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 da Série Game of Thrones

Alunos: Gabriel Augusto Ricardo, João Vitor Mendes

Professor: Zilton Cordeiro Jr.

Abril
2019

IEC - Instituto de Educação Continuada

Pós-Graduação em Ciência dos Dados e Big Data

Projeto Final

Análise de Popularidade no Twitter dos Personagens da Série Game of Thrones

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: Gabriel Augusto Ricardo, João Vitor Mendes

Professor: Zilton Cordeiro Jr.

Abril
2019

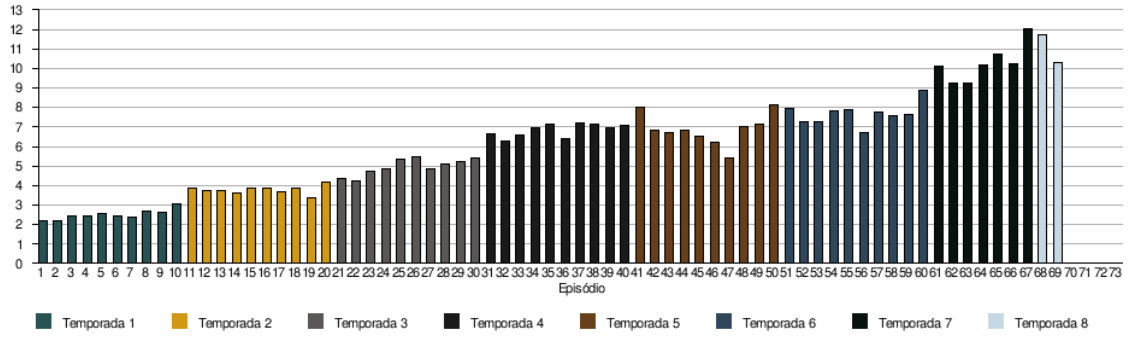
Conteúdo

Abril
2019

1 Introdução

Game of Thrones é uma série de televisão distribuída pelo canal HBO criada por David Benioff, inspirada na série de livros A Song of Ice and Fire, de George R. R. Martin, a série é um fenômeno de audiência no mundo como ser visto na Imagem 1, Game of Thrones está em sua oitava e última temporada(WIKIPEDIA, 2019).

Imagem 1: Audiência Game of Thrones (em milhões)



Fonte: Wikipédia(2019)

O Twitter é uma das maiores redes sociais da atualidade, nele os usuários podem se fazer comentários de vários temas. Com a última temporada de Game of Thrones em andamento, a série está sendo muito comentada no Twitter, gerando assim grande quantidade de dados sobre a série.

Esse trabalho busca comparar a popularidade dos personagens da série Game of Thrones, usando uma amostra de 4000 comentários do Twitter feitos por usuários dessa rede social. Os personagens que serão pesquisados são Brienne, Arya, Cersei, Sansa, Tyrion, Jon, Daenerys, Jaime, Bran, Theon, Samwell e Littlefinger. Os softwares utilizados foram o Knime, Microsoft Office Excel e Microsoft Power BI para fazer a extração, tratamento, análise e apresentação dos dados.

2 Descrição das Atividades

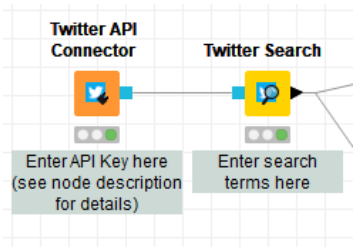
i Extração dos dados e Tratamento dos Dados

Foi utilizado o software KNIME 3.7.1 para fazer a extração dos dados e posteriormente para fazer o tratamento desses dados em várias etapas utilizando os nodos disponíveis na aplicação como o explicado abaixo. A extração e tratamento foi feito para gerar um arquivo que possuísse os personagens do Game of Thrones por coluna junto as colunas de sentimento e autor, o resultado esperado nas linhas é a quantidade de vezes que cada personagem foi citado nas publicações do Twitter extraídas.

O nodo Twitter API Connector foi utilizado para a conexão com o Twiter, a conexão foi feita com as credenciais geradas no Twitter Development.

O nodo Twitter Search foi empregado para realizar a busca dos dados, query do nodo "Game of Thrones", número de linhas 4000, campos buscados Tweet, “User(Screen Name)” e Name.

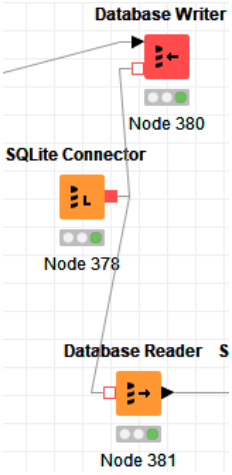
Imagem 2: Nodos Twitter Search e Twitter API Connector



Fonte: ScreenShot software Knime

Os Nodos DataBase Writer, SQLite Connector e Database Reader foram utilizados para armazenar os dados pesquisados anteriormente no Twitter.

Imagem 3: DataBase Writer, SQLite Connector e Database Reader



Fonte: Screenshot software Knime

O nodo Strings to Document foi usado para passar os dados do Twitter para tipo Documento, Campo Tweet usado para o texto do documento, campo user usado para autor do documento.

O nodo Column Filter foi usado para remover as colunas TweetID, Tweet e User.

O nodo Punctuation Erasure foi usado para limpar a pontuação da coluna documento criada.

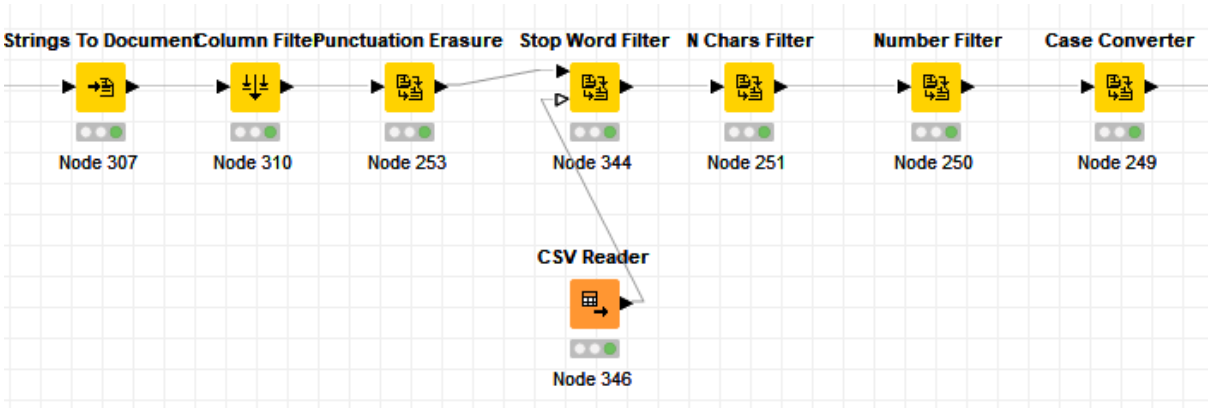
O nodo Stop Word Filter foi utilizado para remover as palavras indesejadas do documento, foi usada a lista de stopWords do Knime e uma lista disponível no site Ranks NL (2019).

O nodo NChars Filter foi utilizado para filtrar os termos que tenham menos que duas letras.

O nodo Number Filter foi usado para filtrar os termos que representam números.

Case Converter foi usado para converter todos os termos do documento para minúsculo.

Imagem 4: Nodos Strings to Document, Column Filter, Punctuation Erasure, Stop Word Filter, NChars Filter, Number Filter e Case Converter



Fonte: Screenshot software Knime

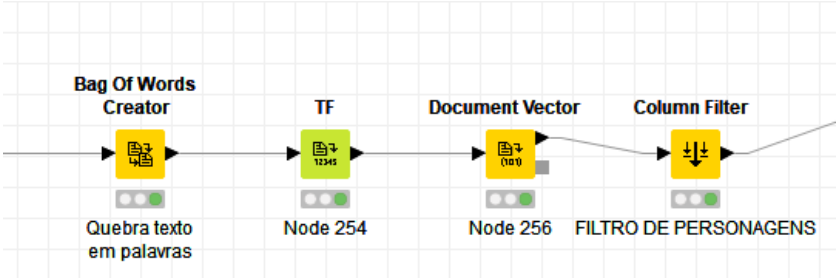
O Nodo Bag of Words foi utilizado para quebrar o texto da coluna documento em uma linha por palavra.

O nodo TF foi utilizado para contar o número que cada termo apareceu no documento.

Document Vector foi utilizado para transformar as linhas que continham palavra em colunas.

O nodo Column Filter foi usado para selecionar apenas as colunas desejadas, essas colunas são a de documento e todas as outras que possuem o nome ou são sinônimo dos personagens da série: Brienne, Arya, Cersei, Sansa, Tyrion, Jon, Daenerys, Jaime, Bran, Theon, Samwell e Littlefinger.

Imagem 5: Nodos Bag of Words, TF, Document Vector e Column Filter

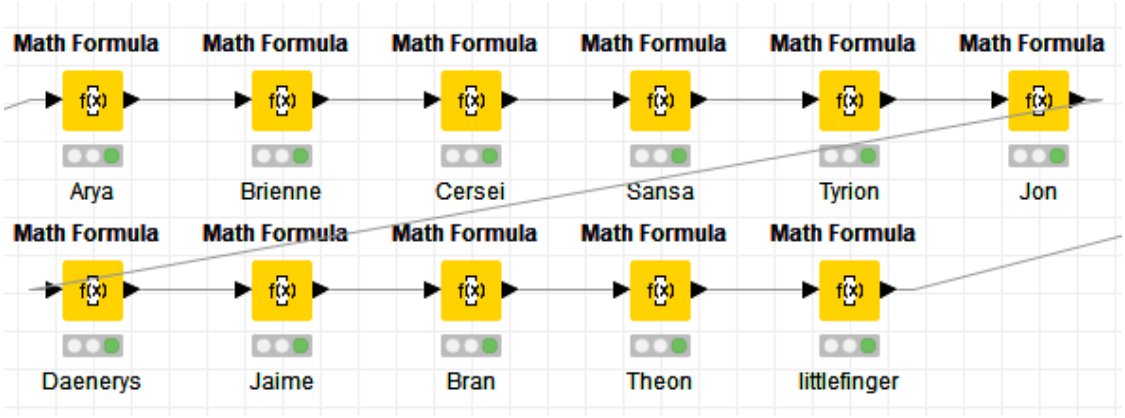


Fonte: Screenshot software Knime

Foi utilizado um nodo Math Formula para cada um dos personagem que esse trabalhado está analisando. O objetivo de cada um desses nodos é somar a contagem de palavras feita entre as colunas que remetem ao mesmo personagem, criando uma nova coluna.

Expressão utilizada para o personagem Arya: $arya$ + $aryagendry$ + $aryax$. Os outros personagens tiveram os nodos com as expressões criadas com a mesma abordagem.

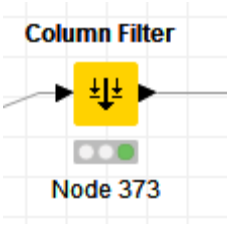
Imagem 6: Nodos Math Formula por personagem



Fonte: Screenshot software Knime

Logo após o nodo Column Filter foi utilizado para filtrar os campos Documento e os campos criados a partir dos nodos anteriores Math Formula.

Imagem 7: Nodo Column Filter

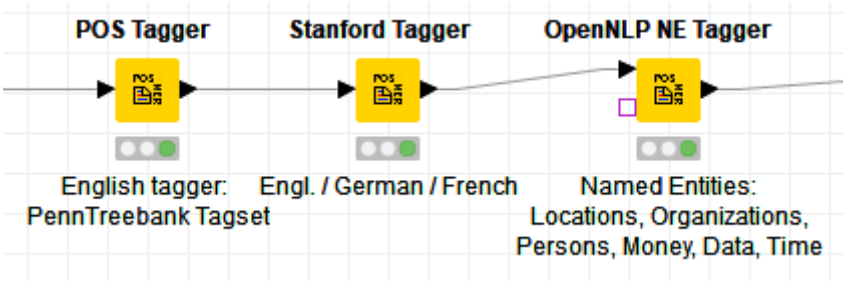


Fonte: Screenshot software Knime

O nodo POS Tagger Utiliza uma gramática para a língua inglesa e através de modelos probabilísticos procura detectar classes de palavras: adjetivos, verbos, sujeitos. Da mesma forma o nodo Stanford Tagger procura detectar classes de palavras (CORDEIRO JR, 2019).

O nodo OpenNLP NE Tagger tenta reconhecer algumas categorias de entidades no texto(CORDEIRO JR, 2019).

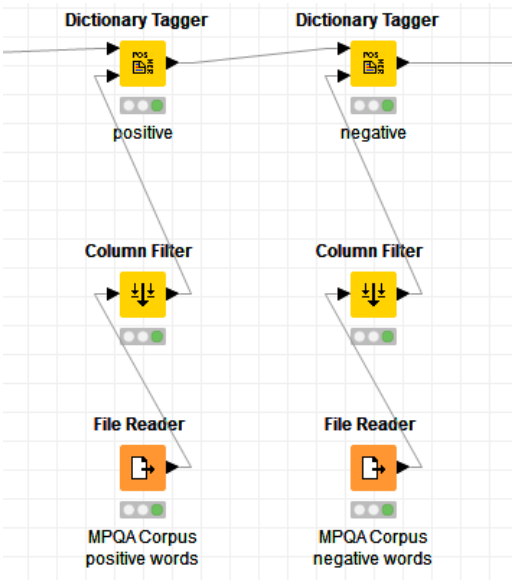
Imagem 8: POS Tagger, Stanford Tagger e OpenNLP NE Tagger



Fonte: Screenshot software Knime

O Nodo Dictionary Tagger cria tags de acordo com um dicionário passado ao nodo, foram utilizados dois nodos um para um dicionário com palavras positivas e um para um dicionário com palavras negativas, ambos dicionários encontrados no documento Recuperação da Informação na Web e em Redes Sociais criado por Zilton Cordeiro Jr foi usado nodo File Reader para ler esse documento.

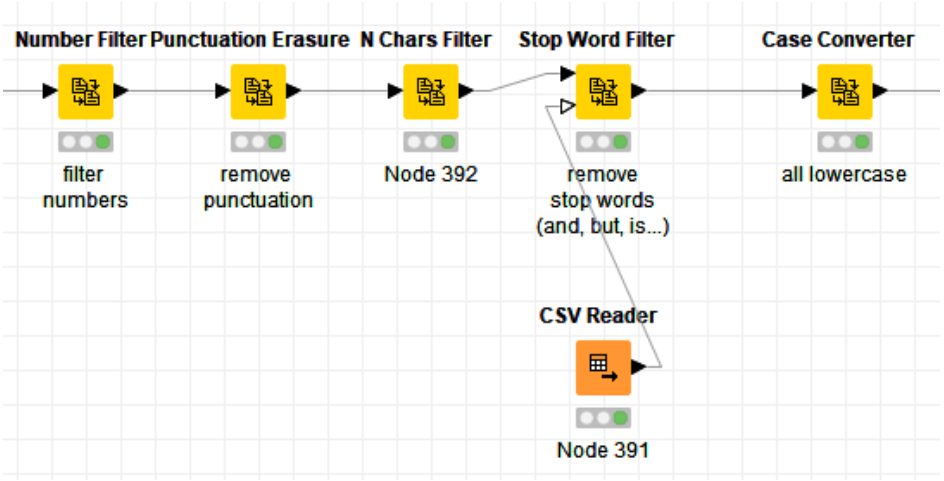
Imagem 9: Nodos Dictionary Tagger e File reader



Fonte: Screenshot software Knime

Foi feito o mesmo processamento de dados que foi feito no começo do tratamento de dados desse documento, pois o documento pré-processado anteriormente se perdeu após o nodo Document Vector.

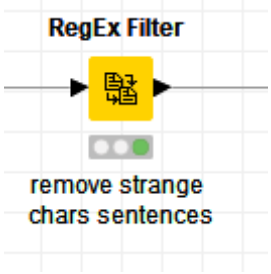
Imagem 10: Number Filter, Punctuation Erasure, N Chars Filter, Stop Word Filter e Case Converter



Fonte: Screenshot software Knime

O Nodo RegEx Filter foi utilizado para remover termos estranhos do documento.

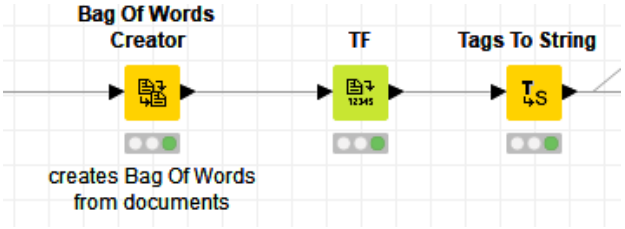
Imagem 10: Nodo RegEx Filter



Fonte: Screenshot software Knime

Os Nodos Bag of Words Creator, TF e Tags to String foram criados respectivamente para quebrar o documento em palavras, contar o numero de palavras por documento e passar as tags de sentimento para string e as que não estivessem classificadas receberam “neutral” como tag. Dessa forma agora existe para cada palavra uma tag positiva, negativa ou neutra, essas estão multiplicadas pelo numero de colunas de atores, posteriormente nesse documento será feito o tratamento dessas duplicidades.

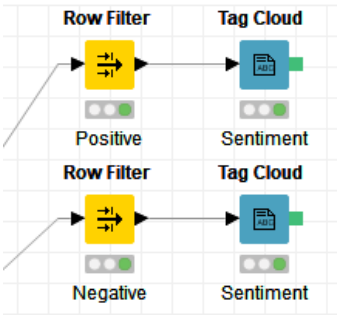
Imagem 12: Nodos Bag of Words Creator, TF e Tags to String



Fonte: Screenshot software Knime

Os dois nodos Row Filter foram utilizados para filtrar a linhas positivas e negativas, para que depois os nodos Tag Cloud gerassem uma nuvem de palavras positiva e negativa sobre a série.

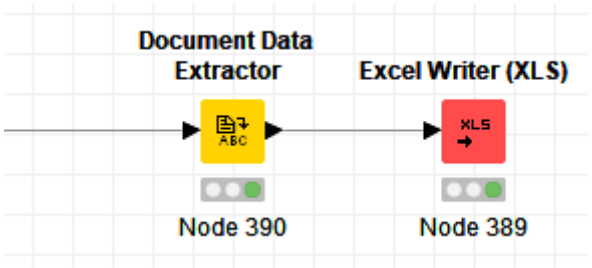
Imagem 13: Nodos Row Filter e Tag Cloud



Fonte: Screenshot software Knime

O nodo Document Data Extractor foi usado para extrair o autor e texto do documento, o nodo Excel Writer foi usado para gerar um arquivo Excel do resultado do tratamento de dados, nesse nodo foi removido via filtro as colunas Document, Tagged_Document e TF abs, o nome do arquivo gerado é “ExtracaoFinal.xlsx”.

Imagem 14: Nodos Document Data Extractor e Excel Writer.



Fonte: Screenshot software Knime

ii Análise dos Dados

A análise dos dados foi feita através do software Microsoft Office Excel 2016.

Foi utilizado o arquivo cópia do gerado anteriormente “ExtracaoFinal.xlsx”.

As linhas em que o dado das colunas abaixo somadas é igual a zero foram removidas, isso corresponde a 93% das linhas do arquivo Excel.

Colunas: Arya_Column;

Brienne_Column;

Cersei_Column;

Sansa_Column;

Tyrion_Column;

Jon_Column;

Daenerys_Column;

Jaime_Column;

Bran_Column;

Theon_Column;

littlefinger_Column.

Foi feito uma “média” da coluna sentimento por autor.

Para isso foi aplicado a formula abaixo em uma nova coluna “Sentimento_Média” que apontava para coluna SENTIMENT.

Formula: =IF(SENTIMENT ="neutral",0,IF(SENTIMENT ="POSITIVE",1,-1))

A formula dá o valor zero(0) para sentimentos neutros, um(1) para positivos e menos um (-1) para negativos.

Por final foi feito a soma desses valores agrupando por autor.

Os autores que tiveram o resultado do agrupamento igual a zero foram classificados como neutro, os que tiveram o sentimento maior que zero foram classificados como positivo, os que tiveram o sentimento menor que zero foram classificados como negativo.

Foi removido as colunas Term, SENTIMENT e Text.

Continuaram no arquivo as colunas Author, Sentimento_Média, Arya_Column, Brienne_Column, Cersei_Column, Sansa_Column, Tyrion_Column, Jon_Column, Daenerys_Column, Jaime_Column, Bran_Column, Theon_Column, littlefinger_Column.

Dando continuidade foi usado a funcionalidade Remove Duplicates, para remover as linhas duplicadas, deixando apenas uma linha por autor.

O Arquivo “PrimeiraAnalise_ExtracaoFinal.xlsx” representa o resultado da análise feita acima, esse está sendo usado para geração de alguns dos gráficos foram apresentados no trabalho.

iii Apresentação dos Dados

Foi utilizado o Software Microsoft Power BI Desktop 2.68 para realizar a apresentação dos dados gerados.

Foi carregado no Power BI o arquivo “ExtracaoFinal.xlsx”, para coluna Term foi removido os caracteres após a string “[“.

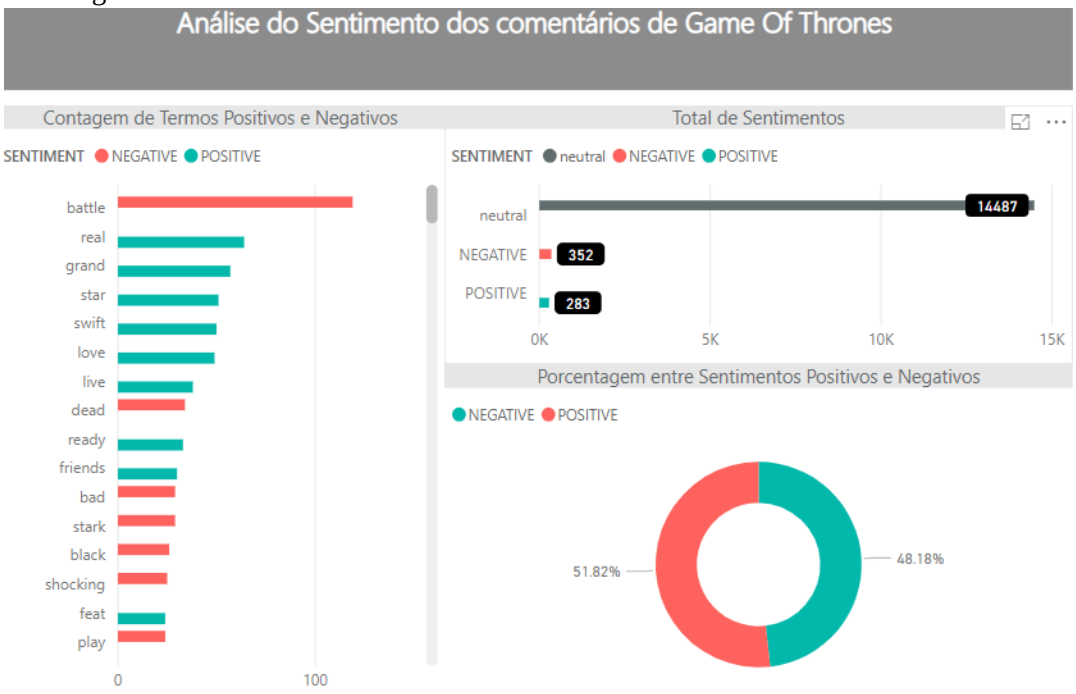
O arquivo “PrimeiraAnalise_ExtracaoFinal.xlsx” foi carregado sem edições no Software.

3 Análise dos Resultados

Foi gerado o relatório “Análise do Sentimento dos comentários de Game Of Thrones” para mostrar a comparação entre palavras neutras, positivas e negativas.

Relatório disponível em:
https://app.powerbi.com/view?r=eyJrIjoiMmRmNWI0OTItMDRmNC00MDczLWJhMzEtNmU1ODc5MzU2ZjUwIiwidCI6IjE0Y2JkNWE3LWVjOTQtNDZiYS1iMzE0LWNjMGZjOTcyYTE2MSIsImMiOjh9

Imagem 15: Relatório Análise do Sentimento dos comentários de Game Of Thrones



Fonte: Screenshot software Power BI

Além da apresentação no PowerBI, como o explicado anteriormente foi gerado duas nuvens de palavras, uma positiva e uma negativa via software Knime onde podemos comparar as palavras com o relatório acima.

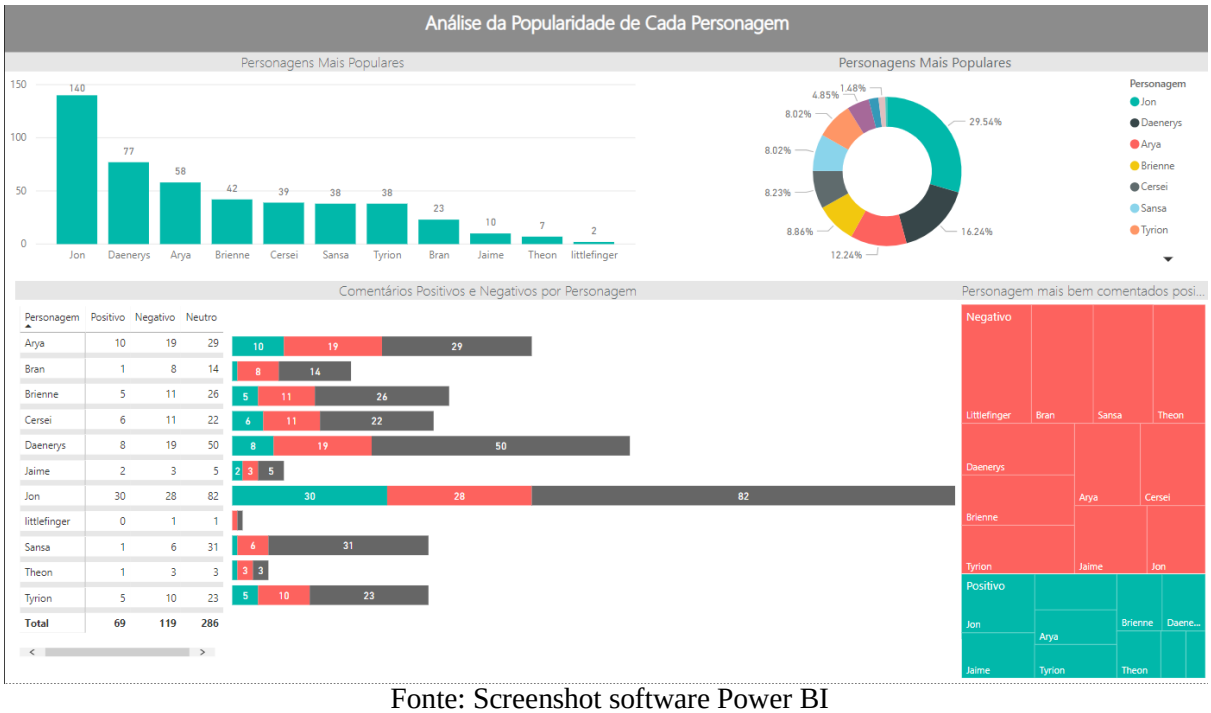
Imagem 16 e 17: Respectivamente nuvem de palavras positiva e nuvem de palavras negativa

Fonte: Screenshot software Knime

Foi gerado o relatório “Análise da Popularidade de Cada Personagem” que mostra quem foram os personagens mais comentados no Twitter.

Relatório disponível em:
https://app.powerbi.com/view?r=eyJrIjoiMzU4ZmQzYWYtNzNiZi00YzdhLTlkODUtMjY4ZWNiZDM5Mjg3IiwidCI6IjE0Y2JkNWE3LWVjOTQtNDZiYS1iMzE0LWNjMGZjOTcyYTE2MSIsImMiOjh9

Imagem 18: Relatório Análise da Popularidade de Cada Personagem



Fonte: Screenshot software Power BI

4 Conclusão

O tema do trabalho foi Análise de Popularidade no Twitter dos Personagens da Série Game of Thrones. Optou-se por realizar um analise objetiva utilizando o Software Knime para a extração e tratamento de dados, o Microsoft Office Excel para realizar a análise dos dados e o software Microsoft Power Bi para realizar a apresentação dos dados. Foram extraídos 4000 Tweets que possuíam "Game Of Thrones" escrito, da rede social Twitter. No tratamento de dados foi extraído o nome dos personagem Brienne, Arya, Cersei, Sansa, Tyrion, Jon, Daenerys, Jaime, Bran, Theon, Samwell e Littlefinger, para eles foi feito uma contagem de qual é o mais citado. Além disso adicionado uma etiqueta a cada palavra dos comentários do Twitter extraídos, para saber se eram termos positivos ou negativos.

A maior dificuldade encontrada no trabalho foi a falta de conhecimento em alguns nodos e a falta de experiência com o software Knime. O software Excel foi utilizado por esse motivo, pois a aplicação Knime demanda muito tempo para a geração do mesmo resultado para quem não tem conhecimento na mesma. A decisão de utilizar o Software Power BI foi pela facilidade e afinidade com o software. A decisão de trabalhar com a série Game Of Thrones foi pela audiência que ela possui, além dos integrantes da grupa gostarem dessa série.

Obteve-se o resultado que o personagem mais comentado é o Jon Snow, a palavra negativa mais comentada foi "battle" e a positiva mais comentada foi "real". O estudo não teve a intenção de generalizar resultados para toda a série, pois foi extraído apenas uma amostra dos dados no dia 27 de Abril de 2019, o objetivo do trabalho é instigar discussões sobre o assunto além de mostrar uma forma de fazer tal análise dos dados.

Bibliografia

AGUIRRE, L. A. Introdução à Identificação de Sistemas, Técnicas Lineares e Não lineares Aplicadas a Sistemas Reais. Belo Horizonte, Brasil, EDUFMG. 2004.

GAME Of Thrones. In: Wikipédia: a enciclopédia livre. Disponível em: <https://pt.wikipedia.org/wiki/Game_of_Thrones > Acesso em: 4 abril 2019.

STOPWORD Lists. In: Ranks NL. Disponível em: <https://www.ranks.nl/stopwords> Acesso em: 4 abril 2019.

CORDEIRO JR, Zilton. Recuperação da Informação na Web e em Redes Sociais. PUC-Minas, Belo Horizonte, Brasil. 2019.

Anexo

ANEXO 1 – Documento de palavras não desejadas em inglês – “stopWordsEN.csv”

ANEXO 2 – Dicionário de dados de palavras negativas em inglês – “corpus-traduzido-negativo.csv”

ANEXO 3 – Dicionário de dados de palavras positivas em inglês – “corpus-traduzido- positivo.csv”

ANEXO 4 – Documento gerado depois do tratamento dos dados no Knime – “ExtracaoFinal.xlsx”

ANEXO 5 – Documento gerado após a análise dos dados no Excel – “primeiraAnalise_ExtracaoFinal.xlsx”

ANEXO 6 – Documento com o Workflow desenvolvido para o trabalho no Knime, o WorkFlow possui a base de dados onde foi salvo os dados extraídos – “TrabalhoFinal.knar”