Tarefa Aula 3 – Exploração de SNA em R

Novembro de 2019

Table of Contents

# Alunos / ID (matrícula)

* [Daniel Campos](mailto:daniel.ferraz.campos@gmail.com) / A57635769
* [Leandro Daniel](mailto:contato@leandrodaniel.com) / A57622988
* [Rodrigo Goncalves](mailto:rodrigo.goncalves@me.com) / A57566093
* [Ygor Lima](mailto:ygor_redesocial@hotmail.com) / A57549661

# Estrutura Analítica

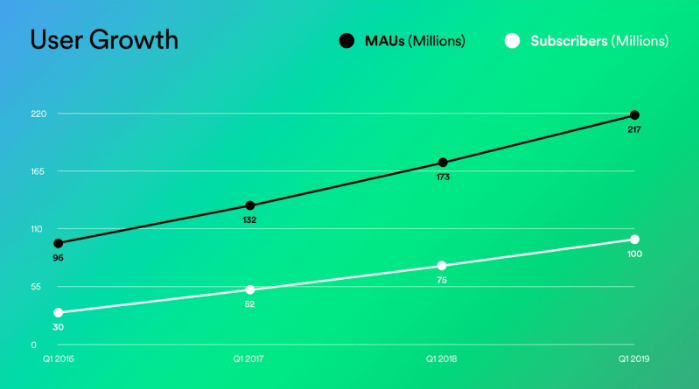
Proponham uma análise/projeto envolvendo SNA e Text Mining que possa ser realizada através da plataforma R e softwares livres contemplando:

* Problema / Oportunidade de Negócio
* Relevância / Motivação
* Escopo: Dados, Processamento, Saída
* Estrutura Estratégica / Operacional
* Indicadores de Avaliação
* Adoção no ambiente empresarial (recorrência, implicação gerencial, etc)
* Se possível: realizem Prova de Conceito

*Cada item acima indicado deverá estar bem claro e detalhado no documento final.*

# Proposta de Análise / Projeto

Em abril deste ano, o maior serviço de streaming de música do mundo, o Spotify, anunciou um grande marco na indústria fonográfica. A plataforma alcançou o incrível número de **100 milhões de usuários pagantes** (na chamada, contas premium). Esse quantitativo cresceu ainda, principalmente após o lançamento na Índia — que aconteceu no fim de fevereiro de 2019. Na ocasião, a plataforma recebeu mais de 1 milhão de usuários em apenas 1 semana. Segundo a empresa, no fim do primeiro mês já se contabilizava o dobro dessa quantidade.



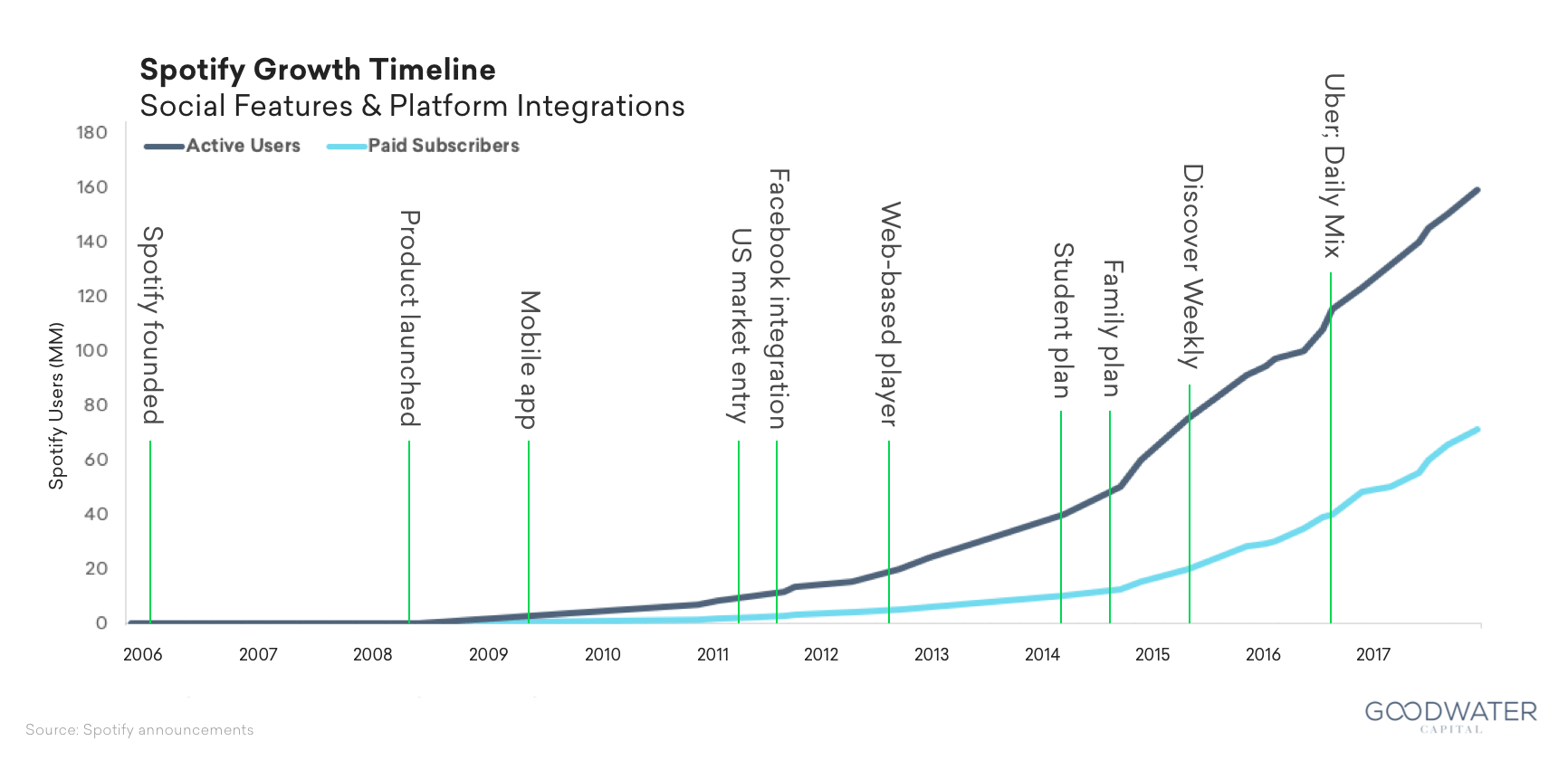
Fonte: <https://newsroom.spotify.com/2019-04-29/spotify-reports-first-quarter-2019-earnings/>

Abaixo listamos mais alguns dos incríveis números do Spotify:

* **217 milhões de usuários** ativos no Spotify;
* **100 milhões de usuários pagantes** do serviço Spotify Premium;
* **50 milhôes de músicas** disponíveis na plataforma;
* **3 bilhões de playlists** criadas e atualmente disponíveis na plataforma;
* O Spotify responde por **30% da receita total gerada pela indústria da música gravada e 42% do mercado de streaming**;
* Os usuários médios ouvem 41 artistas únicos por semana;
* A média de horas gastas ouvindo Spotify por mês é de 25 horas;
* 44% dos usuários ouvem o Spotify diariamente;
* Existem, aproximadamente, 4040 milhões de faixas disponíveis no Spotify;
* Cerca de um terço do tempo de audição do Spotify é gasto em listas de reprodução geradas pelo Spotify, e outro terço nas listas de reprodução geradas pelos usuários;
* O Spotify tornou-se público em abril de 2018, com uma avaliação de US $ 26,5 bilhões até o final do primeiro dia de negociação;
* O valor de mercado do Spotify em meados de maio de 2019 era de US $ 25 bilhões - relativamente inalterado por pelo menos seis meses;
* O maior valor de mercado até o momento é de US $ 35 bilhões em julho de 2018.

Fonte: [Business of Apps](https://www.businessofapps.com/data/spotify-statistics/), 10 de maio de 2019

O gráfico abaixo mostra o crescimento da base de usuários do Spotify, desde o seu lançamento em 2017, pontuada pela introdução de novos recursos e *Features Sociais*. Uma das principais características que fazem do Spotify a plataforma de streaming mais consumida é a **curadoria** de listas de música e indicações para os usuários.

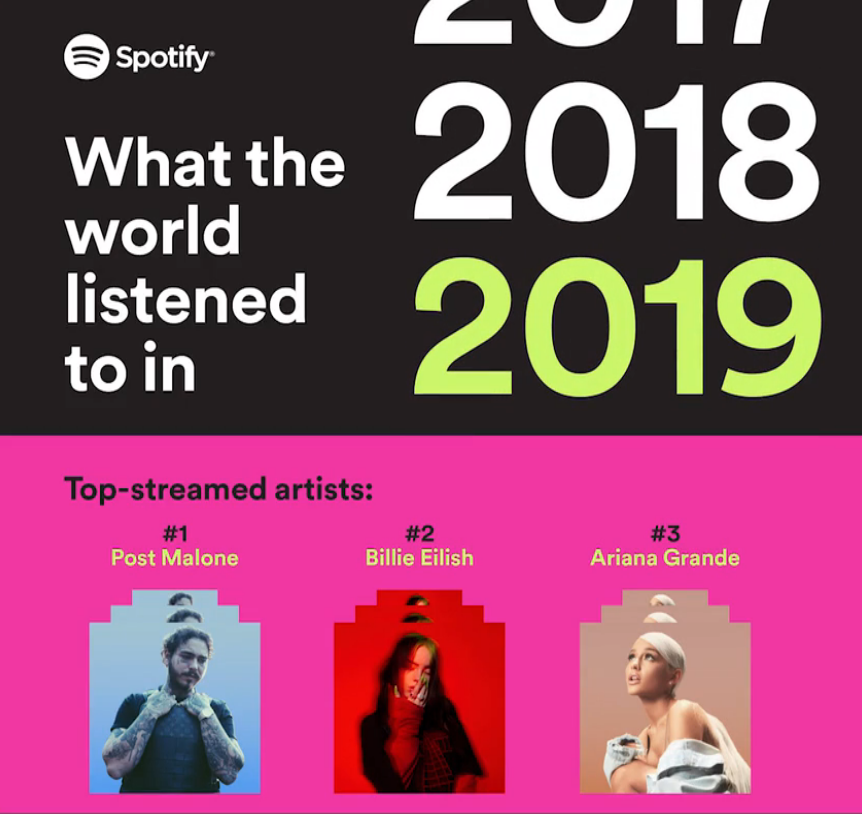


Fonte: <https://www.goodwatercap.com/thesis/understanding-spotify>

## Problema / Oportunidade de Negócio

Como vimos, a indústria da música e, em especial do streaming, vem crescendo de maneira consistente na medida em que novas *Features Sociais*, curadoria e técnicas de integrações da plataforma alavancam novas formas de consumo e compartilhamento entre os usuários.

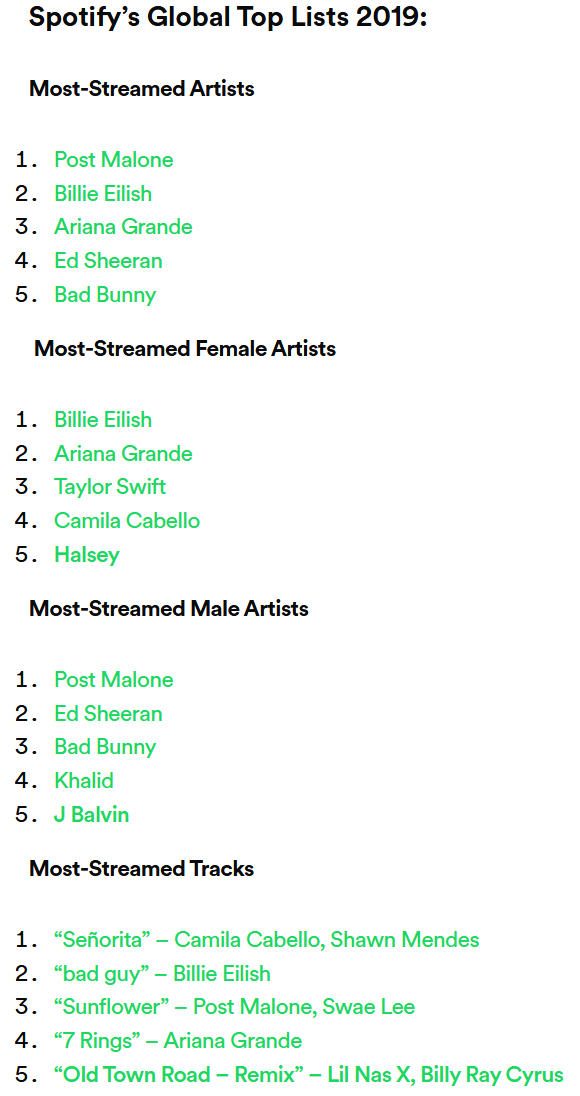
No centro de toda essa efervescência, temos centenas de milhares de artistas gerando conteúdo e competindo entre si para conseguir um lugar ao sol. Acompanhar e entender o que faz um mercado com tanta competitividade de oferta eleger seus artistas prediletos é uma tarefa monumental e virtualmente impossível de ser realizada sem técnicas de *Data Science*.



Top 3 artistas mais ouvidos do mundo na plataforma Spotify

Aplicando as técnicas corretas de *Data Science* e valendo-se dos dados disponíveis sobre os artistas, entendemos que é possível gerar planos táticos e estratégicos para orientar artistas, empresários e agentes da indústria da música.

Se pegarmos, por exemplo, a *top list* de 2019 do Spotify e fizermos uma exploração nas principais músicas e artistas consumidos em 2019, extrairemos *insights* a respeito do estilo de música e teor de letra mais consumidos.



Top lists de 2019 na plataforma Spotify

Diversos cenários de planos estratégicos, curadoria e planejamento de carreira podem se beneficiar com estas análises empregando técnicas de **SNA**\* e **Text Mining**.

## Relevância / Motivação

Com o advento da internet e popularização de serviços de download e streaming, a música digital se tornou imensamente acessível a população, e isso aumentou em muito a variedade de gêneros, artistas e tipos de músicas para a escolha do público em geral.

As plataformas de descoberta de novos conteúdos que tender a ir de acordo com o gosto musical do público tem se tornado um dos principais ativos de diferenciação das plataformas de distribuição de música. Existem diversas formas de se construir sistemas de recomendação, a maior parte deles baseada no histórico de streaming dos usuários.

Pensamos em uma alternativa diferente, que pode até não ser eficiente sob técnicas de machine learning aplicadas no comportamento do público, mas que pode despertar a curiosidade dos fãs de música.

Pretendemos aplicar técnicas de mineração de texto nas letras de diversas músicas e gêneros a fim de identificar a correlação entre artistas, possivelmente revelando padrões de influência entre eles.

Além da identificação de possíveis influências entre artistas e gêneros podemos segmentar as letras das músicas utilizando técnicas de **Análise de Sentimento** e **Topic Modeling**.

A apresentação dos resultados desta análise pode ser um ativo de atração de público ao despertar a curiosidade deles em descobrir de que forma seus artistas favoritos estão correlacionados aos demais atores no cenário musical.

## Escopo: Dados, Processamento, Saída

Para esta tarefa iremos utilizar um dataset extraído do [Kaggle](https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics). Nele, encontram-se disponíveis cerca de 380 mil letras de músicas de uma grande variedade de artistas e gêneros.

Após uma primeira exploração dos dados, notamos a neccessidade de retirar todas as inconsistências e aplicar os seguintes filtros no dataset:

* Valores faltantes;
* Gêneros não definidos;
* Ano anterior a 1970 (poucas observações).

Ficamos com cerca de 237 mil letras de músicas que utilizaremos para analisar diversos aspectos utilizando técnicas de mineração de texto e análise de rede tais como:

* Identificar as palavras e bigramas que são mais utilizadas por artista, gênero e década.
* Analisar a correlação das letras de diferentes artistas e montar uma rede identificando as possíveis influências entre os artistas e gêneros.
* Aplicar análise de sentimento nas letras de músicas e segmentação dos artistas com base nesta análise.

Abaixo uma pequena exploração do dataset selecionado.

**- Uma amostra dos dados contidos no dataset:**

lyrics <- read\_csv('../../data/raw/lyrics.csv',  
 col\_types = cols(.default = col\_character(),  
 index = col\_integer(),  
 year = col\_integer()))  
  
# elimina musicas sem letras, anterioeres a 1970 (dado o baixo número de observações),  
# e sem a identificação de genero musical.  
lyrics <- filter(lyrics,   
 !is.na(lyrics),   
 !(genre %in% c('Not Available', 'Other')),  
 as.integer(year) >= 1970)  
  
# seleciona uma amostra de 15 musicas com os 50 primeiros caracteres da letra.  
temp <- sample\_n(lyrics, size = 15)  
temp$snippet <- paste(str\_sub(temp$lyrics, 1, 50), '...', sep = '')  
temp <- select(temp, -lyrics)  
table\_output <- flextable(as.data.frame(temp))  
table\_output <- fontsize(table\_output, part = "all", size = 9)  
table\_output <- autofit(table\_output)  
table\_output <- theme\_vanilla(table\_output)  
table\_output

| **index** | **song** | **year** | **artist** | **genre** | **snippet** |
| --- | --- | --- | --- | --- | --- |
| 95516 | hollywood-kids | 2007 | bruce-springsteen | Rock | They're the Hollywood kids each and every one With... |
| 85295 | real-one-ice-t | 2006 | 2-live-crew | Hip-Hop | featuring Ice-T Verse 1: Brother Marquis Only the ... |
| 152586 | cold-love | 1993 | donna-summer | Pop | I'm walking the line I'm walking the line Treading... |
| 116645 | here-s-my-heart | 2015 | casting-crowns | Rock | Chorus 1 Here's my heart Lord, here's my heart Lor... |
| 72477 | how-did-we-get-here | 2010 | fat-joe | Hip-Hop | Oh, I dont know how we made it. Crack. Verse 1 [Fa... |
| 273086 | i-m-goin-away | 2007 | graham-coxon | Rock | I'm goin' away, just for how long I can't say I'm ... |
| 226892 | anyway-men-are-from-mars | 2006 | amber | Pop | Do do do do do... You love black and I like red Yo... |
| 95723 | book-of-dreams | 2006 | bruce-springsteen | Rock | I'm standing in the backyard Listening to the part... |
| 114978 | travelin-light | 2007 | del-reeves | Rock | I'm travelin' light travelin' light Just movin' on... |
| 156847 | drastic-actions-second-version | 2006 | bad-religion | Rock | Heard a word, suicide, Not from one, but from thou... |
| 215080 | otherworld | 2010 | eluveitie | Metal | Where ere thou hast been Here or in yon world mani... |
| 96212 | vernichtung-der-nazis | 2006 | c | Rock | Der erste wird getötet, der zweite kommt gleich dr... |
| 252030 | the-silence-thereafter | 2007 | craft | Metal | The feeling as if cold, jagged steel was carving y... |
| 92545 | another-one | 2016 | dvsn | Hip-Hop | She loves the way I love her She don't know what s... |
| 85948 | new-heights | 2009 | a-fine-frenzy | Pop | You are a thunderous thing, wondrous king The face... |

**- Quantidade de músicas e artistas por gênero musical:**

per\_genre <- group\_by(lyrics, genre) %>%  
 summarise(songs = n(),  
 artists = length(unique(artist))) %>%   
 arrange(desc(songs))  
  
table\_output <- flextable(as.data.frame(per\_genre))  
table\_output <- fontsize(table\_output, part = "all", size = 9)  
table\_output <- autofit(table\_output)  
table\_output

| genre | songs | artists |
| --- | --- | --- |
| Rock | 109234 | 3765 |
| Pop | 40465 | 2066 |
| Hip-Hop | 24845 | 1107 |
| Metal | 23759 | 996 |
| Country | 14387 | 370 |
| Jazz | 7971 | 336 |
| Electronic | 7966 | 837 |
| R&B | 3401 | 174 |
| Indie | 3149 | 295 |
| Folk | 2243 | 158 |

**- Quantidade de músicas e artistas por década em análise:**

lyrics$decade <- paste(str\_sub(lyrics$year, 1, 3), '0', sep = '')  
  
per\_decade <- group\_by(lyrics, decade) %>%   
 summarise(songs = n(),  
 artists = length(unique(artist))) %>%   
 arrange(desc(decade))  
  
table\_output <- flextable(as.data.frame(per\_decade))  
table\_output <- fontsize(table\_output, part = "all", size = 9)  
table\_output <- autofit(table\_output)  
table\_output

| decade | songs | artists |
| --- | --- | --- |
| 2010 | 62206 | 7112 |
| 2000 | 164982 | 5412 |
| 1990 | 6782 | 391 |
| 1980 | 1794 | 145 |
| 1970 | 1656 | 144 |

**- Número de músicas por gênero dos top 3 artistas baseado no número de músicas presente no dataset:**

per\_artist <- count(lyrics, genre, artist, sort = TRUE) %>%   
 group\_by(genre) %>%   
 arrange(desc(n)) %>%   
 filter(row\_number() <= 3) %>%   
 arrange(desc(genre), desc(n))  
  
table\_output <- flextable(as.data.frame(per\_artist))  
table\_output <- fontsize(table\_output, part = "all", size = 9)  
table\_output <- autofit(table\_output)  
table\_output

| genre | artist | n |
| --- | --- | --- |
| Rock | elton-john | 676 |
| Rock | bob-dylan | 596 |
| Rock | david-bowie | 570 |
| R&B | babyface | 367 |
| R&B | aretha-franklin | 282 |
| R&B | brian-mcknight | 214 |
| Pop | barbra-streisand | 607 |
| Pop | bee-gees | 591 |
| Pop | american-idol | 568 |
| Metal | fall | 368 |
| Metal | anthrax | 183 |
| Metal | cradle-of-filth | 183 |
| Jazz | ella-fitzgerald | 571 |
| Jazz | dean-martin | 560 |
| Jazz | frank-sinatra | 417 |
| Indie | devendra-banhart | 141 |
| Indie | dar-williams | 102 |
| Indie | the-blood-brothers | 85 |
| Hip-Hop | chris-brown | 628 |
| Hip-Hop | eminem | 578 |
| Hip-Hop | 50-cent | 398 |
| Folk | clannad | 194 |
| Folk | celtic-woman | 125 |
| Folk | gordon-lightfoot | 91 |
| Electronic | bjrthrk | 237 |
| Electronic | david-guetta | 149 |
| Electronic | everything-but-the-girl | 134 |
| Country | dolly-parton | 744 |
| Country | eddy-arnold | 591 |
| Country | bill-anderson | 466 |

## Estrutura Estratégica / Operacional

O NLP (*Natural Language Processing*) é um dos campos mais populares e promissores da inteligência artificial e machine learning. No longo prazo, é provavel que a grade maioria das empresas precisem fazer o uso de técnicas de mineração de dados e estar à frente dos **hot topics** do momento é extremamente estratégico nos preceitos empresariais.

Existem muitos projetos e publicações desmistificando o conceito de NLP. Operacionalmente, quando lidamos com os dados de textos, a abordagem fundamental é dividir o texto em palavras únicas. Daí, a mágica fica com a estatística, que aplica técnicas de contagens e distâncias de palavras na extração de insights relevantes para os negócios. Dentre outras, conseguimos modelar temas e sentimentos ao longo de textos.

Neste projeto, técnicas de topic modeling, TF-IDF, análise de sentimentos e contagens básicas de palavras serão aplicadas com o objetivo que poderá ser estudar causalidades de eventos, rede de relacionamentos de artistas, principais temas por categoria ou artista ou, até mesmo, encontrar correlações para avaliação de hits de sucesso.

## Indicadores de Avaliação

Os principais indicadores a serem usados serão:

* **Betweenss:** Avaliação de proximidade de nós (artistas) na análise de grafos;
* **Cliques:** Avaliação de clusteres de nós pela proximidade gráfica (cliques);
* **Índice de sentimento:** Avaliação de sentimento (usando dicionário de palavras) para entender a positividade/negatividade das mensagens das músicas;
* **TF-IDF:** Avaliação de palavras mais características de cada artista;
* **Frequência de tópicos:** Agrupamento de artistas com base em suas preferências temáticas .

## Adoção no ambiente empresarial (recorrência, implicação gerencial, etc)

Hoje em dia grande parte (senão a maioria) de todos os dados capturados pelas empresas são dados não estruturados, ou seja, gravação de ligações, e-mails, troca de mensagens, etc. E a partir desses dados é possível realizar uma ampla gama de análises, por exemplo o nível de satisfação do cliente com algum produto ou serviço. As vezes uma análise de uma gravação telefônica consegue captar melhor o sentimento do cliente do que se fosse respondida uma pesquisa de satisfação. Dado esse contexto é imprescindível que grandes empresas invistam fortemente nesse tipo de tecnologia e estudos. Atualmente empresas dedicam setores inteiros para a análise de informações desestruturadas no formato de texto falado ou escrito. Com o avanço das técnicas de speech-to-text, a avaliação de textos torna-se ainda mais relevante. Desta forma, entender os processos e métodos de mineração de textos e NLP são estratégias fundamentais para o futuro da maior parte dos negócios!

## Se possível: realizem Prova de Conceito

Iremos desenvolver a prova de conceito utilizando o dataset mencionado nas sessões anteriores como entregável da segunda parte desta disciplina ao Professor Gustavo Mirapalheta.

# Referências utilizadas neste trabalho

* [**Spotify Usage and Revenue Statistics (2019)**](https://www.businessofapps.com/data/spotify-statistics/) – Mansoor Iqbal
* [**Kaggle - 380,000+ lyrics from MetroLyrics**](https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics) – Gyanendra Mishra