



**PUC Minas**

IEC - Instituto de Educação Continuada  
Pós-Graduação em Ciência dos Dados e Big Data

**Recuperação da Informação na Web  
e em Redes Sociais**

# **Análise de interação com a Band envolvendo a Formula 1 no Twitter**

**Aluno:** Clinton Julio Novais Amaral

**Aluno:** Pedro Henrique de Souza Martins

**Professor:** Zilton Cordeiro Jr.

julho  
2021



**PUC Minas**

IEC - Instituto de Educação Continuada  
Pós-Graduação em Ciência dos Dados e Big Data

## **Projeto Final**

# **Análise de interação com a Band envolvendo a Formula 1 no Twitter**

Trabalho apresentado ao Instituto de Educação Continuada (IEC) da pós-graduação em Ciência dos Dados e Big Data da PUC Minas, como requisito parcial para a obtenção de créditos na disciplina de Recuperação da Informação na Web e em Redes Sociais.

**Aluno:** Clinton Julio Novais Amaral

**Aluno:** Pedro Henrique de Souza Martins

**Professor:** Zilton Cordeiro Jr.

julho

# Conteúdo

<b>1</b>	<b>Resumo</b>	<b>1</b>
<b>2</b>	<b>Introdução</b>	<b>2</b>
<b>3</b>	<b>Descrição das Atividades</b>	<b>3</b>
3.1	Processo de Coleta . . . . .	3
3.2	Análise de Sentimentos . . . . .	4
3.2.1	Enriquecimento . . . . .	4
3.2.2	Preprocessamento . . . . .	5
3.2.3	Visualização dos dados . . . . .	6
3.3	Rede de Influência . . . . .	6
3.4	Análise de Similaridade Textual . . . . .	7
3.4.1	Preprocessamento . . . . .	8
<b>4</b>	<b>Análise dos Resultados</b>	<b>9</b>
4.1	Processo de Coleta . . . . .	9
4.2	Análise de Sentimentos . . . . .	10
4.3	Rede de Influência . . . . .	12
4.4	Análise de Similaridade Textual . . . . .	16
<b>5</b>	<b>Trabalhos Futuros</b>	<b>17</b>
	<b>Bibliografia</b>	<b>18</b>

# 1 Resumo

Este trabalho tem como objetivo aplicar todas as técnicas de recuperação, análise de dados não estruturados, mineração de texto e visualização de dados aprendidas no curso de Recuperação de informações na Web e Redes Sociais.

Para o desenvolvimento deste trabalho foi realizada a coleta no Twitter no dia 14 de julho de 2021 para todos os tweets que continham as hashtags 'F1NoBandSports' ou 'F1NaBand'.

A partir dos tweets coletados foi feita a análise de sentimentos, fazendo o enriquecimento dos dados, classificando os verbos, adjetivos e substantivos e identificação de entidades. Com a base de dados também foi modelada uma rede que apresenta a influência dos usuários que fizeram um tweet contendo os termos buscados. Além do mais a partir da coleta foi desenvolvido um fluxo de similaridade para que através de uma consulta fossem retornados os tweets com mais similaridade.

## 2 Introdução

Com a utilização da rede social por grande parte da população, elas têm se tornado grandes fontes de informações, uma vez que, nessas plataformas os usuários compartilham suas experiências, opiniões e sentimentos sobre assuntos diversos, como produtos, serviços, personalidades públicas e grandes eventos. Com isso a extração e análise desses dados possuem grande importância para anunciantes, governos, fabricantes, etc.

Essas plataformas possuem grandes números de usuários que geram um enorme volume de dados diariamente, entre os sites de redes sociais, uma das plataformas mais populares é o Twitter que possui por volta de 186 milhões de usuários. O Twitter fornece uma experiência simplificada para a criação de postagens na web, em que cada tweet possui no máximo 140 caracteres, o que o torna fácil para a produção e consumo de conteúdo. Portanto essa rede social permite com que grandes empresas e pesquisadores consigam coletar os dados e realizar análises em cima desse conteúdo em uma escala muito grande.

Levando em consideração o grande volume de dados e a necessidade das empresas de acompanharem como suas campanhas e marcas estão sendo recebidas pelos usuários, se tornou impraticável que um analista de rede social consiga inspecionar todos este conteúdo gerado por cada usuário.

Essas plataformas possuem grandes números de usuários que geram um enorme volume de dados diariamente, entre os sites de redes sociais, uma das plataformas mais populares é o Twitter que possui por volta de 186 milhões de usuários. O Twitter fornece uma experiência simplificada para a criação de postagens na web, em que cada tweet possui no máximo 140 caracteres, o que o torna fácil para a produção e consumo de conteúdo. Portanto essa rede social permite com que grandes empresas e pesquisadores consigam coletar os dados e realizar análises em cima desse conteúdo em uma escala muito grande.

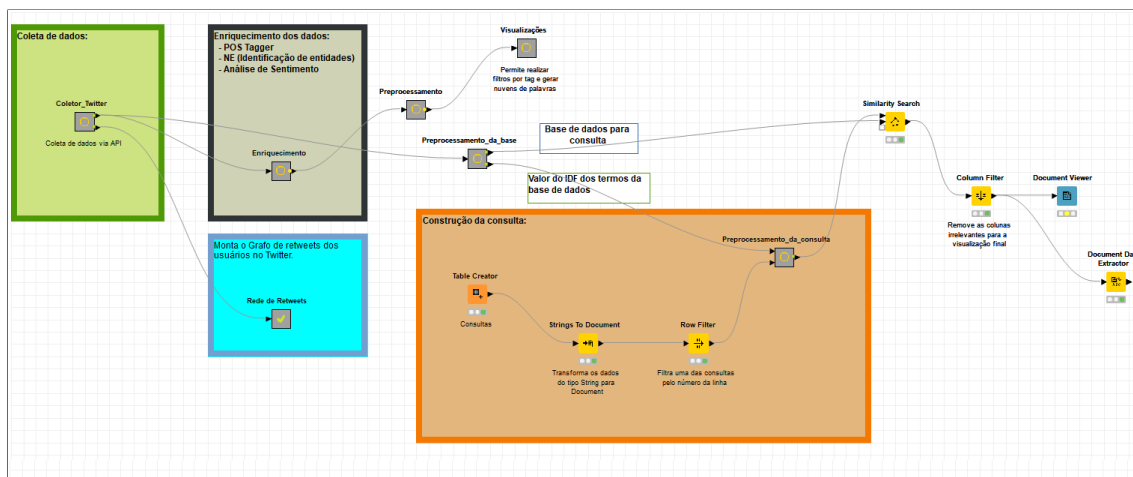
O presente artigo tem como objetivo recuperar os dados gerados pelos usuários do Twitter e identificar as suas reações durante a transmissão brasileira do campeonato mundial de Fórmula 1, descobrindo também os usuários que possuem uma maior influência dentro dos usuários que estão interagindo.

Para realizar essa atividade de análise foi utilizado o Knime [?] que é um software de código aberto fornecendo ferramentas para o desenvolvimento de técnicas de ciência de dados baseados em programação visual o que torna a compreensão mais facilitada dos dados e o fluxo do trabalho de coleta de dados.

### 3 Descrição das Atividades

Para o desenvolvimento do trabalho foi utilizado o Knime, um software gratuito e de código aberto que permite a integração, coleta e análise de dados, juntamente com os recursos providos por ele.

A imagem a seguir mostra uma visão geral do fluxo construído no Knime:



Para cada núcleo de entrega foi criado um metanodo, sub-fluxos de trabalho, e o desenvolvimento de cada um será contextualizado nos subtópicos a seguir:

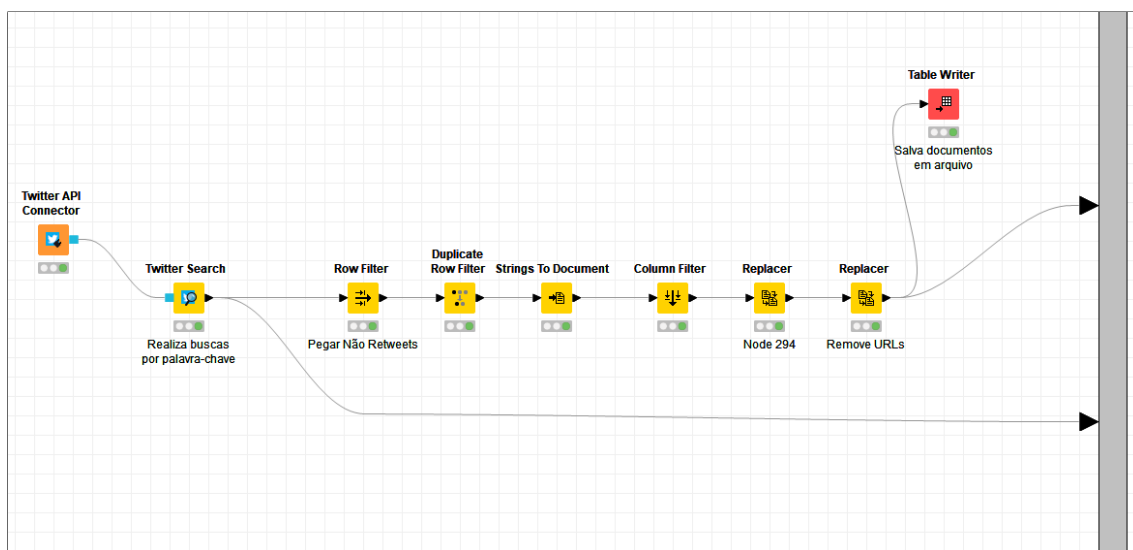
#### 3.1 Processo de Coleta

Para o processo de coleta foi utilizado um nodo do Knime chamado Twitter API Connector que recebe as credencias do Twitter, as credenciais utilizadas foram concedidas pelo professor. Após a conexão com o Twitter foi utilizado o nodo Twitter Search para enviar a API a consulta que gostaríamos de fazer.

O valor colocado no campo query do nodo foi: `'(#F1NoBandSports) OR (#F1NaBand) lang:pt'`, a pesquisa foi feita por registros mistos, tanto recentes quanto populares, o retorno esperado de registros era de 2000 registros, o número de tentativas foi 0 e todos os campos possíveis foram retornados.

Como utilizaríamos o resultado obtido de maneiras diferentes durante o trabalho, o metanodo de coleta tem duas saídas, a superior com um filtro de tweets não retweetados, para não haver duplicidade de informações, e uma saída inferior com os dados sem nenhuma filtragem para serem utilizados posteriormente na parte de criação de redes.

A imagem a seguir mostra o fluxo do metanodo de coleta de dados.



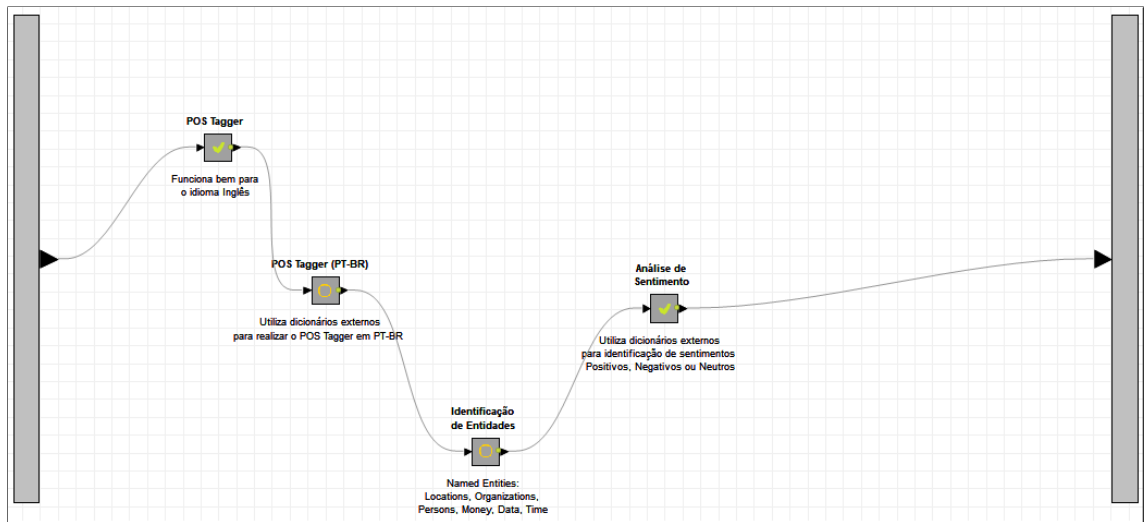
## 3.2 Análise de Sentimentos

Para ser feita a análise de sentimentos, foram criados alguns metanodos para modularizar os tratamentos dos dados. Os subcapítulos a seguir detalharão melhor o que foi feito em cada metanodo.

### 3.2.1 Enriquecimento

No metanodo de enriquecimento foram feitas algumas atribuições as palavras contidas nos tweets coletados. Foi feita a marcação gramatical, tanto de palavras em português quanto de palavras em inglês, a identificação de entidades e a análise de sentimento das palavras.

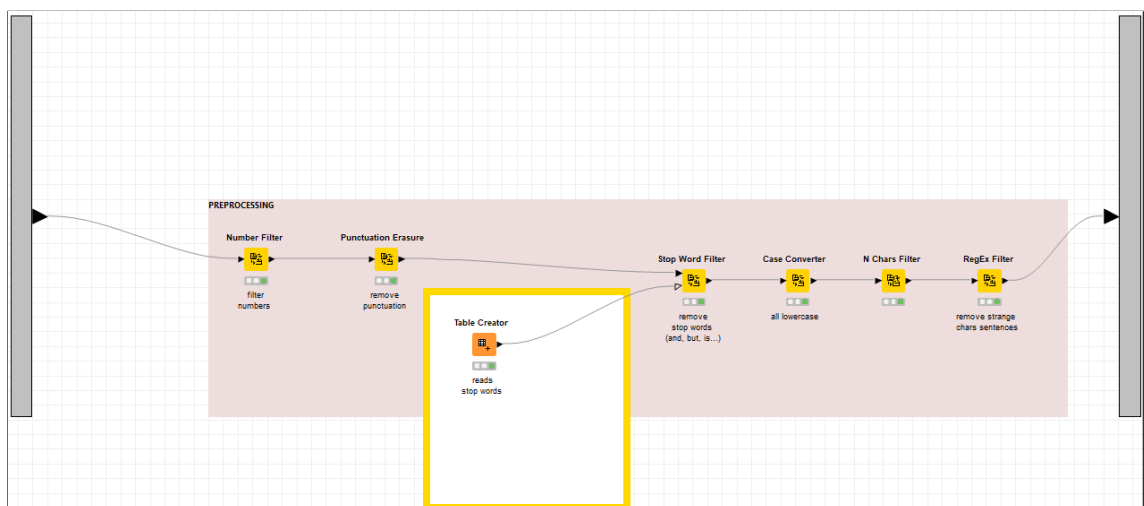
A imagem a seguir mostra o fluxo do metanodo de enriquecimento dos dados.



### 3.2.2 Preprocessamento

Na etapa de processamento foram feitas algumas alterações básicas nos documentos, remoção de pontos, caracteres especiais e palavras conhecidas com 'stop words'. Vale ressaltar que foi criado um nó de 'stop words' inseridas manualmente para desprezar palavras como 'https' e 'rt' que não eram consideradas de forma automática.

A imagem a seguir mostra o fluxo do metanódo de preprocessamento dos dados.

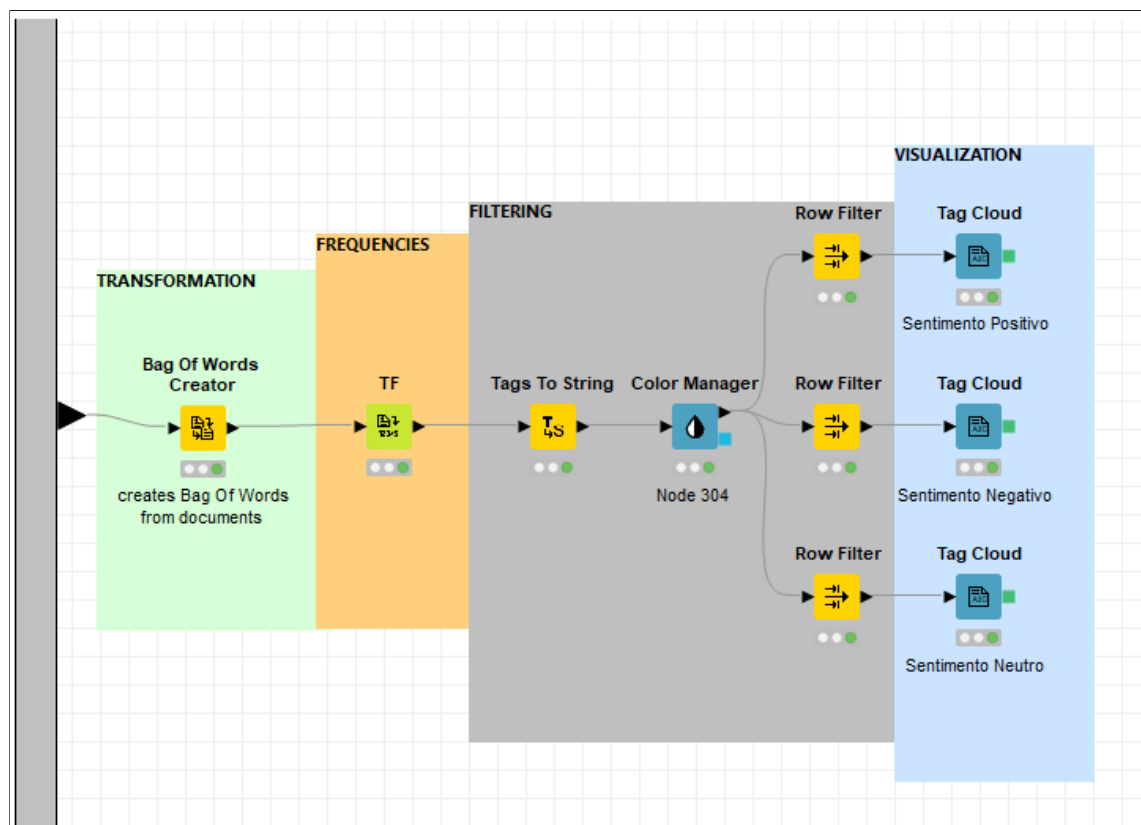




### 3.2.3 Visualização dos dados

No metanodo de visualização foi feita a transformação dos documentos em palavras, foi feito o cálculo da frequência das palavras, e a geração da visualização feita através da nuvem de palavras com uma filtragem por tipo de sentimento, que poderiam ser neutro, positivo ou negativo.

A imagem a seguir mostra o fluxo do metanodo de visualização dos dados.

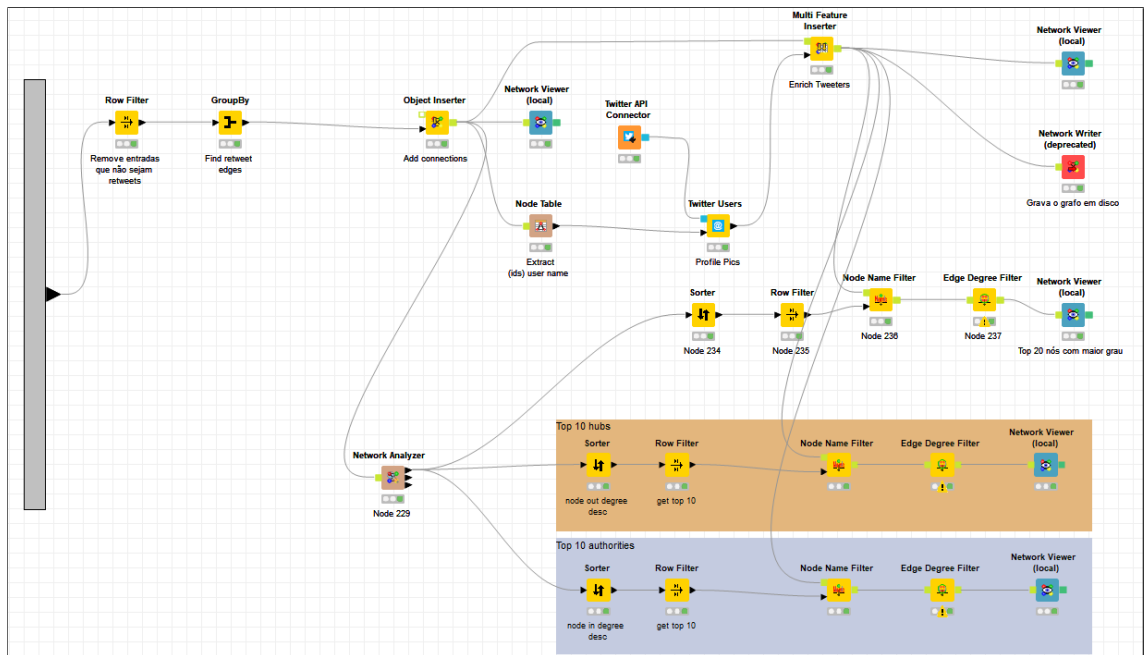


### 3.3 Rede de Influência

A rede de influência foi configurada para criar nós a partir da ação de retwittar outra pessoa no twitter. Foi feito um agrupamento dos tweets pelo campo Retweet from e User. A partir do agrupamento realizado foi utilizado o nó do Knime Object Inserter que cria a rede com seus nós e conexões.

Após a criação da rede foram utilizados alguns outros nós do Knime para visualizar e entender a rede, como por exemplo os 20 maiores nós com maior grau, as top 10 autoridades da rede e os top 10 hubs da rede, além obviamente de uma visão gráfica da rede como um todo.

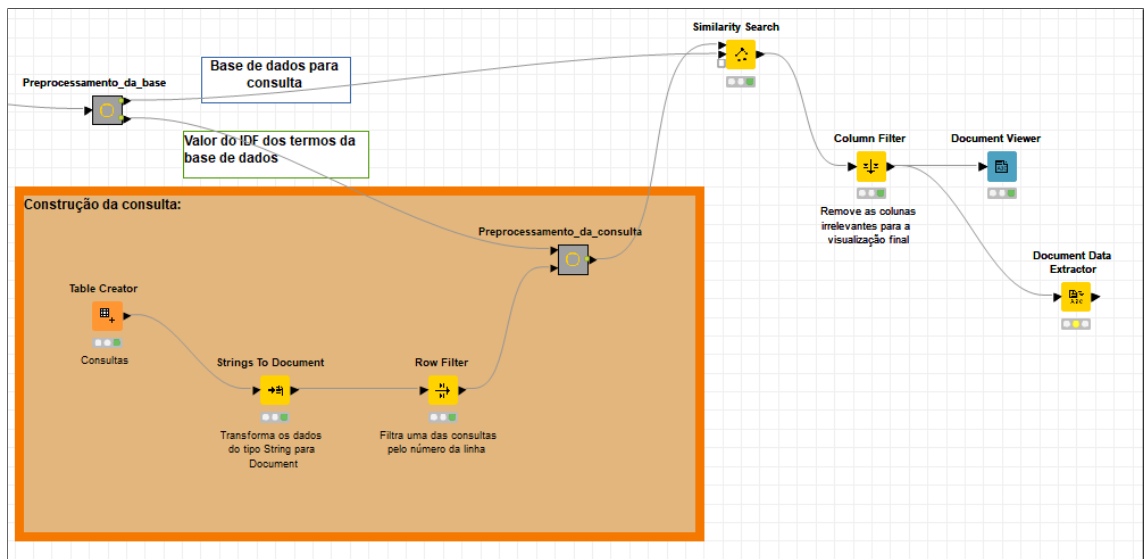
A imagem a seguir mostra o fluxo do metanodo da rede de influência.



### 3.4 Análise de Similaridade Textual

A Análise de Similaridade Textual consistiu em pegar os dados da coleta já filtrada sem os retweets fazer um pré-processamento para que depois pudesse ser feita a busca de similaridade em uma outra consulta a parte, simulando uma consulta a um buscador da WEB. Cabe ressaltar foram utilizadas as mesmas técnicas de pré-processamento para a consulta feita no Twitter e a consulta "simulada".

A imagem a seguir mostra o fluxo da Análise de Similaridade Textual.

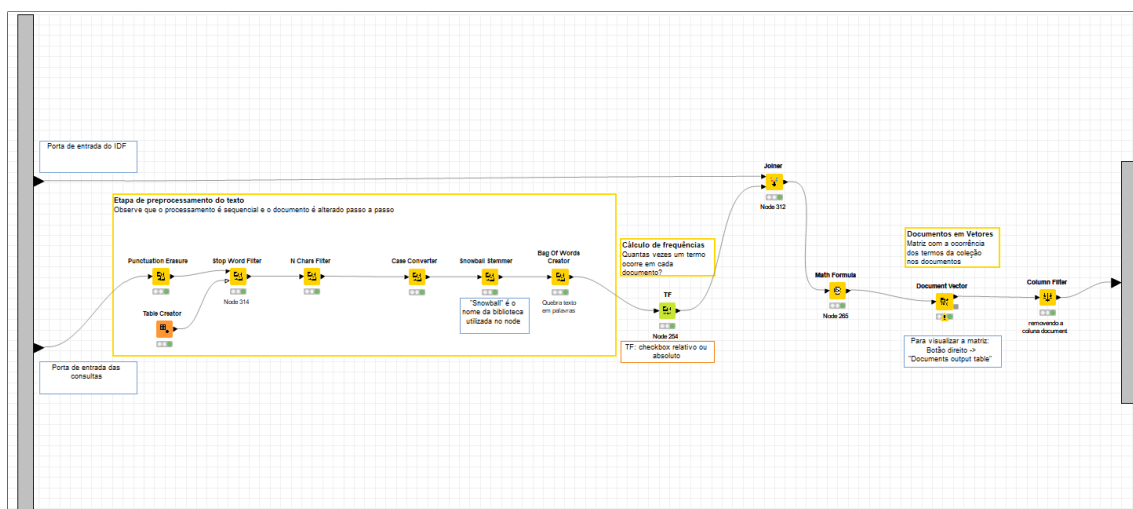


O subtópico a seguir detalhará melhor o que foi feito na etapa de pré-processamento.

### 3.4.1 Pré-processamento

A etapa de pré-processamento serviu para fazer uma espécie de sanitização dos dados, removendo pontuações, 'stop words', utilização de algoritmo de stemming, algoritmo para obter o radical dos termos, calcular a frequência da ocorrência dos termos nos documentos, calcular o IDF, calcular o TF-IDF e gerar a matriz de documentos.

A imagem a seguir mostra o fluxo do metanodo do pré-processamento dos dados.



## 4 Análise dos Resultados

Os resultados obtidos pelos autores ajudaram a entender a base de dados coletada. Nos subtópicos a seguir serão detalhados cada resultado dos núcleos desenvolvidos.

### 4.1 Processo de Coleta

O processo de apesar de estar configurado para 2000 registros retornou apenas 431 registros. O que não foi um grande problema, mas também não foi algo que era esperado pelos autores.

A imagem a seguir mostra o resultado da coleta de dados.

Search results - 0:315:324:317 - Twitter Search (Realiza buscas)

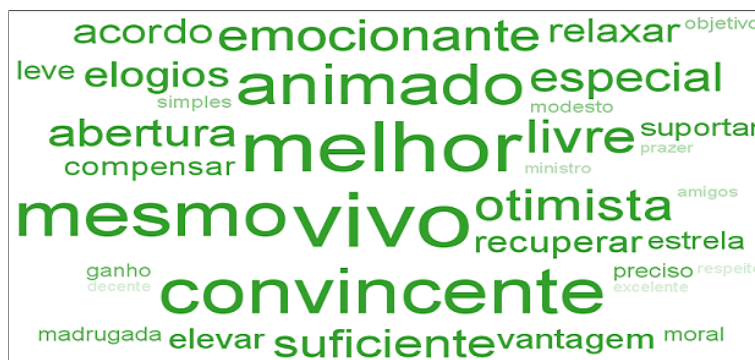
Table "default" - Rows: 431 Spec - Columns: 26 Properties Flow Variables

Row ID	S Tweet	L Tweet ID	S Time	F Favorited
Row0	Confira os horários de transmissão do Grande Prêmio da Inglaterra de Fórmula 1. Em Silverstone a FIA fará o primeiro teste da classificação ser definida em uma corrida sprint. #F1 #F1noBandSports #F1naBand #F1Sprint https://t.co/ru5vwm88nk	1415484988305334...	2021-07-14 22:34:54	0
Row1	Palco do primeiro GP da história do Mundial 71 anos atrás vive outra estreia... #reginaldoem...	1415474556463157...	2021-07-14 21:53:27	1
Row2	RT @bandsports: "É muito risco para pouco ganho" Ricardo Molina falou sobre o uso da corrida classificatória na @F1. #Supermotor #F1noBandSports #F1naBand #Fórmula1 https://t.co/zMKyVtQ7n	1415470009674092...	2021-07-14 21:35:23	2
Row3	#VamoqueVamo! O Final de Semana Promete no GP da Inglaterra! @@ @ @ @ #F1naBand e #F1noBandSports #F1naBand #F1naBand https://t.co/8tUgJ6a9	1415465594309251...	2021-07-14 21:17:50	0
Row4	RT @jogoaberto: A Fórmula 1 revelou que as corridas de classificação não terão pódio, mas... #F1naBand https://t.co/sUyF#2n54W	1415460200329457...	2021-07-14 20:56:24	1

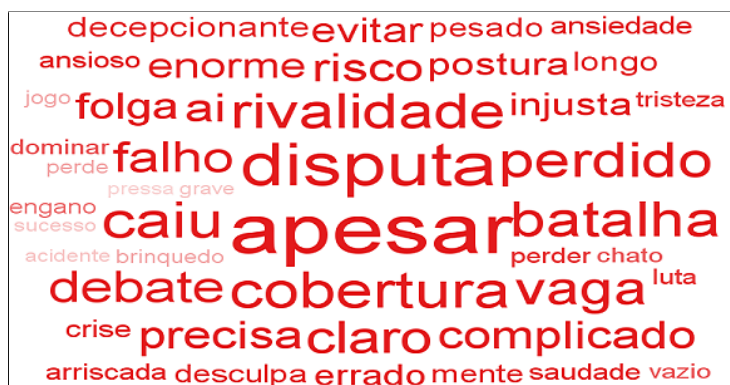
## 4.2 Análise de Sentimentos

Na análise de sentimentos foram geradas três nuvens para as palavras identificadas como tendo sentimento positivo, negativo e neutro.

A imagem a seguir mostra a nuvem de palavras definidas com sentimento positivo.



A imagem a seguir mostra a nuvem de palavras definidas com sentimento negativo.



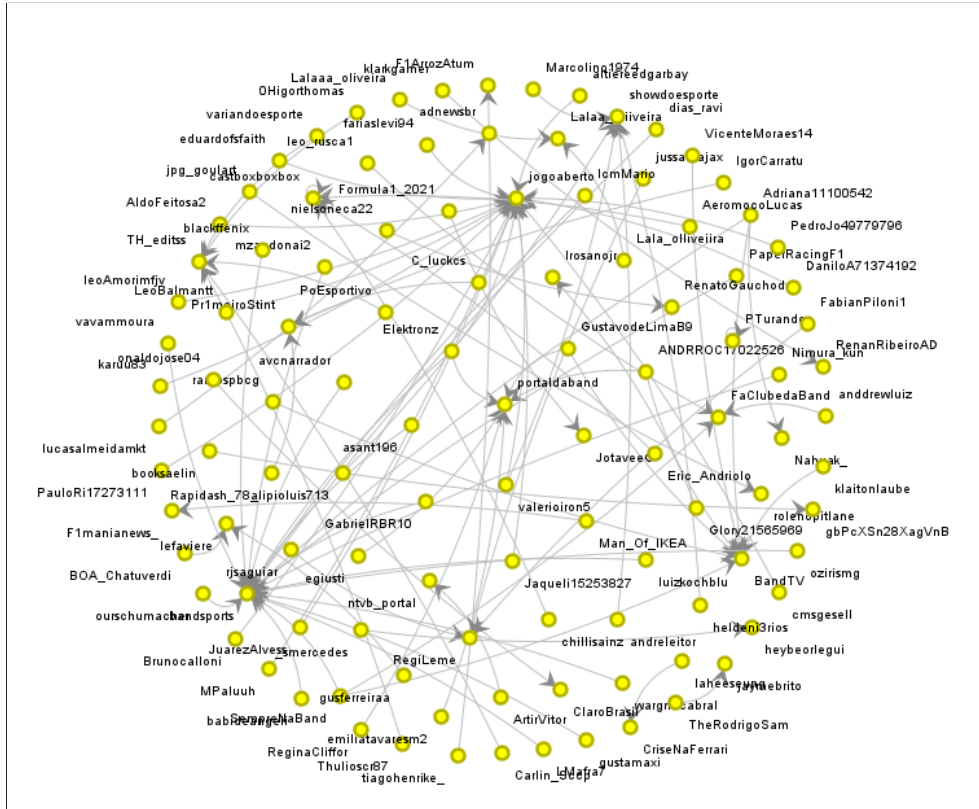
As palavras com sentimento neutro são a maioria. A imagem a seguir mostra a nuvem de palavras definidas com sentimento neutro.

enquanto podcast games tráfego granares batommentar motoros y leonardomundo tcoogscawwag confirmadme sed, mink app substituindo viva alan brunoasfonseca tcoyukqpninmw orlandovidalsp tcojzpxsmwfw afinal sociais tcofmesidkeghanunciado podcasting anunciam charlesleclerc casa troco palco diferente motorsport substituto pessoal revelar rodrigaomeira confirmadas celsomiranda up saber alfaromeo tcoaoodqxoeuw greatbritaingp mibnaespn site tcoatqpsncojbin racer aposta vote santos alfaromeoracing tcoojmimhutau rogersterronismo quenia britshgp ligando novidades tcoqpcxxubfc confere youtube sprintrace assistiu tcoocuepnuyupho steve velociakta manoela porsche cup imagens lewishamilton nfl raceday tcoobbgabdnwod asuwink tcoesbxscorad stranger viciada loucos thlandonomis larissa poderia quais introduzir nele anote vamos tcoofiejpmkhfuthethevoicekids wallpapers lucmonteiro diferentes lembra theditss icatlewis paposfut faltar grandprix amg tcoolzivmanhythings wseries maarianalvs governador adilsreis aspirante italiano tcoewekegkvytnhotdognaespn ovr falou tcoondzmqlfcoji blog acabou pensando senador cronograma prost vc argentinos holandeses tcooutwtgugiew fiaflunoprimevideo canalrafabin transmite rafaelramada vida fracasso sainz julietteprecisar olha argentina hino tcoots encontro opinou brasil saudades treinos programa valteribottas maneira quo gilbertolargadas bottas rclubepara hervedo cblo1 narrar soubemos afastou querem konshal ayrtton sacudir reginaldoleme participaprimeiros havia fia tv hehehe butteroneuro receber clima tcoohhtingfth teremos definir status ferrari provas rolenopitlanemostrar colocadoshaas anunciou fds fase tcohleoxxdip austriangp sauber valsa senna julho acho pontuou motor dezembro eveguimaraes focada trator interna atmosfera molaren atrapalhado gasly ambos acabar transmitir norris restantes teste gente sonoris decidir aposentadoria atrapalhado analisa deixou esperar leclerc presente veja alpine mexicano rever talesrodriguess briga discute engenharia decididomonegasco reclamaram encostou completatransmitem testes horner ep preza ecclestone felizes formulaone possui prova podcast mazepin paddock espanhol estreiam servir equipa galerinha etapa semelhante agenda rodada revelou outubro desempenho madrugace cotado frente tabela companheiro tripla contrato novidade piloto marcada sente carro chamadas segue estreia hamilton temporada band assista faclubedaband ideias interlagos atual thescoborges russellrace sido assumir retomada largada bons semana jogo aberto mercedes lista obter divulgou campeonato equipe verstappen quase grid gp showdoesporte confira devem alfa tempos sprint corridas mostra bandsports portaldaband corrida pilotos romeo raikkonen pra inglaterra britishgp resultados durante domingo disse silverstone bandesportclub williams dias jornalistas ganhou jovem sofrido antiga saiba ricciardo alonso diretor regileme relembrou maiores rasgou hasmann sair impressionou encontrava bandtv marca pontuar piores ritmo bandeirantes elogiou ocon lola revela motores hr alongar rendimento modernizar automotoresp perguntado vporcas assistir largar ajudar aumenta auge rivais novas perdeu dupla esporte evitou viria automobilismopromete confirmou gpdainglaterra frisou botar chega novembro staff eventual tornasse vettel pandemia christian criaram universo possibilidade texto aceitar musical jornalista dobro paulo tmj tema eventos rbatvband afirmou traz desenvolvimento mundo qualquer visto alto produzidas sono chegando previsto estado capricho formato tenha boas helmut mandado tcoedcbjteyps vivo eduvirtual colocou reconheceu usar extra funcionar vendo pregando alain consultor nbanaband ir tcoeshbbljvz categoria aproveitou deveria torcida ganhou curta palpites supermotor esporte estante nba ideia alguem tcohsfobibgtf velocidade assinar fundamentais vamoquevamo necessidade williams racing landononis jotaveeg equipes rbatvbelem tcoqgtvsmiyqx crescido processo liderado tela pista vivendo compartilhar tcojyzcwbmwo narrador avise tcosmodxgvitifique mos mundial tcoorcuzizxthk ganhos saiu pegou vim comigo tcoegfjipkctbtsforuefaeuromim futebol ausgp televisaorj link ansiosos tcoowoonszfk limita tcoopyifsqdnncorhamilton tcoocudplvgvsilverstonecircuit tcookakzfojkd austriangp treino negriesposito aproveito tcoxdociauoly carvalhorenato tcoobokfprtqnpouvir genisonkobe cresce mesma tcoouescsktbzii auto trocar portanto caprichadas renanibeiroad oficial fosse castboxboxbox maluco apps definida pq tcozimkvxqtn brasil urgente silverstonegp post game doido carreira entusiasmante jujubasoaresz playstore tabelas daniel Ricciardo defina dinheiro prperalta to coloca te deixa weekend dividem jantou tcofvqktpmdujregulamento podcastle fica entenda especulado stockcar uso copatruck weekconfira diferenciados grandepremio tcoodvsjbxuw tcoojowrkqgdxr rip dollarbillr iszdsfghd esquadrista brasil vive circuitos globolixo scuderiaferrari posto oficialmente entrava minas energia testa senti eps monza foralula carlossainz fu saive passam nosportv desobedeceu redbullracing cascavel modo toquio tcoeogrbqviek fichas palavra foratite regular nao mobile dela redbull greatbritshgp albuquerque estende chegou nahuac troca tcohapxrqqdus fotos podpah in lulugamepass daniel cantera pronunciamiento gol shsh is euro cobrindo natan ultrapassagem tcojpcwezonc igual hutchgames dae poles glendaacompanhe renovou shumacher islero ficando construtores nepomuceno xingar abaixo cretino pois fora acelerar rba tcoovybqwnhqkkgarantido

### 4.3 Rede de Influência

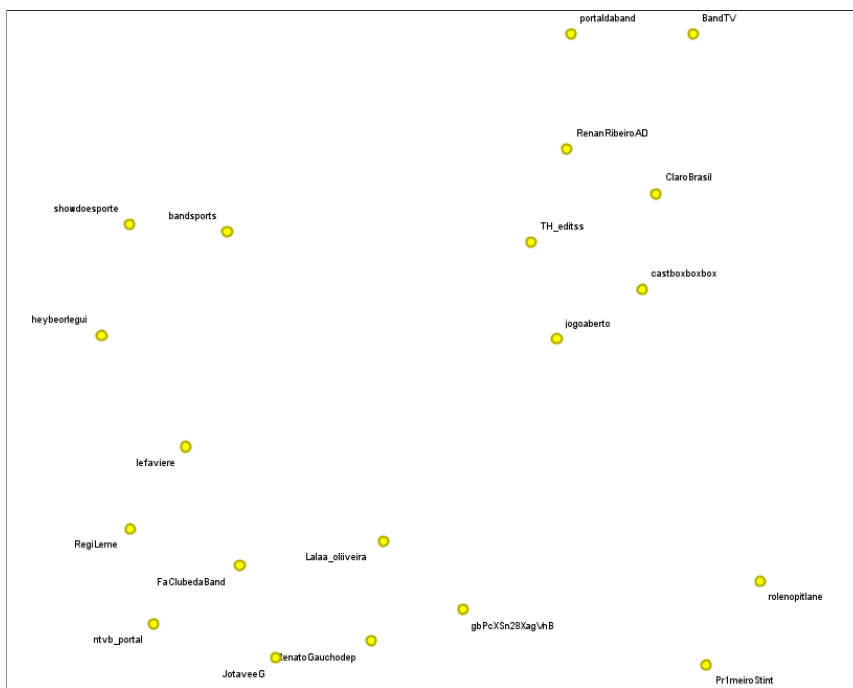
Como mencionado anteriormente, foram criadas algumas redes a partir da base de dados coletada. A imagem a seguir mostra a rede de influência

geral.

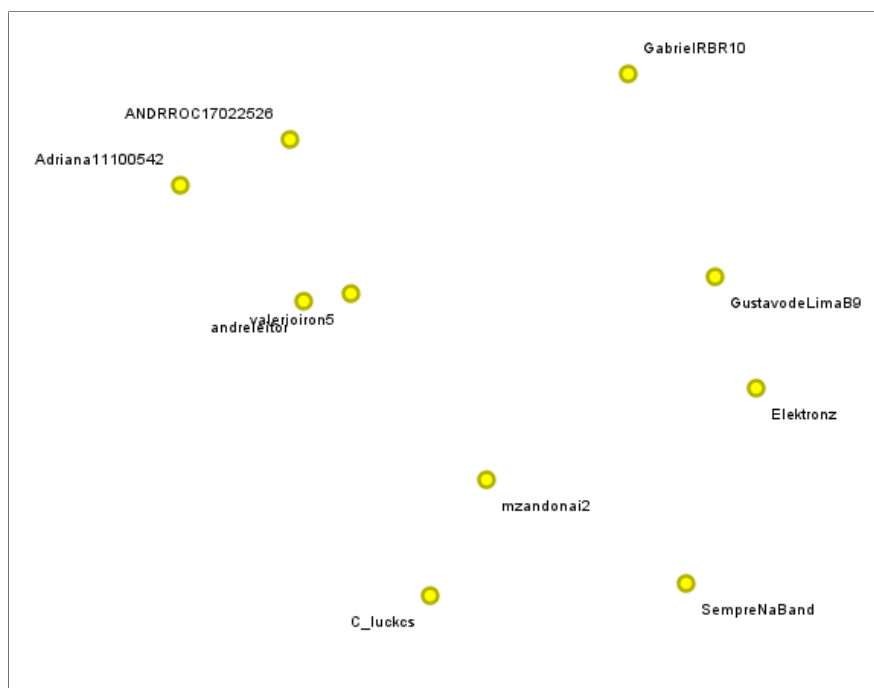


A imagem a seguir mostra os 20 maiores usuários que interagiram na rede.





A imagem a seguir mostra os 10 maiores hubs.

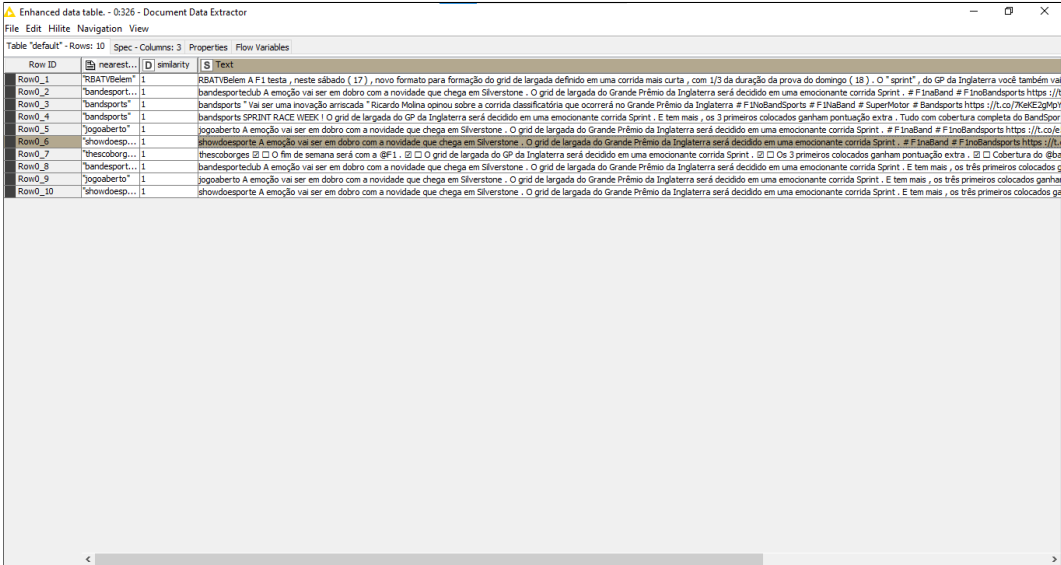


A imagem a seguir mostra as 10 maiores autoridades.



## 4.4 Análise de Similaridade Textual

Para a análise de similaridade textual foi simulada uma consulta com o termo 'desafios da corrida na Inglaterra'. O resultado foi no mínimo curioso, como pode ser visto na imagem a seguir, aparentemente existem alguns tweets iguais que na verdade não são, pois existem pequenas diferenças que de fato os fazem diferentes. Fazendo uma pesquisa paralela foi possível entender que os usuários que fizeram um tweet parecido são usuários pertencentes a Emissora Band, e portanto, disseminam a mesma informação para possivelmente alcançar usuários diferentes da rede social.



Row ID	nearest...	similarity	Text
Row0_1	"RBATVBelem"	1	RBATVBelem A F1 testa , neste sábado ( 17 ) , novo formato para formação do grid de largada definido em uma corrida mais curta , com 1/3 da duração da prova do domingo ( 18 ) . O " sprint " , do GP da Inglaterra você também vai
Row0_2	"bandesport..."	1	bandesportclub A emoção vai ser em dobro com a novidade que chega em Silverstone . O grid de largada do Grande Prêmio da Inglaterra será decidido em uma emocionante corrida Sprint . # F1naBand # F1noBandports <a href="https://t.co/NextEzghp0r">https://t.co/NextEzghp0r</a>
Row0_3	"bandports"	1	bandports " vai ser uma inovação arriscada " Ricardo Molna opinou sobre a corrida classificatória que ocorrerá no Grande Prêmio da Inglaterra # F1naBand # SuperMotor # Bandports <a href="https://t.co/NextEzghp0r">https://t.co/NextEzghp0r</a>
Row0_4	"bandports"	1	bandports SPRINT RACE WEEK 1 O grid de largada do GP da Inglaterra será decidido em uma emocionante corrida Sprint . E tem mais , os 3 primeiros colocados ganham pontuação extra . Tudo com cobertura completa do BandSport
Row0_5	"jogoaberto"	1	jogoaberto A emoção vai ser em dobro com a novidade que chega em Silverstone . O grid de largada do Grande Prêmio da Inglaterra será decidido em uma emocionante corrida Sprint . # F1naBand # F1noBandports <a href="https://t.co/le">https://t.co/le</a>
Row0_6	"showdoesp..."	1	showdoesporte A emoção vai ser em dobro com a novidade que chega em Silverstone . O grid de largada do Grande Prêmio da Inglaterra será decidido em uma emocionante corrida Sprint . # F1naBand # F1noBandports <a href="https://t.co/le">https://t.co/le</a>
Row0_7	"thescooborg..."	1	thescooborgs <input type="checkbox"/> O fim de semana será com a #F1 . <input type="checkbox"/> O grid de largada do GP da Inglaterra será decidido em uma emocionante corrida Sprint . <input type="checkbox"/> Os 3 primeiros colocados ganham pontuação extra . <input type="checkbox"/> Cobertura do @band
Row0_8	"bandesport..."	1	bandesportclub A emoção vai ser em dobro com a novidade que chega em Silverstone . O grid de largada do Grande Prêmio da Inglaterra será decidido em uma emocionante corrida Sprint . E tem mais , os três primeiros colocados g
Row0_9	"jogoaberto"	1	jogoaberto A emoção vai ser em dobro com a novidade que chega em Silverstone . O grid de largada do Grande Prêmio da Inglaterra será decidido em uma emocionante corrida Sprint . E tem mais , os três primeiros colocados ganha
Row0_10	"showdoesp..."	1	showdoesporte A emoção vai ser em dobro com a novidade que chega em Silverstone . O grid de largada do Grande Prêmio da Inglaterra será decidido em uma emocionante corrida Sprint . E tem mais , os três primeiros colocados ganha

## 5 Trabalhos Futuros

Para os trabalhos futuros as atividades desenvolvidas poderiam ser mais aprofundadas e envolver uma maior complexidade. O cruzamento com outras bases de dados, e comparações com outras pesquisas poderiam trazer um debate interessante, como por exemplo, quais eram os sentimentos dos usuários quando utilizando hashtags de interação com a Emissora Globo.

Seria interessante também uma forma de disponibilizar os resultados para que pudesse haver uma interação com o público. Por exemplo uma API para consumir o resultado e mostrar a nuvem de palavras, ou construir uma interação com o fluxo de análise de similaridade textual para simular um buscador WEB.

## **Bibliografia**

KNIME. Disponível em: < <https://www.knime.com/knime>> Acesso em: 19 de julho de 2021.