FEATURE EXTRACTION

A extração de características textuais refere-se ao processo de identificar e representar informações essenciais e distintivas contidas em documentos de texto. Por meio de métodos como o TF-IDF, word embeddings, ou técnicas mais avançadas de processamento de linguagem natural, a extração de características permite transformar texto não estruturado em representações numéricas que podem ser utilizadas para análise, classificação, agrupamento e outras tarefas de mineração de texto. Essas características capturam a essência do conteúdo textual, permitindo que algoritmos de aprendizado de máquina e técnicas de análise de dados tirem insights significativos e tomem decisões informadas a partir de grandes volumes de informações textuais.

TFIDF

O TF-IDF (Term Frequency-Inverse Document Frequency) é usado para avaliar a importância de um termo (ou palavra) em um documento dentro de um conjunto de documentos (corpus). Esse método considera tanto a frequência do termo em um documento quanto sua raridade em todo o corpus, o que ajuda a identificar quais termos são mais discriminativos e informativos para um determinado documento em relação ao conjunto de documentos.

Frequência do Termo (TF - Term Frequency):

A Frequência do Termo é uma medida da frequência com que um termo específico aparece em um documento. Geralmente, é calculada usando a fórmula:

$$TF(t,\ d)\ =\ \frac{Número\ de\ vezes\ que\ o\ termo\ t\ aparece\ no\ documento\ d}{Número\ total\ de\ termos\ no\ documento\ d}$$

Frequência Inversa nos Documentos (IDF - Inverse Document Frequency):

A Frequência Inversa nos Documentos é uma medida da raridade de um termo em todo o corpus. Ela é calculada da seguinte maneira:

$$IDF(t,\ D)\ =\ log\left(\frac{D}{1 + Número\ de\ documentos\ que\ contêm\ o\ termo\ t}\right)$$

O "+1" no denominador é usado para evitar divisão por zero quando um termo não aparece em nenhum documento e D é o número total de documentos.

Pontuação TF-IDF:

A pontuação TF-IDF para um termo em um documento é calculada multiplicando-se a Frequência do Termo (TF) pela Frequência Inversa nos Documentos (IDF) para o mesmo termo:

$$TF - IDF(t,\ d,\ D)\ =\ TF(t,\ d)\ *\ IDF(t,\ D)$$

Após calcular as pontuações TF-IDF para todos os termos em todos os documentos, você obtém uma matriz de características onde cada linha representa um documento e cada coluna representa um termo, e os valores nas células são as pontuações TF-IDF.

O TF-IDF é uma técnica poderosa para a extração de características de texto, pois permite que você identifique termos-chave em documentos e pondere sua importância relativa. No entanto, ele tem algumas limitações, como a falta de consideração de informações semânticas e a sensibilidade a erros ortográficos. Portanto, em algumas tarefas de PLN mais avançadas, métodos baseados em redes neurais, como word embeddings (incorporação de palavras), tornaram-se populares.

WORD EMBEDDINGS BERT

Os word embeddings são representações vetoriais de palavras em um espaço multidimensional, que capturam a semântica e a relação entre palavras com base no contexto em que elas aparecem em um corpus de texto. Uma das mais notáveis e avançadas abordagens em word embeddings é o modelo BERT (Bidirectional Encoder Representations from Transformers), que revolucionou o processamento de linguagem natural.

Ao contrário de métodos anteriores, como o Word2Vec e o GloVe, o BERT utiliza uma arquitetura de rede neural transformer bidirecional para treinar embeddings contextualizados de palavras. Isso significa que, em vez de considerar o contexto de uma

palavra apenas à esquerda ou à direita dela em uma sentença, o BERT considera todas as palavras na sentença, tornando seus embeddings mais ricos e sensíveis ao contexto.

O BERT não é apenas uma ferramenta poderosa para tarefas de PLN, como classificação de texto, tradução automática e resumo de texto, mas também é extremamente útil para a extração de características de texto. Isso ocorre porque os embeddings de palavras gerados pelo BERT capturam nuances semânticas e relações entre palavras de forma muito precisa. Para realizar feature extraction usando o BERT, pode-se utilizar os embeddings de palavras ou mesmo embeddings de frases ou documentos inteiros.

Para extrair características de texto com o BERT, um documento é tokenizado em palavras ou subpalavras, e esses tokens são alimentados no modelo BERT pré-treinado. O modelo retorna os embeddings correspondentes para cada token. Esses embeddings podem ser agregados de várias maneiras (como média, soma, ou outro método) para criar uma representação vetorial do documento ou sentença. Essa representação vetorial pode então ser usada em tarefas de aprendizado de máquina ou análise de dados, permitindo a extração de características ricas e contextuais do texto.

Em resumo, o modelo BERT e outros modelos de embeddings contextuais revolucionaram a extração de características de texto, permitindo que as informações semânticas e contextuais sejam capturadas de forma mais precisa e eficaz do que métodos anteriores. Isso abre um leque de oportunidades para análise de texto avançada e tarefas de PLN que requerem uma compreensão profunda do conteúdo textual.