João Paulo de Almeida Castañón

Análise de Popularidade de Canais do YouTube

João Monlevade

Agosto, 2017

João Paulo de Almeida Castañón

Análise de Popularidade de Canais do YouTube

Monografia apresentada ao Curso de Sistemas de Informação do Departamento de Computação e Sistemas, como requisito parcial para aprovação na Disciplina Trabalho de Conclusão de Curso II.

UNIVERSIDADE FEDERAL DE OURO PRETO – UFOP

INSTITUTO DE CIENCIAS EXATAS E APLICADAS – ICEA

DEPARTAMENTO DE COMPUTAÇÃO E SISTEMAS – DECSI

Orientador: Theo Silva Lins

João Monlevade

Agosto, 2017

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E APLICADAS
COLEGIADO DO CURSO DE SISTEMAS DE INFORMAÇÃO

ANEXO IV - Ata de Defesa

ATA DE DEFESA

Aos 29 dias do mês de agosto de 2017, às 19 horas, na sala C304 do Instituto de Ciências Exatas e Aplicadas, foi realizada a defesa de Monografia pelo aluno **João Paulo de Almeida Castañón,** sendo a Comissão Examinadora constituída pelos professores: Prof. Msc. Theo Silva Lins, Prof. Msc. Helen de Cássia Sousa da Costa Lima e Prof. Dr. Vinicius Soares Fernandes Mota.

O candidato apresentou a monografia intitulada: *"Análise de Popularidade de Canais do YouTube"*. A comissão examinadora deliberou, por unanimidade, pela aprovação do candidato, com nota 6,0 (seis), concedendo-lhe o prazo de 15 dias para incorporação das alterações sugeridas ao texto final.

Na forma regulamentar, foi lavrada a presente ata que é assinada pelos membros da Comissão Examinadora e pelo graduando.

João Monlevade, 29 de agosto de 2017.

Theo Silva Lins

Prof. Msc. Theo Silva Lins

Professor Orientador

Helen de C. S. Costa Lima

Prof. Msc. Helen de Cássia Sousa da Costa Lima

Professor Convidado

Vinicius F. S. Mota

Prof. Dr. Vinicius Soares Fernandes Mota

Professor Convidado

João Paulo de Almeida Castañón

João Paulo de Almeida Castañón

Graduando

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E APLICADAS
COLEGIADO DO CURSO DE SISTEMAS DE INFORMAÇÃO

ANEXO III – Termo de Responsabilidade

TERMO DE RESPONSABILIDADE

Eu, João Paulo de Almeida Cartonian, declaro que o texto do trabalho de conclusão de curso intitulado "Análise de Popularidade de Canais do YouTube" é de minha inteira responsabilidade e que não há utilização de texto, material fotográfico, código fonte de programa ou qualquer outro material pertencente a terceiros sem as devidas referências ou consentimento dos respectivos autores.

João Monlevade, 29 de Setembro de 2017

Assinatura do aluno

Agradecimentos

Agradeço primeiramente a Deus por fazer parte da minha vida e estar comigo através de pessoas incríveis que me ajudaram a escrever mais um capítulo da minha vida.

Agradeço a minha Mãe por estar presente em minhas madrugadas falando "... Paulo vai dormir..., já chega de estudar meu filho! ", mal sabe ela que ainda continuo dormindo tarde. Ao meu Pai, que desde que eu era criança, sempre dá os mesmos concelhos ".... Não aceite nada de estranhos e cuidado com quem anda! ", sempre preocupado com a família e disposto a nos ajudar. Aos meus irmãos, que por mais difícil que foi, compreendiam as minhas ausências nas reuniões de família por estar estudando para as provas e trabalhos que não tinham fim.

Agradeço a minha família por toda a dedicação e paciência para que eu pudesse ter uma caminhada mais fácil e prazerosa durante esses anos.

Agradeço a República Sparramados por ser a minha segunda família. Foram muitos anos de aprendizados, brigas, concelhos, mudanças e muita luta com a imobiliário (que dificuldade viu). Foi com essa família que aprendi o significado de amizade e que ainda existem pessoas que possamos não apenas chamar, mas levar como amigos por toda a vida.

Agradeço a Kiara, uma filha que adotamos na Sparramados e que sempre estava parada na porta dos quartos esperando atenção e fazendo aquela cara de felicidade que só ela sabia fazer. Temos muitas histórias para contar, e só Deus na causa quando ela fugia, misericórdia.

Agradeço aos professores que sempre estiveram dispostos a ajudar e contribuir para meu melhor aprendizado e em especial aos meus orientadores: Alexande Magno, Helen Costa e Theo Lins, que juntos fazem parte de mais um capítulo da minha vida.

Agradeço a UFOP por ter me proporcionado todas as ferramentas que me permitiram chegar ao final desse ciclo de maneira satisfatória.

RESUMO

Este trabalho apresenta um estudo da base de dados de canais do YouTube, identificando canais com o mesmo comportamento de popularidade. O YouTube é uma plataforma de compartilhamento de vídeos que permite o envio de conteúdo personalizado para o canal do YouTube, listando vídeos com base em escolhas de categorias. A interação com os vídeos no canal do YouTube é realizada por meio de comentários, visualizações, avaliações, compartilhamento e adição vídeos na lista de favoritos. Em nossas análies realizadas por meio da correlação de Pearson e regressão linear, observamos que o número de visualizações que um canal do YouTube recebe, está fortemente relacionado com número de likes adicionados ao vídeo. Diante disso, quanto mais visualizações um canal receber, maior será sua popularidade, com mais probabilidade de ser comentando e avaliado. Estas características foram encontradas em canais de diferentes categorias: Entretenimento e Música.

Palavras-chave: caracterização de canais, mídias sociais online, popularidade

ABSTRACT

This work presents a study of the YouTube channel database, identifying channels with the same popularity behavior. YouTube is a video sharing platform that allows you to submit personalized content to the YouTube channel by listing videos based on category choices. Interacting with videos on the YouTube channel is done through comments, views, ratings, sharing, and adding videos to the list of favorites. In our analyzes using Pearson correlation and linear regression, we found that the number of views a YouTube channel receives are strongly related to the number of likes added to the video. Faced with this, the more views a channel receives, the greater its popularity, the more likely it will be to comment and evaluate. These characteristics were found in channels of different categories: Entertainment and Music.

Keywords: channel characterization, online social media, popularity

LISTA DE FIGURAS

LISTA DE TABELAS

Lista de abreviaturas e siglas

OSN	Oline Social Network
API	Application Programming Interface
ID	Identification
JSON	JavaScript Object Notation
ISP	Internet Service Provider
IP	Internet Protocol
GTK	GIMP Toolkit
Qt	Qt Software
WX	WX Code
GLUT	The OpenGL Utility Toolkit
OS X	Operating System
SPYDER	Scientific PYthon Development EnviRonment
PCC	Pearson Correlation Coefficient
REST	Representational State Transfer
HTTP	Hipertext Transfer Protocol

Sumário

1 Introdução

Redes Sociais Online (Oline Social Network - OSN) têm se tornado extremamente populares. Mais de dois terços da população Online global visita ou participa de Redes Sociais e Blogs. Como comparação, se o Facebook fosse um país, este seria o terceiro país mais populoso do mundo e que, domina a internet. Vários tipos de Redes Sociais Online surgiram, incluindo redes de profissionais (ex.,Linkedin), redes de amigos (ex., MySpace, Facebook), redes para o compartilhamento de conteúdos específicos (ex.,Twitter), diários e Blogs (ex.,LiveJournal), fotos (ex., Instagram) e vídeos (ex., YouTube), entre outras OSNs. [Benevenuto, 2010]

Segundo [Andreas, 2009], as mídias sociais onlines são definidas como um grupo de aplicações baseadas na Internet e construídas sobre as bases ideológicas e tecnológicas da Web 2.0, que permitem não apenas a criação, mas também a troca de conteúdo gerado pelo usuário. Para [Benevenuto, 2010], devemos considerar que a popularidade de todas as redes sociais onlines está associada a funcionalidade do usuário criar e compartilhar conteúdo nesses ambientes. Com tanto conteúdo disseminado através das redes sociais online, estatísticas demonstram que esses conteúdos gerados atingem grandes escalas de publicação. Por exemplo, o YouTube tem mais de um bilhão de usuários, quase um terço dos usuários da Internet e, a cada dia, as pessoas assistem a milhões de horas de vídeos no YouTube e geram bilhões de visualizações. [YouTube, 2017]

[Jussara, 2010] ressalta que o crescimento das redes sociais onlines está se tornando tema central de pesquisas para o estudo de vários assuntos da computação, incluindo sistemas distribuídos, padrões de tráfego na Internet, mineração de dados, sistemas multimídia e interação humano-computador.

Alguns cenários onde é aplicada a análise das OSNs, por exemplo, são as empresas, onde é importante saber como os colaboradores se relacionam e organizam. Na área da Ciência, as OSNs podem auxiliar nas pesquisas de propagações de endemias e até mesmo epidemias, bem como serem utilizadas para compreender ou melhorar as formações de grupos. Também pode ser usada como tática de propaganda e marketing para ajudar a divulgar uma determinada marca ou conceito, podendo servir, também, para o estudo de um público alvo relacionado. Segundo [França, 2012], podemos aproveitar uma das grandes questões do início do século XXI, que aborda questões de séries de atentados terroristas, para pensar na identificação e análise de comportamentos de terroristas.

Na visão de [Castells, 2009], a tendência de pessoas se unirem e formarem grupos é uma característica de qualquer sociedade. Esse comportamento é retratado, nos dias atuais, através do avanço das mídias sociais e comunidades online que unem vários usuários ao

redor do mundo. Por meio das interações nas OSNs, os usuários postam suas opiniões, críticas e até mesmo recomendações, que são lidos, disseminados e comentados, de forma quase instantânea, em diversas plataformas na Web 2.0.

O YouTube é uma plataforma de compartilhamento de vídeos que permite que envie conteúdo personalizado para o canal do YouTube, que lista vídeos com base em escolhas de categorias. O YouTube apresenta tópicos de comentários no canal, vídeos gerenciados através de estatísticas na própria plataforma e um contador que permite acompanhar quem está assistindo os vídeos. Chad Hurley, Steve Chen e Jawed Karim, ex-funcionários do site de comércio on-line PayPal, são fundadores da plataforma YouTube, lançado oficialmente em junho de 2005, [Souza, 2016]. O YouTube permite que qualquer pessoa no mundo inteiro possa visualizar os vídeos compartilhados no canal do YouTube.

O YouTube é o segundo site mais acessado do mundo, com mais de dois bilhões de visitantes únicos a cada semana [Alexa, 2017]. É incomparável como uma plataforma para hospedagem e compartilhamento de conteúdo de vídeo, e pode ser facilmente integrado com outras plataformas de mídia social. Com o YouTube é possível criar conteúdo, ao mesmo tempo que oferece um aspecto social poderoso. Os canais do YouTube também permitem uma grande quantidade de personalização e oferecem oportunidades para consolidar a marca, o produto, o serviço, entre outros, em todas as plataformas. A plataforma de compartilhamento de vídeos fornece uma poderosa ferramenta de análise para todos os usuários, portanto, é fácil acompanhar a quantidade de exibições que você está recebendo, quais vídeos geram a maior propagação e quais países e/ou dados demográficos contribuem mais para sua contagem de visualizações.

1.1 Definição do Problema

Naturalmente, as pessoas desenvolveram maneiras para aumentar a sua visibilidade no YouTube através do aumento da popularidade de seus vídeos. Através da grande popularização dessa mídia, análises de comportamento focam em observar a popularidade de vídeos no Youtube.

Esta análise foca no estudo da popularidade de canais. E pretende-se compreender como cada vídeo contribui para a popularidade do canal, isso é interessante porque relata como provedores de serviço, como Youtube, escolhem seus modelos de interação.

1.2 Identificação dos Objetivos

Neste trabalho, os objetivos propostos se concentraram em caracterizar a popularidade de canais do YouTube e analisar se existem canais com o mesmo comportamento de popularidade.

1.2.1 Objetivo Geral

O objetivo geral do trabalho é compreender fundamentais propriedades da popularidade de canais do YouTube. Um estudo aprofundado desta popularidade será necessária para compreender relações e padrões temporais de todas estas métricas.

1.2.2 Objetivos Específicos

Para atingir o objetivo geral do trabalho os objetivos específicos serão:

- Desenvolver um crawler para obter dados públicos dos canais do YouTube;
- Baixar os dados dos vídeos via API do YouTube e acessar o ID de cada canal;
- Realizar triagem das estatísticas diárias para uma amostragem de vídeos;
- Análise e sumarização dos resultados.

1.3 Estrutura da Monografia

O presente trabalho está organizado da seguinte forma. No Capítulo 2 apresenta os fundamentos teóricos necessários para o desenvolvimento deste trabalho, tais como definição, características, técnicas de coleta de dados para análise e as tecnologias utilizadas. No Capítulo 3 descreve a revisão da literatura, que são trabalhos relacionados à analises de dados em mídia sociais e a popularidade de vídeos no YouTube. O Capítulo 4 descreve as metodologias utilizadas para a criação da nossa base de dados e apresenta a descrição dos dados coletados. O Capítulo 5 apresenta o método proposto para analisar o comportamento da popularidade dos canais da base de dados e, por fim, as análises realizadas. A conclusão e trabalhos futuros presente no Capítulo 6.

2 Conceitos Gerais

Neste capítulo serão descritos alguns conceitos importantes utilizados no trabalho. Esta seção define o que são OSNs e as principais características, as técnicas e coletas de dados em OSNs e para finalizar este Capítulo são mostradas as tecnologias utilizadas na realização deste trabalho.

2.1 Definição e Características das Redes Sociais

O aumento de usuários com smartphones e tablets com acesso à internet permite que as pessoas permaneçam conectadas grande parte do dia, aumentando a quantidade de informações que são disponibilizadas na Internet. Segundo pesquisa realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE, 2013), baseado em dados da Pesquisa Nacional por Amostra de Domicílio (Pnad), revelou o crescimento da importância dos smartphones e tablets na utilização da internet. Considerando o uso somente desses equipamentos, houve um acréscimo de 7,2 milhões no número de pessoas que utilizaram a internet em 2013 [Valor, 2017]. Grande parte do acesso e disponibilização de conteúdo se deve à popularização das mídias sociais. Com tanta popularização e enorme quantidade de contéudo disponível, o termo Rede Social Online ou OSN é uma estrutura que possibilita que pessoas ou comunidades se comuniquem principalmente através de qualquer mídia de comunicação. Alguns autores definem OSNs como:

Para [Wasserman, 1994], "uma rede social é um conjunto de atores que pode possuir relacionamentos uns com os outros."

Para [Ellison, 2007], "Serviço Web que permite indivíduos(1) construir perfis públicos ou semi – públicos dentro de um sistema, (2) articular uma lista de outros usuários com os quais compartilham conexões e (3) visualizar e percorrer suas listas de conexões e outras listas feitas por outros no sistema. A natureza e nomenclatura dessas conexões podem variar de local para local."

Para [Tomaél, 2007], "Um conjunto de pessoas (ou organizações ou outras entidades) conectadas por relacionamentos sociais, motivadas pela amizade, relações de trabalho ou troca de informação."

Com base nestas definições, as redes sociais podem ser relacionadas ao dia a dia das pessoas, como por exemplo: família, trabalho, colegas e amigos. Existem várias redes sociais online disponíveis na Web 2.0, que variam de acordo com seus objetivos primários.

Segundo [Franca, 2014], existem muitas OSNs com diferentes objetivos, onde:

- Colaboração: está relacionada às redes sociais colaborativas, sites em que é importante a interação de diferentes usuários compartilhando informações a fim de atingir um objetivo comum. Como exemplo destacam-se a Wikipedia[1] e Digg[2];

- Comunicação: está relacionada ao fenômeno da conversação entre pessoas e o modo como essa conversa é percebida por seus participantes, que podem participar de forma direta, através da realização de comentários e produção de conteúdo, ou indireta, compartilhando e divulgando conteúdo e, consequentemente, ajudando a promover discussões. Podem-se citar, neste contexto, os blogs e microblogs, as redes sociais online (OSNs) e os fóruns. Como exemplo, é possível mencionar WordPress[3] , Twitter[4], Facebook[5] e GoogleGroups[6], respectivamente.

- Multimídia: refere-se aos componentes audiovisuais que ficam além do texto puro e simples como fotos, vídeos, podcasts e músicas. Alguns exemplos dessas mídias, respectivamente, são Flickr[7], YouTube[8] e Lastfm[9].

- Entretenimento: diz respeito aos conteúdos que geram um mundo virtual favorecendo o desenvolvimento da "gamificação", ou seja, ambientes focados em games online ou ainda atividades que podem ser transformadas em algum tipo de competição, nos quais seus usuários se juntam com o objetivo de jogarem juntos ou compartilharem informações a respeito do tema. Como exemplo podem ser citados o Second Life[10].

Na visão de [Franca, 2014], o rápido crescimento destas mídias sociais estão se tornando um problema de Big Data, onde precisamos tratar um volume grande de dados para que diferentes análises sejam viáveis.

2.2 Técnicas de Coletas de Dados em Redes Sociais

Com a popularização das OSNs, a obtenção de dados em larga escala se tornou possível e inúmeras áreas da computação começaram a realizar coletas de dados. De acordo com [Benevenuto, 2010], as principais OSNs disponibilizam interfaces ou serviços de captura parcial ou total da sua base de dados. Com o aumento no interesse de pesquisas na área de análises de mídias sociais, tornou-se indispensável o conhecimento

1 <https://www.wikipedia.org/>
2 <https://digg.com>
3 <https://br.wordpress.com/>
4 <https://twitter.com/?lang=en>
5 <https://www.facebook.com/>
6 <https://groups.google.com/>
7 <https://www.flickr.com/>
8 <https://youtube.com/>
9 <https://www.last.fm/>
10 <https://secondlife.com//>

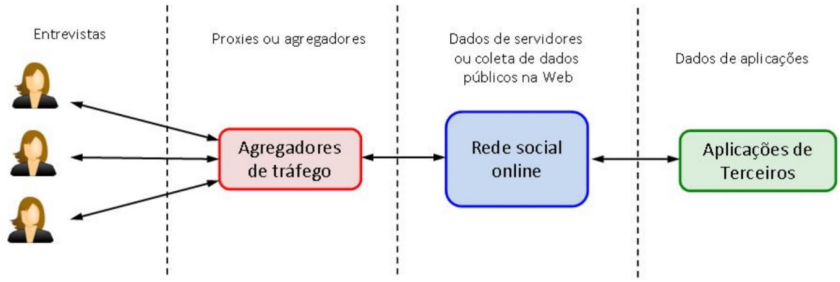

Figura 1 – Possíveis pontos de coleta de dados segundo Benevenuto 2010

das diversas formas de coleta de dados. Segundo [Benevenuto, 2010], existem diferentes áreas de pesquisas e diversas formas de coletas de dados em redes sociais online e cabe ao pesquisador identificar qual o procedimento é mais adequado para o estudo e/ou pesquisa que está sendo realizada. Na Figura 1 são observados os possíveis pontos para a coleta de dados de OSNs, que variam de entrevistas até dados de aplicações de terceiros.

2.2.1 Dados dos Usuários

Segundo [Benevenuto, 2010], uma forma simples de analisar o comportamento de uma rede social online é baseada em entrevistas. Uma técnica simples para coletar dados em mídias digitais, que consiste em obter dados diretos dos usuários por meio de entrevistas, a fim de coletar o comportamento dos usuários dentro das OSNs. O questionário é um dos procedimentos mais utilizados para obter informações, apresentando uma alta confiabilidade. Pode ser elaborado formulários online, garantindo o anonimato e contendo perguntas que atendam as finalidades específicas de uma pesquisa. A entrevista é um método flexível de obtenção da informação que requer um bom planejamento. Ela se sobressai em relação ao questionário, pois relata uma maior quantidade de informações sobre o entrevistado. As entrevistas estruturadas são técnicas mais simples e populares de coleta de dados em OSNs.

2.2.2 Dados de Ponto Intermediário

[Benevenuto, 2010] diz que existem técnicas comuns utilizadas para coletar dados de pontos de agregação de tráfego na rede. A primeira consiste em coletar os dados que passam por um provedor de serviços Internet (ISP) e filtrar as requisições que correspondem a acessos às redes sociais online. A segunda consiste em coletar dados diretamente de uma agregador de redes sociais.

2.2.3 Servidor Proxy

Um Servidor Proxy é um sistema de computador ou uma aplicação que permite máquinas de um rede privada acessar uma rede pública, sem ter a necessidade de uma ligação direta com esta. Um servidor proxy funciona como intermediário entre um navegador da Web (browser) e a Internet. Na visão de [Benevenuto, 2010], os servidores proxys podem ser utilizados com basicamente três objetivos:

1. Melhorar o desempenho na Web armazenando uma cópia das páginas da Web utilizadas com mais frequência;

2. Bloquear o acesso à paginas onde é necessário ter uma lista de endereços ou palavras que devem ser bloqueadas;

3. Compartilhar a conexão com a Internet quando existe apenas um IP (Internet Protocol ou Protocolo de Internet) disponível, assim eles ajudam a melhorar a segurança porque filtram o conteúdo da Web e softwares mal-intencionados. Esse servidores são usados para determinar uma parte na rede, onde computadores estão no local de uma mesma localização geográficas.

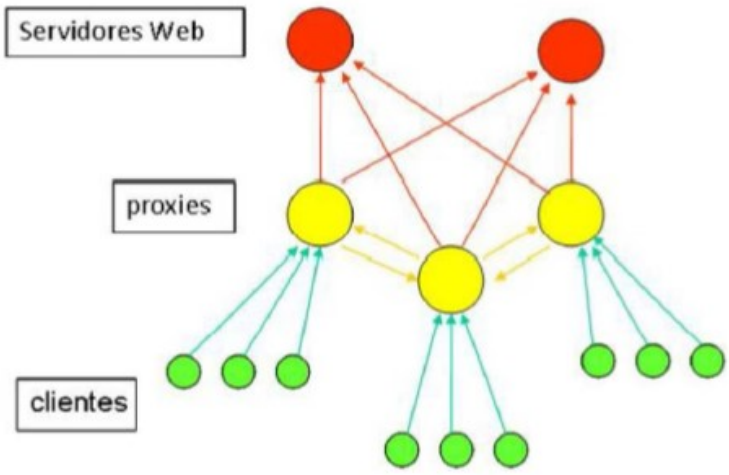

Figura 2 – Exemplo de um servidor proxy intermediando o tráfego entre clientes e servidores - Fonte: Benevenuto (2010)

A Figura 2 ilustra como um servidor proxy funciona para agregar tráfego de seus clientes. Tais servidores são utilizados para delimitar parte da rede, onde esses computadores estão em uma mesma localização geográfica [Benevenuto, 2010].

2.2.4 Agregadores de Rede Social

Agregação de rede sociais são sistemas que possibilita acesso a várias redes sociais simultaneamente, através de um portal único. O processo de coleta de conteúdo de múltiplos serviços de redes sociais são sistemas que permitem acesso a várias redes sociais simultaneamente, através de um portal único com o objetivo de poupar tempo e facilitar a vida de seus utilizadores [Benevenuto, 2010].

Tecnicamente, os agregadores são ativados através de APIs (Application Programming Interface ou em português Interface de Programação de Aplicativos) fornecidas pelas redes sociais. Para a API dar acesso à ações de um usuário que solicita de outra plataforma, o usuário terá que dar permissão para a plataforma de agregação social, especificando ID (Indetificador) de usuário e senha das mídias sociais para ser distribuído, e isto acontece através de uma interface única. Através de uma interface única os usuários podem utilizar diversas funcionalidades de cada rede social na qual se encontra conectado, publicando textos, visualizando imagens e vídeos, verificando as atualizações disponíveis. A Figura 3 descreve o esquema de interação entre os usuários, um sistema agregador e algumas OSNs. Alguns exemplos de agregadores como: Hootsuite é um sistema de gestão de mídia social que faz integração com Twitter, Facebook, Linkedin, Google+, Foursquare, MySpace, WordPress e TrendSpottr Mixi. Flipboard agregador de notícias que integra com o Facebook, Twitter, Tumblr, Instagram e Linkedln. O Cliqset permite mesclar informações de redes sociais e compartilhá-las de forma simples.

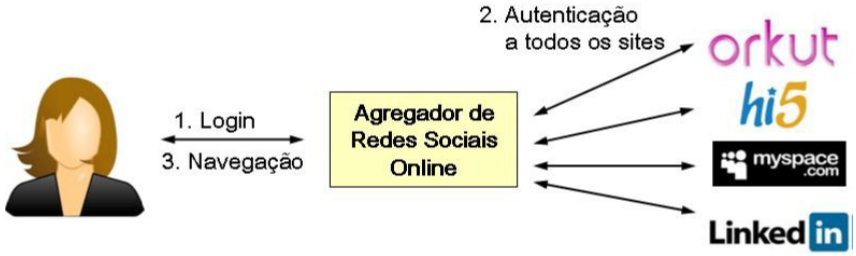

Figura 3 – Ilustra um usuário interagindo a múltiplas redes sociais a partir de um agregador - Fonte:Benevenuto (2010)

2.2.5 Dados de Servidores de Redes Sociais

Para [Benevenuto, 2010] os servidores de OSNs são os locais mais adequados para a coleta de dados, porém esta coleta é muito difícil devido a política de segurança e privacidade que as OSNs oferecem a seus usuários em se obter dados diretamente deste tipo de servidores. Uma estratégia comum consiste em visitar páginas de redes sociais com o uso de uma ferramenta automática, que definimos como crawler ou rôbo, e coletar sistematicamente informações públicas de usuários e objetos. Tipicamente, os elos entre

usuários de uma rede social online podem ser coletados automaticamente, permitindo que os grafos de conexões entre os usuários sejam reconstruídos.

2.2.5.1 Coleta por Amostragem

As mídias sociais online são interpretadas por grafos onde os nodos são os atores (perfis de usuários) e arestas são relacionamentos entre esses atores. É sempre mais interessante coletar o grafo inteiro de uma rede social online para evitar que a coleta seja tendenciosa a um grupo de usuários da rede. [Benevenuto, 2010]. Uma técnica utilizada na coleta é conhecida como *Snowball* (bola de neve), que segue a abordagem da busca em largura. Inicialmente você começa pelo nodo raiz e explora todos os nodos vizinhos, então para cada um desses nodos próximos é explorado os seus vizinhos em diante até que todos os nodos alcançáveis pela busca em largura sejam atingidos [Benevenuto, 2011], como ilustra a Figura 4.

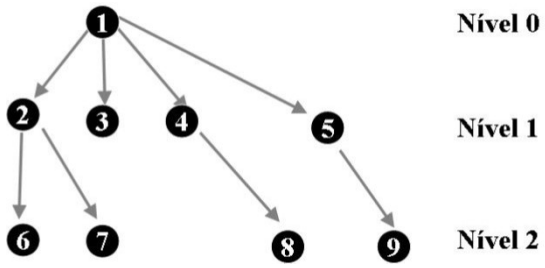

Figura 4 – Busca em largura em redes sociais - Fonte:Benevenuto (2010)

2.2.5.2 Coleta em Larga Escala

Seguindo uma metodologia mais adequada, precisamos coletar o grafo inteiro, o que envolve a coleta de grafos compostos por milhões de nodos e bilhões de arestas [Benevenuto, 2011]. A coleta de grandes bases de dados de OSNs envolve a construção de coletores distribuídos em diversas máquinas, sendo necessário para evitar que os servidores de redes interpretam a coleta como um ataque a seus servidores [Benevenuto, 2010]. Segundo [Benevenuto, 2011], uma estratégia para realizar a coleta seria migrando a lista de usuários para uma máquina mestre, enquanto o rastreamento é distribuído entre máquinas escravas. A máquina escrava solicita a máquina mestre o usuário disponível e retorna os dados deste usuário. Desta forma a máquina mestre mantém a consistência da lista e garante que o usuário seja rastreado apenas uma vez. A Figura 5 ilustra essa estratégia.

Figura 5 – Exemplo de coleta feita de forma distribuída - Fonte:Benevenuto (2010)

2.2.5.3 Coleta por Inspeção de IDs

Para a coleta de uma base de dados de uma OSNs, é importante coletar o grafo inteiro, incluir a rede completa e não somente uma porção dela. A técnica de inspeção por ID é utilizada quando não é possível coletar apenas uma parte da rede, ela é adequada para redes com identificação sequencial de usuários[Santos, 2014]. Com este tipo de identificação é necessário apenas percorrer todos os IDs, sem ter que coletar IDs novos [Benevenuto, 2010].

2.3 Tecnologias Utilizadas

2.3.1 API

A API é uma Interface de Programação de Aplicativos, consistindo em rotinas e padrões de programação que permitem a construção de aplicativos viabilizando uma melhor utilização para os usuários. Ela é composta de diversas funções nativas acessíveis através da comunicação com linguagens de programação, que permite a utilização de recursos menos evidentes do software/sites. A sua funcionalidade é dada através da interligação de códigos com as funcionalidades da API, definindo comportamentos específicos de determinados objetos em uma interface, como por exemplo, busca de dados dos usuários, artigos, informações publicadas pelo site, etc. APIs diferentes estão presentes em navegadores, aplicativos implementados em várias linguagens com diversas finalidades. É um recurso oferecido por algumas OSNs que possibilitam a coleta de dados públicos por meio de aplicativos, oferecendo dados em padrões estruturados para evitar problemas no

processamento de dados.

2.3.1.1 YouTube Data API V3

O YouTube é de propriedade do Google, portanto, para se inscrever em uma conta do YouTube, é preciso de uma conta do Google. Depois de fazer isso, basta acessar o YouTube e fazer login com suas credenciais do Google, que será direcionado pelo processo de configuração do canal. É possível personalizar o tipo de atividade que é compartilhado publicamente, o layout do canal, o tipo de conteúdo que deseja exibir no canal, o título, obra de arte e ícone do canal. Também pode se inscrever, apresentar e comentar os vídeos de outros canais, para que possa facilmente interagir com outras organizações com as quais possa ter interesses ou audiências partilhadas.

A API[11] de dados do YouTube permite a incorporação de funções normalmente executadas no site do YouTube em seu próprio site ou aplicativo. A Tabela 1 identifica os diferentes tipos de recursos que pode recuperar usando a API.

[11] <https://developers.google.com/youtube/v3/getting-started>

Método	**Descrição**
activity	Contém informações sobre uma ação que determinado usuário executou no site do YouTube. Ações do usuário que são informadas em feeds de atividades incluem a classificação de um vídeo, o compartilhamento de um vídeo, a marcação de um vídeo como favorito, a publicação de um boletim do canal etc.
channel	Contém informações sobre um canal simples do YouTube.
channelBanner	Identifica o URL que será usado para definir uma imagem recém-enviada como imagem do banner de um canal.
guideCategory	Identifica uma categoria que o YouTube associa aos canais com base em seu conteúdo ou outros indicadores, como a popularidade. As categorias de guia servem para organizar canais de modo que os usuários do YouTube possam encontrar com mais facilidade o conteúdo que procuram. Embora os canais possam ser associados a uma ou mais categorias de guia, não é certeza que eles estejam em uma delas.
playlist	Representa uma playlist simples do YouTube. Uma playlist é um conjunto de vídeos que podem ser visualizados em sequência e compartilhados com outros usuários.
playlistItem	Identifica um recurso, como um vídeo, que faz parte de uma playlist. O recurso playlistItem também contém detalhes que explicam como o recurso incluso é usado na playlist.
search result	Contém informações sobre um vídeo, um canal ou uma playlist do YouTube que corresponde aos parâmetros de pesquisa especificados em uma solicitação da API. Embora indique um recurso exclusivamente identificável (como um vídeo), um resultado de pesquisa não tem seus próprios dados persistentes.
subscription	Contém informações sobre a inscrição de um usuário do YouTube. Uma assinatura notifica o usuário quando novos vídeos são adicionados a um canal ou quando outro usuário executa uma das várias ações no YouTube, como o upload ou a classificação de um vídeo ou comentários sobre um vídeo.
thumbnail	Identifica imagens em miniatura associadas a um recurso.
video	Representa um vídeo simples do YouTube.
videoCategory	Identifica uma categoria que foi ou pode ser associada a vídeos enviados.

Tabela 1 – Recursos que podem ser recuperados usando a API

A API de dados do Youtube atual é a versão 3.0. Nesta versão, é possível buscar por informações de vídeos e canais com base nos seus IDs[12]. Entretanto, não é possível buscar por usuários que comentaram em um determinado vídeo ou mesmo por comentários de um vídeo. Isso pode ser uma grande limitação quando se pensa em análises utilizando o Youtube. Mesmo assim, dado um id de um determinado vídeo e requisições feitas em determinados intervalos de tempo, é possível conhecer a progressão de visualizações de um vídeo e acompanhar sua divulgação em outra rede social, correlacionando esses dados e estudando a interatividade entre as redes.

Para esta versão de API, todas as operações que funcionam com canais usam IDs de canal exclusivamente como meio de identificar esses canais. O ID de um canal de usuário específico do YouTube é idêntico tanto na v2 quanto na v3 da API, simplificando as migrações entre as versões. Esta dependência completa de IDs de canais pode ser desconcertante para os desenvolvedores que estavam acostumados a transferir nomes de usuários do YouTube para métodos de API, mas a v3 foi projetada para lidar com canais com e sem nomes de usuário legados da mesma maneira, e isso significa usar IDs de canais em qualquer lugar.

Para utilizar a API do YouTube, é necessário a criação de uma aplicação na Google, no site https://console.developers.google.com/. Após a criação, é necessário a ativação da API do Youtube v3.0. Para a consulta de dados de vídeos ou vídeos de um determinado canal, utiliza-se o protocolo HTTPS. Uma URL exemplo para uma consulta é: https://www.googleapis.com/youtube/v3/videos?part=statistics&id=ZKugnwXU5_s&key={Key da sua aplicação}.

Nesse caso, serão retornadas as estatísticas do vídeo cujo ID foi passado como parâmetro, no formato JSON. Mais informações e exemplos de requisições estão disponíveis em https://developers.google.com/youtube/v3/docs/.

2.3.1.2 Implementação da Autenticação OAuth 2.0

O YouTube Data API é compatível com o protocolo OAuth 2.0[13] para autorizar acesso a dados particulares de usuários. Conforme estudo do [OAuth, 2017], a lista abaixo explica alguns dos principais conceitos do OAuth 2.0:

- Na primeira tentativa do usuário em usar a funcionalidade em seu aplicativo que exige que o usuário esteja conectado a um Google Account or YouTube account, seu aplicativo inicia o processo de autorização OAuth 2.0.
- O aplicativo redireciona o usuário ao servidor de autorização do Google. O link para a página especifica o scope do acesso que seu aplicativo está solicitando para a conta

[12] <https://developers.google.com/youtube/v3/guides/working_with_channel_ids>
[13] <https://developers.google.com/youtube/v3/guides/authentication>

do usuário. O scope especifica os recursos que seu aplicativo pode recuperar, inserir, atualizar e excluir ao agir como usuário autenticado.

- Se o usuário consentir em autorizar seu aplicativo para acessar os recursos, o Google retorna um token a seu aplicativo. Dependendo do tipo do aplicativo, ele valida ou token ou troca-o por um tipo de token diferente.

O Google lida com a autenticação e consentimento do usuário e retorna um código de autorização. O aplicativo usa esse código, além de seu client_id e client_secret, para receber um token de acesso, que pode ser usado para autorizar solicitações de API em nome do usuário. Nessa etapa, o aplicativo também pode solicitar um token de atualização para receber um novo token de acesso quando o token de acesso recebido antes expirar. [OAuth, 2017].

2.3.1.3 Usando API KEYS

API Keys[14] é um identificador exclusivo gerado por um console. O uso de uma chave de API não requer uma ação ou consentimento do usuário. As chaves da API não concedem acesso a nenhuma informação da conta e não são usadas para autorização. API Keys são usadas quando o aplicativo está sendo executado em um servidor e acessando um dos seguintes tipos de dados:

- Dados que o proprietário de dados identificou como públicos, como um calendário público ou blog.
- Dados pertencentes a um serviço do Google, como Google Maps ou Google Translate. (Limitações de acesso podem ser aplicadas.)

2.3.2 JSON

JSON[15] (JavaScript Object Notation - Notação de Objetos JavaScript) é uma formatação leve de troca de dados, um modelo para armazenamento e troca de informações no formato de texto, possuindo um aspecto de dados simples de serem interpretados. Está baseado em um subconjunto da linguagem de programação JavaScript, Standard ECMA-262 3a Edição -Dezembro - 1999. Apesar de muito simples, tem sido bastante utilizado por aplicações Web devido a sua capacidade de estruturar informações de uma forma mais compacta, tornando mais rápido a interpretação e gerações dessas informações.

A representação da informação é dada de forma simples, onde para cada valor representado, atribui-se um nome (ou rótulo) que descreve o seu significado. O rótulo é

[14] <https://support.google.com/cloud/answer/6158857>
[15] <http://www.json.org/json-pt.html>

descrito entre aspas duplas, seguido de dois ponto e do valor do atributo. Esses valores podem ser identificados por três tipos: númerico, booleano e string. A partir dos dados básicos podemos construir tipos complexos como objetos e arrays.

Em JSON, os dados são apresentados desta forma:

Um objeto é um conjunto desordenado de pares nome/valor. Um objeto começa com { (chave de abertura) e termina com } (chave de fechamento). Cada nome é seguido por : (dois pontos) e os pares nome/valor são seguidos por , (vírgula). A Figura 6 mostra um exemplo da estrutura em JSON.


```
{
 "videos": [
  {
   "id": "7lCDEYXw3mM",
   "snippet": {
    "publishedAt": "2012-06-20T22:45:24.000Z",
    "channelId": "UC_x5XG1OV2P6uZZ5FSM9Ttw",
    "title": "Google I/O 101: Q&A On Using Google APIs",
    "description": "Antonio Fuentes speaks to us and takes questions on working with Google APIs and
    "thumbnails": {
     "default": {
      "url": "https://i.ytimg.com/vi/7lCDEYXw3mM/default.jpg"
     },
     "medium": {
      "url": "https://i.ytimg.com/vi/7lCDEYXw3mM/mqdefault.jpg"
     },
     "high": {
      "url": "https://i.ytimg.com/vi/7lCDEYXw3mM/hqdefault.jpg"
     }
    },
    "categoryId": "28"
   },
   "statistics": {
    "viewCount": "3057",
    "likeCount": "25",
    "dislikeCount": "0",
    "favoriteCount": "17",
    "commentCount": "12"
   }
  }
 ]
```


Figura 6 – Exemplo de estrutura em JSON - Fonte: Google Developers V3

2.3.3 Python

Pyhton[16] é uma linguagem de programação orientada a objeto interativo interpretado. Em outras palavras, ela têm o propósito de produzir código fácil de ser interpretado, fornecendo estrutura de dados de alto nível, tais como tuplas, lista e dicionários(ou matrizes associativas), classes, exceções, gerenciamento automático de memória, entre outros. Podemos destacar características como:

- baixo uso de caracteres especiais, o que torna a linguagem muito parecida com pseudo-código executável;

[16] <http://www.python.org>

- a biblioteca padrão possui uma grande variedade de extensões adicionais para todo tipo de aplicação;
- o uso de identação para melhor legibilidade do código;
- quase nenhum uso de palavras-chave voltadas para a compilação;
- coletor de lixo para gerenciar automaticamente o uso da memória;
- etc.

Python foi criada por Guido Van Rossum em 1990. Ela suporta múltiplos paradigmas de programação, possibilitando fazer muitas coisas com poucas linha de código. A biblioteca padrão é imensa, incluindo módulos para processamento de texto e expressões regulares, protocolos de rede, acesso aos serviços do sistema operacional, acesso a banco de dados, criptografia, interface gráfica etc. Ela também é uma linguagem livre e multiplataforma, ou seja, os programas desenvolvidos em uma determinada plataforma podem ser executados em outras plataformas sem nenhuma modificação.

A página oficial oferece o acesso a pacotes de instalação, documentação, comunidades, artigos, etc.

2.3.4 IPython

O IPython[17] é um ambiente computacional interativo, no qual você pode combinar execução de código, texto, matemática, gráficos entre outros.

Uma das características mais úteis do Python é o seu intérprete interativo. Ele permite testes muito rápidos sem a sobrecarga de criar arquivos de teste, como é típico na maioria das linguagens de programação. No entanto, o intérprete fornecido com a distribuição Python padrão é um pouco limitado para uso interativo estendido.

O objetivo do IPython é criar um ambiente abrangente para computação interativa e exploratória. Para suportar esta meta, o IPython tem três componentes principais:

- Um shell interativo aprimorado do Python;
- Um modelo de comunicação desacoplado de dois processos, que permite que vários clientes se conectem a um kernel de computação, principalmente o notebook baseado na Web;
- Uma arquitetura para computação paralela interativa.

O shell interativo do IPython (ipython) tem os seguintes objetivos, entre outros:

[17] <http://ipython.org/ipython-doc/3/overview.html>

1. Forneçer um shell interativo superior ao padrão do Python. O IPython tem muitos recursos para preenchimento de tabulações, introspecção de objetos, acesso a shell de sistema, recuperação de histórico de comandos em sessões e seu próprio sistema de comandos especiais para adicionar funcionalidade ao trabalhar interativamente. Ele tenta ser um ambiente muito eficiente para o desenvolvimento de código Python e para a exploração de problemas usando objetos Python (em situações como a análise de dados).

2. Servir como um intérprete embutido, pronto para usar para seus próprios programas. Um shell interativo IPython pode ser iniciado com uma única chamada de dentro de outro programa, proporcionando acesso ao namespace atual. Isso pode ser muito útil tanto para fins de depuração quanto para situações em que uma combinação de processamento em lote e exploração interativa é necessária.

3. Oferece uma estrutura flexível que pode ser usada como o ambiente base para trabalhar com outros sistemas, com Python como a linguagem de ponte subjacente. Especificamente ambientes científicos como Mathematica, IDL e Matlab inspirou seu design, mas ideias semelhantes podem ser úteis em muitos campos.

4. Permitir teste interativo de ferramentas de ferramentas gráficas encadeadas. O IPython tem suporte para controle interativo e não-bloqueante de aplicativos GTK, Qt, WX, GLUT e OS X através de bandeiras de threading especiais. O shell Python normal só pode fazer isso para aplicativos Tkinter.

2.3.5 SPYDER

O SPYDER [18](Scientific PYthon Development EnviRonment) é um ambiente de desenvolvimento interativo para a linguagem Python com recursos avançados de edição, testes, depuração, focado em computação científica. O SPYDER contém um editor multilingaguem, um console Python interativo e um visualizador de documentação, além de gerenciamento de variáveis e de arquivos.

2.3.5.1 NumPy

NumPy [19] é um pacote fundamental para computação científica com Python. Para além das suas utilizações científicas, NumPy também pode ser utilizado eficientimente como um recipiente multi-dimensional de dados genéricos. Isso permite que o NumPy integre, de forma transparente e rápida, com uma ampla variedade de bancos de dados.

[18] <https://pythonhosted.org/spyder/>
[19] <http://www.numpy.org/>

2.3.5.2 SciPy

A biblioteca SciPy[20] é um dos pacotes básicos que compõem a pilha SciPy. Ele fornece muitas rotinas de fácil utilização e eficiênica em trabalhos numéricos, tais como rotinas para integração numérica e otimização.

2.3.5.3 Pandas

Pandas[21] é um pacote Python que fornece estruturas de dados rápidas, flexíveis e expressivas projetadas para tornar o trabalho com dados "relacionais"ou "rotulados", fácil e intuitivo. Ele tem como objetivo ser o bloco de construção fundamental de alto nível para fazer análises gráficas de dados práticos, reais em Python. Além disso, tem o objetivo mais amplo de se tornar a mais poderosa e flexível ferramenta de análise / manipulação de dados de código aberto disponível em qualquer idioma. Ele já está bem no seu caminho para este objetivo. Pandas é bem adequado para muitos tipos diferentes de dados:

- Dados tabulares com colunas de tipo heterogêneo, como em uma tabela SQL ou planilha do Excel
- Dados de séries temporais ordenados e não ordenados (não necessariamente de frequência fixa).
- Dados de matriz arbitrária (homogeneamente tipados ou heterogêneos) com rótulos de linha e coluna
- Qualquer outra forma de conjuntos de dados observacionais / estatísticos. Os dados realmente não precisam ser rotulados para serem colocados em uma estrutura de dados pandas

[20] <http://www.scipy.org/scipylib/index.html>
[21] <http://pandas.pydata.org/pandas-docs/stable/>

3 Revisão Preliminar da Literatura

Neste capítulo serão descritos alguns estudos importantes utilizados no trabalho.

Os trabalhos encontrados na literatura têm tratado das análises dos conteúdos gerados pelo YouTube. De modo geral estes trabalhos visão avaliar os padrões de crescimento da popularidade do YouTube. Observamos que o YouTube é composto de métricas que contribuem para o crescimento da sua popularidade. Os trabalhos da literatura nortearam nossas análises do comportamento das métricas de popularidade, onde consideramos a métrica de visualização como a métrica básica de popularidade dos vídeos.

Esta seção apresenta 3 estudos sobre a popularidade dos vídeos no YouTube e as métricas utilizadas.

[Chatzopoulou, 2010] analisou os principais fatores que tornam o YouTube a maneira fundamental para os usuários promover a si mesmo, produtos ou serviços. Naturalmente as pessoas e/ou empresas tem desenvolvido diversas formas para ganhar visibilidade no YouTube através da popularidade de seus vídeos.

O estudo procurou entender as principais propriedades da popularidade dos vídeos no YouTube. Baseado neste estudo, diferentes aspectos da popularidade foram levantados com a finalidade de entender os padrões temporais e de relacionamento de todos os vídeos. Mediante o uso de cinco métricas de popularidades (número de visualizações, número de comentários, número de classificações, média das classificações, número de favoritos) que, embora, reflita no grau de popularidade de cada vídeo, é o número de visualizações que é amplamente considerado como uma métrica básica para a popularidade dos vídeos.

Através da coleta dos dados por meio de um crawler com as métricas já definidas, o trabalho procurou estudar a relação do número de visualizações de cada vídeo com as demais métricas por meio de um modelo de regressão linear. Através do estudo das principais propriedades da popularidade de vídeo do YouTube, foi descoberto que quatro deles estão altamente relacionados com o número de visualizações, com o número de comentários, com as classificações dos vídeos e com os favoritos.Também foi analisada as propriedades temporais da popularidade dos vídeos. Nesta etapa foi mostrada como as métricas de popularidade evolui diariamente e semanalmente.

Foi observando as tendências sob um longo tempo para entender como as métricas de popularidade aumentava e diminuía ao logo dos meses ou mesmo anos. Para os estudos diários e semanais dos padrões de popularidade, utilizaram as medias de uploads dos vídeos e as medias de comentários sob o regime de 24 horas por dia, durante 7 dias na semana. Neste acompanhamento foi feito um estudo mais aprofundado que mostrou alguns

padrões de comportamento durante as semanas. Dentre esses comportamentos foi notado que vídeos orientado ao entretenimento recebem bastante visualizações durante os finais de semanas enquanto vídeos orientado a educação, como fazer e notícias seguem como mais acessados durante os dias das semanas.

Dado que o YouTube é globalmente acessível, fica difícil realizar conclusões sobre o tempo e sobre a localização dos usuários. Baseado neste estudo, várias direções foram identificadas no sentido de caracterizar os vídeos mais populares. Nestas direções, foi descoberto que o comportamento dos usuários influencia diretamente na popularidade dos vídeos. As análises realizadas nas evoluções das métricas de popularidade mostraram, durante o tempo, quais seriam as métricas a seguir.

[Wattenhofer, 2012] levantou três perguntas para enfatizar um ponto de vista gráfico a respeito da plataforma YouTube: 1- O que pode ser considerado através da topologia completa dessa rede social? Como pode ser comparada com outras redes sociais? 2- Como os usuários se conectam e interagem entre si? Qual a relação entre os gráficos sociais explícitos e implícitos que descrevem a interação entre as publicações na plataforma e os comentários? 3- Como se constrói a popularidade no YouTube? Qual a correlação entre a popularidade topológica de um usuário e a popularidade do conteúdo?

Por meio da análise estatística e ajuste dos dados amostrais da Distribuição Normal de probabilidade formada pelos dados, obtidos em 2011 e manipulados de forma confidencial, encontrou 50% de usuários extremamente populares, com muitos inscritos, enquanto os outros 50% dos usuários tem um inscrito ou comentário. Relacionando os graus de inscritos com o de comentários, encontra-se expoentes de escala 1,55 e 1,44, que se diferem da maioria das redes sociais, que foram medidos entre 2 e 3.

As análises mostram como o Youtube é uma rede social que se diferencia das demais. Os autores observaram que o Youtube é uma rede social de importante plataforma de divulgação de conteúdo e que sua popularidade está mais relacionada com a máxima popularidade que um conteúdo pode atingir. Através da manipulação dos dados utilizados, entende-se que a popularidade do YouTube está na quantidade de inscritos. Tipicamente, os usuários aumentam o número de uploads à medida que ganham mais inscritos, se tornando assim, mais populares. Mas também há o fato que alguns usuários utilizam o YouTube para difundir conteúdo independentemente de serem populares – eles não têm pessoas inscritas. Entretanto, esse comportamento pode ser objeto de estudo de próximas pesquisas.

O que pode ser observado é que as análises da base de dados permitem entender que a popularidade no YouTube se constitui não só apenas da quantidade de inscritos, mas existem reciprocidade entre os inscritos dignos de atenção. Descobriu-se que a natureza orientada por conteúdo do YouTube diferencia-se das redes sociais tradicionais em termos de comportamentos de interação do usuário. Por meio das análises de inscritos e de

comentários, encontrou-se uma dicotomia de atividades "sociais"e "conteúdo"dentro do mesmo sistema. Do ponto de vista da popularidade, os vídeos mais visualizados recebem mais popularidade.

[Filipa, 2010] apresentou um paralelo dos vídeos mais vistos no youtube com suas categorias, tempo de duração, número de visualizações, comentários relacionados, curtidas e os canais. O estudo trata esses fatores como características importantes que influenciam diretamente na popularidade de um vídeo no canal do YouTube.

A internet permitiu que os utilizadores começassem a participar da disseminação do conteúdo online, deixando de ser um mero receptor e assumindo, também, o papel de emissor. O conteúdo online passou, assim, a ser publicado, partilhado, discutido e votado pelos utilizadores.

A Web evoluiu tanto, que hoje em dia é composta majoritariamente pelas redes sociais, blogues, plataformas de compartilhamento de conteúdo, entre outros. A difusão da internet entre a população trouxe consigo um notável aumento de serviços de streaming de vídeo, como é o caso do YouTube. O crescente sucesso do vídeo on-line está relacionado com o fato de ser possível um upload rápido de muita informação. Neste sentido, os padrões de acesso desde a publicação do conteúdo indicam a popularidade que este poderá alcançar a longo prazo. Um fator de elevada importância na popularidade de um canal do YouTube são os comentários que permitem grupos de discussões por utilizadores de gostos semelhantes.

Uma das principais características do vídeo on-line realçadas por [Filipa, 2010], é o fato de muitos dos vídeos publicados serem criados pelos utilizadores. O Youtube é uma plataforma on-line de publicação e divulgação de vídeos criada em 2005. As redes sociais são importantes numa primeira fase de divulgação do conteúdo, quando ele ainda é restrito a um pequeno número de utilizadores. Quanto mais popularidade o vídeo alcançar logo após a sua publicação, maior é a probabilidade do vídeo manter-se ou tornar-se popular no futuro. Segundo dados de 2010, a medida em que o vídeo obtém mais visualizações é, em média, na primeira semana após ter sido publicado.

Os estudos sobre os comportamentos dos vídeos mais vistos no canal do YouTube apresentaram resultados que concluem que as principais características que influenciam a popularidade de um vídeo são a sua categoria, duração, número de visualizações, tags escolhidas quando publicado e, por fim, o utilizador que o publica, que tem um papel fundamental no nível de popularidade que ele poderá atingir.

Por meio dos trabalhos relacionados, observamos que segundo [Chatzopoulou, 2010]

4 Coleta de Dados

Neste capítulo será descrito as metodologias utilizadas para criação da base de dados. Na seção 4.1 serão apresentas as técnicas utilizadas para a extração dos dados do Youtube, descrevendo a aplicação desenvolvida. A seção 4.2 descreve a base de dados coletada. Na seção 4.3 descreve os atributos de popularidade de um canal.

Este trabalho tem o objetivo de analisar se existem canais com o mesmo comportamento de popularidade.

4.1 Arquitetura proposta para a Coleta de Dados

Com a popularidade das OSNs, aumentou o interesse em pesquisas para o entendimento sobre os padrões e comportamento dos usuários e conteúdos disponibilizados pelas OSNs. Sendo assim, a obtenção de dados das OSNs demandam diferentes tipos de coletas, que variam de acordo com o interesse do tipo de dados a ser coletado. Para analisar o comportamento dos canais do YouTube, foi necessário primeiro criar uma base de dados através de informações dos canais do YouTube e de cada vídeo que compõe o canal, que estejam disponíveis publicamente e que pudessem ser coletadas através da API[1] do YouTube. Essa API utiliza o protocolo OAuth 2.0, que fornece uma forma padronizada de acessar os dados protegidos. Ele proporciona autorização específica para várias aplicações. O aplicativo desenvolvido seguiu a política e os termos de uso de dados da API do Youtube, garantindo a privacidade dos usuários e respeitando a conduta deste estudo.

Antes de utilizar a API, é necessário ter uma conta cadastrada para acessar a página de desenvolvedores. Acessando a página de desenvolvedores, é possível registrar, autenticar-se com a conta do Google e cadastrar a aplicação em desenvolvimento. Após o cadastro da aplicação, é necessário gerar atributos importantes: *Key, Client Secret e Acess Token*, que devem ser sigilosos por motivos de segurança. A autenticação é um processo importante, pois é através dessas chaves que serão realizadas a captura dos dados. O *Access Token* é exclusivo do usuário que está desenvolvendo e deve ser armazenado de forma segura, mas nem sempre é necessário obtê-lo, visto que o YouTube permite fazer algumas requisições não autenticadas. Para obter informações de um usuário privado é necessário criar um aplicativo que requisite a autorização do usuário registrando, ou seja, o seu *Access Token* privado.

Para criar a base de dados proposta, foi coletado dados das métricas proveniente dos vídeos que estão relacionadas com os canais mais populares no Youtube, com base

[1] <https://developers.google.com/youtube/>

em uma lista disponível na página da Socialblade[2]. Utilizou-se esta lista para dar inicio a coleta. Para este trabalho selecionamos 3 categorias diferentes: Música; Entretenimento e Comédia. Para cada categoria selecionamos 10 canais do YouTube. Para cada categoria selecionados os 5 primeios canais com maior número de inscritos e outros 5 canais com maior número de visualizações, totalizando 30 canais selecionados para este trabalho. De acordo com [Chatzopoulou, 2010] as categorias que geram mais insights para análise de dados são as categorias de Música, Entretenimento e Comédia. Na seção 4.1.2 será descrito como foi desenvolvido o crawler utilizado na coleta de dados.

4.1.1 Uso de Quotas

Esta API possui algumas limitações por quotas[3]. O número de quotas é para garantir que os desenvolvedores usem o serviço conforme pretendido e não criem aplicativos que reduzam injustamente a qualidade do serviço ou limitam o acesso para os outros desenvolvedores. Porém esta limitação não foi um obstáculo para este estudo. Visto que, não foram necessárias utilizar da capacidade total de requisições que a API oferece. [Quotas, 2017]

Obervamos que Google calcula o uso da cota atribuindo um custo a cada solicitação, mas o custo não é o mesmo para cada solicitação. Segundo o estudo do uso da API V3 do YouTube, [Quotas, 2017], dois fatores principais influenciam o custo da cota de uma solicitação:

1. Diferentes tipos de operações têm diferentes custos de quotas.
   - A operação de leitura simples que recupera apenas o ID de cada recurso devolvido tem um custo de aproximadamente de 1 unidade.
   - A operação de gravação tem um custo de aproximadamente 50 unidades.
   - O envio de um vídeo tem um custo de aproximadamente 1600 unidades.
2. Dependendo de quantas partes do recurso são recuperadas por cada solicitação, as operações de leitura e gravação usa várias quantidades de quotas diferentes. As operações insert e update gravam dados e também devolvem um recurso. Por exemplo, a inclusão de uma playlist tem um custo de quota de 50 unidades para a operação de gravação, além do custo do recurso devolvido da playlist.

   Uma solicitação da API que devolve dados de recursos deve especificar as partes dos recursos recuperadas pela solicitação. Sendo assim, cada parte adiciona cerca de 2 unidades ao custo de quota da solicitação. Portanto, uma solicitação videos.list que só recupera a parte snippet de cada vídeo pode ter um custo de 3 unidades. No

[2] <https://socialblade.com/youtube/>
[3] <https://developers.google.com/youtube/v3/getting-started#quota>

entanto, uma solicitação videos.list que recupera todas as partes de cada recurso pode ter um custo de aproximadamente 21 unidades de quota.

Considerando essas regras, o [Quotas, 2017] diz que o número de solicitações de leitura, gravação ou envio que o aplicativo pode enviar por dia sem exceder sua quota devem ser levadas em consideração. Por exemplo, para uma quota diária de 5.000.000 unidades, o aplicativo poderá ter qualquer um dos seguintes limites aproximados:

- 1.000.000 operações de leitura, sendo que cada uma recupera duas partes de recursos.
- 50.000 operações de gravação e 450.000 operações de leitura adicionais, sendo que cada uma recupera duas partes de recursos.
- 2.000 envios de vídeo, 7.000 operações de gravação e 200.000 operações de leitura, sendo que cada uma recupera três partes de recursos.

4.1.2 Crawler desenvolvido

O aplicativo foi desenvolvido na linguagem de programação Python versão 2.7.3, utilizando a interface SPYDER versão 3.0 e o console de compilação IPython versão 2.7.13, em conjunto com os recursos da API do YouTube versão 3.0. Utilizou-se os príncipios REST (Representational State Transfer), onde o módulo cliente faz uma requisição HTTP (Hypertext Transfer Protocol), contendo origem, destino e tipo de busca, e o módulo servidor responde com um objeto JSON. O desenvolvimento do crawler foi executado em uma máquina local utilizando a plataforma Linux/Ubuntu 15.10, com processador Intel Core i7 CPU @ 2.20GHz x 8, sistema de 64-bit, 5.7 GB de memória e 771.7 GB de disco.

O aplicativo consistiu de funcionalidades específicas para a coleta de dados. Desta forma, foram executadas nas respectivas ordens as seguintes funcionalidades do crawler: Buscar os dados do vídeo através da propriedade id_video, onde, através do retorno em JSON coletou-se o id_channel por meio da função checa_token; utilizou-se um laço de repetição para coletar as diversas métricas dos vídeos do canal, fazendo requisições da função checa_token para coletar dados da próxima página, que conforme a lista de vídeos coletados, foi necessário passar o token da próxima página; da mesma forma foi utilizada a função durationToSeconds, que conforme a lista de vídeos coletados, foi necessário converter o tempo do vídeo para segundos; as lista de vídeos coletadas foram unificadas em uma arquivo final. O arquivo final apresentou os dados em formatos .CSV, contendo os atributos do vídeo dos canais definidos para a análise.

4.2 Base de dados coletada

Inicialmente foram coletados alguns atributos relevantes para a criação da base de dados proposta, esses atributos contém métricas dos vídeos dos canais do Youtube. Dessa maneira, o aplicativo desenvolvido extraiu 10.225 IDs de vídeos dos 30 canais do YouTube, sendo necessário coletar as IDs dos canais para realizar a busca de informações. Esta coleta foi realizada em dias diferentes devido as limitações de quotas obtidas no desenvolvimento deste trabalho, assim a primeira etapa da coleta foram obtidos os IDs dos vídeos e em uma segunda etapa foram executas as outras funcionalidades do aplicativo para obter os IDs dos canais. Desta forma, como o período de busca das informações foi demorado alguns vídeos deixaram de ser públicos e passou a ser privada. Devido a isto, alguns vídeos não permitiram a coleta das informações, sendo aproveitados então, 8.898 registros de vídeos. Os atributos extraídos inicialmente foram:

- id_vídeo: Define o id do vídeo.
- id_channel: Define o id do canal a qual pertence o vídeo.
- channelTitle: Fornece o título do canal.
- categoria: Retorna a qual categoria pertence o vídeo.
- comentário: Quantidade de comentários gerados no vídeo.
- temp: Fornece o tempo de duração do vídeo.
- dislike: Mostra a quantidade de avaliações “Não Gostei” do vídeo.
- like: Mostra a quantidade de avaliações “Gostei” do vídeo.
- views: Fornece a quantidade de visualizações geradas no vídeo.

A base de dados coletados ainda não apresentava conclusões valiosas capazes de diferenciar o comportamento dos canais do YouTube. Assim, desenvolveu-se uma aplicação em Python para calcular algumas informações importantes, de modo que pudessem ser analisadas. Esses atributos serão apresentados no capítulo 5. Durante as requisições realizadas pelo crawler no momento da extração dos dados de cada vídeo, o crawler recebeu as seguintes respostas às requisições de páginas: "Não disponíveis"e "Vídeos recuperados com sucesso". A Tabela 2 resume pelo número de vídeos para cada tipo de resposta às requisições descritas anteriormente.

Tipo de resposta às requisições	**# de Vídeos**
Não disponíveis	1.327
Vídeos recuperados com sucesso	8.898
TOTAL	10.225

Tabela 2 – Resumo do número de vídeos retornado durante a obtenção da base. Retornamos todos os vídeos disponíveis em cada canal por meio da sua ID.

4.3 Caracterização dos Atributos

Os vídeos do YouTube pertencem a diferentes tipos de categorias, e portanto, apresentam contéudos específicos para determinados públicos alvo. Sendo assim, neste seção será analisado os vários atributos que refletem o comportamento da popularidado dos canais do Youtube.

Esse grupos serão discutidos a seguir.

4.3.1 Atributos dos Canais

Inicialmente foi coletada uma base de dados de canais do Youtube pela API contendo 30 canais que apresentavam: ID do canal, número de views, número de seguidores, número de comentários. A partir dessa base, foram coletados 10.225 vídeos como amostra para a realização da análise de popularidade dos canais.

Em seguida, por meio da API de dados do Youtube, a partir do ID dos 30 canais selecionados, foram obtidos para cada canal as features descritas na Tabela 3. Do total de 30 canais, foram obtidos os dados de 8.898 vídeos, os outros 1.327 vídeos foram inspecionados e foi constatado que estes vídeos foram removidos pelo Youtube por violações de copyright, foram removidos pelos proprietários dos canais, não estavam disponíveis ou são privados.

Nome	**Descrição**	**Tipo**
id_channel	Identificação do canal ou id de canal.	categórica
title	Título do canal.	categórica
published	Data de publicação ou criação do canal.	categórica
# views	Total do número de visualizações ou views do canal.	numérica
# comments	Total do número de comentários do canal.	numérica
# subscriber	Total do número de assinantes do canal.	numérica
# videos	Total do número de vídeos pertencentes ao canal.	numérica

Tabela 3 – Descrição das features dos canais.

4.3.2 Atributos dos Vídeos

A partir dos 30 canais, foram selecionados os 10.225 vídeos dado em ordem decrescente do número de views/popularidade pela API de dados do Youtube. No total, foram obtidos 8.898 vídeos. Destes vídeos, foram obtidos todas as features descritas na Tabela 4.

Nome	**Descrição**	**Tipo**
id_video	Identificação do vídeo ou ID do vídeo.	categórica
title_video	Título do vídeo.	categórica
id_channel	Identificação do canal o qual o vídeo pertence.	categórica
category_id	Identificação da categoria do vídeo.	categórica
# likes	Total do número de likes do vídeo	numérica
# dislikes	Total do número de dislikes do vídeo.	numérica
# favorite	Total do número de usuários que marcaram o vídeo como favorito.	numérica
# views	Total do número de visualizações ou views do vídeo.	numérica

Tabela 4 – Descrição das features dos vídeos.

4.3.3 Importância dos Atributos

Fizemos uma breve introdução do YouTube no início da pesquisa e analisamos vários recursos importantes sobre ele. Depois de assistir a um vídeo, os usuários podem dar feedback de várias maneiras. Eles podem postar um ou mais comentários de texto em um vídeo, avaliar o vídeo marcando como "Gostei", "Não Gostei"ou adicioná-lo ao seu conjunto de vídeos favoritos. O YouTube permite, no máximo, uma vez para classificar um vídeo e adicionar um vídeo ao seu conjunto favorito. Esses feedbacks são medidos por várias métricas de popularidade no YouTube, como contagem de #visualizações, #comentários, #favoritos, #likes e #dislikes.

Um proprietário do vídeo pode bloquear comentários e/ou classificações conforme sua necessidade. Embora os usuários possam visualizar essas métricas na página do vídeo, a API do YouTube fornece aos pesquisadores acesso através de uma interface de programação.

Neste estudo, nos concentramos em quatro métricas fornecidas pelo YouTube, número de visualizações, número de comentários, número de dislikes e o número de likes. Embora todas essas métricas refletem o grau de popularidade de cada vídeo, viewcount é amplamente considerado como a métrica básica de popularidade de vídeo [Chatzopoulou, 2010]. Para cada vídeo, também gravamos sua categoria. No momento da coleta, o YouTube tinha 16 categorias. Observamos que as catergorias Música, Comédia e Entretenimento, representados respectivamente pelos IDs de categoria 10, 23 e 24, são as três maiores categorias, representando 88,06% dos vídeos em nosso conjunto de dados, como mostra na Tabela 5.

Id_Category	**Descrição**	**Frequência/Vídeos**	**Porcentagem**
1	Film & Animation	67	0,75%
10	Music	2046	22,99%
15	Pets & Animals	5	0,06%
17	Sports	8	0,09%
19	Travel & Events	6	0,07%
20	Gaming	395	4,44%
22	People & Blogs	37	0,42%
23	Comedy	2660	29,89%
24	Entertainment	3130	35,18%
25	News & Politics	28	0,31%
26	Howto & Style	56	0,63%
27	Education	63	0,71%
28	Science & Technology	19	0,21%
29	Nonprofits & Activism	1	0,01%
30	Movies	5	0,06%
43	Shows	372	4,18%

Tabela 5 – Descrição e frequência das categorias dos vídeos.

5 Caracterização de Canais

Aproveitamos a API de dados do YouTube para medir as métricas em relação à popularidade de vídeos no YouTube. Sendo assim, neste capítulo será analisado os vários atributos que refletem o comportamento dos canais no YouTube, para identificar as características comuns entre os canais do YouTube. Foram considerados os quatros grupos de atributos: visualização, likes, dislikes e comentários.

Neste capítulo será descrito as metodologias utilizadas para caracterização do comportamento da popularidade entre os canais do YouTube. Na seção 5.1 serão apresentas as técnicas utilizadas para identificar as relações entre as métricas de popularidade. A seção 5.2 descreve as análises dos resultados da base de dados coletadas.

5.1 Métricas de Avaliação

Existe um conjunto de métodos estatísticos que visam estudar a associação entre duas ou mais variáveis aleatórias. Dentre tais métodos, a teoria da regressão e correlação ocupa um lugar de destaque por ser o de uso mais difundido.

Usando nossa base de dados, analisamos as correlações entre diferentes métricas de popularidade. Nesta seção, primeiro analisamos as correlações *Pairwise*(pares) entre as quatro métricas de popularidade. Em seguida, apresentamos nosso modelo de regressão linear. Finalmente, estudamos as razões entre o número de visualizações e o restante das métricas.

5.1.1 PCC - Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson (PCC) ou definido como r de Pearson mede o grau da correlação linear entre duas variáveis quantitativas. Duas variáveis apresentam uma correlação linear quando os pontos se aproximam de uma reta. Essa correlação pode ser positiva (para valores crescentes de X, há uma tendência a valores também crescentes de Y) ou negativa (para valores crescentes de X, a tendência é observarem-se valores decrescentes de Y). As correlações lineares positivas e negativas encontram-se ilustradas na Figura 7. [Larson, 2010].

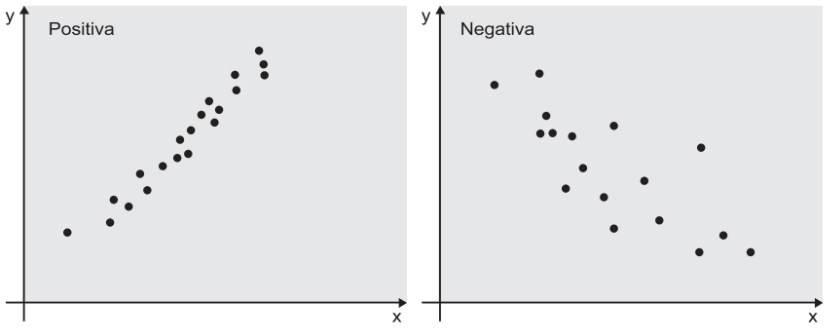

Figura 7 – Correlações Lineares Positivas e Negativas

Segundo [Larson, 2010], o coeficiente de correlação entre duas variáveis em um conjunto de dados é igual a sua covariância dividida pelo produto de seus desvios padrão individuais.

O coeficiente de correlação de uma base de dados é definido pela seguinte fórmula 5.1, onde S_x e S_y são os desvios padrão da amostra, e S_{xy} é a covariância da amostra. [Larson, 2010].

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \tag{5.1}$$

Da mesma forma, o coeficiente de correlação populacional é definido como segue a fórmula do item 5.2, onde α_x e α_y são os desvios padrão da população, e α_{xy} é a covariância da população. [Larson, 2010].

$$\rho_{xy} = \frac{\alpha_{xy}}{\alpha_x \alpha_y} \tag{5.2}$$

Observamos que, quando o coeficiente de correlação estiver próximo de 1, indica que as variáveis estão positivamente relacionadas e o gráfico de dispersão cai quase ao longo de uma linha reta com inclinação positiva. Para -1, mostra que as variáveis estão negativamente relacionadas e o gráfico de dispersão quase cai ao longo de uma linha reta com inclinação negativa. E para zero, indicaria uma relação linear fraca entre as variáveis. [Larson, 2010].

A covariância entre duas variáveis pode ser estimada pela equação 5.3, que representa uma medida do grau e do sinal da correlação onde, S_{xy} é a covariância amostral entre as variáveis X e Y; $\bar{X}$ e $\bar{Y}$ são as médias aritméticas de cada uma das variáveis; n é o tamanho da amostra; X_i e Y_i são as observações simultâneas das variáveis. [Larson, 2010]

$$\frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) \tag{5.3}$$

5.1.2 Regressão Linear

Apenas por meio da visualização do diagrama de dispersão é possível observar a existência de uma relação funcional entre as duas variáveis. Após identificar que a correlação linear entre duas variáveis é significativa, a próxima etapa é determinar a equação da linha que melhor modela os dados. [Larson, 2010].

A equação que descreve uma reta de regressão entre X e Y é dado pela seguinte formúla:

$$Y = \alpha + \beta X + \epsilon \tag{5.4}$$

A fórmula do item 5.4 define que Y é a variável dependente, X é a variável independente, enquanto α e β são os coeficientes do modelo e denota os erros ou resíduos da regressão.

Os coeficientes α e β da reta são calculados através dos dados observados e fornecidos pela amostra, obtendo uma reta estimativa conforme:

$$\hat{y}_i = a + bx_i \tag{5.5}$$

Para a fórmula do item 5.5 do coeficiente da reta, a é definico como a estimativa do coeficiente $\alpha\,(\tilde{\alpha} = a)$; b é a estimativa de $\beta\left(\tilde{\beta} = b\right)$; $\hat{y}_i$ é o valor esperado da variável dependente e x_i é o valor observado para variável independente. [Larson, 2010]

[Larson, 2010] considera a equação de uma reta de regressão permite que use a variável independente X para fazer previsões para a variável dependente Y. Um exemplo de uma reta de regressão está ilustrada na Figura 8.

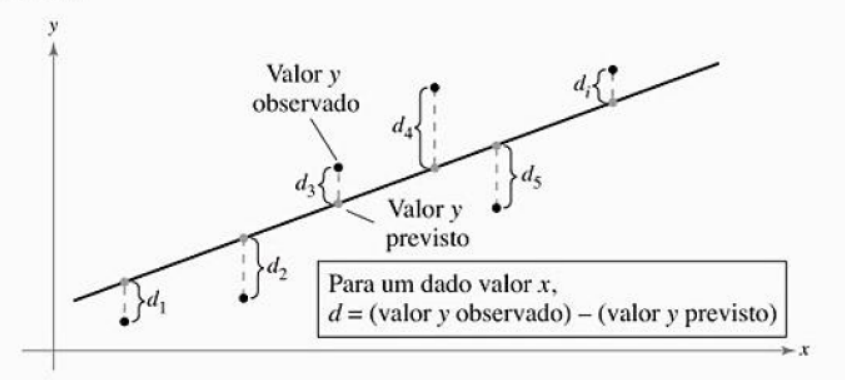

Figura 8 – Exemplo de uma reta de regressão, fonte: Larson (2010)

5.2 Resultados Experimentais

Realizamos testes de correlação de Pearson para as categorias de Música, Entretenimento e Comédia a fim de identificar possíveis relacionamentos entre as variáveis: #viewCount, #likeCount, #dislikeCount e #commentCount. Como vemos na Figura 9a, 9b e 9c, à medida que aumenta o coeficiente de #viewCount aumentam na mesma proporção a frequência de interações com #likeCount, #dislikeCount e #commentCount para a base de todos os vídeos.

Category	**PCC (#viewCount)**	**r-Squared**
Entretenimento	#likeCount	0.977
Música	#likeCount	0.961
Comédia	#likeCount	0.783

Tabela 6 – Correlação forte das categorias agrupadas por viewCount.

Correlações fortes são observadas entre as métricas. Calculamos o Coeficiente de Correlação de Pearson entre a contagem de visualizações, #viewCount, com cada uma das variáveis indicadoras de métricas: #likeCount, #dislikeCount e #commentCount para a base de vídeos extraída. O Coeficiente de Correlação de Pearson é a métrica mais comum para medir a dependência entre duas quantidades [Chatzopoulou, 2010]. Um resumo do cáculo Coeficiente de Correlação de Pearson está incluído na Tabela 6. Como primeiro passo, para identificar as categorias mais influentes, observamos o comportamento da quantidade de visualizações, #viewCoutn, comparado com as métricas #likeCount, #dislikeCount e #commentCount. O valor do Coeficiente de Correlação de Pearson, entre uma escala de 0 a 1, demostrou valores acima r = 0,9 para duas categorias: Entretenimento e Música.

O coeficiente indica quanto da variação total é comum aos elementos que constituem os pares analisados. Evidentemente, quanto mais próximo de 1 for o coeficiente, maior será a validade da regressão.

Para as categorias estudadas, inspecionamos o resultado do coeficiente listado na Tabela 6 e observamos a seguinte situação:

Entretenimento: 1 - 0,977 = 0,023, ou seja, 2% de variância

Música: 1 - 0,961 = 0,039, ou seja, 3% de variância

Comédia: 1 - 0.783 = 0,217, ou seja, 21% de variância

Assim, pode-se dizer que as catergorias de Entretenimento e Música apresentam, respectivamente, apenas 2% e 3% da variância da regressão não depende das variáveis

estudada. Para a categoria de Comédia a variâcia de regressão não dependente da outras variáveis foi de 21%, o que demonstrou baixa correlação entre a métrica #viewCount com as demais métricas. As categorias Entretenimento e Música estão altamente correlacionado com a métrica #likeCount.

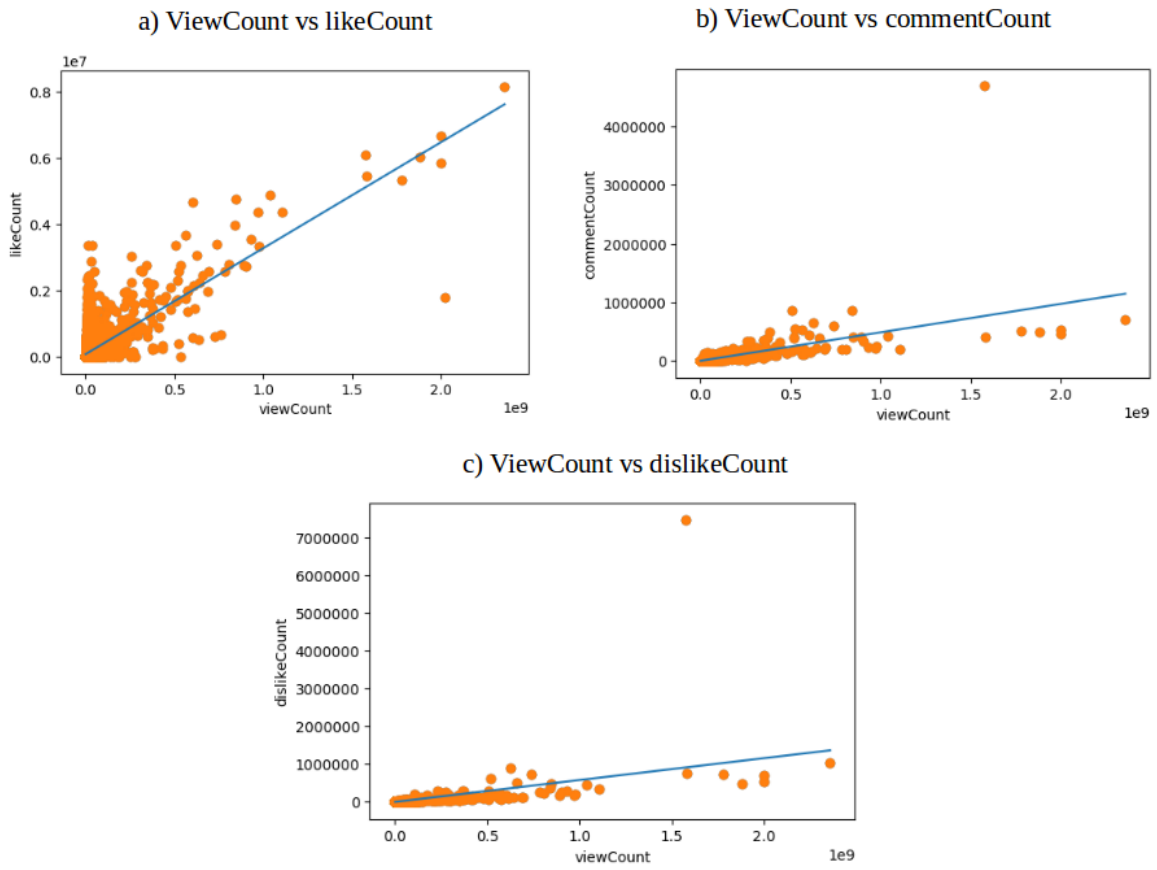

Figura 9 – a) Curva de dispersão e correlação de Pearson entre as variáveis coeficiente de #viewCount e #likeCount. b) Curva de dispersão e correlação de Pearson entre as variáveis coeficiente de #viewCount e #commentCount. c) Curva de dispersão e correlação de Pearson entre as variáveis coeficiente #viewCount e #dislikeCount. As linhas diagonais indicam a tendência da correlação entre as variáveis.

Para visualizar essas correlações, usamos a nossa base de dados de 8.898 vídeos e traçamos os valores das métricas de popularidade em pares, como mostrado na Figura 9. Podemos observar a tendência linear para as correlações entre contagem de visualizações, #viewCount, com as métricas #likeCount, #dislikeCount e #commentCount. A correlação entre contagem de visualização e o número de avaliação como likeCount demonstra uma forte relação.

5.2.1 Coeficiente de Correlação de Pearson e os Canais do YouTube

Os experimentos e a análise dos resultados obtidos são divididos basicamente em três categorias: 1) Entertainment, 2) Music e 3) Comedy. Para as nossas análises, consideramos o coeficiente de correlação com valores aproximados ou acima de r = 0,96, que é um valor próximo de 1.

Portanto, para a base de dados extraída do YouTube, consideramos 96% de depência para o coeficiente de correlação entre as variáveis de métricas da popularidade: #viewCount, #likeCount, #dislikeCount e #commentCount. A Tabela 7 mostra apenas as categorias e os canais que obtiveram o coeficiente de correlação (r-Squared) próximo ou acima de r= 0,96.

Category	**Canal**	**Métrica**	**r-Squared**
Entertainment	Baby Big Mouth	#likeCount	0.975
Entertainment	Baby Big Mouth	#dislikeCount	0.990
Entertainment	Baby Big Mouth	#commentCount	0.949
Entertainment	Get Movies	#likeCount	0.988
Entertainment	Get Movies	#dislikeCount	0.985
Entertainment	Get Movies	#commentCount	0.962
Entertainment	Ryan ToysReview	#likeCount	0.968
Music	KatyPerryVEVO	#likeCount	0.969
Music	JustinBieberVEVO	#likeCount	0.954

Tabela 7 – ViewCount altamente relacionado com as métricas de popularidade.

A Tabela 7 mostra os resultados do coeficiente de correlação entre o atributo #viewCount, como o índice de popularidade, com os outros indicadores, os atributos de #likeCount, #dislikeCount e #commentCount. Para os canais listados, observamos que o resultado do coeficiente de correlação apresenta valores acima de r = 0,9, demonstrando,

assim, forte relação entre estes atributos. Observamos em nossas análies que nenhum valor resultante dos cálculos do coeficiente de correlação apresentou valores acima de 0,9 para a categoria Comedy.

Este resultado pode ser observado pela análise empírica de vários vídeos populares. Não é inacreditável ver um vídeo extremamente popular com um grande número de comentários, likes e favoritos ao mesmo tempo.

Na Figura 10, examinamos o comportamento dos índices de popularidade mediante crescimento do número de visualizações. Surpreendentemente, as categorias Entertainment e Musica são próximos de 0.960, resultado do valor para o coeficiente de correlação, enquanto a categoria Comedy obteve o valor de r = 0,783 para o coeficiente de Pearson. Isso significa que um vídeo tende a receber as reações/respostas dos usuários (um comentário, uma classificação e adicionado à lista favoritos) a cada vez que o número de visualizações aumenta.

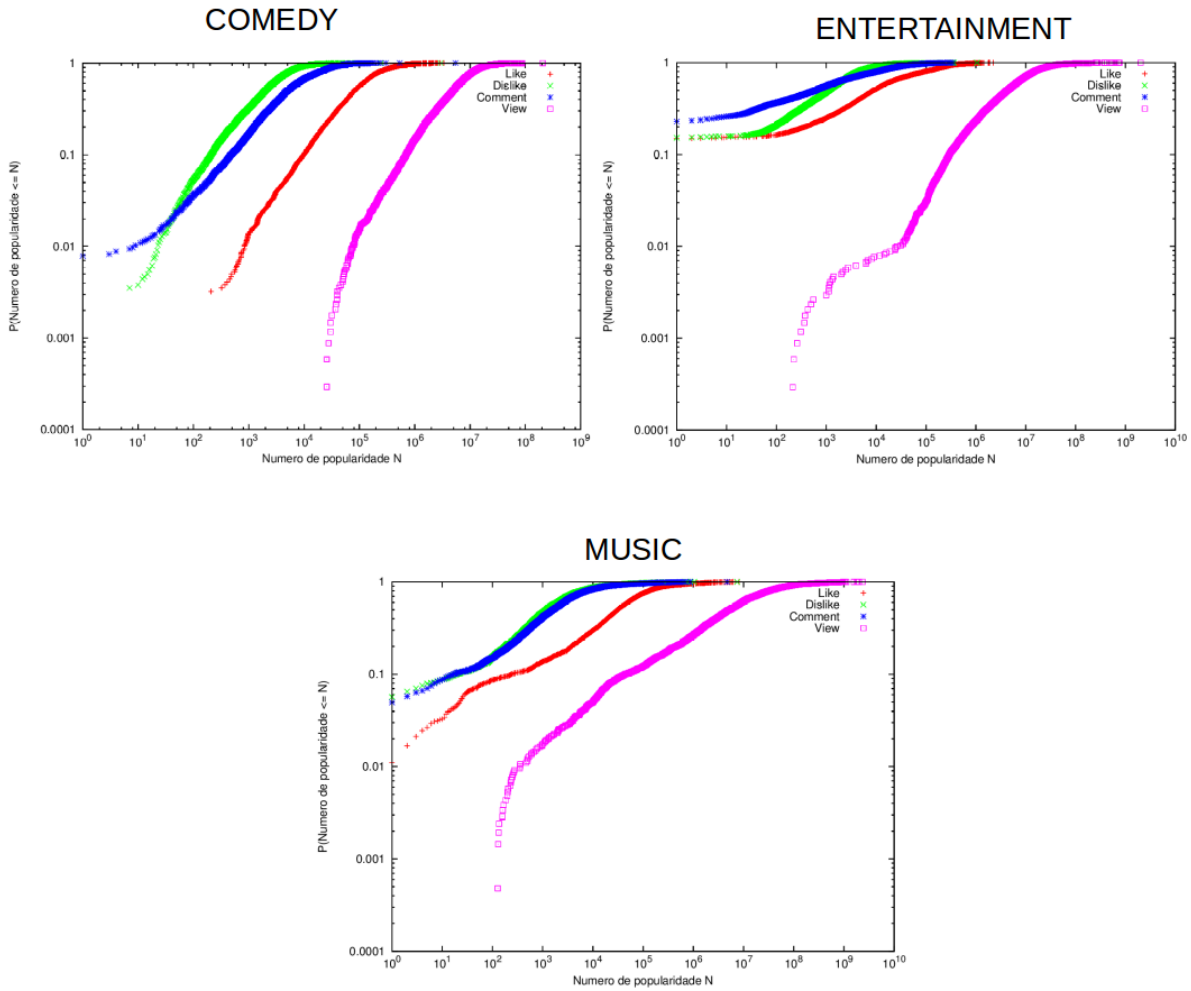

Figura 10 – Valores estatísticos em cada grupo de popularidade para os vídeos em diferentes categorias de um canal.

Podemos observar que o comportamento das visualizações para as categorias Entertainment e Music começam a evoluir após atingir 1000 visualizações. Isso sugere duas coisas interessantes. Primeiro, responder a um vídeo é uma indicação de uma forte reação, já que comentar um vídeo faz mais esforço do que simplesmente vê-lo. Além disso, a resposta exige que o usuário faça o login no YouTube, enquanto que assistir a um vídeo não requer o login. Em segundo lugar, a probabilidade de um usuário fazer um comentário, dar uma classificação e adicionar à lista de favoritos são igualmente prováveis. Observe que não sabemos se as ações são tomadas pelo mesmo usuário, mas pode ser uma direção futura interessante.

Nesta capítulo, queremos compreender qual o desejo de uma "resposta ativa"dos usuários para um canal, especificamente, sob as seguintes métricas:

- A) **#commentCounts**: Definido como o número de comentários de um vídeo.

  Representa o desejo dos usuários de responder ao vídeo deixando um comentário.

- B) **#likeCounts**: Definido como o número de avaliações de um vídeo.

  Expressa o desejo do usuário de se tornar um "fã"do vídeo.

- C) **#viewCounts**: Definido como o número de visualizações de um vídeo.

  Demonstra o desejo do usuário de assistir o vídeo.

Como a evolução dos números de #viewCount é geralmente maior do que outras métricas, contamos ações a cada 1000 visualisações.

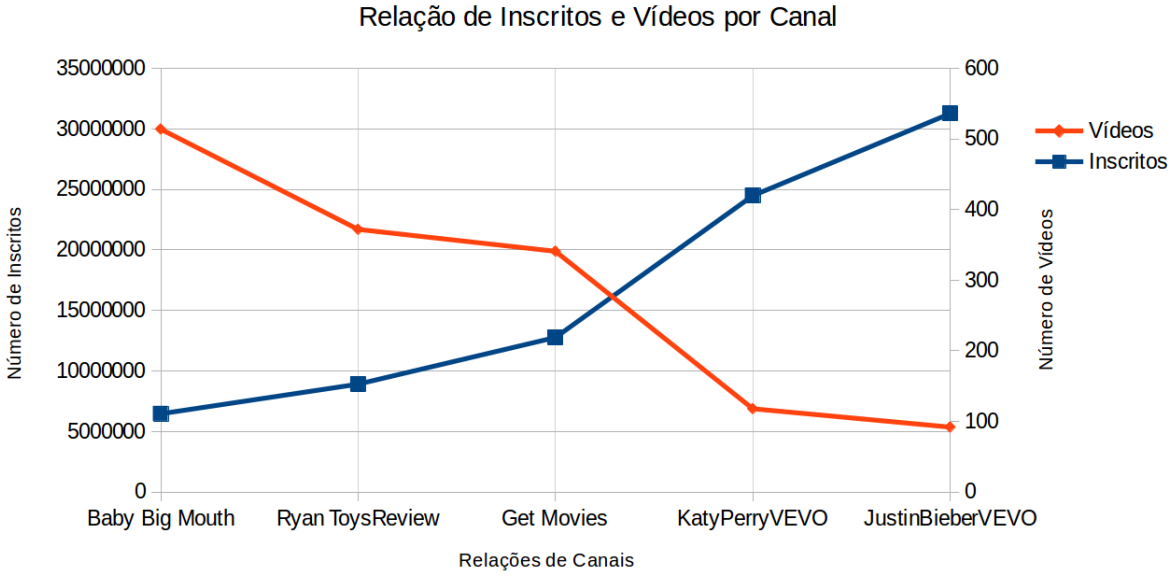

Figura 11 – Distribuição da quantidade de inscritos e vídeos por canal

O número de inscritos e a quantidade de vídeos distribuídos entre os canais se mostrou muito importante, conforme observado na Figura 11. Observamos que a quantidade de vídeos que compõe um canal não está diretamente relacionado com a quantidade de inscritos do mesmo canal. Desta forma, nota-se, pelo gráfico, que o canal Baby Big Mouth, pertencente a categoria de entretenimento, apresenta a menor quantidade de inscritos, mas com o maior número de vídeos disponível em seu canal. Neste sentido, observamos que o canal JustinBieberVEVO, da categoria de Música, possui um comportamento diferente, com o maior número de inscritos, porém com a menor quantidade de vídeos disponíveis em seu canal. Isso mostra que, embora os canais não possuem uma relação direta do número de inscritos com a quantidade de vídeos, ainda apresentam comportamento semelhantes entre as principais métricas de popularidade, visualização, like, dislike e comentários.

Quando comparamos os resultados dessas análises, podemos verificar que a popularidade dos canais é composta pelo conjunto de vídeos disponíveis por ele e que as relações das principais métricas de popularidade estão altamente correlacionadas (#viewCount, #commentCount, #likeCount, #dislikeCount), onde duas categorias diferentes, Entertainment e Music, apresentaram comportamentos semelhantes a medida que as visualizações aumentavam. Assim, aqueles que obtêm mais visualizações tendem a ter mais influência na popularidade dos vídeos, com mais probabilidade de ser comentado, avaliado e até mesmo adicionado na lista de favoritos.

6 Conclusão e Trabalhos Futuros

O trabalho apresentou um método para identificar se existem canais no YouTube que apresentam o mesmo comportamento de popularidade. Através de uma base de dados coletada do YouTube via API, fizemos uma análise do comportamento das métricas de popularidade: #viewCount, #likeCount, #commentCount e #dislikeCount. Utilizando a nossa base de dados sumarizada e organizada nas categorias: Comédia, Entretenimento e Música, fizemos uma caracterização, revelando aspectos comportamentais semelhantes.

Utilizamos a técnica de correlação, denominada Coeficiente de Correlação de Pearson, que foi capaz de apresentar de forma eficaz a correlação entre as métricas de popularidade. Apenas as categorias de Entretenimento e Música mostraram forte correlação entre as métrias de #viewCount e #likeCount. A categoria de Entreteminamento, em uma escala de 0 a 1, apresentou r = 0,977 de forte correlação. A categoria de Música, na mesma escala de 0 a 1 de forte correlação, motrou r = 0,961, equanto a categoria de Comédia apresentou correlação de r = 0,783.

As análises realizadas mostraram alguns padrões de comportamento. Dentre eles pode-se citar os canais orientados às categorias de entretenimento e música, que apresentam uma média de r= 0,970 de forte correlação, o que colabora para aumento dos likes. Diante disso, quanto mais visualizações um canal recebe maior será sua popularidade, com mais probabilidade de ser comentando, avaliado e até mesmo adicionado na lista de favoritos. Estas características foram encontradas em canais de diferentes categorias.

Como trabalhos futuros, pretendemos realizar uma coleta de dados de canais em larga escala e além disso, analisar apenas canais que apresentem suas estatísticas de vídeos disponíveis com intuito de identificar diferentes tipos de categorias que apresentam o mesmo comportamento de popularidade.

Referências

Alexa. *The top 500 sites on the web*, @ONLINE. Disponível em:<http://www.alexa.com/topsites>. Acessado em 4 de maio de 2017. 10

Benevenuto, F. *Redes sociais online: Técnicas de coleta, abordagens de medição e desafios futuros.* Tópicos em Sistemas Colaborativos, Interativs, Multimidi, Web e Banco de Dados, pages 41–70. 2010. 9

Benevenuto, F.; Almeida, J.; Silva, A. *Coleta e análise de grandes bases de dados de redes sociais online.* Jornadas de Atualizaçãoo em Informática (JAI), pages 11–57. 2011. 17

Castells, M. *Communication power.* Oxford, New York: Oxford University Press, (ISBN 978-0-19-956-701-1): pages 571. 2009. 9

Ellison, B; Nicole, B. E. *Social network sites: Definition, history, and scholarship.* Journal of Computer-Mediated Communication, 13(1):210–230. 2007. 12

Filipa, A. J.; Valentim, H. S.; Mário, J. C. *Os vídeos mais vistos no youtube: Uma possível caracterização.* INTERNET LATENT CORPUS JOURNAL. 2010. 29

Franca, T.; Faria, F; Rangel, F. M.; Farias, C. M.; Oliveira, J. *Big social data: Princípios sobre coleta, tratamento e análise de dados sociais.* Anais do SBBD. Tópicos em Gerenciamento de Dados e Informações. 2014. 12, 13

Chatzopoulou, G.; Sheng, G.; Faloutsos, G. *A first step towards understanding popularity in youtube.* INFOCOM IEEE Conference on Computer Communications Workshops. 2010. 27, 29, 31, 36, 41

Jussara, A.; Gonçalves, M.; Benevenuto, F.; Pereira, A.; Rodrigues, T.; Almeida, V. *Characterization and analysis of user profiles in online video sharing systems.* Journal of Information and Data Management, 1(2):261. 2010. 9

Kaplan, A. M.; Haenlein, M. *Users of the world, unite! The challenges and opportunities of Social Media.* volume 53. ESCP Europe, 2009. 9

Larson, F. *Estatística Aplicada.* 4º Edição. *Capítulo 09: Correlação e Regressão.* pages 412-438. 2010. 38, 39, 40

Wattenhofer, M.; Wattenhofer, R.; Zhu, Z. (2012). *The youtube social network.* Sixth International AAAI Conference on Weblogs and Social Media (ICWSM 2012). 2012. 28

OAuth, YouTube. Conseguir credenciais de autorização, @Online. Disponível em: <https://developers.google.com/youtube/registering_an_application>. Acessado em 1 de ago. de 2017. 21, 22

Quotas, Youtube. *Uso de Quotas*. Disponível em: https://developers.google.com/youtube/v3/getting-started#quota>. Acessado em 1 de ago. de 2017. 31, 32

Santos, D. C. *Coleta automatizada e análise de dados em fanpages do facebook*. UFPR, 2014. 18

Souza, E. J. M. *Narrativas pessoais na internet: seriam os youtubers um novo modelo de narrador?*. Intercom – Sociedade Brasileira de Estudos Interdisciplinares da Comunicação. 2016. 10

Tomaél, M. I. *Redes sociais, conhecimento e inovação localizada. Informação & Informação*. 12(1esp). 2007. 12

Valor, IBGE. *IBGE: Mais de 50% usam celular e tablete para acessar a internet*. @ONLINE.<http://www.valor.com.br/ brasil/4027294/ibge-mais-de-50-usam-celular-e-tablet-para-acessar-internet>. Acessado em 4 de abril. de 2017. 12

Wasserman, S. *Social network analysis: Methods and applications*. volume 8. Cambridge university press. 1994. 12

YouTube. *Encontre rapidamente as estatísticas e os vídeos que você está procurando*. @ONLINE. <https://www.youtube.com/yt/press/pt-BR/ statistics.html>. Acessado em 23 de julh0 de 2017. 9