Modelo para o processamento de Linguagem Natural

Chat-BTG

Objetivo da Sprint

Lematização

Modelo Word2Vec

AGENDA

Retrospectiva Sprint 2

Lematização

Modelo Word2Vec

Resultados

Cronograma

01

Retrospectiva Sprint 2

Retrospectiva da Sprint 2

Entendimento da base

Análise descritiva

Pré-processamento

Bag of Words

02

Lematização

Pré-processamento

- Colocar letras minúsculas e *tokenização*
- Retirada de *stop words*
- *Stemming*
- *Lematização*

Lematização

-Processo de transformação de palavras para sua forma base (derivação inversa)

-O processo foi feito a partir da **biblioteca spaCy**

Estou : Estar

Taxas : Taxa

Gosto : Gostar

Altíssima : Alta

Remuneração : Remunerar

Abreviações

- Convertemos abreviações para a melhoria da base de dados

```
'vc': 'voce',
'vcs':'voce',
'Vc': 'voce',
'pq': 'porque',
'tbm': 'tambem',
'q': 'que',
'td': 'tudo',
'blz': 'beleza',
'flw': 'falou',
'kd': 'cade',
'to': 'estou',
'mt': 'muito',
'cmg': 'comigo',
'ctz': 'certeza',
'jah': 'ja',
```

```
'naum': 'nao',
'ta': 'esta',
'eh': 'e',
'vlw': 'valeu',
'p': 'para',
'qnd': 'quando',
'msm': 'mesmo',
'fzr': 'fazer',
'agr': 'agora',
'btgpactual': 'btg',
'pactual': 'btg',
```

Emojis

-Corversão de emojis para a melhoria da base de dados

```
pipeline('Eu gosto de investir nesse banco 🚀 ')
```

```
['gostar', 'investir', 'em esse', 'banco', 'foguete']
```

Nova Tabela

- Nova tabela com coluna de frases lematizadas

	autor	texto	sentimento	tokens_lema
0	v8_capital	Confira os resultados dos nossos fundos no mês...	NEUTRAL	[confira, o, resultado, de o, nosso, fundo, me...
1	winthegame_of	A Alvarez & Marsal estará conosco no Sportainm...	NEUTRAL	[Alvarez, Marsal, estara, conosco, sportainmet...
2	marta_bego	#Repost btgpactual with make_repost · · · Entend...	NEUTRAL	[repost, btg, With, makerepost, entenda, o, im...
3	Imviapiana	Minuto touro de ouro	POSITIVE	[minuto, touro, ouro]
4	vanilson_dos	@ricktolledo Sim	NEUTRAL	[Ricktolledo, sim]

03

Modelo Word2Vec

Vetorização

-Processo de transformar dados textuais em representações numéricas

	tokens_lema	vec1	vec2	vec150	sentimento
0	['confira', 'o', 'resultado', 'de o', 'nosso',...	0.262776	-0.351657	-0.315743	0
1	['Alvarez', 'Marsal', 'estara', 'conosco', 'sp...	0.029775	-0.371140	-0.509014	0
2	['repost', 'btg', 'With', 'makerepost', 'enten...	0.222636	-0.364943	-0.431190	0
3	['minuto', 'touro', 'ouro']	0.019667	0.067234	-0.002506	1
4	['Ricktolledo', 'sim']	0.304922	-0.007242	-0.183279	0
...	...	...	...	...	...
12169	['um', 'noite', 'encontro', 'muito', 'conhecim...	0.385639	-0.163126	-0.428794	0
12170	['erro', 'financeiro', 'eliminar', 'antes', 'd...	0.501189	-0.143322	-0.398472	0
12171	['estar', 'muito', 'grato', 'todo', 'esforco',...	0.574576	0.287958	-0.245109	1
12172	['dorsodamaocomdedoindicadorapontandoparaadire...	0.304480	0.013746	-0.196458	0
12173	['btg', 'Morning', 'call', 'nao', 'este', 'mai...	0.503737	0.093182	-0.268835	-1

-Realizada também a conversão númerica da coluna de sentimento

Modelo Word2Vec

- Técnica de processamento
- Representação de palavras como vetores
- Representação da similaridade entre palavras

Modelo Word2Vec

- Biblioteca Gensim
- Modelo CBOW (palavras circundantes -> palavra-alvo)
- Modelo próprio, utilizando 150 vetores

05

Resultados

Naive Bayes

-Algoritmo de aprendizado supervisionado utilizado para classificação do dados

-Utiliza a teoria de probabilidade condicionais

Accuracy: 0.5552361396303901

CatBoost

-Outro algoritmo de aprendizado supervisionado

-Utiliza técnicas de gradiente (gradiente boosting)

-Se destaca em lidar com dados categóricos

```
Acurácia de treinamento: 0.9519457849881918
Acurácia de teste: 0.7232032854209446
```

06

Cronograma

Etapa	Status	Sprint
Entendimento de Negócios	Feito	1
Mapeamento de Persona	Feito	1
Modelo de Bag of Words	Feito	2
Modelo utilizando Word2Vec e Análise de sentimento	Feito	3
Pré-processamento e desenvolvimento da proposta	Em progresso	4
Solução Final em formato MVP	A iniciar	5

Muito obrigado pela atenção!

Estamos disponíveis para maiores esclarecimentos.