

Análise resultados obtidos

DataSet

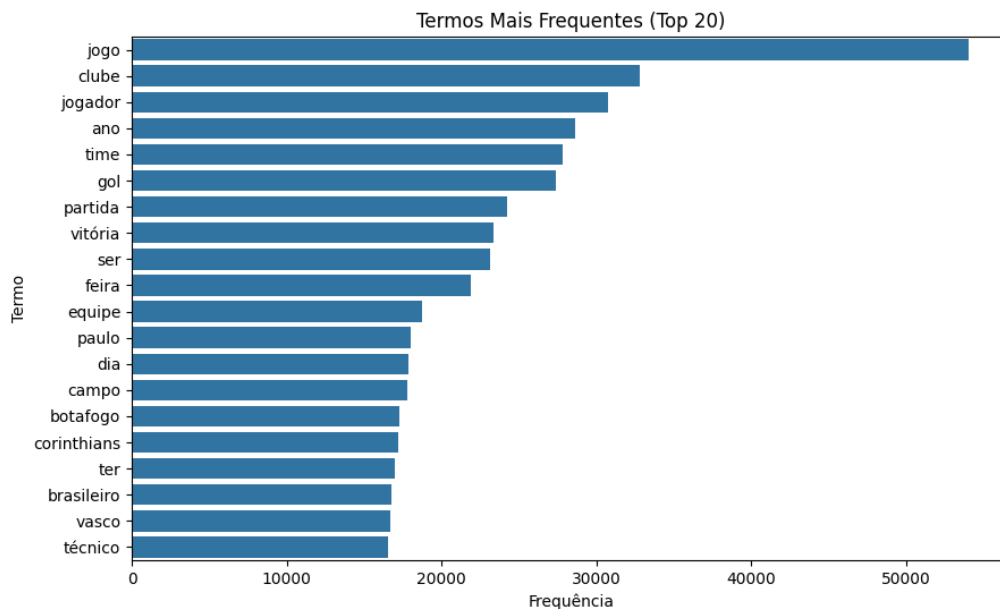
DataSet: o DS é representado no arquivo combined_news_data.csv, que combina em média 1400 notícias por time de todos os times (20) da série A brasileira, sendo assim tendo em média 28.000 notícias. Foram mantidos todos os esquemas de limpeza do último trabalho e adicionado uma variação adicional que exclui tais palavras:

```
words_to_remove = ['ge', 'globo', 'sportv', 'canal', 'podcast', 'clique', 'sequin', 'siga', 'sigar', 'facebook', 'GEPR', '+ GEPR', 'whatsapp']
```

Algoritmos aplicados

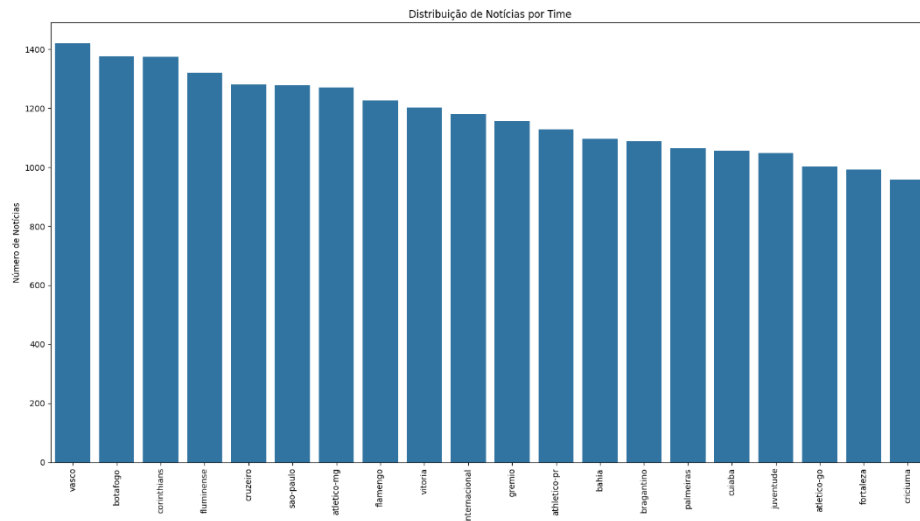
Bag Of Words:

Top 20 termos mais frequentes



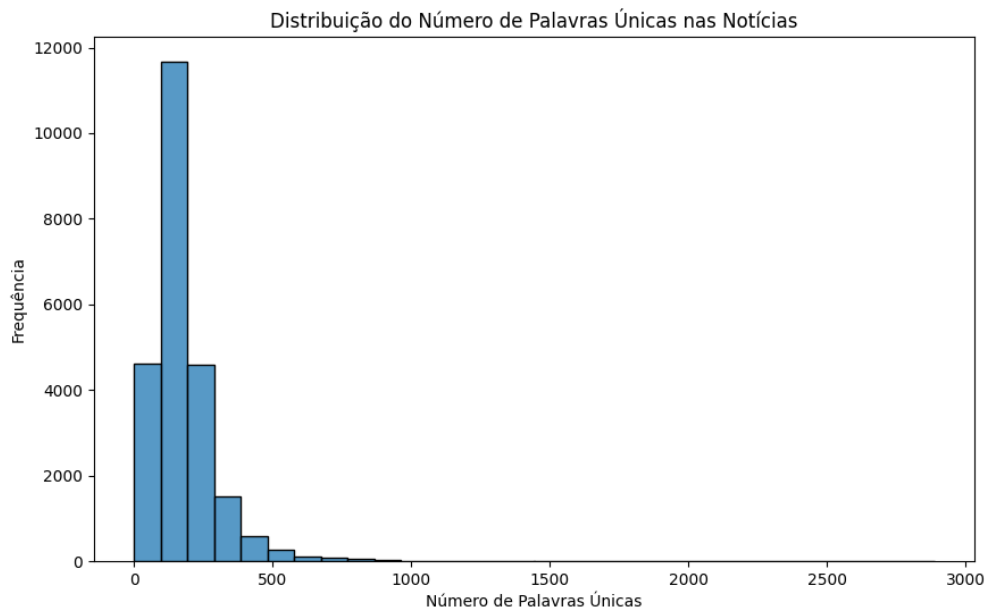
Podemos analisar que por mais que as palavras mais frequentes sejam generalistas (sirvam para os mesmos times), algoritmos mais simples como por exemplo o BoW, se saem bem clusterizando e classificando os times

Distribuição de Notícias por Time



Apenas para mostrar como o dataset é bem equilibrado!

Distribuição do Número de Palavras Únicas nas Notícias



A distribuição mostra que a maioria das notícias possui um número baixo de palavras únicas, concentrando-se entre 0 e 500 palavras. Isso sugere que muitas notícias são breves ou repetem termos comuns, típico em textos esportivos. O que parece não atrapalhar o resultado do algoritmo em BOW

Matriz de Confusão

Cluster Predito	athletico-pr	athletico-go	athletico-mg	bahia	botalogo	bragantino	corinthians	crucima	cruzeiro	cuiaba	flamengo	fluminense	fortaleza	gremio	internacional	juventude	palmeiras	sao-paulo	vasco	vitoria
athletico-pr	1059	0	0	2	6	0	0	0	0	1	7	7	1	3	2	0	0	0	5	1
athletico-go	33	891	111	32	57	24	25	31	113	101	68	58	64	61	64	29	38	58	48	19
athletico-mg	9	30	1064	2	0	12	2	0	36	16	5	9	11	3	1	2	4	2	11	5
bahia	0	0	0	1040	3	4	2	0	1	1	7	3	1	0	0	2	0	2	1	6
botalogo	2	0	0	0	1295	0	0	0	0	0	4	0	1	0	0	0	7	0	1	0
bragantino	0	0	0	0	2	1019	5	0	0	2	0	6	1	0	0	1	0	3	2	0
corinthians	0	0	0	0	0	2	1325	0	2	3	2	0	1	0	1	0	0	1	0	0
crucima	0	1	1	0	0	2	0	910	1	2	0	0	0	1	2	1	0	0	0	1
cruzeiro	13	3	23	2	4	7	1	2	1068	6	13	14	2	4	18	3	5	14	6	4
cuiaba	7	63	38	1	2	9	2	5	39	898	9	5	9	1	3	44	0	1	1	4
flamengo	1	0	2	2	1	0	2	0	0	1	1087	3	1	1	0	0	0	0	0	2
fluminense	1	3	7	9	3	2	1	6	7	4	11	1201	4	2	1	11	0	0	5	3
fortaleza	0	0	1	2	0	0	0	1	4	1	1	0	690	1	0	1	0	0	0	2
gremio	0	0	0	0	0	0	0	0	0	0	0	0	0	1077	0	1	0	0	1	1
internacional	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1085	7	0	0	0	0
juventude	3	4	20	0	0	5	0	2	2	12	2	3	14	2	3	996	1	0	0	5
palmeiras	0	0	1	1	0	0	0	0	6	0	3	0	1	0	0	0	1006	4	1	1
sao-paulo	0	0	1	0	0	0	0	0	0	0	0	4	0	0	0	1	2	1192	0	2
vasco	0	0	0	0	3	0	7	0	0	0	1	3	0	0	0	0	0	0	1336	0
vitoria	0	7	2	3	0	3	2	2	1	9	7	5	2	0	0	9	1	2	2	1147

A matriz de confusão mostra que o modelo obteve alta precisão para muitos times, com a maioria das notícias corretamente classificadas ao longo da diagonal principal. No entanto, alguns times apresentam confusão significativa, como Atlético-GO e Cuiabá, que têm várias previsões incorretas para outros times, sugerindo características linguísticas similares entre suas notícias. Essa confusão pode indicar uma limitação do BoW em capturar contextos específicos dos times, já que ele não considera a ordem das palavras.

Precisão do BoW

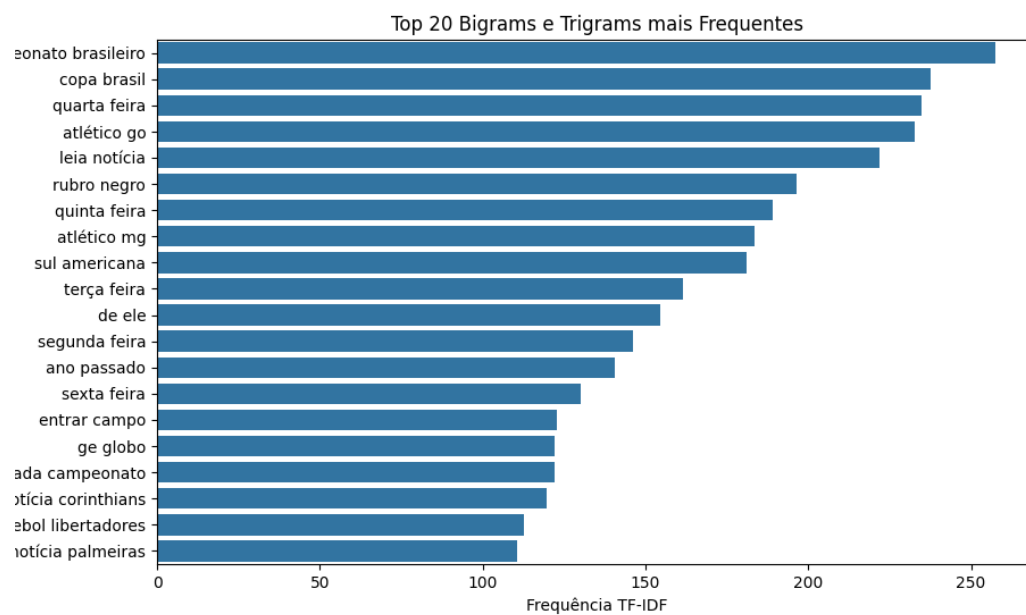
```

Run bagOfWords x
C:\Users\lomelo\AquaProjects\untitled\venv\Scripts>python main.py
Acurácia com 400 clusters: 0.91
Process finished with exit code 0

```

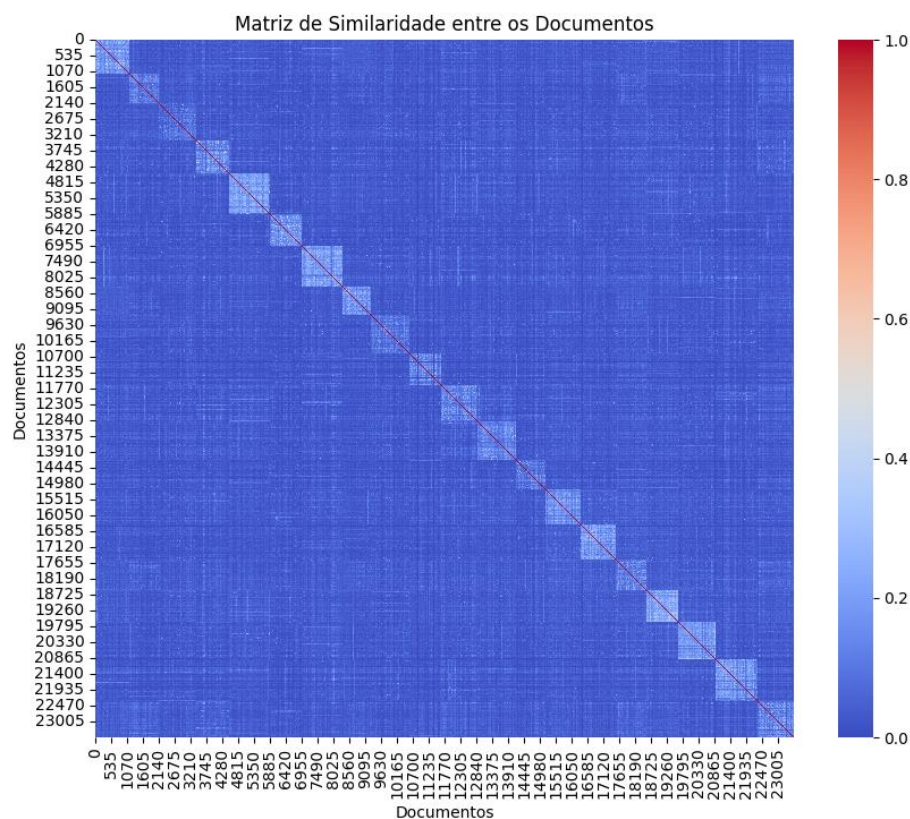
A precisão de 91% é surpreendentemente alta para um modelo simples como o Bag of Words, indicando que as notícias contêm palavras suficientemente distintas para diferenciar a maioria dos times.

TF-IDF:



A análise é semelhante ao que printamos do bow primeiro, termos generalistas são mais frequentes, geralmente relacionados a competições disputadas pelas equipes

Matriz de Similaridade entre documentos



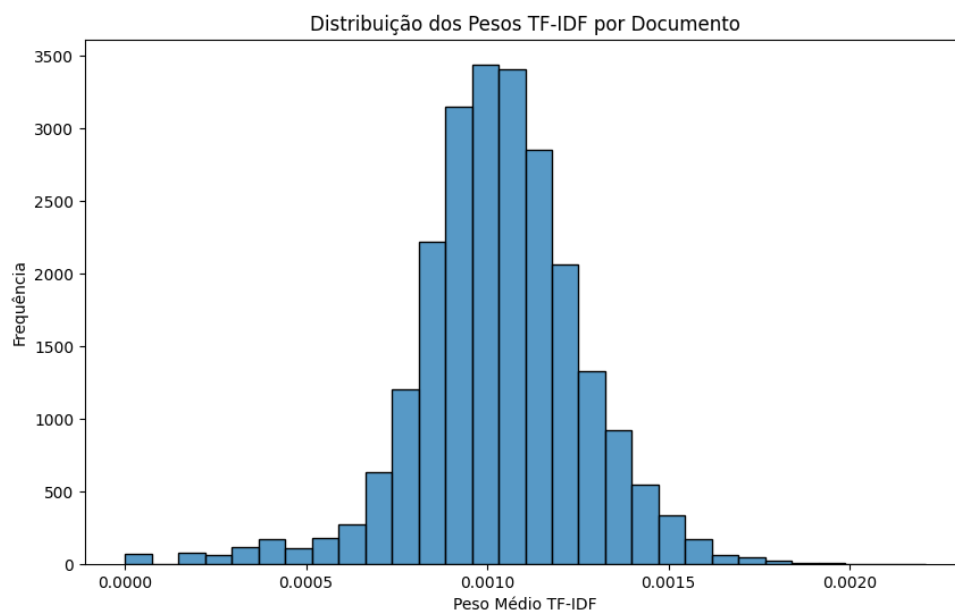
A matriz de similaridade revela uma baixa similaridade geral entre as notícias, o que indica diversidade temática no dataset. Os blocos mais claros ao longo da diagonal sugerem a formação de clusters, onde as notícias têm maior similaridade interna, provavelmente agrupando conteúdos relacionados ao mesmo time. Como o código mapeia os clusters ao time mais comum em cada grupo, esses blocos reforçam que as notícias sobre um time são mais semelhantes entre si do que com notícias de outros times.

Matriz de confusão

		Matriz de Confusão																			
Cluster Predito	athletico-pr	1092	1	3	1	2	0	0	0	5	2	1	0	1	0	0	0	0	0	0	0
	athletico-go	0	965	5	2	5	5	0	4	4	19	13	2	10	5	4	52	0	1	1	5
	athletico-mg	21	19	1216	22	10	17	17	18	121	70	62	53	34	18	20	33	36	25	23	38
	bahia	1	0	1	1002	0	0	0	0	2	0	4	0	1	0	0	0	0	0	0	0
	botafoogo	0	1	0	0	1322	0	0	0	0	0	2	2	2	0	0	1	10	6	0	0
	bragantino	0	3	0	0	2	1038	1	0	0	5	0	0	0	0	0	2	0	0	0	0
	corinthians	2	1	0	0	0	0	1323	0	24	1	2	0	2	0	0	0	3	2	0	0
	criciuma	1	0	1	0	0	0	0	924	0	0	0	0	0	0	0	1	0	0	0	0
	cruzeiro	4	2	26	3	3	5	9	3	1100	16	19	14	22	6	8	6	11	7	5	1
	cuiaba	0	0	2	0	0	0	0	0	0	925	0	0	0	0	0	1	0	0	0	0
	flamengo	7	2	5	2	14	17	12	4	9	9	1078	17	3	6	1	3	7	23	15	10
	fluminense	0	0	5	0	5	0	0	0	4	0	1	1168	7	1	0	0	0	0	5	0
	fortaleza	0	0	1	5	0	0	1	1	0	1	1	0	900	1	0	0	0	0	2	0
	gremio	0	0	0	0	1	0	0	0	0	0	0	1	1	1098	0	0	0	0	0	0
	internacional	0	0	0	1	0	0	0	0	0	0	0	0	0	19	1146	19	0	0	0	1
	juventude	0	2	0	0	0	0	0	0	0	0	1	0	0	1	0	923	0	0	0	1
	palmeiras	0	0	0	1	0	0	0	0	3	0	4	0	0	0	0	0	997	3	1	2
	sao-paulo	0	0	0	0	0	0	0	0	0	0	8	4	0	0	0	0	0	1209	3	0
	vasco	0	0	5	0	5	0	6	0	2	2	4	14	2	1	0	0	0	0	1340	0
	vitoria	0	6	1	57	7	7	5	5	6	7	27	46	8	0	1	7	0	3	25	1145
		athletico-pr	athletico-go	athletico-mg	bahia	botafoogo	bragantino	corinthians	criciuma	cruzeiro	cuiaba	flamengo	fluminense	fortaleza	gremio	internacional	juventude	palmeiras	sao-paulo	vasco	vitoria

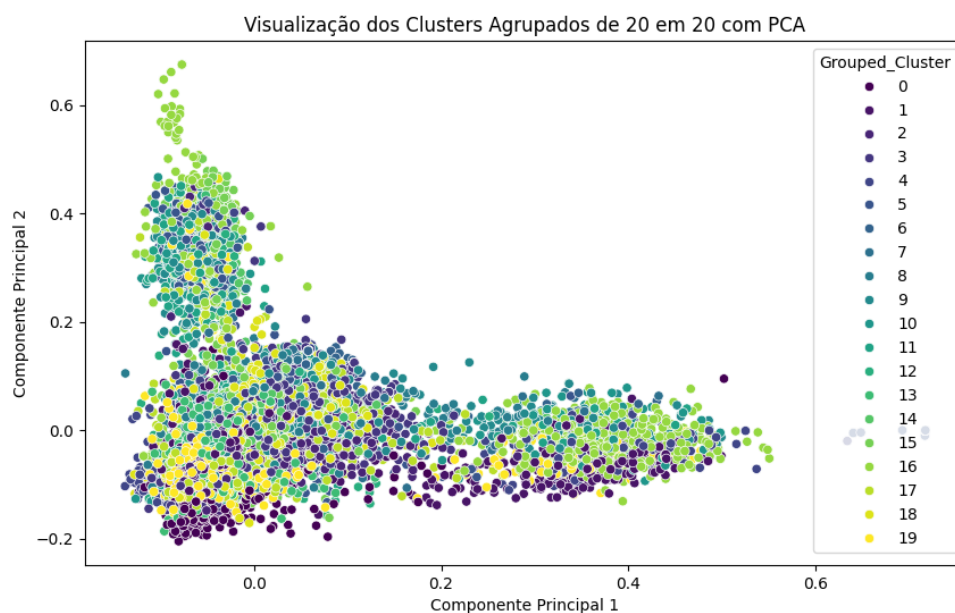
A matriz de confusão mostra que o algoritmo de clusterização conseguiu agrupar as notícias de forma consistente para a maioria dos times, como evidenciado pelos altos valores na diagonal. No entanto, há alguns erros de classificação, especialmente entre times como Athletico-MG e Flamengo, sugerindo semelhanças nos conteúdos que podem confundir o modelo. É curioso que esse mesmo tipo de problema aconteceu para o BOW, mas com times diferentes, provavelmente pois os algoritmos tratam os dados de maneira diferente, mas o dataset tem o mesmo “tipo” de vícios.

Distribuição dos Pesos TF-IDF por Documento



A distribuição dos pesos TF-IDF apresenta uma forma normal, com a maioria dos documentos concentrada em uma faixa média de peso, em torno de 0,001. Esse padrão indica que, em média, os termos não possuem um peso excessivamente alto ou baixo, sugerindo uma diversidade razoável de conteúdo e vocabulário entre os documentos, é curioso tendo em vista que o dataset é desbalanceado quando se trata da quantidade repetida de termos.

Visualização dos Clusters Agrupados de 20 em 20 com PCA



O gráfico de PCA revela uma dispersão clara dos clusters de notícias, com várias densidades e distribuições ao longo dos componentes principais, indicando diferentes

níveis de similaridade entre os grupos de documentos. Parece sugerir que alguns tópicos podem compartilhar termos relevantes, dificultando a separação completa dos temas.

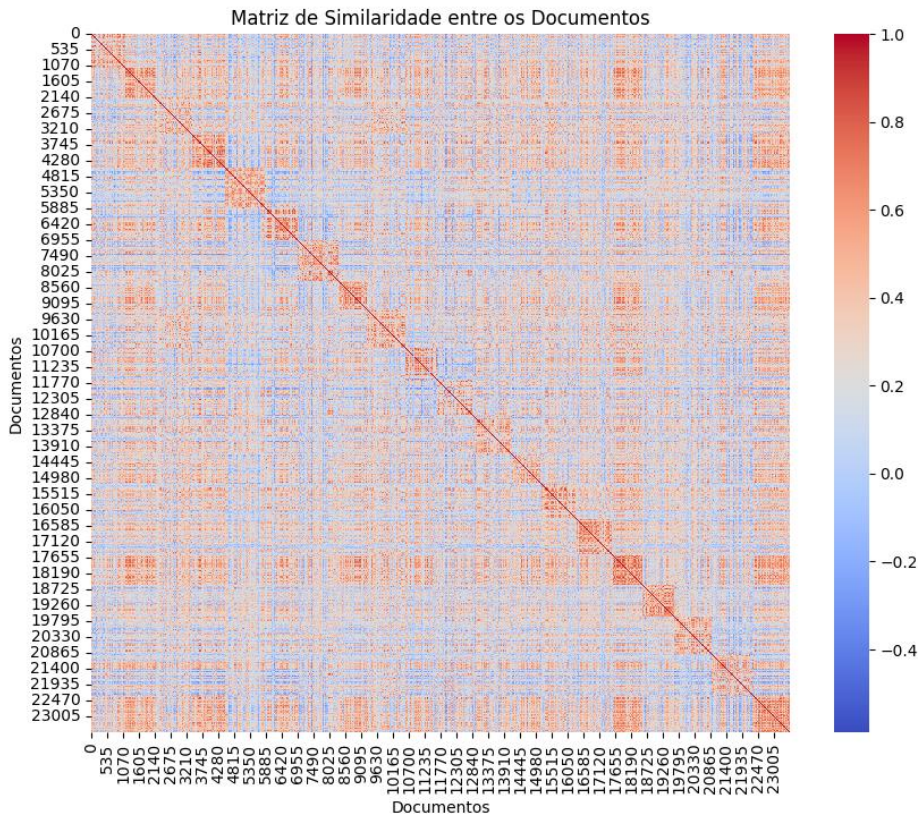
Acuracia TF-IDF

```
Run  tf-idf x
C:\Users\lomelo\AquaProjects\untitled\venv\Scripts\python.exe
Acurácia com 400 clusters: 0.93
```

A acurácia de 93%, mostra que o tf-idf foi eficiente na classificação de notícias por time, sugerindo uma boa captação das características textuais específicas de cada grupo.

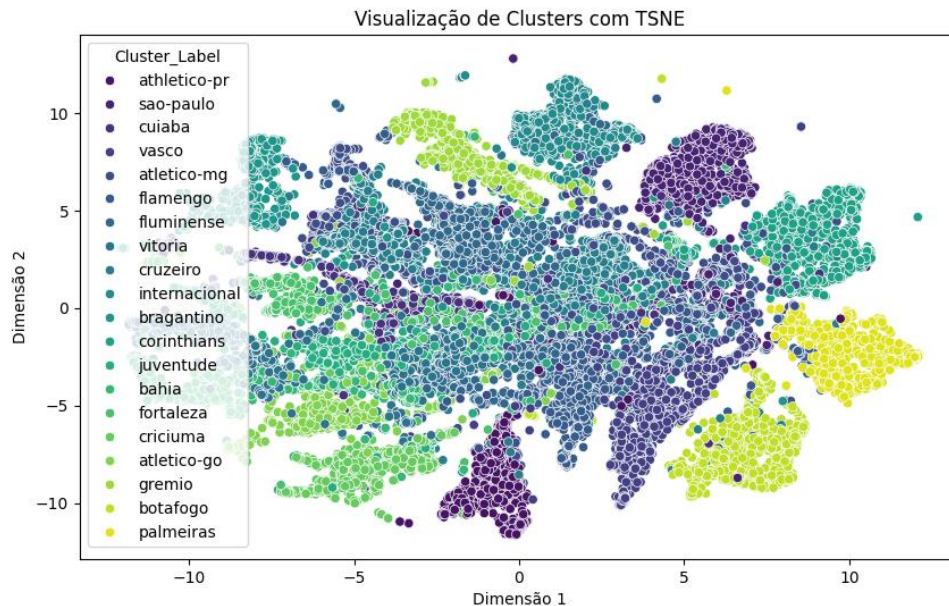
Embeddings com Conv2Word:

Matriz de Similaridade entre os Documentos



A matriz de similaridade entre os documentos revela que a diagonal vermelha intensa representa a autocorrelação perfeita de cada documento consigo mesmo, com similaridade igual a 1, o que é esperado. Fora da diagonal, há uma predominância de tons mais claros e azulados, sugerindo que muitos documentos possuem baixa similaridade entre si, indicando diversidade temática entre as notícias. Em algumas áreas da matriz, são visíveis agrupamentos de tons vermelhos, o que sugere a existência de grupos de documentos mais similares entre si, mesmo em clusters diferentes.

Visualização de Clusters com TSNE



Essa revela algumas informações interessantes sobre a distribuição dos dados e o comportamento do modelo. Os pontos representando diferentes times tendem a se agrupar de acordo com suas similaridades temáticas, o que sugere que o modelo de Word2Vec conseguiu capturar bem o contexto de cada time ao longo dos documentos. Observa-se uma separação relativamente clara entre alguns clusters, indicando que certos times possuem características distintas nas notícias, enquanto outros grupos se misturam um pouco mais, o que pode refletir uma semelhança temática entre as notícias desses times.

Matriz de Confusão entre Clusters e Times

Matriz de Confusão entre Clusters e Times

Cluster	athletico-pr	athletico-go	athletico-mg	bahia	botafoogo	bragantino	corinthians	criciuma	cruzeiro	cuiaba	flamengo	fluminense	fortaleza	gremio	internacional	juventude	palmeiras	sao-paulo	vasco	vitoria
athletico-pr	974	0	6	0	2	13	1	7	3	3	4	8	1	6	1	0	3	5	1	
athletico-go	0	815	11	5	0	1	0	5	3	14	3	2	4	0	1	27	0	1	2	21
athletico-mg	39	25	843	20	21	31	30	7	156	49	50	74	46	31	29	23	21	30	22	31
bahia	4	6	16	886	1	3	5	12	12	5	12	10	14	5	1	6	0	4	6	95
botafoogo	0	0	2	0	1247	0	0	0	0	0	7	2	3	0	0	0	10	6	1	0
bragantino	7	8	1	2	1	947	1	1	9	15	1	3	4	0	5	4	3	1	0	1
corinthians	3	5	9	4	2	0	1235	1	1	12	11	7	4	4	0	0	0	2	6	0
criciuma	1	3	0	1	0	0	0	833	0	0	0	1	3	0	0	4	0	0	0	1
cruzeiro	21	24	244	23	17	15	12	14	971	23	30	42	33	23	32	23	6	23	17	20
cuiaba	11	24	3	2	8	5	8	14	7	822	8	10	6	11	8	28	0	3	6	7
flamengo	34	23	41	28	30	19	38	18	48	28	1000	43	26	39	25	13	45	25	55	22
fluminense	11	10	41	9	28	15	15	5	28	10	71	1093	10	9	11	15	7	18	50	24
fortaleza	4	4	8	26	1	7	3	5	8	20	0	2	783	2	2	3	0	0	3	29
gremio	0	3	5	1	1	2	1	0	0	0	3	3	1	1012	4	9	0	3	2	1
internacional	3	3	2	3	0	0	0	0	0	3	1	2	0	0	1046	7	0	1	0	1
juventude	1	18	6	5	0	5	0	29	8	14	2	4	4	3	0	834	0	0	1	53
palmeiras	0	0	0	1	1	1	0	0	4	0	5	0	1	0	0	0	956	4	0	0
sao-paulo	3	0	14	2	1	0	10	0	1	4	4	2	3	7	5	0	15	1150	0	15
vasco	3	1	2	2	7	0	12	2	4	0	11	12	3	3	0	0	1	0	1241	1
vitoria	9	30	17	76	8	25	3	12	13	35	5	5	37	6	5	51	0	5	3	880

A matriz revela que o modelo conseguiu agrupar muitos documentos corretamente, com forte concentração de valores na diagonal, indicando uma correspondência consistente entre os clusters e os times. No entanto, algumas linhas apresentam dispersão de valores fora da diagonal, sugerindo similaridade temática ou conteúdo sobreposto entre essas notícias. A presença de números elevados fora da diagonal em times específicos também indica que o modelo teve dificuldade em separar alguns clubes, possivelmente devido a temas ou estilos de escrita semelhantes.

Acuracia

```

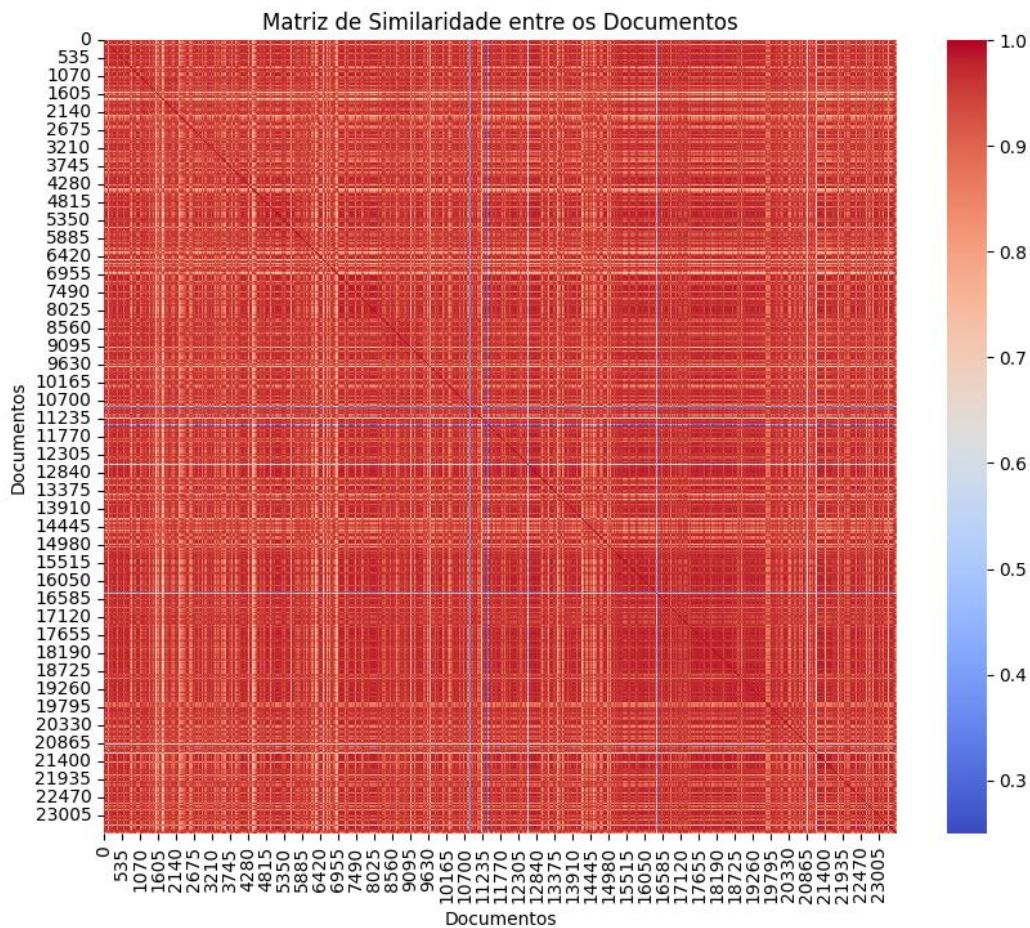
Run embeddings x
C:\Users\lome\AquaProjects\untitled\venv\Scripts\
C:\Users\lome\AquaProjects\untitled\venv\Lib\site
warnings.warn(
Acurácia com 400 clusters: 0.83
Process finished with exit code 0

```

A acurácia de 83% indica que o modelo Word2Vec com KMeans conseguiu capturar bem as similaridades entre as notícias dos times, mesmo com um modelo mais robusto, que leva mais em consideração o contexto.

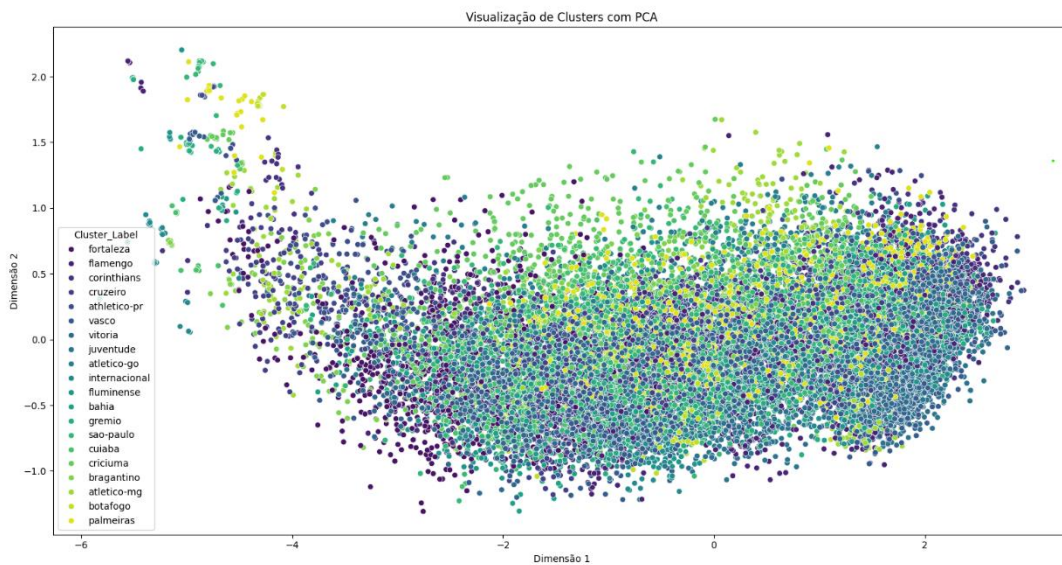
Embeddings com BERT:

Matriz de similaridade entre os documentos



A matriz de similaridade mostra que a maioria dos documentos tem alta similaridade entre si, o que indica que os textos compartilham muitos temas ou termos em comum. Os padrões quadriculados visíveis na matriz sugerem a existência de subgrupos ou clusters naturais de documentos, possivelmente agrupados por temas específicos, como times ou partidas e campeonatos semelhantes.

Visualização de Clusters com PCA



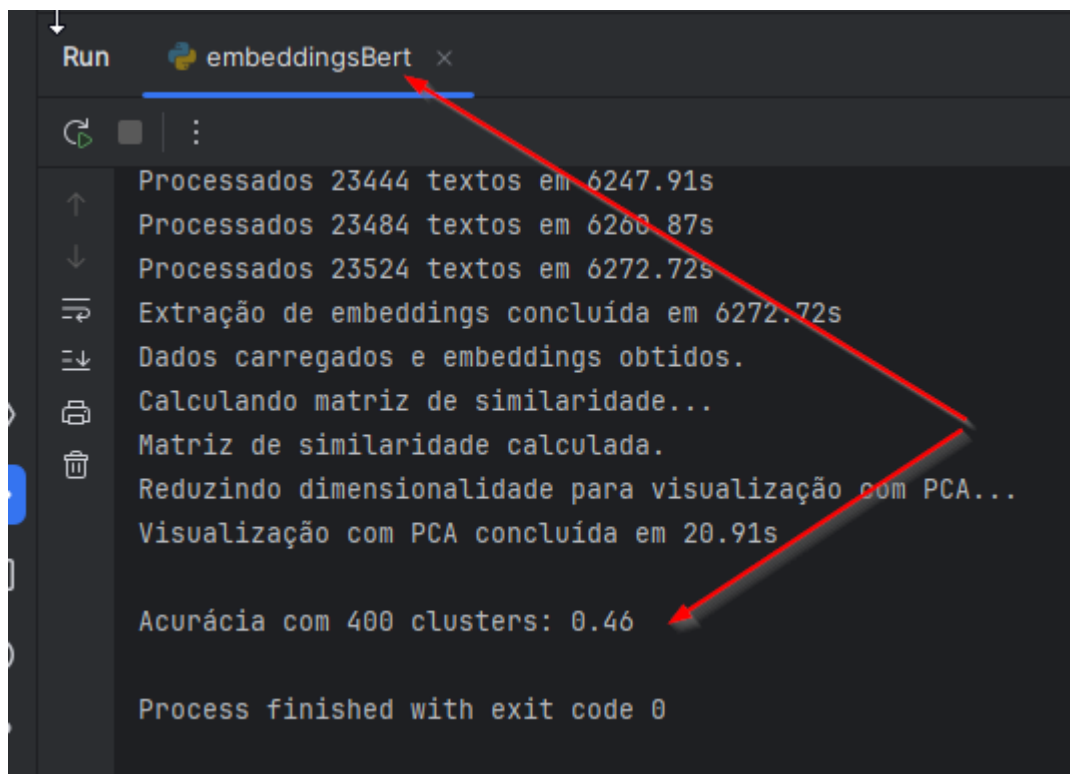
A visualização dos clusters com PCA revela que os dados têm uma distribuição contínua, com os pontos se espalhando de forma densa em algumas áreas e mais dispersa em outras, sugerindo diferenças de conteúdo entre algumas notícias. Os pontos coloridos representam os diferentes times, e vemos que, embora haja alguma separação, muitos times estão misturados, o que indica que os clusters não estão completamente isolados, e definidos

Matriz de Confusão entre Clusters e Times

		Matriz de Confusão entre Clusters e Times																			
Cluster		athletico-pr	athletico-go	athletico-mg	bahia	botafogo	bragantino	corinthians	criciuma	cruzeiro	cuiaba	flamengo	fluminense	fortaleza	gremio	internacional	juventude	palmeiras	sao-paulo	vasco	vitoria
	athletico-pr	458	26	109	68	46	50	27	47	85	50	72	58	42	76	82	35	27	29	50	58
	athletico-go	35	488	98	16	11	94	7	55	57	58	13	13	87	9	22	8	1	17	18	39
	athletico-mg	27	33	177	26	17	27	8	19	48	18	39	30	32	19	17	7	3	2	26	19
	bahia	46	11	22	458	42	54	55	8	56	37	57	36	15	39	28	36	23	41	52	59
	botafogo	22	4	16	17	929	14	27	2	19	10	36	11	3	16	8	12	7	8	12	14
	bragantino	8	26	17	9	5	393	0	14	12	33	14	7	17	3	9	2	8	8	9	4
	corinthians	20	8	15	15	22	13	1012	10	17	22	28	56	11	17	11	11	21	9	44	21
	criciuma	41	112	80	32	5	45	5	960	24	90	7	25	130	16	13	17	4	8	15	22
	cruzeiro	49	40	109	40	22	52	25	30	415	48	62	67	60	56	56	70	22	21	106	42
	cuiaba	56	53	80	21	16	30	18	57	53	298	15	27	42	47	62	35	17	10	17	57
	flamengo	80	16	89	66	47	43	44	16	83	29	479	95	40	55	65	39	38	24	118	40
	fluminense	10	13	77	17	20	19	16	4	35	24	65	614	46	42	53	17	8	15	27	13
	fortaleza	44	60	100	36	9	56	6	38	40	58	38	33	258	25	36	25	6	2	23	22
	gremio	54	17	51	49	25	36	18	22	49	32	75	53	36	485	113	51	16	8	41	39
	internacional	44	15	71	24	50	32	27	17	40	55	63	45	48	75	382	98	9	18	50	49
	juventude	62	23	63	30	52	28	32	15	69	92	44	50	54	107	147	489	33	16	75	81
	palmeiras	5	1	7	5	7	11	7	2	13	8	12	9	3	4	2	3	790	34	15	3
	sao-paulo	12	11	22	19	14	21	11	6	26	11	27	22	2	13	2	7	16	989	15	9
	vasco	28	25	51	24	17	39	17	18	112	43	55	43	28	32	34	45	11	8	672	63
	vitoria	27	20	17	124	20	32	12	19	27	41	26	27	39	20	38	41	4	12	35	549

A matriz mostra que a maioria dos clusters possui uma alta concentração de notícias corretamente atribuídas ao time correspondente, o que indica que o algoritmo conseguiu identificar padrões específicos para muitos dos times. No entanto, há alguns clusters com sobreposição considerável, como Athletico-PR e Internacional, sugerindo que as notícias desses times podem compartilhar tópicos ou linguagem similar.

Acuracia



```
Run embeddingsBert x
Processados 23444 textos em 6247.91s
Processados 23484 textos em 6260.87s
Processados 23524 textos em 6272.72s
Extração de embeddings concluída em 6272.72s
Dados carregados e embeddings obtidos.
Calculando matriz de similaridade...
Matriz de similaridade calculada.
Reduzindo dimensionalidade para visualização com PCA...
Visualização com PCA concluída em 20.91s

Acurácia com 400 clusters: 0.46

Process finished with exit code 0
```

A acurácia de 46% indica que o modelo tem um desempenho limitado para diferenciar entre as notícias dos times, possivelmente devido à similaridade de conteúdo ou linguagem entre elas. Ao considerar o BERT como um modelo bem mais encorpado e contextualizado.

Considerações Finais

Nosso grupo decidiu traçar um meio semelhante para avaliar a aplicabilidade dos algoritmos, todos eles são aplicados em modelos de classificação, com clusterização, com 400 Clusters. Sobre o mesmo DS, com as mesmas métricas sendo extraídas, dado esse contexto... É claro que modelos mais robustos vão perdendo-se entre a similaridade do tema proposto em cada matéria, sendo porque os times se enfrentam, entre si e as matérias refletem esses confrontos, seja porque tudo versa sobre o mesmo foco com labels diferentes para cada um dos times, mudando apenas seus nomes, então ficou claro para nós que algoritmos mais simples, que atribuem pesos as palavras únicas, ou separam elas para análise, e desconhecem um pouco do contexto, tiveram uma efetividade quase que perfeita, a revelia de algoritmos contextuais.