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 Popularidade no Twitter dos Personagens do Filme Os Vingadores

Aluno: Ana Carolina de Albuquerque Santos (1334073)
Aluno: Claudia Pereira Gonçalves (1350005)
Professor: Zilton Cordeiro Jr.

Julho
2021

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

Projeto Final

Análise de Popularidade no Twitter dos Personagens do Filme Os Vingadores

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: Ana Carolina de Albuquerque Santos

Aluno: Claudia Pereira Gonçalves

Professor: Zilton Cordeiro Jr.

Julho
2021

Conteúdo

1 Introdução

Os Vingadores são um grupo de super-heróis de história em quadrinhos publicados nos Estados Unidos pela editora Marvel Comics. O grupo também aparece em adaptações da Marvel para cinema, desenho animado e jogos eletrônicos. (WIKIPEDIA, 2021).

Figura 1 - Fonte: Diario de Pernanbuco

O Twitter é uma das maiores redes sociais da atualidade, que permite aos usuários enviar e receber atualizações de diversos assuntos. A sequência de filmes e seriados do filme Vingadores da Marvel, vêm sendo assunto do momento. No ano de 2019, o filme bateu recorde de audiência ultrapassando o filme "Avatar" com US 2,790 bilhões em bilheteria. No ano de 2021 iniciou o lançamento dos seriados e novos lançamentos de filme, baseados em cada personagem da Saga, como exemplo seriado Loki e o filme da Viúva Negra, o que gerou um grande volume de dados no twitter neste ano.

Esse trabalho tem como objetivo comparar a popularidade dos personagens da saga os Vingadores, utilizando uma amostra de 4000 comentários do Twitter feitos por usuários dessa rede social. Os personagens que serão pesquisados são Thor, Loki, Homem de Ferro, Capitão America, Homem de Ferro, Hulk e Viúva Negra. O software utilizado foi o Knime para fazer a extração, tratamento, análise e apresentação dos dados.

2 Descrição das Atividades

Foi utilizado o software KNIME 4.3.1 para fazer a extração dos dados, o tratamento e visualização das informações coletadas em várias etapas utilizando os nodos disponíveis na aplicação conforme explicado nos fluxos abaixo. A etapas de extração, tratamento e enriquecimento dos dados foram feitas com o objetivo de gerar um arquivo, o qual classifica-se o texto coletado em tags, conforme repassado para o dicionário de tags, alimentado por uma tabela identificadora de palavras chave.

Primeiramente montamos um fluxo principal no KNIME, conforme a Figura 2, para coletar e minerar os dados, montar as análises de sentimentos e de popularidade, ambas visualizadas por nuvem de palavas. Para análisar o comportamento da rede relacionada aos dados coletados, montamos o grafo de retweets dos usuários do Twitter.

Fluxo

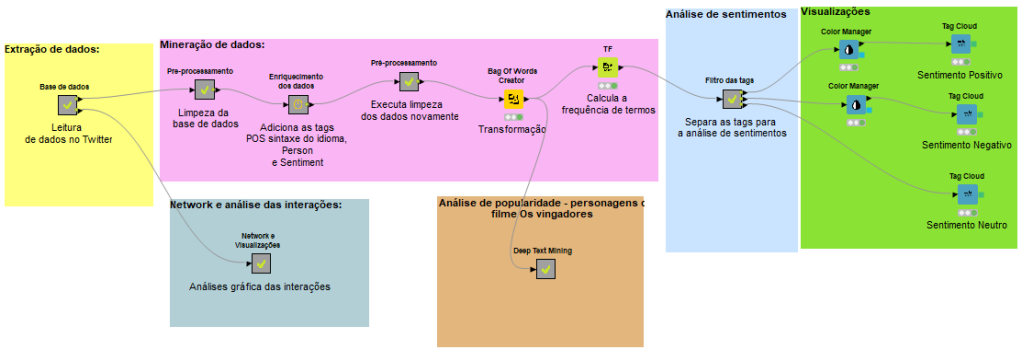

Figura 2 - Fonte: ScreenShot software Knime

A primeira etapa do fluxo, **Extração dos dados** conforme a Figura 3, inserimos em um meta-nodo todas as fases necessárias para a extração dos dados do Twitter. Esse meta-nodo gerou informações para duas saídas: fluxos de análise de sentimentos, popularidade e fluxo análise de comportamento da rede relacionado aos dados coletados.

Extração de Dados

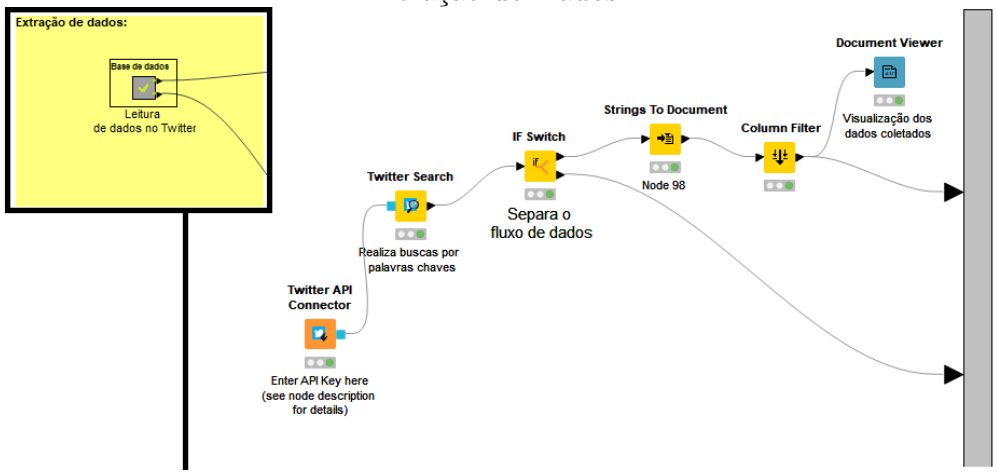

Figura 3 - Fonte: ScreenShot software Knime

Na etapa de Extração dos dados trabalhamos com os seguintes nodos:

O nodo **Twitter API Connector** foi utilizado para a conexão com o Twiter. Essa conexão foi feita com as credenciais geradas no Twitter Development. Em seguida o nodo **Twitter Search** foi empregado para realizar a busca dos dados utilizando a seguinte query do nodo **("homem de ferro") OR ("Hulk") OR ("Thor") OR ("viúva negra") OR ("viuva negra"OR("Loki") OR ("Capitão América") OR ("Capitao America")**, considerando a coleta de 4000 twitters. Os dados coletados foram transformados em tipo documento utilizando o nodo **Strings to Document**. Neste, para cada linha criou-se um documento anexado a essa linha. O Campo Tweet foi utilizado para o texto do documento, o campo user foi utilizado utilizado para autor do documento. Posteriormente o nodo **Column Filter** foi inserido para filtrar somente a coluna Documento, selecionando a opção *textdocument*, a qual corresponde aos textos coletados do twitter. O nodo **if swtich** separou fluxo de extração dados, de forma a que na fase descrita acima, os dados foram utilizados para alimentar o fluxo referente a mineração dos dados. Os dados brutos foram carregados no fluxo análise de comportamento da rede.

Após coletar as informações essas foram repassadas para as etapas seguintes. O fluxo de mineração dos dados consiste nas etapas de pré-processamento, enriquecimento dos dados, transformação dos dados e cálculo da frequência de termos.

Esse fluxo foi implementado para preparar os dados para a análise de sentimento e popularidade dos personagens coletados e recebe as informações do Fluxo de Extração dos dados, no formato Documento. Essa etapa é ilustrada na Figura 4.

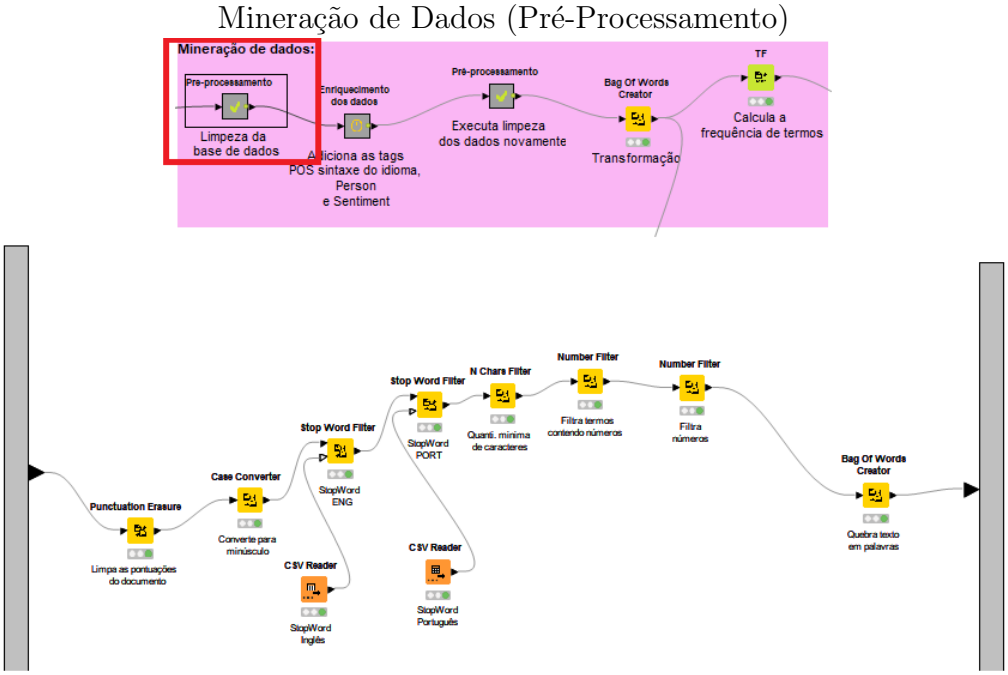

Figura 4 - Fonte: ScreenShot software Knime

O meta-nodo Limpeza dos dados, na fase de mineração dos dados, tem como objetivo: limpar as pontuações do documento **Punctuation Eraser**, modificar as palavras para minúsculo **Case converter**, eliminar palavras (Português e Inglês) que poucam acrescentam no objetivo da análise **Stop Word Filter**, eliminar palavras com caracteres menor ou igual a 3 **N Chars Filter** e eliminar numeros contidos nas frases e numeros puros com o **Number Filter**.

Após os dados no formato Documento passarem pelo fluxo Limpeza de dados, estes foram inseridos no fluxo de enriquecimento de dados conforme a Figura 5. O fluxo de enriquecimentos dos dados adiciona as tags POS sintaxe do idioma (inglês e português), Person (personagens) e Sentiment (positivo, negativo) nas informaçoes coletadas.

Mineração de Dados (Enriquecimento)

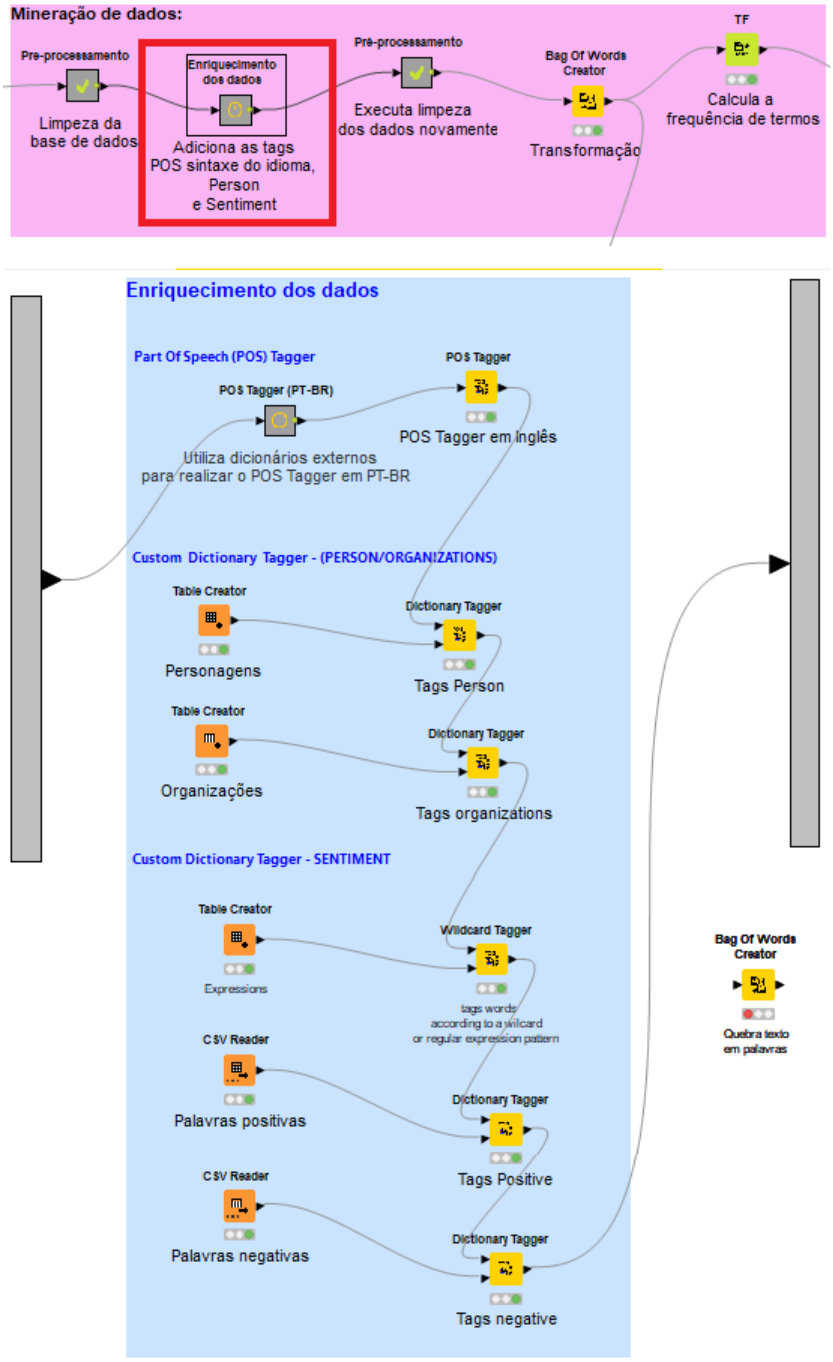

Figura 5 - Fonte: ScreenShot software Knime

A primeira parte do enriquecimento dos dados consistiu em utilizar um meta-nodo POS **TAGGER PT-BR**, conforme a Figura 6, para identificar e adicionar tags em palavras classificadas como adjetivos, substantivos e verbos no idioma português. Para adicionar o tag sintaxe para palavras em inglês, foi configurado o nodo **POS Tagger** nativo do Knime.

Mineração de Dados (POS Tagger)

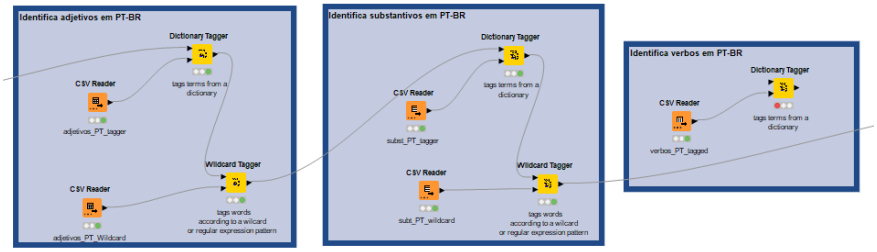

Figura 6 - Fonte:ScreenShot software Knime

Após a inserção das tags do tipo **POS Tagger**, referente a sintaxe dos idiomas das palavras, as próximas etapas consistiram em inserir as tags referentes aos personagens (PERSON) e as organizações (ORGANIZATION), utilizando como referência os dicionários **Dicionary tagger** inseridos nas Figura 7 e Figura 8.

Table Creator - Dicionário Organizações

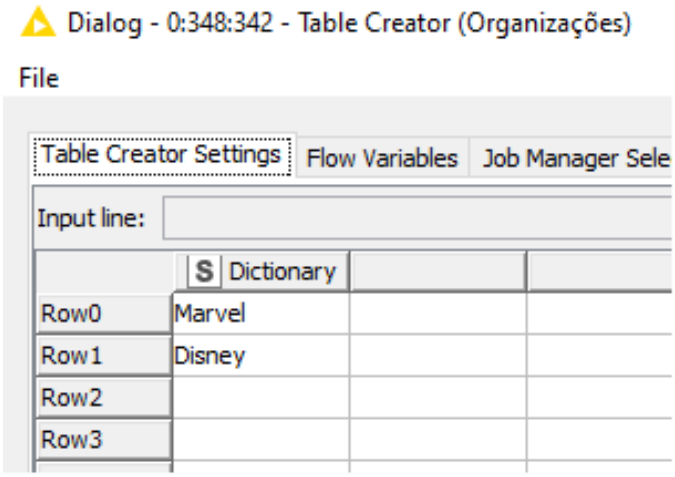

Figura 7 - Fonte:ScreenShot software Knime

Table Creator - Dicionário Personagens

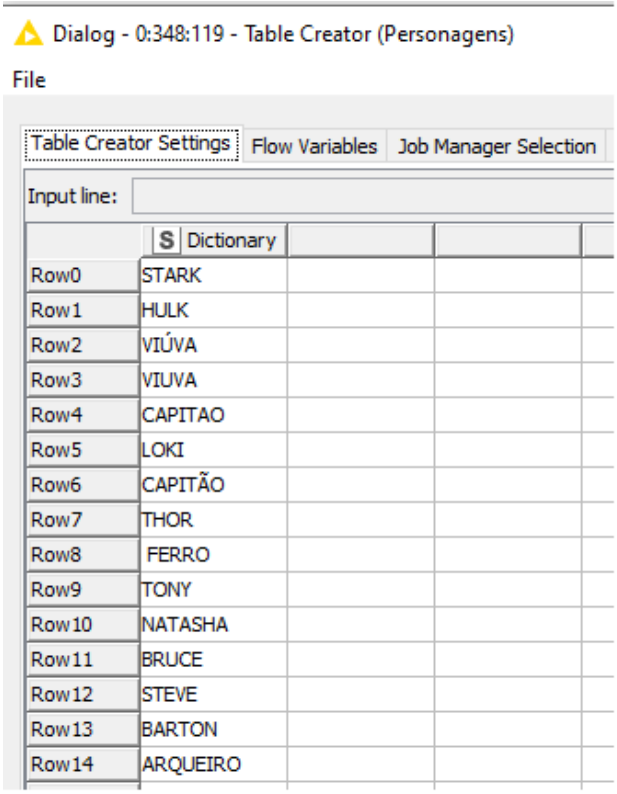

Figura 8 - Fonte:ScreenShot software Knime

As palavras que identificam os sentimentos positivos (POSITVE) e negativos (NEGATIVE), foram inseridas em dicionários **Dicionary tagger**, da mesma forma que explicado acima, para o mapeamento as palavras coletadas dentro do contexto da análise. O nodo **Bag of Words Creator** foi utilizado para criar a coluna *Term* referente aos termos de palavras com suas respectivas tags.

Mineração de Dados (Transformação)

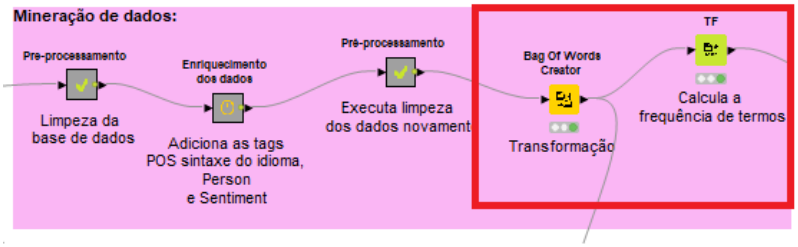

Figura 9 - Fonte:ScreenShot software Knime

As informações gerados pela etapa anterior foram repassadas para os fluxos de análise de sentimentos conforme a Figura 10 e análise de popularidade dos Personagens, conforme Figura 11.

Para ambos os fluxos foram calculados os indices de similaridade das palavras com o nodo **TF**, as tags foram transformados em Strings e inseridas em uma nova coluna **Tags TO String**. Posteriormente, as tags referente a cada análise foram filtradas com o nodo **Row Filter**. Assim foram geradas uma Word-Cloud para cada conjunto de palavras com significado positivo, negativo, e neutro. No caso do sentimento Neutro , os termos que não coincidiram com as palavras fornecidas no dicionário de tags, na etapa de enriquecimento, foram classficados como termos neutros utilizando o nodo **Tags TO String**.

Análise de Sentimentos

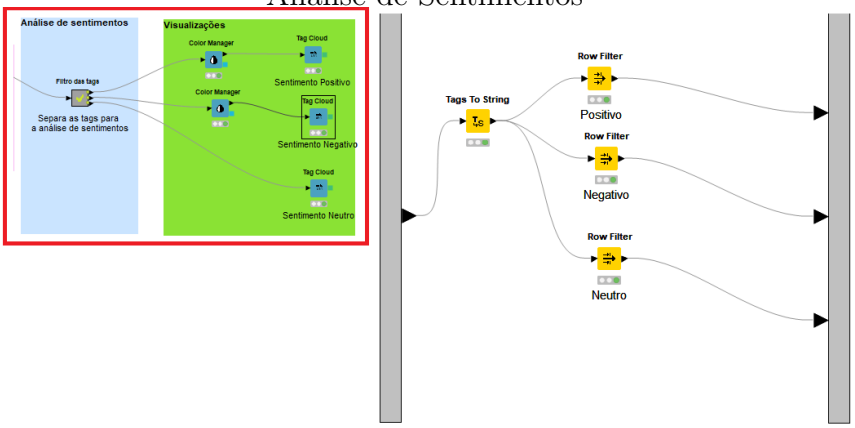

Figura 10 - Fonte:ScreenShot software Knime

Análise de Popularidades dos Personagens

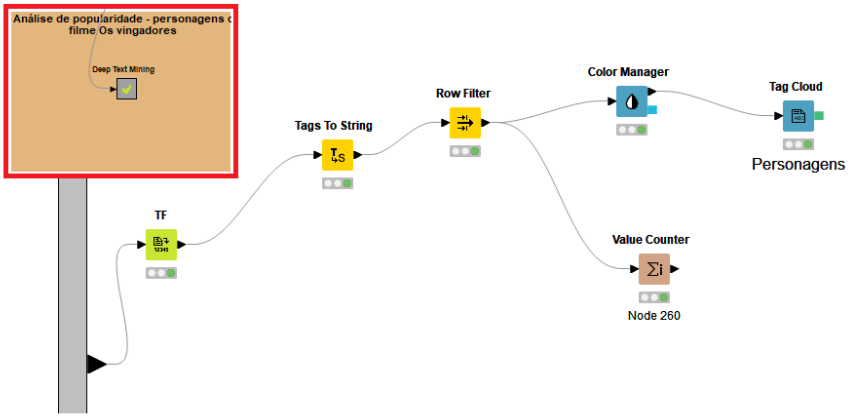

Figura 11 - Fonte:ScreenShot software Knime

Na última etapa de análise foi feita o feito o fluxo para a análise de iterações dos tweets e retweets coletados, conforme ilustrado na Figura 12.

Network e Análise das Interações

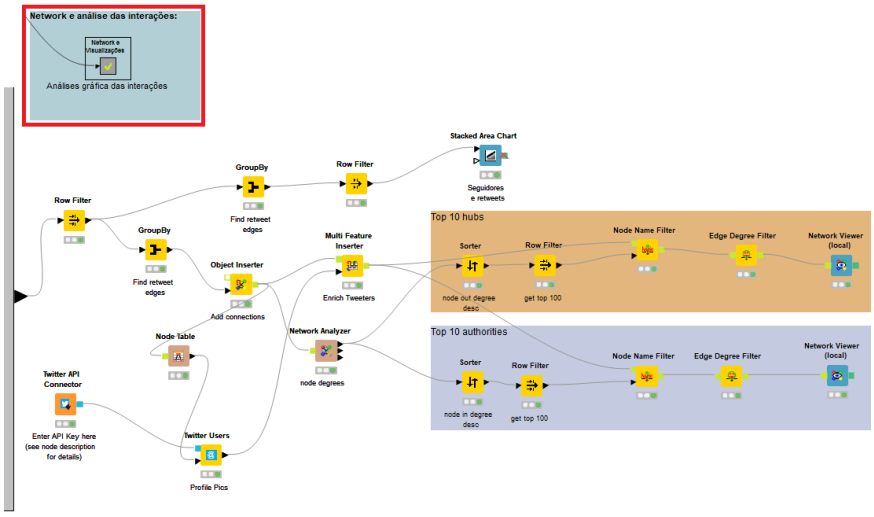

Figura 12 - Fonte:ScreenShot software Knime

3 Descrição dos Resultados

Os resultados para a análise de sentimentos Positivo, Negativo e Neutro, para os tweets coletados são apresentados nas word-cloud das Figuras 13, 14 e 15, respectivamente. Nesta pode ser verificado que associação de tags e filtro das mesmas foram feitas corretamente, porém o dicionário inserido para a classificação de tags deveria ter sido mais direcionado para o propósito do trabalho, de forma a evitar possíveis viéses.

Análise de Sentimentos: Positivo

Figura 13 - Fonte:ScreenShot software Knime

Análise de Sentimentos: Negativo

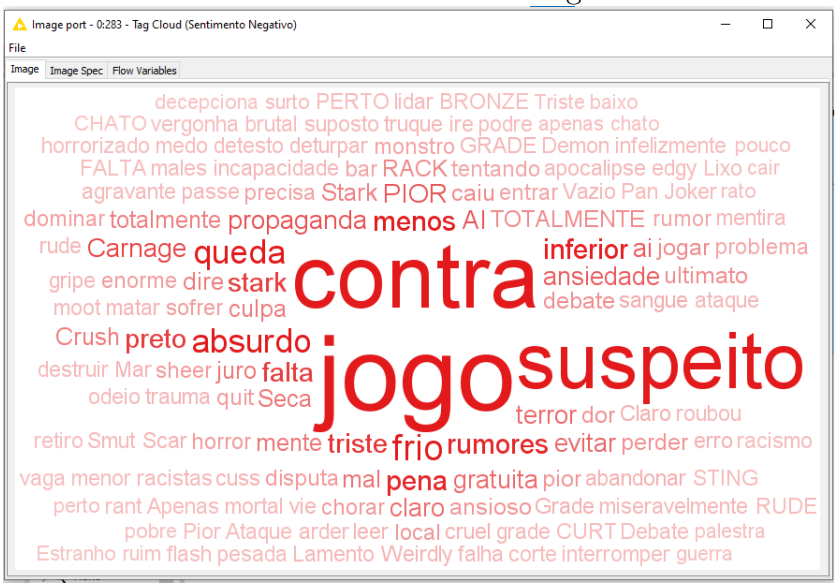

Figura 14 - Fonte:ScreenShot software Knime

Análise de Sentimentos: Neutro

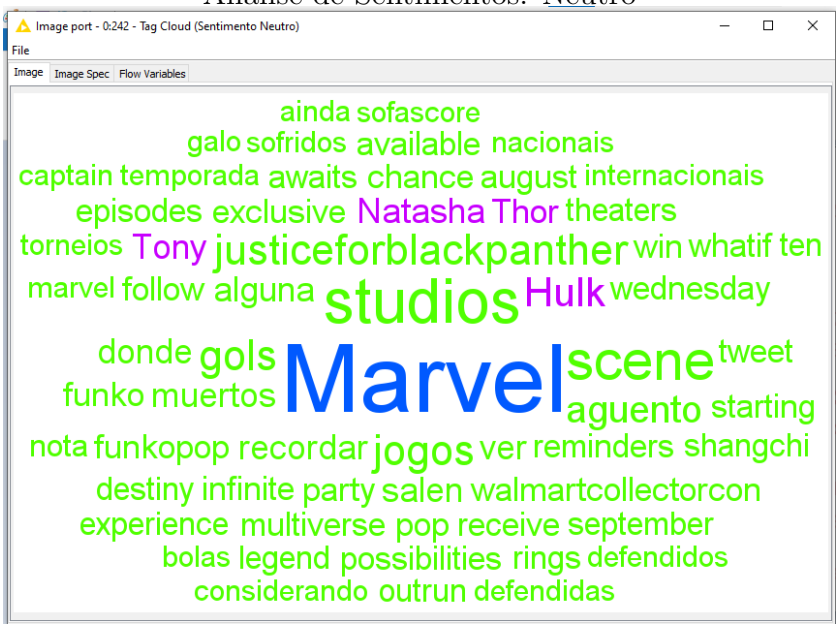

Figura 15 - Fonte:ScreenShot software Knime

Os resultados para a análise de popularidade dos personagens Vingadores, para os tweets coletados é apresentado Word-Cloud da Figura 16. Nesta Figura pode ser observado que os personagens mais populares foram Loki e Hulk. As palavras com sentimento negativos mais comentados foram Jogo, contra e suspeito. Na análise de sentimentos positivo, as mais comentadas foram personagens.

A palavra personagens pode ter sido a mais comentada pela personificação e representação dos personagens em quadrinhos (super, heróis e vilões etc), bem como referência geral de todos os personagens relacionados aos filmes e séries da Marvel.

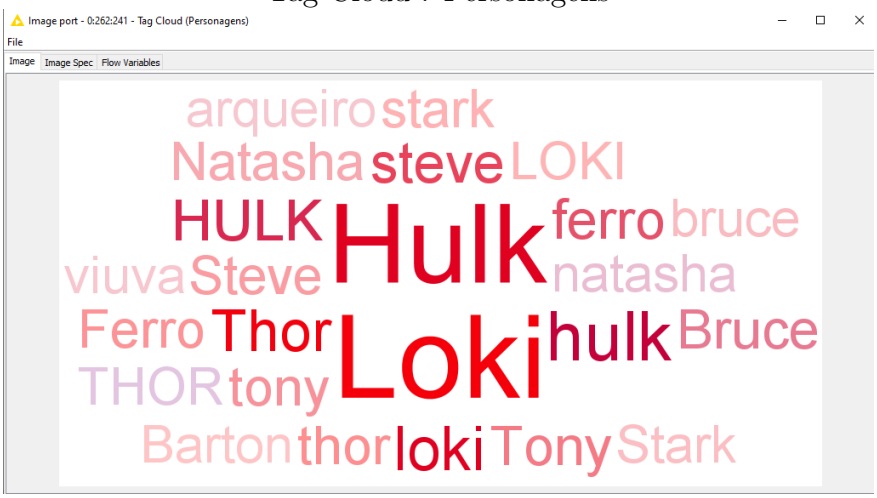

Figura 16 - Fonte: ScreenShot software Knime

Neste figura pode ser observado que o personagem Loki teve o número alto de comentários. A popularidade deste personagem, pode ser associada a decisão da emissora da Disney em lançar a série solo antes 09/06/2021, com apenas 6 episódios. O último episódio da série Loki bateu recorde na plataforma Disney+.

A alta popularidade do Hulk foi ocasionada por problemas de coleta no dados. Isto se deve ao fato da ocorrência de ambiguidade dos termos utilizados para a coleta. O termo Hulk faz referência a um jogador de futebol do time atlético mineiro, bem como referência ao jogador de luta livre Hulk Hogan. Na Figura 17 foi exemplificado o problema descrito na coleta dos tweets.

Analise tweets (Hulk)

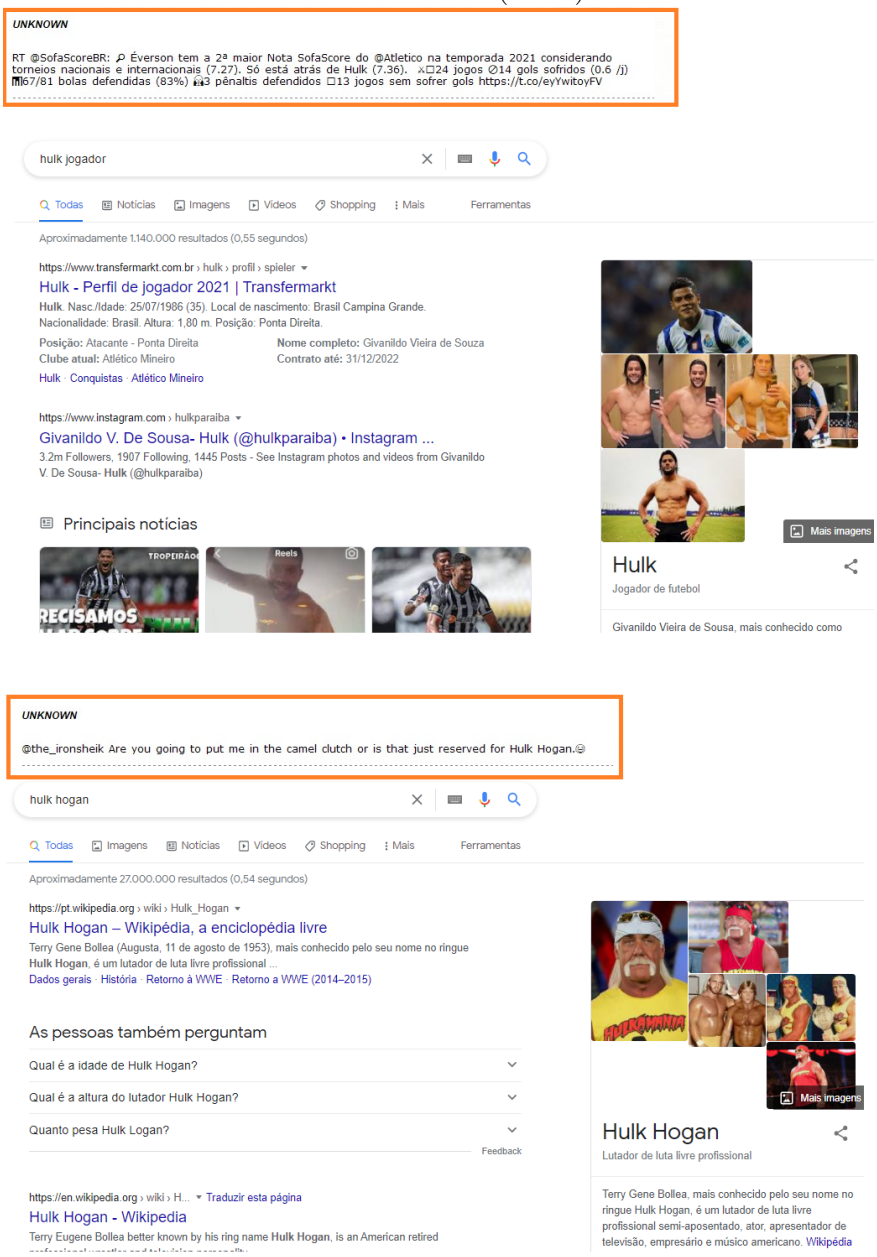

Figura 17 - Fonte: ScreenShot software Knime e Pesquisa Google

A utilização de palavaras mais genéricas, como por exemplos Marvel, Advengers, Vingadores,resultariam em coleta mais efetivas, dentro do contexto desejado. Implicando em maior similaridade com os termos referentes aos personagens da Marvel.

4 Considerações Finais

O tema do trabalho foi principalmente para analisar a Popularidade no Twitter dos Personagens da Saga Vingadores. Optou-se por realizar uma análise objetiva utilizando somente o Software Knime para o tratamento, análise e a apresentação dos dados.

Foram extraídos 4000 Tweets que possuíam ("Marvel","vingadores", e o nome dos personagens selecionados) escrito, da rede social Twitter. No tratamento de dados foi extraído o nome dos personagem Tony Stark (Homem de ferro), Bruce (Hulk), Natasha (Viúva negra), Steve (Capitão américa), Loki, Thor, Barton (Arqueiro), para eles foi feito uma contagem de qual é o mais citado.

A maior dificuldade encontrada no trabalho foi a falta de conhecimento em alguns nodos e a falta de experiência com o software Knime, a ideia principal era integrar o Knime com um banco de dado para atualização em tempo real das informações, além da falta de conhecimento de gráficos mais apresentáveis e mais claros para análise dos resultados finais.

A decisão de trabalhar com a Saga foi pela audiência e destaque que ela possui na mídia, além da história que envolve o universo Universo Cinematográfico Marvel e HQ (História em quadrinhos).

Bibliografia

KNIME. Disponível em: ¡ https://www.knime.com/knime¿ Acesso em: 1 Agosto 2021.

HUPDATA Data Analysis Solutions. KNIME: Como otimizar a performance de seus workflows. Disponível em: ¡https://hupdata.com/otimizando-workflows-knime/¿ Acesso em: 1 Agosto 2021.

Vingadores. In: Wikipédia: a enciclopédia livre. Disponível em: ¡https://pt.wikipedia.org/wiki/Vingadores ¿ Acesso em: 1 Agosto 2021.

Hulk (futebolista). In: Wikipédia: a enciclopédia livre. Disponível em: ¡https://pt.wikipedia.org/wiki/Hulk (futebolista) ¿ Acesso em: 1 Agosto 2021.

Hulk Hogan. In: Wikipédia: a enciclopédia livre. Disponível em: ¡https://pt.wikipedia.org/wiki/Hulk Hogan¿ Acesso em: 1 Agosto 2021.