Relatório do PGC II (2020.3)

# Sistema de coleta de dados na internet para medição de reputação de empresas em arquitetura distribuída

Aluno: Lucas Andrade Cioffi  
Orientador: Francisco Isidro Massetto

UFABC – 2020

# 1 – Introdução

A imagem pública de uma pessoa, grupo, empresa ou organização pode ser um de seus assets mais valiosos. Nos dias atuais, a grande maioria das empresas presentes na bolsa de valores possuem uma presença em redes sociais. Redes sociais têm um impacto em todas as esferas de performance do negócio. (Paniagua & Sapena, 2014). Empresas com reputações boas têm mais facilidade em encontrar e manter clientes e em anunciar seus produtos. O oposto é válido para empresas com reputações ruins (Herbig & Paul, 1993). A percepção do consumidor sobre a imagem da marca é fundamental na decisão de compra ou a escolha de uma empresa em favor de outras (Ataman & Ulengin, 2003).

A imagem pública também é imensamente importante para indivíduos famosos, influencers, organizações e afins. Tendo isso em vista, seria interessante um indivíduo ou organização poder monitorar e ter um indicador de qual a sua reputação on-line e como ela está evoluindo com o passar do tempo.

No cenário cibernético atual, os espaços de discussão onde as reputações de um grupo ou indivíduo possa ser discutido são extremamente numerosos e volumosos. Monitorar todos manualmente por menções à um indivíduo ou organização e transformar isso em indicadores de reputação seria extremamente custoso ou praticamente impossível.

# 2 – Justificativa

Hoje existem ferramentas de monitoração de reputação ou de presença digital, mas a maioria delas são pagas (ReputationVIP, Mention, Alerti) e as gratuitas não fornecem uma visão completa ou um histórico da reputação, apenas alertas sobre conteúdos novos publicados (Google Alerts, Talkwalker).

Em uma pesquisa com executivos de relações públicas, um dos desafios relacionados a redes sociais que foi frequentemente citado é a falta de controle sobre o que as pessoas podem fazer ou dizer – isso inclui a propagação de informações falsas, criticismo e ações por parte de grupos ativistas (DiStaso & McCorkindale, 2011). É claro que nem todas essas ameaças podem ser contidas, mas poder detectá-las o mais rápido possível é interessante.

Portanto seria útil à sociedade como um todo a existência de uma ferramenta open-source, gratuita, modular, extensível e de fácil utilização que permitisse a monitoração, acompanhamento e geração de indicadores sobre a reputação de um indivíduo, organização, tema ou afins.

# 3 – Objetivos

O objetivo é desenvolver um sistema distribuído capaz de monitorar menções ou termos disponíveis na internet a respeito de uma empresa, pessoa ou instituição, enviar estes dados para um sistema de processamento e disponibilizar para um usuário final uma gama de informações, como por exemplo: termos e correlações mais frequentes, popularidade, sentimentos mais frequentes e afins. O usuário também poderá fazer buscas personalizadas. O objetivo é fornecer ao usuário uma visão automatizada do que está sendo falado sobre a sua empresa, pessoa ou instituição. Tais informações podem apoiar na tomada de decisões relacionadas à marketing, negócios, comunicações e afins.

Tal sistema deve ser open-source, gratuito, modular, escalável e de fácil utilização.

# 4 – Metodologia

Para a criação desse sistema pretendo utilizar um conjunto de sistemas distribuídos escaláveis e gratuitos já existentes aliados a scripts, programas e configurações especiais para o resto das funcionalidades. O sistema pode facilmente ser hospedado na nuvem, se o usuário desejar.

A coleta de dados será feita por meio de robôs especializados e que podem ser executados de maneira assíncrona e escalonável. O armazenamento e processamento dos dados coletados será feito por um banco de dados não relacional, para facilitar a captura de dados variados.

Durante o PGC 1, comecei a criar o sistema utilizando o Splunk (<https://www.splunk.com/>), que é um banco de dados não relacional com capacidades avançadas de processamento de dados e geração de relatórios. O Splunk é escalonável por design: servidores adicionais podem ser adicionados à sua configuração para auxiliar no armazenamento, coleta ou processamento dos dados. Diversos aplicativos e módulos também estão disponíveis para download gratuitamente, caso necessário. O Splunk também é gratuito para até 500MB ingeridos na ferramenta por dia, que é o suficiente para uma organização de pequena escala.

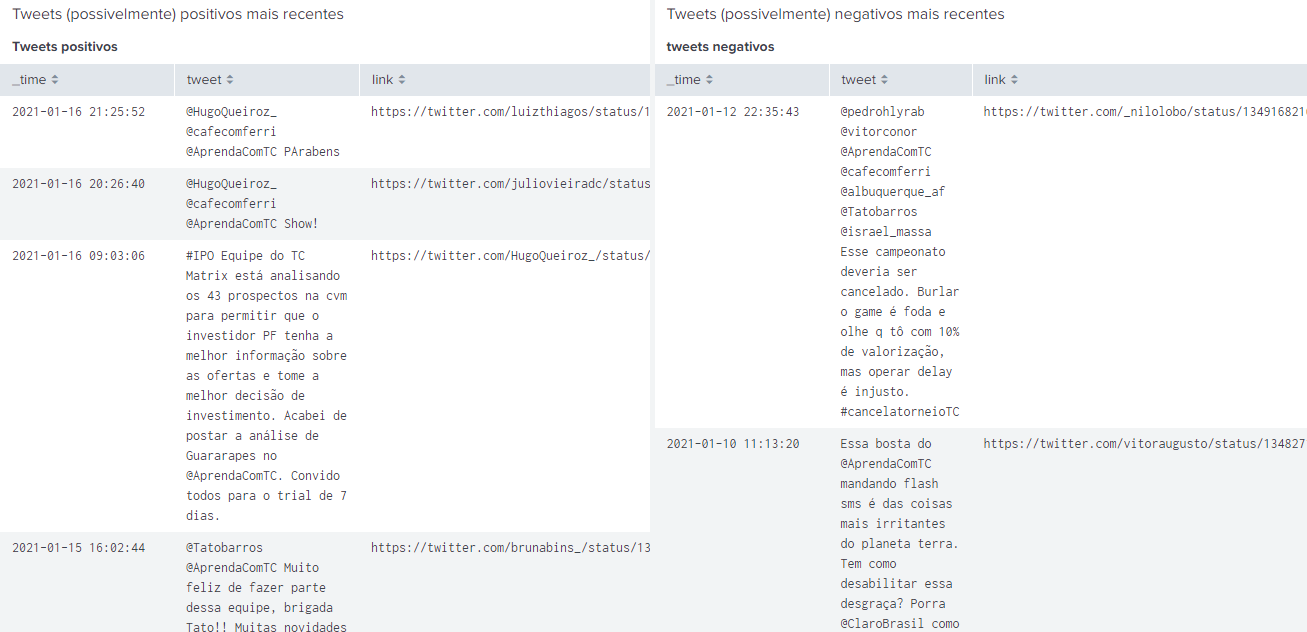
Durante o PGC 1 criei um robô em Python especializado na coleta de informações pelo Twitter. Este robô utiliza a biblioteca “twint” para realizar buscas no Twitter. O robô realiza uma busca por termos pré-definidos pelo usuário e coleta todos os tweets que contenham o termo. Os dados então são enviados para o Splunk usando a biblioteca oficial do Splunk para Python.

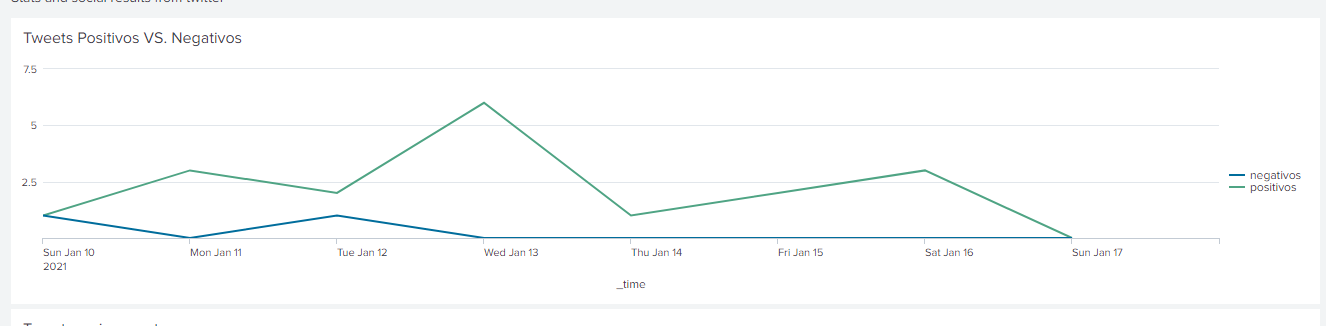
Durante o PGC 2 criei um manual de instalação completo para a solução, melhorei o coletor de informações do Twitter em diversos pontos, busquei mais referências bibliográficas para o projeto (inclusas nesse relatório) e atualizei as bibliotecas utilizadas e empacotei todo o projeto e disponibilizei em um repositório no Github (<https://github.com/lacioffi/UniversalBrandMonitor>).

O Splunk é responsável por armazenar os dados e processá-los. Quando o Splunk recebe os dados, que são salvos como “eventos” distintos, ele realiza um processamento inicial para catalogar os campos do evento, como por exemplo: conteúdo do tweet, usuário responsável, citações, likes, etc. Após esse processamento eles ficam disponíveis para busca e processamentos adicionais.

O processamento dos dados e geração de resultados é feito por meio de “queries”, que são as buscas feitas em SPL (Search Processing Language), uma linguagem própria do Splunk extremamente poderosa. A linguagem é capaz de realizar filtragens, cálculos estatísticos, procedimentos sobre os dados, gerar gráficos e mais. Para o projeto em questão, utilizei-a para gerar uma lista de tweets sobre uma empresa específica que podem ser “positivos” e uma lista de tweets “negativos”. A filtragem é feita comparando as palavras do tweet com uma lista de palavras “positivas” como “legal”, “sensacional”, “show” e “negativas” como “lixo”, “problema” e “erro”.







# 5 – Cronograma

O cronograma restante é o seguinte:

PGC III: Criação de indicadores numéricos para medir reputação digital baseado nos dados extraídos até o momento. Extração de mais inteligência à partir dos dados coletados (ex: usando PLN). Criação de coletor para obter dados de outras fontes além do Twiter. Geração de relatórios exportáveis para apresentação ao usuário. Geração do relatório final de PGC.

# 6 - Referências Bibliográficas

Paniagua, Jordi, and Juan Sapena. "Business performance and social media: Love or hate?." *Business horizons* 57.6 (2014): 719-728.

Herbig, Paul, and John Milewicz. "The relationship of reputation and credibility to brand success." *Journal of consumer marketing* (1993).

Ataman, Berk, and Burç Ülengin. "A note on the effect of brand image on sales." *Journal of Product & Brand Management* (2003).

DiStaso, Marcia W., Tina McCorkindale, and Donald K. Wright. "How public relations executives perceive and measure the impact of social media in their organizations." *Public relations review* 37.3 (2011): 325-328.

|  |
| --- |
|  |
|  |
|  |  |