Título: Classificando tópicos no dataset de artigos da BBC

Sessão 1: Dataset: BBC (artigos da BBC) https://storage.googleapis.com/dataset-uploader/bbc/bbc-text.csv

Este documento visa analisar o banco de dados de artigos da BBC e como um classificador para suas 5 categorias possíveis funciona, além de como pode melhorar. Citado no artigo [1] para uso comparativo com outros datasets em um Cluster de documentos, este dataset conta com 2126 valores únicos e 5 categorias rotuladas, sendo a primeira coluna chamada category, contendo uma string para as categorias catalogadas, e a segunda chamada text, contendo o título e o corpo dos artigos concatenados.

Para encontrar esse dataset, eu fui ao site do Kaggle, pesquisei text categorization e cliquei no filtro de Notebooks e no de Datasets, assim encontrando o notebook “BBC text categorization”. Após encontrar esse notebook, fui até a sessão Input, vi que tinha um artigo científico que usou o Dataset, baixei ele e o usei para fazer minhas análises.

**Section 2: Classification pipeline**

Nesta seção, o pipeline de classificação foi definido para pré-processar o texto bruto e treinar modelos de machine learning. O texto foi limpo, removendo-se palavras irrelevantes, stopwords e aplicando lematização com a biblioteca WordNet. O vetor de palavras foi obtido com a técnica Bag-of-Words (BoW), onde a presença de palavras específicas é considerada crucial para a categorização.

Os modelos de classificação incluem Regressão Logística e Naive Bayes, treinados para distinguir entre as categorias dos artigos. Utilizando BoW, supõe-se que palavras individuais têm importância independente para a classificação, o que faz sentido ao identificar palavras-chave que são distintivas de cada categoria. Por exemplo, a palavra "match" é frequentemente associada à categoria de esporte. No entanto, essa abordagem pode ser explorada maliciosamente; se um documento contiver deliberadamente palavras de categorias erradas, o classificador poderá ser confundido, resultando em uma categorização errônea.

**Processos executados**:

* Pré-processamento: tokenização, remoção de stopwords e lematização;
* Vetorização: Bag-of-Words para converter o texto em vetores de características;
* Modelos: Regressão Logística e Naive Bayes.

**Section 3: Evaluation**

Os modelos foram avaliados utilizando várias métricas, incluindo acurácia balanceada, devido à natureza potencialmente desbalanceada do dataset. Como o dataset não possui uma divisão fixa de treino e teste, as divisões foram feitas aleatoriamente, mantendo-se a estratificação para garantir uma representação justa de todas as categorias.

Para cada modelo, foram feitas múltiplas iterações com embaralhamentos diferentes. Os resultados mostraram que o Naive Bayes teve uma ligeira vantagem em termos de precisão geral em relação à Regressão Logística. No entanto, observou-se que certos termos possuem um impacto exagerado, devido ao seu uso frequente em contextos fora de suas categorias principais. Por exemplo, a palavra "government" pode aparecer tanto em artigos de negócios quanto de política, causando confusão.

Matriz de confusão e relatórios de classificação foram gerados para analisar os erros de classificação. O modelo Naive Bayes teve melhor desempenho para categorias claramente distintas, como esporte e tecnologia, enquanto houve mais confusão entre negócios e política.

**Section 4: Dataset size**

Nesta seção, foi realizada uma avaliação do impacto do tamanho do dataset sobre os erros de treino e teste. Diversos níveis de amostragem foram usados para verificar a variabilidade da performance dos modelos conforme o número de exemplos disponíveis aumentava. Ao avaliar erros em múltiplas proporções do dataset, observou-se que tanto Regressão Logística quanto Naive Bayes mostram uma diminuição nos erros à medida que o tamanho do conjunto de dados aumenta, indicando que mais dados podem melhorar a performance.

No entanto, é importante considerar que, para algumas categorias, o aumento do número de exemplos não necessariamente melhora a acurácia, devido à sobreposição de tópicos. Em termos de viabilidade comercial, aumentar o dataset para além de um certo ponto pode ser difícil, especialmente se as categorias compartilharem muitas palavras em comum.

**Conclusão**: Embora um aumento no tamanho do dataset tenha contribuído para reduzir o erro de classificação, a taxa de melhoria diminui significativamente à medida que mais dados são adicionados, sugerindo um limite prático para esse aumento.

**Section 5: Topic analysis**

A análise de tópicos revelou insights importantes sobre a estrutura dos dados e o desempenho dos classificadores. Através do uso de nuvens de palavras e análise de tópicos, observou-se que certas categorias possuem palavras predominantes, como "government" em política e "technology" em tecnologia, o que ajudou na classificação. Entretanto, a sobreposição de tópicos entre negócios e política aumentou o erro nesses grupos, sugerindo a possibilidade de confusão.

Além disso, um classificador de duas camadas foi testado. Nessa abordagem, os documentos foram primeiro classificados por tópicos principais (usando modelos de análise de tópicos como LDA), e depois encaminhados para classificadores mais especializados para cada tópico. Este método reduziu os erros em categorias com mais sobreposição de palavras, como negócios e política, resultando em uma melhora geral da performance.

**Conclusão**: A análise de tópicos demonstrou que certas categorias possuem vocabulário muito distinto, permitindo uma boa classificação. No entanto, a sobreposição de tópicos em categorias relacionadas dificulta a tarefa de classificação, especialmente com métodos simples como Bag-of-Words.

[1] D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006. [PDF] [BibTeX].