classificações, auxiliando empresas a entender rapidamente a percepção dos consumidores sobre seus produtos.

BASE DE DADOS

A base de dados foi extraída da plataforma Kaggle e tem o seguinte contexto:

Esse conjunto de dados contém avaliações de comidas feitas por usuários da Amazon, incluindo tanto o texto livre escrito pelo cliente quanto informações estruturadas como o identificador do produto, data, usuário e, principalmente, a nota atribuída (de 1 a 5 estrelas). O objetivo principal é explorar a relação entre o conteúdo textual da avaliação e a nota correspondente, possibilitando a criação de modelos de predição.

O conjunto de dados (base de dados original) Reviews.csv possui um total de 568.454 avaliações, abrangendo mais de 74.000 produtos e cerca de 256.000 usuários distintos. Cada entrada inclui, entre outros campos:

- Text: o comentário textual do usuário;
- Score: a nota atribuída pelo usuário (variando de 1 a 5);
- Summary: um resumo curto da avaliação;
- UserId e ProductId: identificadores do usuário e do produto.

ATÉ DUPLA (de 1 ou 2 alunos no grupo)

Aqui, use **obrigatoriamente** o arquivo Cria base de dados Treino e Teste – DUPLA.ipynb para obter mensagens aleatórias da base de dados original. Ao rodar esse notebook, serão criados dois arquivos no seu computador no mesmo diretório que salvou este notebook. Esses novos arquivos, com extensão csv, terão o username que colocar ao rodar o notebook. Por exemplo,

dados_treino_DUPLA_username.csv
dados_teste_DUPLA_username.csv

ATÉ TRIO (de 1 a 3 alunos no grupo)

Aqui, use **obrigatoriamente** o arquivo Cria base de dados Treino e Teste – ATE_TRIO.ipynb para obter mensagens aleatórias da base de dados original. Ao rodar esse notebook, serão criados dois arquivos no seu computador no mesmo diretório que salvou este notebook. Esses novos arquivos, com extensão csv, terão o *username* que colocar ao rodar o notebook. Por exemplo,

dados_treino_ATE_TRIO_username.csv
dados_teste_ATE_TRIO_ username.csv

ETAPAS DO PROJETO

Para entregar um projeto de sucesso, você deve seguir os seguintes passos:

1. Criação das base de dados de treinamento e de teste

Usando o notebook de acordo com a quantidade de alunos no grupo, crie os conjuntos de dados necessários para contruir seu classificador a partir do Teorema de Bayes.

2. Montando SEU classificador Naive-Bayes (Boot)

Use o arquivo **Projeto1_Template.ipynb** disponibilizado no Blackboard como template para construir o Projeto 1 conforme demanda abaixo.

Nesta etapa do projeto, use o arquivo TREINAMENTO **dados_treino_....csv**, já que o objetivo aqui é ensinar o seu classificador quais são as palavras mais comuns (frequentes) nas mensagens de **cada** categoria.

Nesse caso, seu código deve conter preferencialmente:

- ✓ Limpeza de mensagens removendo os caracteres: enter, :, ", ', (,), etc.
- ✓ Proposta de outras limpezas/transformações que não afetem a qualidade da informação.
- ✓ Suavização de Laplace: link1 (com leitura até **antes** da seção "Creating a naive bayes classifier with Monkeylearn") e link2.

3. Verificando a *performance*

Nesta seção, use o arquito TESTE **dados_teste_....csv**, já que seu objetivo aqui é testar a qualidade do seu classificador (seu Boot).

Para tanto, você deve extrair as seguintes contagens:

- ✓ Porcentagem de verdadeiros positivos (Ex: mensagens relevantes e que são classificadas como relevantes)
- ✓ Porcentagem de falsos positivos (Ex: mensagens irrelevantes e que são classificadas como relevantes)
- ✓ Porcentagem de verdadeiros negativos (Ex: mensagens irrelevantes e que são classificadas como irrelevantes)
- ✓ Porcentagem de falsos negativos (Ex: mensagens relevantes e que são classificadas como irrelevantes)
- ✓ Acurácia (mensagens corretamente classificadas, independente da categoria)

4. Análise Qualitativa da Performance do Classificador

A avaliação qualitativa dos percentuais obtidos no modelo de classificação é essencial para entender seu desempenho e possíveis limitações. Ao comparar as métricas de acurácia, precisão (útil em

contextos onde falsos positivos são mais problemáticos, como diagnóstico médico) e *recall* (ou sensibilidade, quando perder um positivo é mais problemático do que um falso positivo, como na detecção de fraudes), podemos identificar padrões de acertos e erros, ajudando na interpretação dos resultados. Nesse caso, faça:

- ✓ Um comparativo qualitativo sobre os percentuais obtidos para que possa discutir a *performance* do seu classificador.
- ✓ Explique como devem ser tratadas as mensagens com dupla negação (“Não acho que ele não esteja certo”) e sarcasmo (“Que ótimo, meu voo foi cancelado!”).
- ✓ Proponha um plano de expansão. Por que eles devem continuar financiando o seu projeto?
- ✓ Proponha diferentes cenários de uso para o classificador Naive-Bayes. Pense em outros cenários sem intersecção com este projeto.
- ✓ Sugira e explique melhorias reais no seu classificador com indicações concretas de como implementar (não é preciso codificar, mas indicar como fazer. Indique material de pesquisa sobre o assunto).

5. Qualidade do Classificador a partir de novas separações das mensagens entre Treinamento e Teste

Um importante passo no aprendizado de máquina é trabalhar com uma boa base de dados para o treinamento e teste do seu classificador. Entretanto, é razoável pensar que a divisão de dados utilizada no seu Classificador representa uma entre muitas possíveis combinações em dividir o total de mensagens em treinamento e em teste.

Assim sendo, aqui o objetivo é avaliar como as mensagens contidas na base de dados de treinamento podem interferir numa melhor ou não tão boa classificação das mensagens contidas na base de teste.

Nesse caso, faça:

- ✓ Junte todas as mensagens do **Treinamento** e do **Teste** em único *dataframe* (DUPLA: 2400; até TRIO: 3800) e separe as mensagens, de forma aleatória **mantendo as proporções**, em 10 grupos de mesmo tamanho. **Obs.: Apenas aqui seu grupo poderá usar alguma biblioteca que possua um comando já pronto que realiza essa separação na base de dados (procure no google "StratifiedKFold in python")**;
- ✓ Para cada grupo, faça os itens de 2 a 3 descritos no tópico **Etapas do projeto** utilizando esse grupo como teste e todos os outros como treinamento e guarde os percentuais de acertos (= soma da diagonal principal do cruzamento entre verdadeiro rótulo com rótulo sugerido pelo classificador);
- ✓ Repita os dois passos acima 10 vezes.

Construa um histograma com esses percentuais de acertos e discuta o resultado do histograma refletindo sobre possíveis vantagens ou desvantagens sobre construir um Classificador considerando uma única vez a divisão da base de dados em treinamento e em teste.

Ciência dos dados - ENGENHARIA 2025 / 2º. Semestre

Profa. Maria Kelly Venezuela e Prof. Márcio F. Stabile Jr.

REGRAS

1. O Projeto 1 é em até DUPLA. No caso de TRIO, terá base de dados e rubrica diferentes para seguir. Se o grupo for em DUPLA e usar a base de dados com três categorias, poderá ser bonificado pela rubrica.

2. O projeto será corrigido conforme os critérios da rubrica.

3. Use os **notebooks** disponibilizados no Blackboard.

4. Os entregáveis deverão ser colocados no Blackboard:

   ✓ Arquivos notebooks com o código para obter as mensagens e com código do classificador, seguindo layout dos notebooks disponibilizados na pasta Projeto 1.

   ✓ Arquivos csv treinamento e teste.

   A estrutura do documento deve ser clara e de fácil compreensão da linha de raciocínio. Nesse caso, o notebook não deve haver excesso de impressões não discutidas de variáveis e de *dataframe*.

   Aconselhamos fazer uma análise geral e, após finalizada, salve com outro nome, limpe seu IPython Notebook apenas com os resultados relevantes e melhore seu texto.

ENTREGAS

As entregas deverão ser feitas via Blackboard, nos locais relacionados à atividade. Caso etapas sejam atrasadas, haverá desconto conforme disponível no cronograma.

Ciência dos dados - ENGENHARIA 2025 / 2º. Semestre

Profa. Maria Kelly Venezuela e Prof. Márcio F. Stabile Jr.

CRONOGRAMA

DATA	Finalização:
27/08 (quarta)	Cadastro do grupo no Blackboard: ✓ DUPLA ou TRIO formado para PROJETO 1.
29/08 (sexta)	Deve estar no Blackboard até 23h59: ✓ Leiam a seção CRIAÇÃO DAS BASE DE DADOS DE TREINAMENTO E DE TESTE para fazer este item. ✓ Anexar no Blackboard os arquivos: **dados_treino_....csv** e **dados_-teste_....csv** contendo as bases de treinamento e teste**.**
19/09 (sexta) **FINAL**	Deve estar no Blackboard até às 23h59 com as seguintes evidências: ✓ Arquivos Excel **dados_treino_....csv** e **dados_teste_....csv** contendo a as mensagens de treinamento e teste. ✓ Arquivo **Projeto1 Template.ipynb** com o código do classificador e análise dos resultados, seguindo *layout* descrito nesse notebook.

Ciência dos dados - ENGENHARIA 2025 / 2º. Semestre

Profa. Maria Kelly Venezuela e Prof. Márcio F. Stabile Jr.

RUBRICA

NÍVEL	DESCRIÇÃO
I	Não entregou Entregou, mas não conseguiu gerar a base de dados de treinamento e teste corretamente
D	Entregou; A base de dados foi gerada corretamente, mas o classificador apresenta falhas graves. Existem rotinas para cálculos de probabilidades, mas as fórmulas ou cálculos estão incorretos, ou a implementação não funciona. **Qualquer outra rotina (comandos) feitos a mais serão desconsiderados na correção, caso os erros acima ocorram no projeto.**
C	Entregou; Limpou: \n, :, ", ', (,), etc A base de dados foi gerada corretamente e o classificador funciona, mas a análise da performance não está completa ou apresenta erros. Pequenos erros na suavização de Laplace **E** no Naive Bayes estão presentes. (Ex: não usar a frequência correta das palavras, esquecer da priori, entre outros) **Qualquer outra rotina (comandos) feitos a mais serão desconsiderados na correção, caso os erros acima ocorram no projeto.**
B	Entregou; Limpou: \n, :, ", ', (,), etc. O classificador funciona bem e a análise crítica da performance está bem feita, utilizando métricas adequadas. No entanto, há um pequeno erro na suavização de Laplace **OU** no Naive Bayes, mas não em ambos. **Qualquer outra rotina (comandos) feitos a mais serão desconsiderados na correção, caso os erros acima ocorram no projeto.**
CASO SEU PROJETO SE ENQUADRE EM ALGUM DOS NÍVEIS ACIMA, ENTÃO OS ITENS AVANÇADOS SERÃO IGNORADOS; **SENÃO, SEU NÍVEL SERÁ PELA CONTAGEM DE ITENS AVANÇADOS:** **B+ : 3 itens** **A : 4 ou 5 itens** **A+ : 6 ou 7 itens**	IMPLEMENTOU outras limpezas e transformações que não afetem a qualidade da informação contida nas mensagens, mas tendem a melhorar na classificação das mensagens. Ex: stemming, lemmatization, stopwords.
	CONSIDEROU arquivo com três categorias na classificação das variáveis **(OBRIGATÓRIO PARA TRIO, sem contar como item avançado)**
	CONSTRUIU o cálculo das probabilidades corretamente utilizando bigramas E apresentou referência sobre o método utilizado.
	DOCUMENTOU bem o código, com explicações claras para cada etapa do processo, incluindo o raciocínio por trás das decisões de modelagem e das transformações de dados.
	PROPÔS diferentes cenários para Naïve Bayes fora do contexto do projeto (pelo menos dois cenários diferentes, exceto aqueles já apresentados em sala pelos professores: por exemplo, filtro de spam)
	REFLETE criticamente sobre os resultados obtidos, identificando limitações do modelo e sugerindo possíveis melhorias ou diferentes abordagens com indicações concretas de como implementar (indicar como fazer e indicar material de pesquisa).
	FEZ o item 5 (**Qualidade do Classificador a partir de novas separações das mensagens entre Treinamento e Teste**) **(OBRIGATÓRIO para conceitos A ou A+)**