Lucas Machado de Oliveira

Caracterização de Conteúdo de Rádio Transmitido pela Internet

Belo Horizonte

2017/1

Lucas Machado de Oliveira

Caracterização de Conteúdo de Rádio Transmitido pela Internet

Apresentado como relatório final da disciplina de Monografia em Sistemas de Informação II do Curso de Bacharelado em Sistemas de Informação da UFMG

Orientador:
Olga Nikolaevna Goussevskaia - Departamento de Ciência da Computação

Co-orientador:
Fabrício Benevenuto de Souza - Departamento de Ciência da Computação

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Belo Horizonte

2017/1

Resumo

O aumento na adoção da internet abriu caminho para uma nova onde de aplicações e mudou drasticamente o modelo de negócio de diversos serviços. O modelo tradicional de radiodifusão, que utiliza ondas eletromagnéticas para transmissão de conteúdo, também foi afetada. Nos dias de hoje, transmissão de rádio baseada na Internet já é uma realidade.

Uma vantagem da rádio por internet sobre rádio por ondas é a possibilidade de se acessar estações de rádio de outros países, algo que não seria possível devido ao alcance limitado das ondas de rádio. Outra vantagem é o baixo custo da transmissão de conteúdo de rádio pela Internet. Além disso, a coleta e análise de dados de usuário, possibilita a geração de conteúdo personalizado e melhor colocação de propaganda pode ser feita pelas estações.

O principal objetivo deste trabalho consiste na coleta e caracterização deste tipo de serviço de difusão de conteúdo. Para isso, serão coletados dados do SHOUTcast, um serviço gratuito para a transmissão de áudio criado em 2004 pela Nullsoft. Objetiva-se a coleta de uma quantidade relevante de dados e os trabalhos de caracterização serão direcionados em dois aspectos: fatores relacionados a audiência e conteúdo de propaganda, uma vez que o modelo financeiro de estações de rádio depende de propagandas.

Palavras-chaves: radiodifusão, áudio, rádio, internet, propaganda.

Abstract

The increase in Internet adoption opens a new wave of applications and changes the business models of a lot of services around us. The traditional model of radio stations to transmit information through electromagnetic waves was impacted too. Nowadays, internet-based radio streaming is already a reality.

An advantage of internet radio over radio waves is the possibility of accessing a radio station from different countries, something that was not possible before due to the limited reach of electromagnetic waves of radio. Another advantage is that a streaming Internet radio station can be very cheap to set up. Moreover, if user-generated data is collected and analyzed, more personalized content and better advertisement placement can be performed by the radio stations.

The main focus of this project consists on the collection and characterization of this type of content broadcasting service. For this, data will be collected from SHOUTcast's service, which is a free to use audio broadcasting service developed in 2004 by Nullsoft. The aim is to collect a relevant amout of data and the characterizing efforts will be directed in two aspects: factors related to audience and factors related to advertisement content, which directly impacts the financial model of this type of service.

Keywords: broadcasting, audio, radio, internet, advertisement.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O aumento na utilização da Internet possibilitou uma nova onda de aplicações e mudou drasticamente modelos de negócio e diversos serviços usados por nós. O modelo tradicional de estações de rádio, que utiliza de ondas eletromagnéticas para transmissão de informação, também foi afetado.

Com o surgimento da Internet, um novo modelo de transmissão de conteúdo de rádio emergiu (5, 20). Também conhecido como webcasting, transmissão de rádio pela Internet tornou possível não somente que rádios já existentes pudessem transmitir conteúdo em uma nova plataforma, mas também auxiliou no surgimento de inúmeras estações de rádio independentes (25).

Uma vantagem da rádio por internet sobre rádio por ondas é a possibilidade de se acessar estações de rádio de outros países, algo que não seria possível devido ao alcance limitado das ondas de rádio. Outra vantagem é o baixo custo da transmissão de conteúdo de rádio pela Internet. Além disso, a coleta e análise de dados de usuário, possibilita a geração de conteúdo personalizado e melhor colocação de propaganda pode ser feita pelas estações.

Um dos serviços disponíveis na internet para broadcasting de rádio é o SHOUTcast. Esse sistema é um sistema multiplataforma de transmissão de áudio pela internet criado pela Nullsoft em 1998. Ele utiliza os formatos MP3 e AAC para codificação de áudio e o protocolo HTTP para transmissão. É um serviço gratuito e os softwares para cliente e servidor estão disponíveis livre de qualquer cobrança e para múltiplas plataformas, como Windows, Linux, Mac OSX e Solaris. Em Maio de 2014, o SHOUTcast já somava mais de 50.000 estações de rádios.

Neste trabalho, a proposta se foca na coleta e caracterização de informação provida pela API do SHOUTcast, com o objetivo de extrair informações relacionadas aos hábitos dos ouvintes de rádio, bem como informações relacionadas a estilos mais ouvidos, faixas mais tocadas, estações mais ouvidas, entre outros. Informações como essas são impor-

tantes, por exemplo, para empresas interessadas em anunciar seus produtos e/ou serviços em estações de rádios, para artistas e gravadoras interessadas em saber quantas vezes suas faixas são tocadas, bem como quais estilos estão sendo mais apreciados em um dado período.

2 Contextualização e Trabalhos Relacionados

Trabalhos sobre streaming de rádio pela internet foram conduzidos utilizando diferentes perspectivas. Alguns estudos focam na análise de popularidade de artistas (4, 10). Alguns focam o estudo na experiência do usuário e qualidade de serviço (15, 17, 18). Outros trabalhos se preocupam em analisar ferramentas de recomendação musical (13, 24) e geração automática de playlists baseadas em análise de streams de rádio (8, 9, 16, 24).

O aparecimento de vários serviços musicais baseados na web, mudou drasticamente a dinâmica da indústria musical. Faria eta al. (10) perceberam que a forma tradicional de se medir a popularidade de um artista, baseado somente no número de vendas de discos e número de reproduções na rádio, não é mais suficiente. Por conta dista, eles propuseram uma metodologia para comparação da popularidade de artistas, baseada na performance em diferentes mídias digitais, tais como a Internet, bem como em mídias tradicionais, como TV e rádio.

Bellogın et al. (4) também avaliaram a popularidade de artistas considerando plataformas musicais baseadas na web e também redes sociais voltadas para a música. Estudos foram feitos de forma a entender a relação entre índices de popularidade e seus rankings. Uma base de dados com 1312 artistas de diferentes plataformas (como EchoNest, Last.fm e Spotify) foi utilizada. Os resultados mostraram que popularidade é mais sensível a dimensões temporais do que dependentes de serviços. Este estudo mostrou também que popularidade é um sinal estável na maioria dos índices, uma vez que ele muda pouco ao longo do tempo.

Estudos feitos para análise de serviços de rádio por internet foram conduzidos por Melendi et al. (18); os autores consideraram o tráfego entre dispositivos e diferentes elementos relacionados ao comportamento do usuário, como consumo de recursos, qualidade dos dados transmitidos, etc. Outra abordagem proposta por Melendi (17) foi o desenvolvimento de uma ferramenta de simulação para realizar avaliações de rádios trans-

mitidas pela Internet. A ferramenta de simulação foi validada usando um serviço real e os resultados apresentaram alta confiabilidade em situações reais.

Lee et al. (15) analisaram comportamento de usuários em serviços comerciais de música na nuvem. Foi feita uma pesquisa com 198 respostas, explorando critérios de seleção, padrões de uso, limitações percebidas e previsões futuras dos usuários. Os resultados revelaram uma ligeira preferência por serviços de streaming: aproximadamente 25% dos usuários escutavam conteúdo de streaming bem como músicas possuídas, enquanto 25% ouviam músicas por streaming.

Turnbull et al. (24) exploraram o uso de rádios personalizadas para promover a descoberta de música por artistas locais. Eles estudaram um serviço web, chamado `MegsRadio.fm`, que cria streams de músicas de artistas locais e conhecidos, baseando-se em artistas-semente, tags, locais e localização. Os resultados revelaram que usuários do serviços se tornaram mais inclinados a ouvir música de artistas locais.

Aizenberg et al. (1) utilizaram playlists publicamente disponíveis de milhares de estações de rádio na internet, para criar um conjunto de dados em larga escala e desenvolver um modelo de filtragem colaborativa probabilística. Grant et al. (13) desenvolveram um sistema de recomendação de estações de rádio na internet utilizando dados históricos coletados do SHOUTcast. Maillet et al (16) propuseram uma abordagem para a geração de playlists orientáveis a partir de tags de músicas coletadas de streams de estações de rádio profissionais. Chen et al (8) propuseram uma modelagem de playlists como cadeias de Markov, geradas a partir do algoritmo LME (Latent Markov Embedding, em inglês) de aprendizado de máquina, usando streams de rádios online como conjunto de treino.

3 Desenvolvimento do Trabalho

Neste capítulo serão descritas informações sobre a base de dados e o processo de coleta.

3.1 O Serviço de Streaming do SHOUTcast

SHOUTcast é um software para streaming de mídia pela Internet. Este serviço foi desenvolvido, originalmente, pela Nullsoft, mas hoje é mantido pela Radionomy (21). O SHOUTcast possibilitou o surgimento de diversas estações de rádio na Internet, e estas rádios estão listadas em um diretório no site do serviço (22).

O SHOUTcast provê duas formas de se transmitir conteúdo de áudio: um deles, utilizando servidores deles próprios e outro a partir da instalação e configuração do software em um servidor dedicado. O SHOUTcast mantém controle de cada estação sento transmitida e oferece algumas funcionalidades, tais como: relatórios detalhados do público e audiência, injeção automatizada de propaganda, conteúdo geograficamente sensível e serviços de monetização.

O SHOUTcast foi escolhido como o tema deste trabalho principalmente por causa do grande número de estações de rádio que utilizam seu serviço. Segundo o site do serviço, mais de 60000 estações estão sendo transmitidas, o que gera uma grande quantidade de informação a ser coletada e analisada. Além disso, este conta com uma API muito bem documentada e fácil de se utilizar.

o serviço também possui uma maneira bem específica de se identificar propagandas, que é necessária caso a estação de rádio deseje utilizar do serviço de injeção automática e monetização. Para que este serviço funcione, arquivos de áudio contendo propaganda devem ser configurados de maneira específica, que sera explicada em maiores detalhes na próxima seção.

Neste trabalho, foram coletadas informações disponibilizadas pela API de Diretório de

Rádios (Radio Directory API, em inglês) do SHOUTcast [1]. Informações mais detalhadas dos dados coletados também serão explicados na próxima seção.

3.2 Descrição dos Dados Coletados e Processo de Coleta

A API do SHOUTcast disponibiliza acesso à alguns dos dados de sua base através de chamadas HTTP (2). A informação retornada vem em forma de um arquivo XML, contendo algumas informações sobre as rádios, conforme mostrado na Figura 1 abaixo:

```
<stationlist>
 <tunein base="/sbin/tunein-station.pls"/>
 <station name=".977 The Hitz Channel-[SHOUTcast.com]" mt="audio/mpeg" id="9907" br="128"
  genre="Pop Rock Top 40"ct="The Fray - You Found Me" lc="4670"/>
 <station name="HOT FM - Lebih Hangat Daripada Biasa : HOT fm-[SHOUTcast.com]"
  mt="audio/mpeg" id="120149" br="24" genre="Malaysia
  Malay" ct="LELAKI IDAMAN MELLY_GOESLOW " lc="3961"/>
 <station name="S K Y . F M - Absolutely Smooth Jazz - the world's smoothest jazz 24 hours a day-[SHOUTcast.com]"
  mt="audio/mpeg" id="1264" br="96" genre="Soft Smooth Jazz"
  ct="Oli Silk - De-stress Signal" lc="3507"/>
 <station name="Groove Salad: a nicely chilled plate of ambient beats and grooves. [SomaFM]-[SHOUTcast.com]"
  mt="audio/mpeg" id="6687" br="128" genre="Ambient Chill"
  ct="Verbrilli Sound - Descender" lc="2680"/>
 <station name=".977 The 80s Channel-[SHOUTcast.com]" mt="audio/mpeg" id="6803"
 br="128" genre="80s Pop Rock" ct="Starship - Nothing`s gonna stop us now (1987)" lc="2192"/>
 <station name="The Alex Jones Show-[SHOUTcast.com]" mt="audio/mpeg" id="5516" br="32" genre="Talk"
  ct="Refeed: Hour 1 (Listen by phone 512-646-5000)" lc="1987"/>
</stationlist>
```

Figura 1: Exemplo de arquivo XML retornado pela API do SHOUTcast

Conforme mostrado na figura acima, a API retorna informações como nome da rádio (`name`), número identificador (`id`), qualidade medida em bit rate (`bt`), gênero da rádio (`genre`), tipo de mídia (`mt`), conteúdo sendo transmitido (`ct`) e número de ouvintes no momento (`lc`).

Existem diversos serviços disponíveis na API, porém, para este trabalho, foram utilizados dois serviços: o `Get Random Stations` e o `Get Station by Keyword Search`. A utilização de cada um será explicada nos parágrafos seguintes.

Para a construção do conjunto de dados, foi desenvolvido um coletor em Python, que faz requisições ininterruptamente ao serviço do SHOUTcast, conforme Figura 2 abaixo. Este coletor ficou hospedado em uma máquina local e em um serviço na nuvem, de modo a garantir que não ocorra interrupção na coleta.

O algoritmo desenvolvido faz requisições ininterruptamente ao serviço, com intervalos de tempo pequenos o suficiente para que sejam detectadas mudanças na programação da

[1] Radio Directory API do SHOUTcast: http://wiki.shoutcast.com/wiki/SHOUTcast_Radio_Directory_API

rádio.

O coletor, após fazer a requisição e receber o arquivo XML de resposta, converte os dados para uma estrutura de dados e grava as informações em um arquivo .CSV. Além das informações retornadas pelo SHOUTcast, que foram descritas anteriormente, o coletor adiciona também um timestamp, de forma a possibilitar uma análise temporal dos dados.

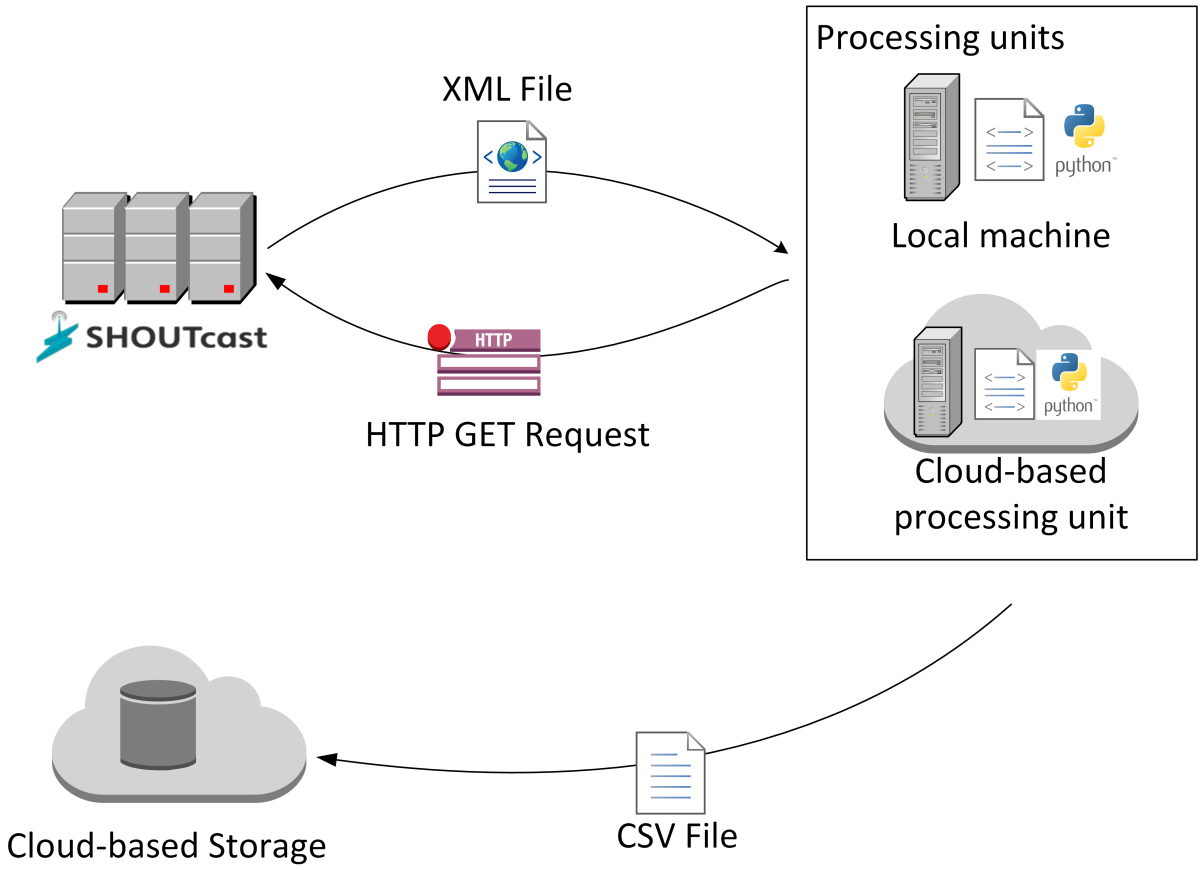

Figura 2: Arquitetura do coletor

O primeiro passo no processo de coleta foi identificar quais estações de rádio possuem algum conteúdo caracterizado como propaganda. Para isso, foi utilizado o serviço `Get Random Stations` citado anteriormente. Este serviço retorna uma rádio aleatória do diretório do SHOUTcast. Por padrão, o serviço retorna somente uma rádio, mas é possível inserir um parâmetro `limit` na chamada HTTP, que define o limite máximo de rádios a ser retornada. Ao definir este parâmetro com um número arbitrariamente grande, o serviço retornará todas as rádios sendo transmitidas naquele momento.

O processo de coleta foi feito em dois momentos: um entre 12 e 31 de Dezembro de 2016 e outro entre 3 e 21 de Abril de 2017, totalizando 27 dias de coleta. Os dados foram

coletados com um intervalo de 45 segundos entre as requisições. A base de dados completa conta com 25.827.411 de registros, com 75.000 estações de rádios.

Para detectar um conteúdo caracterizado como propaganda, o SHOUTcast exige que o arquivo de áudio seja configurado de uma maneira bem específica: é necessário configurar o metadado do arquivo, definindo o nome do artista e nome da música com o texto `Advert:`. Desta forma, é possível identificar quando uma rádio tocou uma propaganda.

3.3 Limitação dos Dados

Apesar de a API do SHOUTcast fornecer uma excelente oportunidade de estudo das atividades de rádios online, existem algumas limitações relativas à base de dados.

Primeiramente, existe um atraso na atualização do conteúdo que está sendo transmitido. Isto foi observado após escutar o stream da programação de 10 rádios utilizando o software `ff2mpeg` (11). A observação durou 3 dias, entre os dias 26 e 28 de Dezembro, e foi detectado que, em alguns casos, mesmo após a música ou outro conteúdo ter finalizado no stream de áudio, a API ainda retornada o nome do conteúdo anterior como o que está sendo transmitido no momento. Ou seja, mesmo com a mudança na programação, a API não reflete esta mudança instantaneamente. Vale lembrar que isso foi observado em apenas algumas rádios.

Em segundo lugar, a API não fornece informações de localização geográfica de usuários e/ou estações de rádio. Esta é uma informação importante, quando se pretende analisar colocação de anúncios (3) e a descoberta correta do fuso horário da programação de rádio.

Em terceiro lugar, foi detectada certa descontinuidade na transmissão da programação. Isto, em alguns momentos, foi causado por conta de bloqueios do SHOUTcast por causa de muitas requisições de acesso ao mesmo tempo. Por isso, a intervalo, na segunda coleta, foi aumentado para 45 segundos. Em outros momentos, as interrupções foram causadas por erros de timeout ou até mesmo indisponibilidade do serviço. Para detectar esses error, o script foi atualizado para gravar um "erro de coleta" toda vez que ocorresse uma interrupção. Estes erros resultaram em um total de 2% da base de dados.

Finalmente, a API não possui um número de identificação único para cada rádio, uma vez que rádios podem desconectar e reconectar ao sistema e, neste caso, elas retornam com um ID diferente. Segundo a API (2), só é garantida a singularidade de um ID por no máximo 1 dia. Isso prejudicou a identificação de estações de rádio. O problema foi

atenuado utilizando o nome da estação como identificador.

4 Resultados e Discussão

De forma a obter um melhor entendimento de como se comporta a audiência de uma estação de rádio e o que a influencia, fizemos algumas análises da audiência das estações na base de dados, de acordo com as seguintes características: média de audiência por hora, gênero das estações de rádio, conteúdo sazonal, formatos de mídia e bit rates.

Além disso, analisamos também como é a distribuição do conteúdo de propaganda na base de dados.

Finalmente, modelamos a base como um problema de classificação, de modo a entender quais características da base mais influenciam a audiência.

4.1 Caracterização da audiência no SHOUTcast

4.1.1 Análise geral da audiência

Para se ter um entendimento geral da base de dados, primeiramente analisamos audiência geral das estações. Figura 3 mostra a CDF (Cumulative Distribution Function) da audiência média por dia, segmentados em três grupos: "Mean Audience", representando a audiência média de todos os registros da base de dados; "Max Audience", representando a audiência máxima de todas as rádios da estação; e "Top100 Mean Audience", representando a audiência média das 100 rádios mais ouvidas.

Podemos ver que existe um número muito pequeno de estações com alta audiência e um número grande de estações com baixa audiência. De fato, cerca de 90% das estações de rádio possuem uma audiência média e máxima menor ou igual a 10 ouvintes. Dado que o SHOUTcast é um serviço gratuito, qualquer pessoa pode criar uma estação, o que explica a grande maioria de estações com baixa audiência.

Apesar disso, 1% (cerca de 750) e 1,5% (cerca de 1.125) das rádios coletadas possuem uma audiência média e máxima maior ou igual a 100 ouvintes diários, respectivamente.

Além disso, se ampliarmos o gráfico para as 100 rádios mais populares, podemos ver que 99% possuem uma média maior ou igual a 100 ouvintes, e 40% tem uma audiência média maior ou igual a 1000 ouvintes.

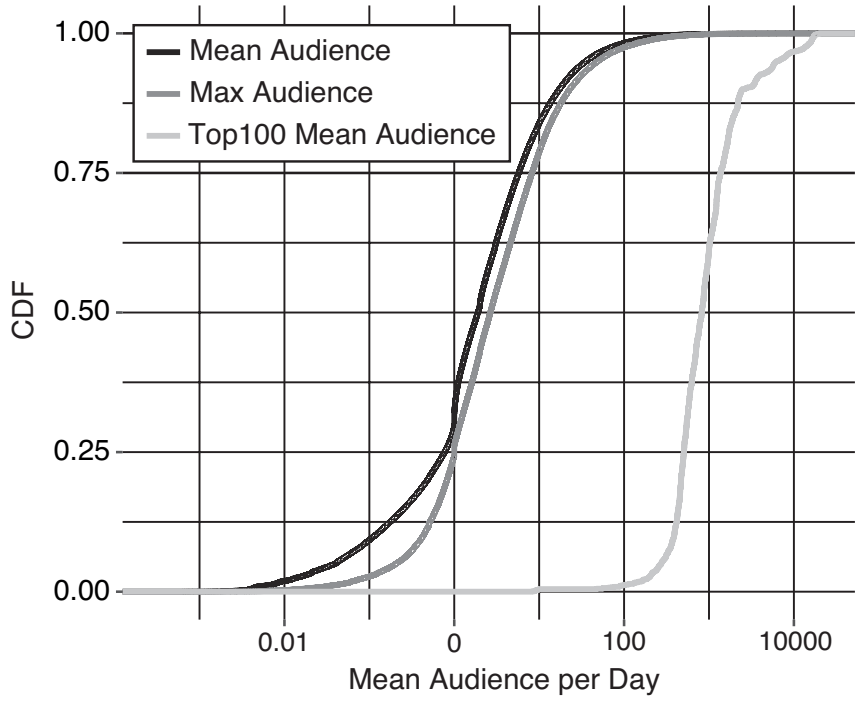

Figura 3: CDF da audiência média diária das estações de rádio no SHOUTcast

Nas subseções seguintes, avaliamos algumas outras características que podem afetar o tamanho da audiência de uma rádio.

4.1.2 Gêneros musicais

Nesta subseção, focamos em analisar a distribuição dos gêneros musicais entre as estações na base de dados. Figura 4 mostra uma lista com os 15 gêneros mais frequentes, em termos de número de estações. Conforme esperado, Pop é o gênero mais frequente. Outros gêneros populares presentes na base de dados são Talk, Rock, Gospel, Blues e Dance.

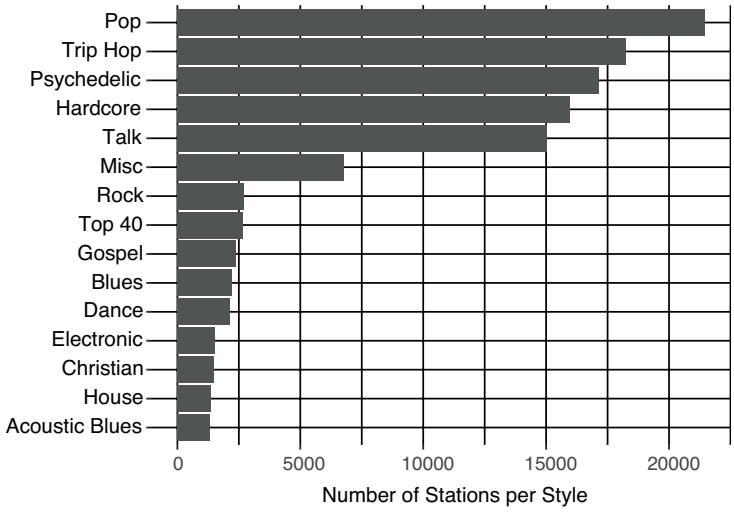

Figura 4: 15 gêneros mais populares, em termos de número de rádios

Figura 5 mostra a lista dos 15 gêneros mais populares, pela média de audiência das estações de rádio. Podemos ver que alguns dos estilos presentes na Figura 4 continuam presentes na Figura 5. Contudo, algum estilos populares caem para o final da lista, como Pop e Rock, enquanto algumas outras desaparecem da lista, como Gospel, Christian e Electronic. O gênero Alternative não só aparece, como se situa no topo da lista. Isso nos mostra que alguns gêneros conseguem atrair mais ouvintes, mesmo não sendo muito frequentes no SHOUTcast.

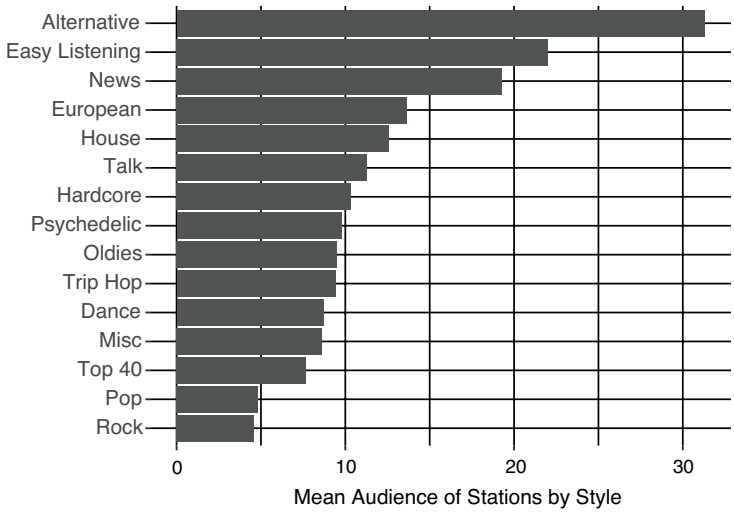

Figura 5: 15 gêneros mais populares, em termos da média de audiência das estações

Isso sugere que, apesar de estilos como Pop e Rock serem mais comuns no SHOUTcast, a alta concorrência, causada pelo grande número de estações, podem causar uma baixa

audiência em várias rádios. Por outro lado, gêneros menos populares tendem a atrair ouvintes recorrentes, possivelmente pela falta de opções de estações de rádio transmitindo estes estilos musicais.

4.1.3 Conteúdo sazonal

Na coleta realizada em Dezembro de 2016, detectamos uma dinâmica interessante na audiência de conteúdo sazonal. Figura 6 nos mostra as 15 estações de rádio com maior média de audiência, e seu respectivo gênero musical, durante os 3 primeiros dias da coleta (24 a 26 de Dezembro), enquanto a Figura 7 no mostra a mesma informação, mas considerando o intervalo de tempo entre 27 e 31 de Dezembro. O aumento substancial da audiência da rádio "Merry Christmas" pode ser causada principalmente por conta da época do ano em que a coleta foi realizada, que aconteceu durante a época de natal. Além desta rádio, outras 2 rádios com temática natalina apareceram no top 15.

Podemos ainda detectar a influência de conteúdo sazonal na audiência, comparando as Figuras 6 e 7. É possível observar que todas as estações natalinas que atingiram o pico durante os dias 24, 25 e 26 de Dezembro, não aparecem mais a listagem das 15 rádios mais ouvidas entre 27 e 31 de Dezembro. Durante o primeiro período, a rádio mais ouvida possuía média de audiência quase duas vezes maior que a segunda rádio, e a lista continha 3 rádios com tema relacionado ao Natal, ao passo que no segundo período, todas as estações natalinas desaparecem da lista. Isso nos mostra que conteúdo sazonal pode, de fato, influenciar grandemente o número de ouvintes de uma rádio.

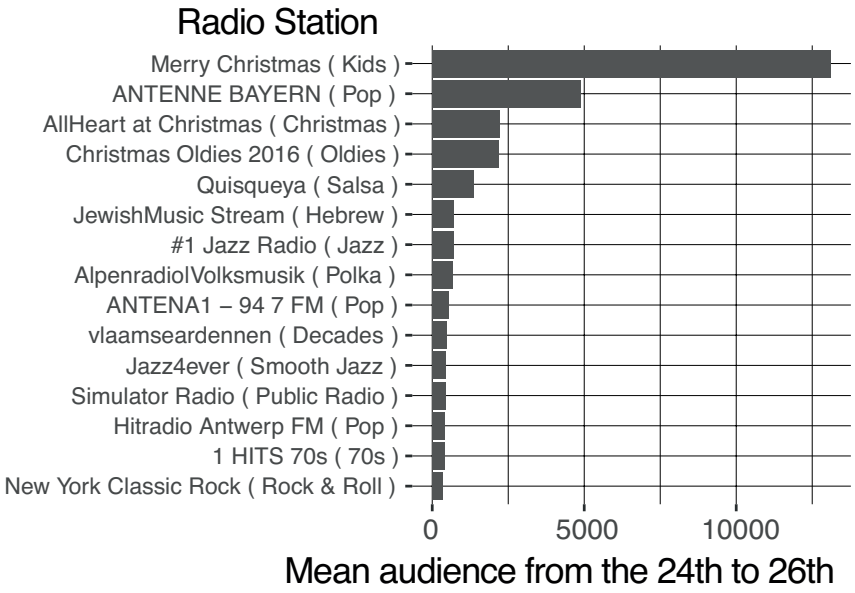

Figura 6: 15 estações de rádio mais ouvidas entre 24 e 26 de Dezembro e seu respectivo gênero

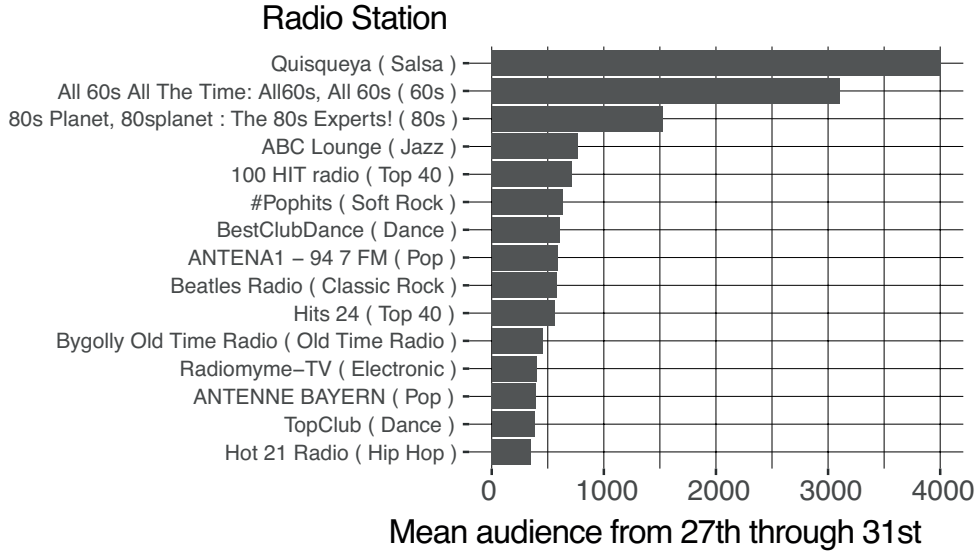

Figura 7: 15 estações de rádio mais ouvidas entre 27 e 31 de Dezembro e seu respectivo gênero

Para corroborar o pressuposto anterior, analisamos o comportamento da audiência das 3 rádios natalinas que apareceram no top 15. Na Figura 8, mostramos a variação da audiência de todas as rádios temáticas presentes na lista das 15 mais ouvidas durante o Natal. É possível notar que, para todas elas, a audiência atingiu seu pico entre os dias 24

e 27 de Dezembro e, após isso, caiu significativamente.

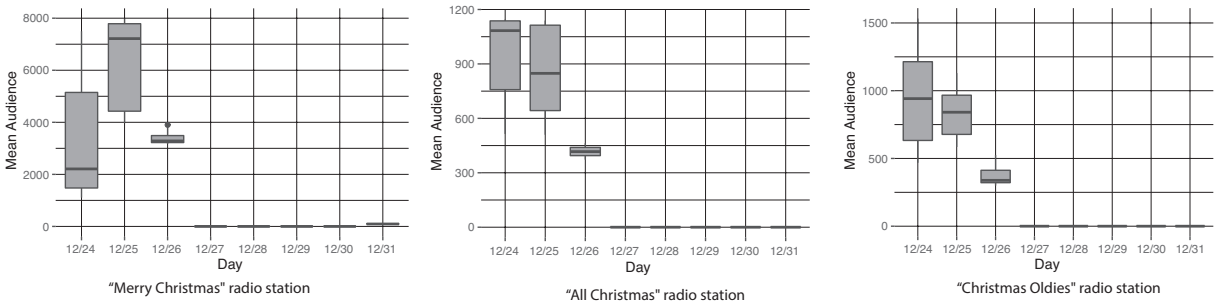

Figura 8: Variação da audiência das 3 rádios natalinas, durante o período entre 24 e 31 de Dezembro.

Baseando na análise anterior, conduzimos um novo experimento para verificar como esse comportamento acontece em outras plataformas de conteúdo musical. Spotify Charts (23) é um website que compila uma lista das músicas mais ouvidas durante um período selecionado no Spotify, que é hoje um dos serviços de streaming de música mais populares. Analisamos os dados do mesmo período da nossa coleta (24 a 31 de Dezembro) e os resultados são mostrados na Tabela 1. Podemos ver que o número de stream de músicas natalinas tem seu pico durante o período de Natal, diminuindo consideravelmente após isso, mantendo o comportamento notado na base de dados do SHOUTcast.

Tabela 1: Spotify streams durante a semana do Natal.

Dia	**Total # streams**	**# Christmas songs**	**Percentage**
12/24/16	215,904,792	58,758,925	27.21%
12/25/16	185,238,748	54,390,411	29.36%
12/26/16	147,661,069	74,808,54	5.06%
12/27/16	160,904,262	1,511,056	0.93%
12/28/16	167,354,900	956,865	0.57%
12/29/16	169,152,480	0	0%
12/30/16	173,237,247	0	0%
12/31/16	208,083,869	0	0%

De fato, conteúdo sazonal aumentou a audiência não só de rádios temáticas, mas a audiência como um todo de toda a base de dados. Conforme pode ser observado na Figura 9, a audiência média total atingiu um pico durante o período de Natal, diminuindo

logo em seguida, atingindo um novo pico durante a época das celebrações de Ano Novo. Isso nos mostra que, de fato, existe uma tendência maior dos ouvintes de ouvirem rádio durante datas festivas.

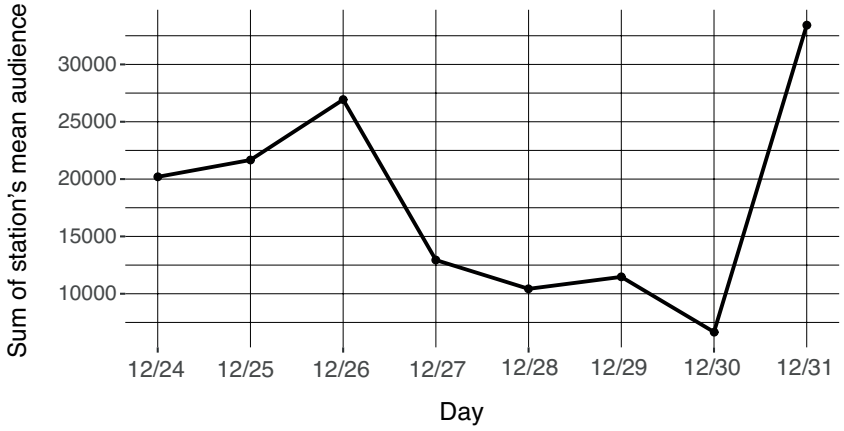

Figura 9: Audiência geral da base de dados durante o período entre 24 e 31 de Dezembro de 2016.

4.1.4 Formatos de mídia e bit rates

O serviço do SHOUTcast permite que estações de rádio transmitam conteúdo em diferentes bit rates e dois formatos diferente de mídia (MPEG e AACP). O bit rate influencia a qualidade do stream de áudio: quanto maior o bit rate, melhor a qualidade, o que também incorre em um maior consumo de banda de Internet.

As Figuras 10 e 11 apresentam a distribuição dos diferentes bit rates, de acordo com o número de estações de rádio e a média da audiência, respectivamente. É possível perceber que, apesar da grande maioria das rádios preferirem por transmitir o conteúdo em uma qualidade de 128bps, estações transmitindo à 120bps tendem a ter uma média de audiência maior. Isso pode indicar uma preferência, por parte dos usuários, por streams de dados com menor consumo de banda.

A forte adoção por parte dos ouvintes por bit rates menores pode estar ligada ao consumo de banda de Internet. Usuários podem estar interessados em um melhor custo-benefício entre qualidade e consumo, principalmente quando o acesso é feito a partir de conexões lentas ou limitadas, como por exemplo através de telefones celulares.

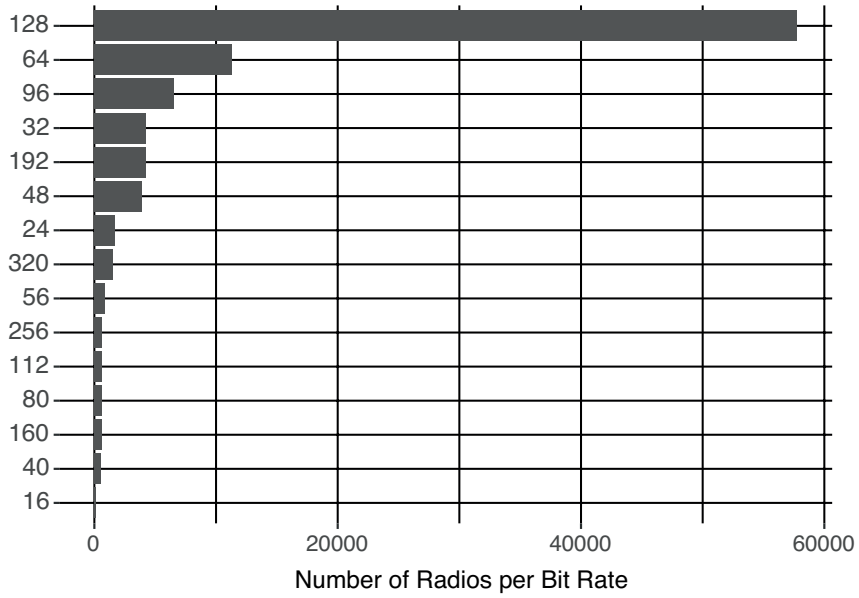

Figura 10: Frequência de bit rates, por número de estações de rádio.

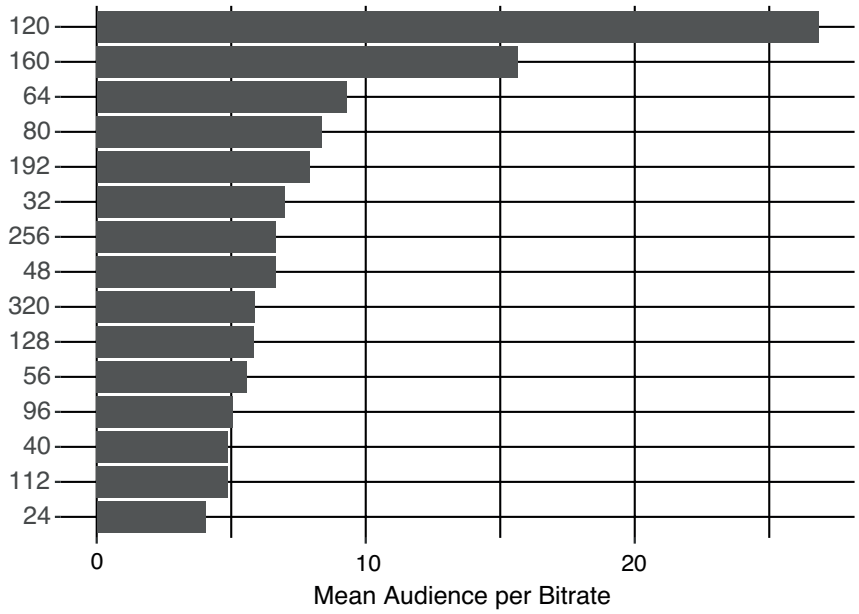

Figura 11: Bit rates, por média de audiência.

Nos dados coletados, existe um número considerável de estações transmitindo à mais de um bit rate simultaneamente. Apesar de o serviço tratar essas rádios como rádios distintas, é possível verificar que se trata de uma mesma rádio através do nome da rádio e do conteúdo, que é o mesmo independente da qualidade do stream. Esta decisão é compreensível caso uma estação pretende atingir uma audiência mais ampla, que inclui

usuários que acessam através de celulares com largura de banda de Internet limitada, e também usuários que acessam por computadores com conexões de Internet de banda larga mais rápidas.

Se tratando dos formatos de mídia, o serviço do SHOUTcast suporta somente 2 codecs de áudio: MP3 e AAC. Segundo Brandenburg et al. (6), MP3 (abreviação de MPEG-1 Layer 3) e ACC (abreviação de MPEG-2 Advanced Audio Coding) são ambos encoders de áudio, com o último sendo uma versão avançada do primeiro. Apesar disso, o MP3, que foi desenvolvido em 1991, ainda é mais amplamente utilizado do que o AAC no SHOUTcast.

A Figura 12 apresenta o número de estações e média de audiência para cada um dos dois formatos. A grande maioria das rádios transmite usando o MP3, por ser mais difundido. Apesar disso, quando se trata da média de audiência, a diferença não é tão grande. Este fato pode ser explicado, novamente, pela quantidade de rádios usando o MP3 e que tem audiência muito baixa.

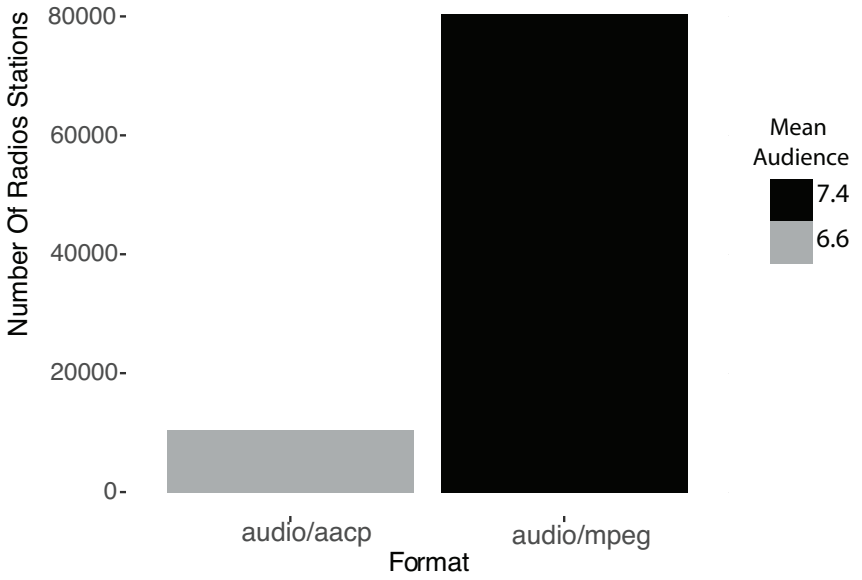

Figura 12: Número de estações de rádio e média de audiência por formato de mídia.

4.2 Propagandas no SHOUTcast

Posicionamento de anúncios em qualquer canal de mídia é sempre um desafio (conteúdo, tempo, frequência, etc.), uma vez que um mal posicionamento pode causar uma experiência de usuário insatisfatória. No caso de rádios transmitidas pela Internet, isso não é diferente. Propaganda é, frequentemente, a principal fonte de renda deste modelo

de negócio, desta forma é importante tanto para estações de rádio quanto para anunciantes a otimização da entrega de conteúdo de propagandas.

O serviço do SHOUTcast provê às rádios transmitidas através de seu sistema, um serviço de injeção automática de propaganda, levando em consideração variáveis como localização geográfica e hora do dia para melhor posicionar os anúncios. Para que esse serviço funcione, arquivos de áudio de propaganda devem ser configurados de uma maneira específica: os metadados devem ser configurados de forma que o nome do artista e da faixa sejam definidos com o texto "Advert:" (2). Por conta deste padrão, o trabalho de se encontrar anúncios na base de dados é bem simples, bastando encontrar este texto nas entradas do nome da faixa.

De forma a melhor compreender como anúncios influenciam a audiência de uma estação de rádio, a base de dados foi divida em dois grupos: Ad e NoAd. O primeiro representa aqueles conteúdos classificados como propaganda, enquanto o segunda representa todos os outros tipos de conteúdo. Uma dada estação de rádio contém um número X de conteúdo Ad e Y de conteúdo NoAd. A razão destes doi números nos dá uma ideia da proporção entre os conteúdos tocados em uma estação de rádio. Uma razão $>$ 1 significa que a rádio contem mais propaganda do que conteúdo, enquanto a razão < 1 representa o contrário. Uma razão igual a 1 representa que a rádio toca o mesmo número de propagandas e conteúdo.

A figura 13 apresenta a CDF da razão Ad/NoAd na base de dados. Como pode ser observado, 38,1% de todas as rádios não transmitem nenhum anúncio, e cerca de 62,8% das estações tem uma razão ≤ 0.01, o que significa que a grande maioria das estações de rádio optam por não incluir muita propaganda em suas programações. Dado que o SHOUTcast é um serviço gratuito e aberto, existe um grande número de estações independentes, que não são grandes o suficiente para atrair interesse de anunciantes, o que pode explicar a baixa ocorrência de anúncios na maioria da base de dados. Apesar disso, 3% das estações de rádio (cerca de 2.250) tem uma razão de ≥ 1, o que representa que mais da metade de todo o conteúdo dessas rádios, é propaganda.

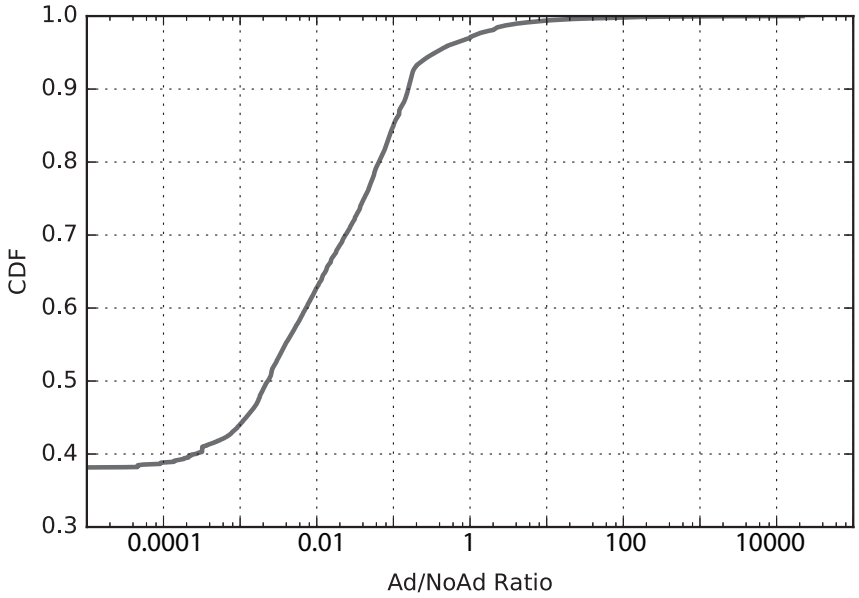

Figura 13: Distribuição da razão Ad/NoAd entre as estações de rádio

A Figura 14 ilustra como essa relação entre conteúdo e propaganda correlaciona com a audiência média de uma estação. Como pode ser visto, estações com audiência alta, possuem baixa razão Ad/NoAd. De fato, a média da razão na base de dados é igual a 0.0029, o que nos diz, conforme dito anteriormente, que a maioria das rádios está transmitindo mais conteúdo do que propaganda.

Para complementar a nossa análise, calculamos a correlação de Pearson entre a audiência e a razão Ad/NoAd, com um valor igual a $-1.298452e-05$. A correlação foi detectada com um intervalo de confiança de 95%. Apesar da correlação ser muito pequena, o valor negativo indica que quanto maior a relação Ad/NoAd, menor a audiência, o que significa que aumentar o número de propagandas pode, de fato, impactar negativamente na audiência. Na Figura 14 podemos ver que as estações com alta audiência, possuem razão Ad/NoAd muito baixa e, por outro lado, estações com a razão muito alta, possuem audiência próxima de 0.

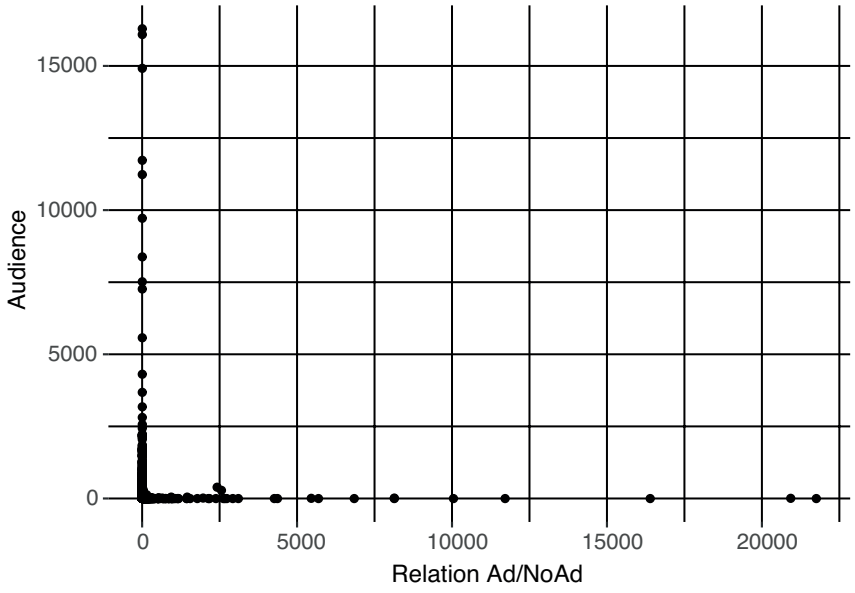

Figura 14: Razão Ad/NoAd por média de audiência.

4.3 Importância dos atributos

Nas seções anteriores, analisamos o comportamento da audiência das estações de rádio de acordo com algumas características, como estilo musical, época do ano, formato de mídia, etc. Estas análises nos ajudaram a ter uma visão geral dos dados coletados. Para melhor entender o que influencia a audiência, utilizamos a base de dados para criar um classificador, utilizando de técnicas de aprendizado de máquina, para calcular a importância dos atributos.

Para isso, modelamos os dados como um problema de classificação, com a audiência como alvo do modelo. Para facilitar a computação, discretizamos a audiência em quartis (Muito Baixa, Baixa, Alta e Muito Alta). As características utilizadas no modelo foram: estilo musical, número de anúncios por hora, número de conteúdo por hora, bit rate, tipo de mídia, hora do dia e dia da semana. Para treinar o classificador, utilizamos o algoritmo de Extremely Randomized Trees, conforme proposto por Geurts et al. (12). Este método é baseado no método de Random Forest, que opera criando um número arbitrário de arvores de decisão em tempo de treino e gerando na saída, a classe que é a moda de todas as classes de cada uma das árvores da floresta. O primeiro algoritmo de Random Forests foi criado por Tin Kam Ho (14), sendo posteriormente estendido por Breiman et al. (7).

Este algoritmo utiliza o índice de impureza Gini para o cálculo das divisões durante

o treino. De acordo com Breiman et al. (7):

> every time a split of a node is made on variable M, the Gini impurity criterion for the two descendant nodes is less then the parent node. Adding up the Gini decreases for each individual variable over all trees in the forest, gives a fast variable importance that is often very consistent with the permutation importance measure.

Para implementação deste algoritmo, foi utilizado o pacote de aprendizado de máquina scikit-learn (19). Este pacote, implementado em Python, conta com diversos algoritmos de aprendizado (o que inclui problemas de classificação, regressão, agrupamento, redução de dimensionalidade, etc.) prontos para uso, bastando somente a modelagem e pré-processamento dos dados.

A Figura 15 apresenta um gráfico com a importância de cada uma das características, conforme calculado pelo algoritmo descrito acima. As barras representam a importância de cada atributo da floresta, juntamente com sua variância entre-árvores.

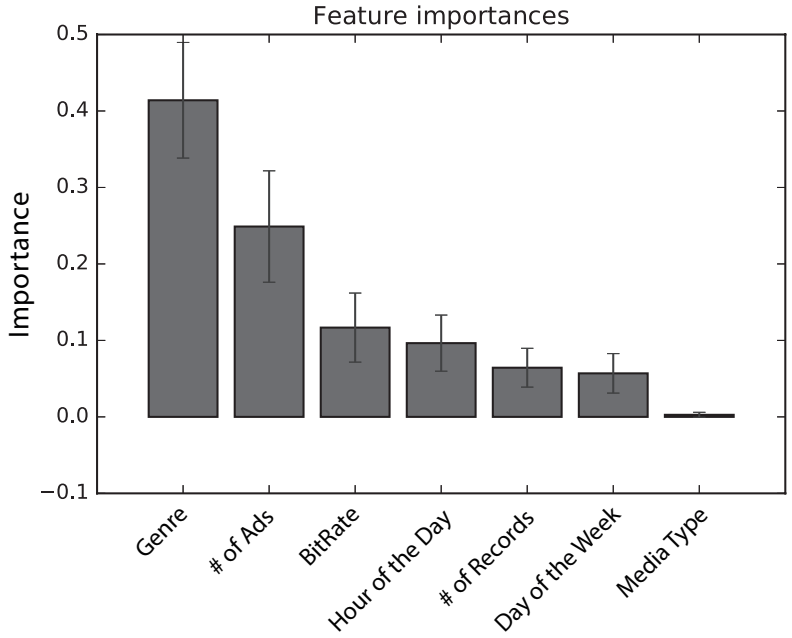

Figura 15: Features relevance

Na figura acima, podemos perceber que o atributo que mais influencia a audiência de uma rádio é o gênero musical, com um índice Gini de mais de 0,5, seguido pelo número de anúncios por hora. O formato de mídia é o atributo menos impactante na audiência.

5 Conclusões

Neste trabalho, coletamos, analisamos e apresentamos uma caracterização do diretório de rádios do SHOUTcast.

O processo de coleta resultou em uma base de dados com mais de 25.000.000 de registros com mais de 75.000 rádios coletadas.

É possível perceber como a Internet possibilitou o surgimento de inúmeras estações de rádio online, não só de estações que já transmitiam conteúdo por métodos tradicionais, mas também de estações independente, que puderam surgir graças a acessibilidade e baixo custo da tecnologia.

Além disso, é interessante perceber como a existência de serviços de broadcasting pela Internet possibilitou o acesso a conteúdo musical de várias partes do mundo, fazendo com que artistas locais pudessem ser descobertos por pessoas de vários lugares do planeta.

O processo de coleta se mostrou bastante desafiador, devido a alguns problemas que, ao acontecerem, demandavam um reinício no processo de coleta. Muito problemas só puderam ser detectados após alguns dias de coleta, o que causou atrasos no processo. Apesar disso, estes problemas fizeram com que fosse possível entender de maneira mais completa como o serviço se comporta e como são os dados retornados pelo serviço.

Pudemos perceber que, apesar de a grande maioria das rádios contarem com uma audiência baixa, algumas conseguem atingir números bem altos, muitas vezes $\geq$ 4.000 ouvintes por hora. Caracterizamos ainda diversos atributos da base de dados, como distribuição da audiência e de estilos musicais, dinâmica de propaganda e conteúdo sazonal, assim como distribuição de bit rates e formatos de áudio utilizados.

Além disso, pudemos perceber também como conteúdo sazonal impacta no número de ouvintes de uma estação de rádio. Nós conjectamos que, por conta disso, palavras de pesquisa e tags podem ser usadas estrategicamente para impulsionar o número de ouvintes. A detecção de influências sazonais nas tendências da audiência pode fornecer

informações valiosas para conteúdo mais personalizado e melhor colocação de propaganda em estações de rádio na Internet.

Ao final, analisamos em que medida esses atributos afetam o tamanho do público de uma rádio. Nós classificamos a importância dos atributos para prever o tamanho do público de acordo com seu coeficiente de Gini. O gênero de música e o número de anúncios por hora revelaram-se os atributos mais relevantes para prever o tamanho do público no SHOUTcast, seguido da taxa de bits e da hora do dia. Ao contrário da taxa de bits, o formato da mídia não apresentou relevância para a previsão do público.

Acreditamos que as informações apresentadas neste trabalho podem ser bastante valiosas tanto para estações de rádio quando para anunciantes nesse serviço, uma vez que, para maximizar o retorno econômico de investimentos em rádios transmitidas pela Internet, é importante saber como se comporta o consumo deste serviço por parte do usuário.

Pretendemos também, em uma data posterior, disponibilizar os dados coletados para a comunidade.

Referências

1 N. Aizenberg, Y. Koren, and O. Somekh. Build your own music recommender by modeling internet radio streams. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 1–10, New York, NY, USA, 2012. ACM.

2 S. API. Shoutcast api, Jan. 2017.

3 S. P. Bauman, K. Schmidt, and D. Preuss. Location based, content targeted online advertising, Oct. 5 2006. US Patent App. 11/539,109.

4 A. Bellogın, A. P. de Vries, and J. He. Artist popularity: do web and social music services agree. In *Int. Conf. on Weblogs and Social Media (ICWSM), Boston*, 2013.

5 D. Black. Internet radio: a case study in medium specificity. *MEDIA CULTURE AND SOCIETY*, 23(3):397–408, 2001.

6 K. Brandenburg. Mp3 and aac explained. In *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society, 1999.

7 L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

8 S. Chen, J. L. Moore, D. Turnbull, and T. Joachims. Playlist prediction via metric embedding. In *18th ACM SIGKDD*, 2012.

9 S. Chen, J. Xu, and T. Joachims. Multi-space probabilistic sequence modeling. In *19th ACM SIGKDD*, 2013.

10 F. L. de Melo Faria, D. M. B. Paiva, and Á. R. Pereira. *ArtistRank – Analysis and Comparison of Artists Through the Characterization Data from Different Sources*, pages 60–76. Springer International Publishing, Cham, 2016.

11 ffmpeg. ffmpeg, Jan. 2017.

12 P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

13 M. Grant, A. Ekanayake, and D. Turnbull. Meuse: Recommending internet radio stations. In *ISMIR*, pages 281–286, 2013.

14 T. K. Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

15 J. H. Lee, Y.-S. Kim, and C. Hubbles. A look at the cloud from both sides now: An analysis of cloud music service usage. In *Proceedings of the 16th International Society for Music Information Retrieval Conference. New York: ISMIR*, 2016.

16 F. Maillet, D. Eck, G. Desjardins, and P. Lamere. Steerable playlist generation by learning song similarity from radio station playlists. In *ISMIR*, 2009.

17 D. Melendi, R. García, X. G. Pañeda, S. Cabrero, and V. García. Performance evaluation of different architectures for an internet radio service deployed on an fttx network. *International Journal of Business Data Communications and Networking (IJBDCN)*, 6(2):46–68, 2010.

18 D. Melendi, M. Vilas, R. Garcia, X. G. Paneda, and V. Garcia. Characterization of a real internet radio service. In *32nd EUROMICRO Conference on Software Engineering and Advanced Applications (EUROMICRO'06)*, pages 356–363, Aug 2006.

19 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

20 C. Priestman. *Web radio: radio production for internet streaming.* Gulf Professional Publishing, 2002.

21 Radionomy. Radionomy, Jan. 2017.

22 SHOUTcast. Shoutcast, Jan. 2017.

23 Spotify. Spotify, Jan. 2017.

24 D. R. Turnbull, J. A. Zupnick, K. B. Stensland, A. R. Horwitz, A. J. Wolf, A. E. Spirgel, S. P. Meyerhofer, and T. Joachims. Using personalized radio to enhance local music discovery. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pages 2023–2028, New York, NY, USA, 2014. ACM.

25 T. Wall. The political economy of internet music radio. *Radio Journal: International Studies in Broadcast & Audio Media*, 2(1):27–44, 2004.