Projeto de Machine Learning - Etapa 3

Análise de sentimento de reviews
Pré-processamento

Larissa Lewartoski Wong

Conjuntos de dados envolvidos no pré-processamento

O dataset original foi copiado para dois conjuntos distintos, um com apenas Score 1 e 5 e outro com os Scores 1, 2 e 5.

	HelpfulnessNumerator	HelpfulnessDenominator	Score	Text
0	1	1	5	I have bought several of the Vitality canned d...
1	0	0	1	Product arrived labeled as Jumbo Salted Peanut...
4	0	0	5	Great taffy at a great price. There was a wid...
6	0	0	5	This saltwater taffy had great flavors and was...
7	0	0	5	This taffy is so good. It is very soft and ch...
...	...	...	...	...

	HelpfulnessNumerator	HelpfulnessDenominator	Score	Text
0	1	1	5	I have bought several of the Vitality canned d...
1	0	0	1	Product arrived labeled as Jumbo Salted Peanut...
3	3	3	2	If you are looking for the secret ingredient i...
4	0	0	5	Great taffy at a great price. There was a wid...
6	0	0	5	This saltwater taffy had great flavors and was...
...	...	...	...	...

Etapas do pré-processamento

- Não foi necessário tratar valores ausentes;
- Atributos que não seriam utilizados foram removidos;
- Caracteres especiais foram removidos e as letras foram colocadas em lower case;
- Foram realizados: tokenização das palavras, remoção de stopwords e stemming;
- Divisão entre conjunto de treinamento e de teste;
- Undersampling para diminuir a diferença entre as classes.

Dados após stemming e remoções de stopwords

- Original

  Right now I'm mostly just sprouting this so my cats can eat the grass. They love it. I rotate it around with Wheatgrass and Rye too

- Snowball Stemmer

  right im most sprout cat eat grass love rotat around wheatgrass rye

- Porter Stemmer

  right im mostli sprout cat eat grass love rotat around wheatgrass rye

Obrigada