Atividade 02 - Análise Exploratória

Relatório

1 - Tema:

SteamDE: Análise de *reviews* de jogos na plataforma Steam.

2 - Equipe:

- Daniel de Viveiros Inácio, 1923820, tredeneo, daniel.060798@alunos.utfpr.edu.br, BSI, UTFPR;
- Eduardo Darrazão, 1906399, eduponto21, darrazao@alunos.utfpr.edu.br, BSI, UTFPR;
- https://gitlab.com/eduponto21/steam-de

3 - Obtenção e processamento de dados:

Os dados contém *reviews* (majoritariamente em inglês) de jogadores de 48 jogos diferentes na plataforma Steam entre os dias 20/12/2010 até 16/02/2019, onde para cada *review*, temos os dados: data da postagem, título do jogo, *review* em si, se o autor do *review* recomenda ou não o jogo, quantas horas o autor jogou o jogo, quantas pessoas marcaram o *review* como engraçado, quantas pessoas marcaram o *review* como útil, e se o jogo estava em *early access* ou não.
Disponível em: https://www.kaggle.com/datasets/luthfim/steam-reviews-dataset

A limpeza dos dados se deu em três etapas:

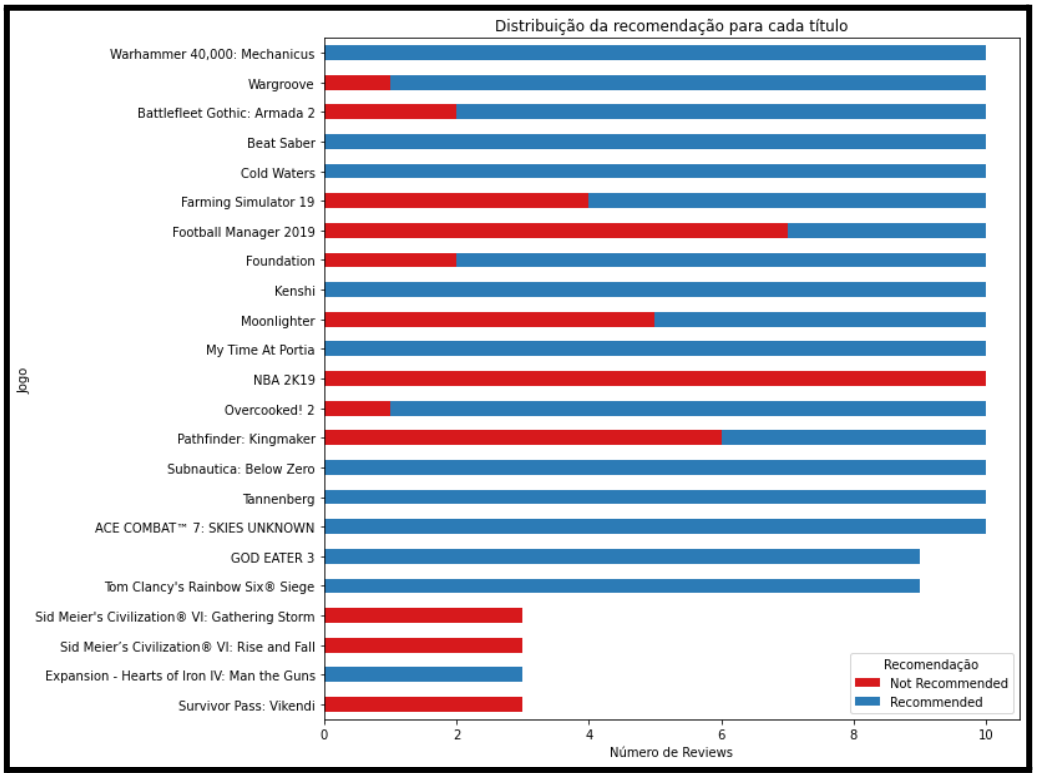
- Inicialmente removemos todas entradas com *reviews* nulos e vazios;
- Depois utilizamos a biblioteca *FastText* para classificar em que língua os *reviews* estavam escritos, dos quais mantemos somente os em inglês;
- Aplicamos técnicas de limpeza de texto para facilitar e melhorar análises futuras, como por exemplo: remover contrações, *tokenizar* as descrições, transformar tudo em letra minúscula, remover pontuações, remover *stopwords*, inferir *POS tags*, e por fim, lematizar as palavras.

Temos duas observações, primeiro que algumas palavras aparentemente não foram lematizadas corretamente, possivelmente devido a inacurácia das *POS tags* em casos específicos. Segundo que alguns reviews ficaram vazios após o passo de remoção de *stopwords,*os quais foram removidos. No começo haviam 434.891 entradas, das quais restaram 399.085 após os procedimentos citados.

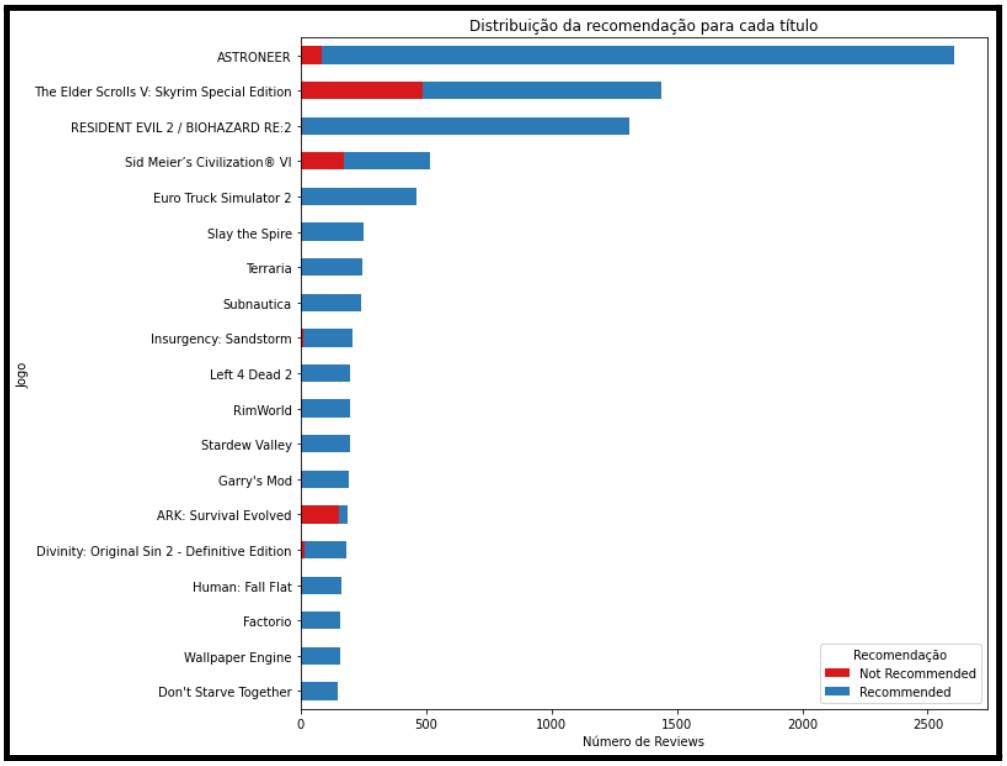
4 - Cobertura e distribuição dos dados:

Dentre os 48 jogos, podemos separá-los em três grupos referentes a quantidade de *reviews,* e ao mesmo tempo verificar a distribuição de recomendações dos usuários para os respectivos jogos:

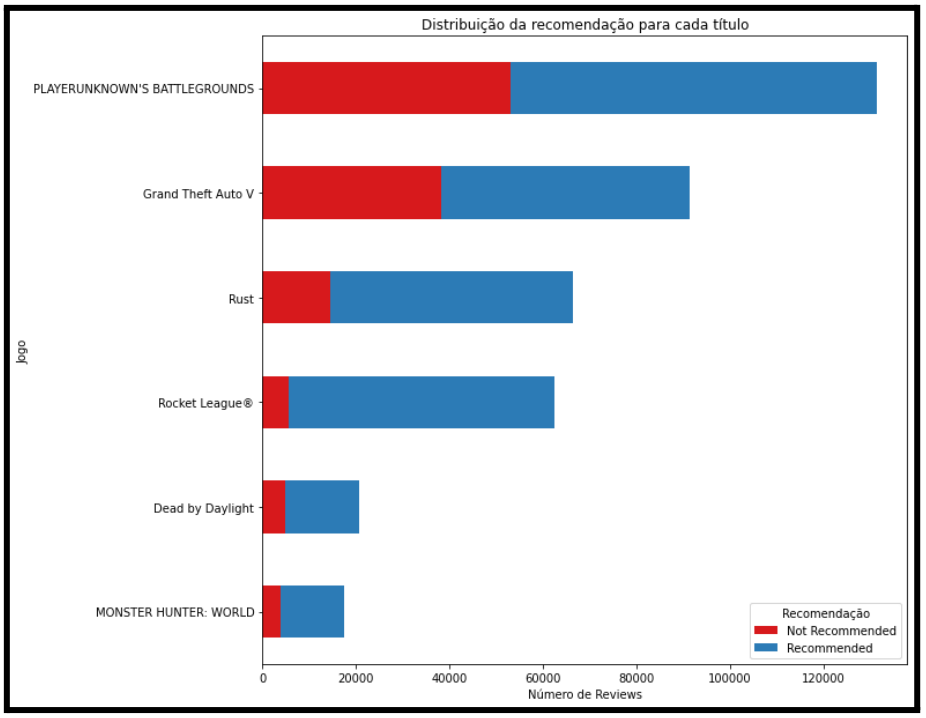
- Grupo A: Jogos com até 10 *reviews*.

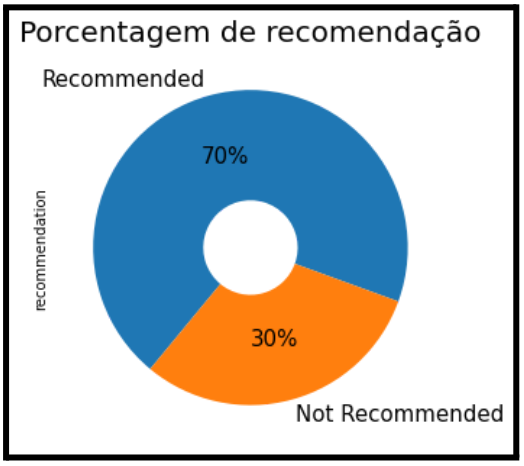


- Grupo B: Jogos entre 10 e 10.000 *reviews*.

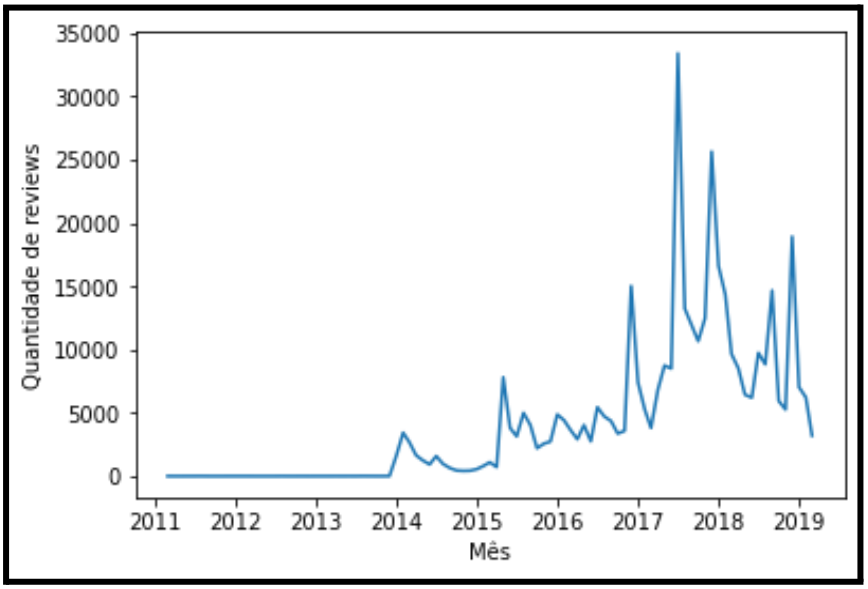


- Grupo C: Jogos com mais de 10.000 *reviews*.



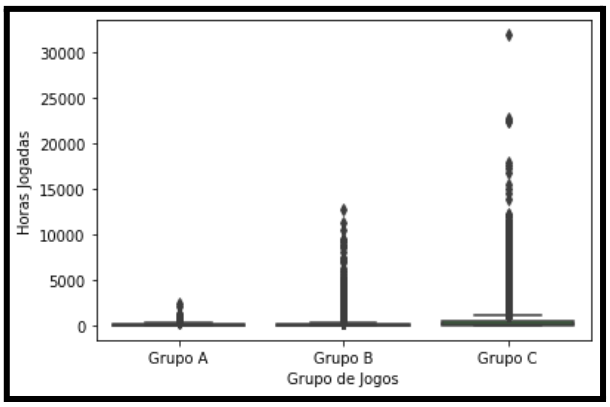


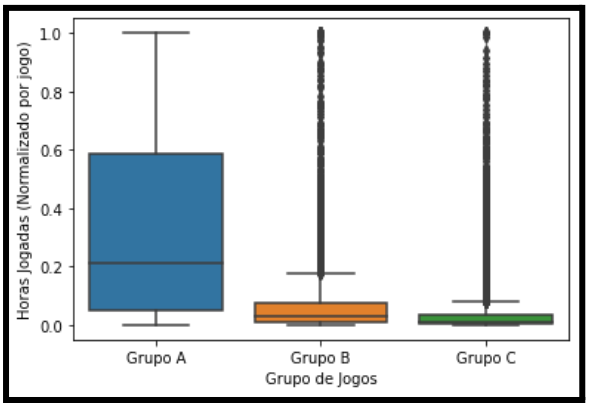
Ao todo, temos que 70% dos *reviews* recomendam os jogos que estão avaliando, enquanto 30% não os recomenda.

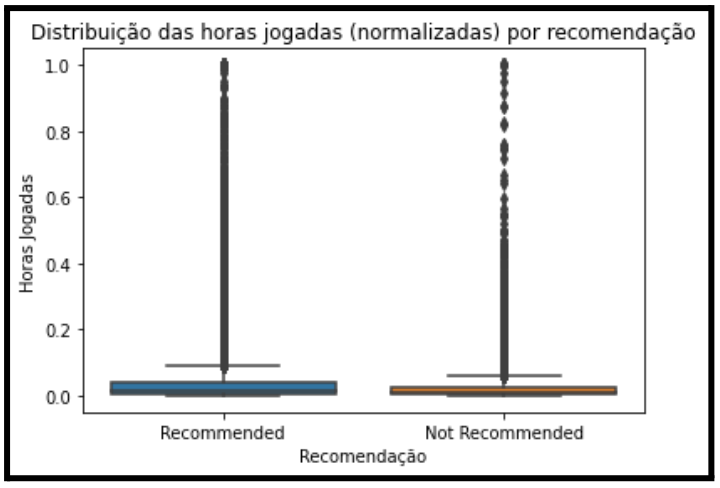


Na distribuição dos *reviews* agrupados por mês, podemos perceber que a maioria dos comentários foram criados de 2015 em diante, e os picos possivelmente podem estar relacionados com o lançamento de jogos ou grandes atualizações.

Os jogos com mais *reviews* tendem a ser jogados por mais horas (absolutas, esquerda), contudo, ao normalizar (por cada jogo, direita), percebemos que quanto mais pessoas jogam o mesmo jogo (levando em consideração somente as pessoas que jogaram e avaliaram tais jogos), menos horas são jogadas, em média.
Esse resultado possivelmente é devido ao fato que jogos mais populares atraem mais atenção das pessoas, levando-as a comprarem e comentarem sobre os jogos mais rapidamente, mesmo que não tenham investido muitas horas no jogo previamente.
Como no grupo A temos jogos com pouquíssimos *reviews* (ao comparar com os demais), provavelmente os retiraremos de análises futuras, visto que não apresentam informações suficientes para estudo, podendo assim considerá-los *outliers*.





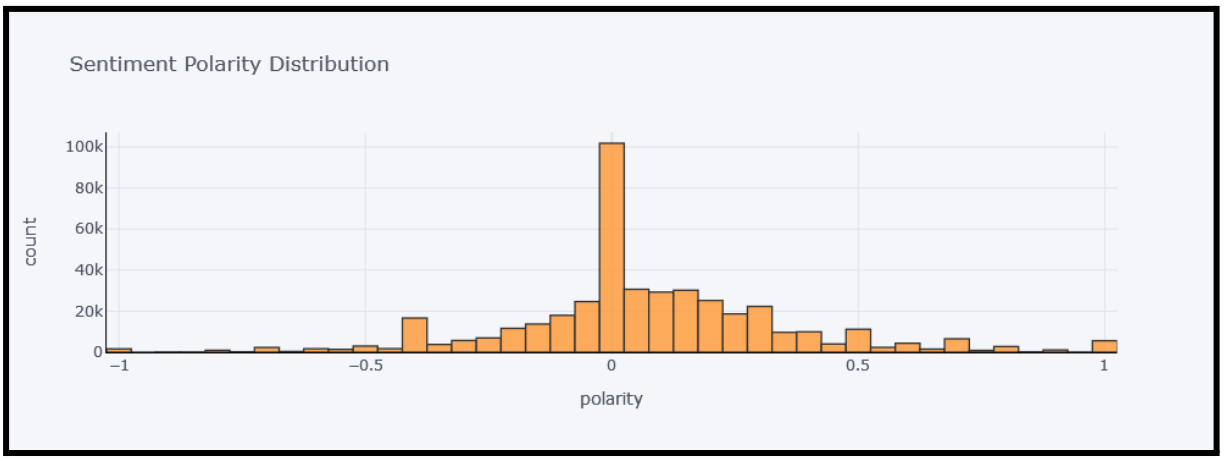


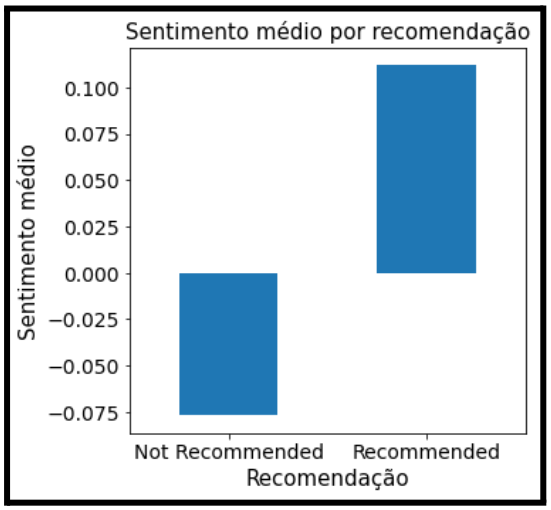
E é perceptível que pessoas que recomendaram os jogos tendem a jogá-los por mais horas.

5 - Análise Exploratória:

5.1 - Análise de Sentimentos:

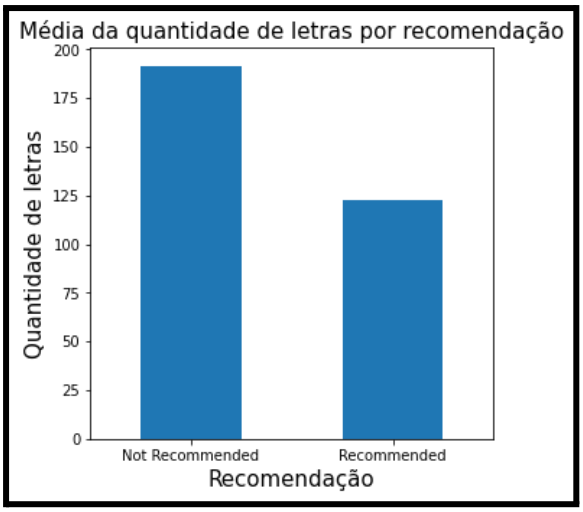
Realizamos a análise de sentimentos dos *reviews*, para ter uma melhor compreensão dos dados em questão. Na distribuição de polaridade verificamos que a maior parte são comentários positivos, assim como existem muito mais extremos positivos do que negativos.

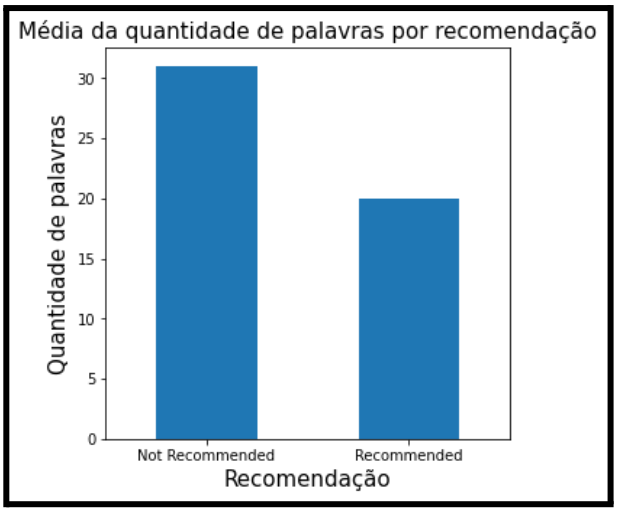


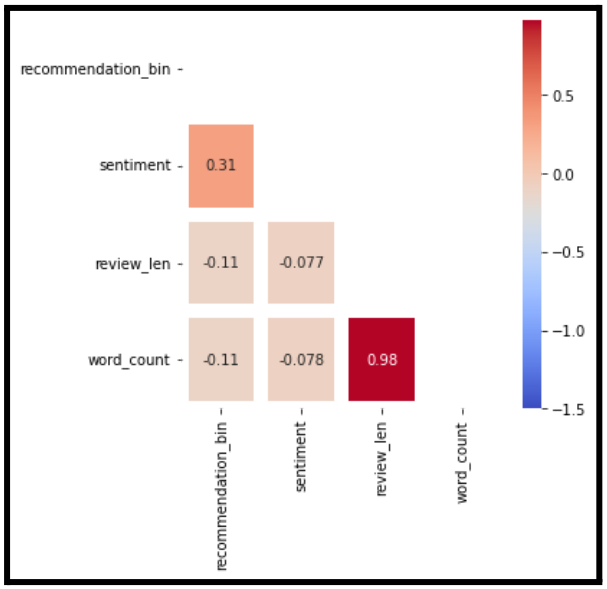


Observamos também como se comporta o sentimento entre reviews recomendando e os que não estão. Apesar de as médias serem próximas de zero (possivelmente um efeito de muitos *reviews* serem neutros), percebemos que os que recomendaram foram levemente positivos e os que não recomendaram, levemente negativos.

Outra *feature* dos *reviews* que podemos comparar entre os que recomendaram e os que não, é as relações de tamanho de textos nos *reviews* e a quantidade de palavras, onde verificamos que *reviews* recomendando tendem a ser significativamente menores.





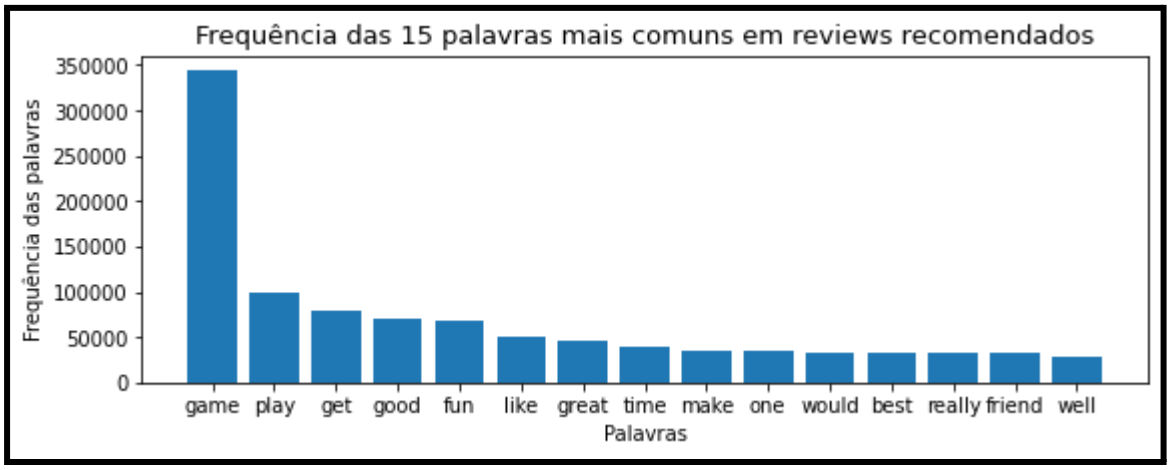


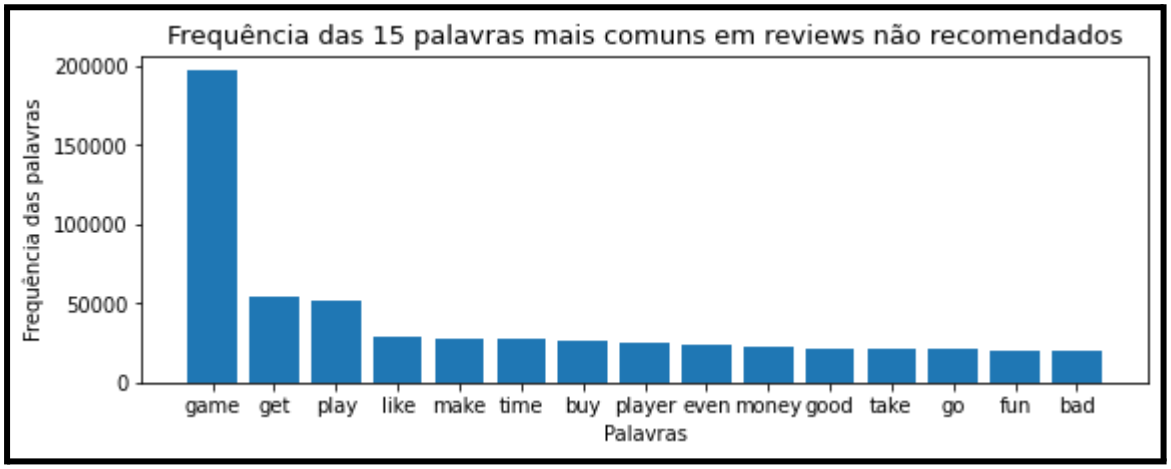
“Mas será que isso realmente faz sentido?” Essa pergunta é o que a correlação nos responde. Para isso, transformamos a recomendação em um binário: 1 para recomendado, 0 para não. Assim vemos que ser um *review* recomendando um jogo tem impacto positivamente no sentimento do texto, assim como negativamente no seu tamanho, logo, confirmando nossas presunções.

5.2 - Análise de Tópicos:

Com o intuito de explorar sobre o que os *reviews* estão comentando, destacamos as palavras mais comuns, e posteriormente procuramos possíveis diferenças entre os que estão recomendando os jogos e os que não. Apesar de uma alta intersecção, observamos a presença das palavras *like, great, best, friend* somente nos recomendados, e *buy, money, player, bad* somente nos não recomendados.

Entre as palavras 'positivas', além de adjetivos de qualidade positivos há 'amigo', ressaltando o aspecto positivo da socialização intrínseca aos jogos. Já entre as 'negativas', há palavras como 'comprar' e 'dinheiro', evidenciando características de jogos muitas vezes vistas como ruins pela comunidade, como *pay to play* e *pay to win*, assim como adjetivo de qualidade negativo, 'ruim'.



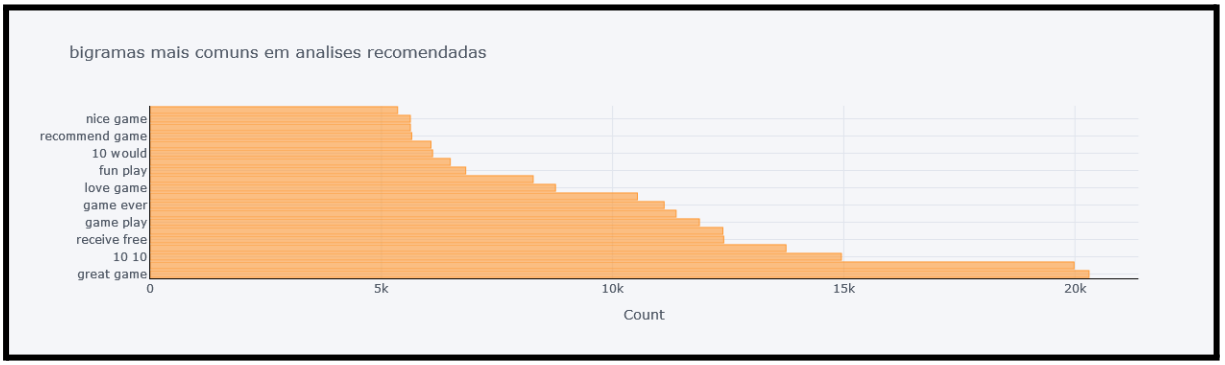


Ampliando o escopo, exploramos os bi-trigramas dos *reviews* recomendando e não recomendando os jogos separadamente. Essa análise traz novas perspectivas para ambos os lados.

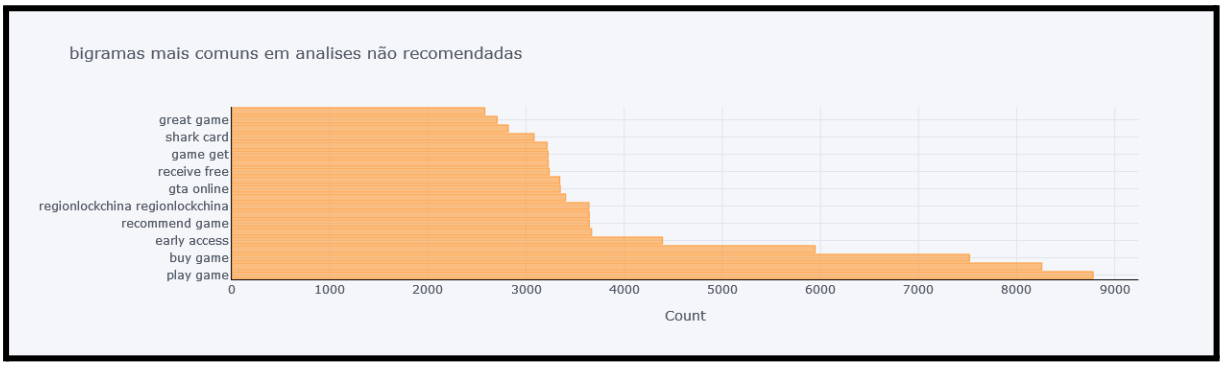
Nos *reviews* que estão recomendando o jogo em questão, percebemos que ter recebido o jogo em si ou algo dentro do jogo de graça é algo bem abordado, assim como o fato de certo jogo ser do estilo *battle royale.* É reforçado o aspecto de jogar com amigos, assim como a expressão comum na comunidade de '10/10, would buy again'.

Já nos que não estão recomendando o jogo, vemos críticas ao jogo ‘Grand Theft Auto V’, como o *shark card*, uma *feature* considerada *pay to win*, e possíveis reclamações quanto a versão online do jogo. Há também indícios de críticas a telas de carregamento, as quais acabam sendo muito longas em alguns jogos, e reclamações a respeito de *hackers* e *cheaters*, que é uma temática negativa comum em jogos onlines. Também nesses reviews há tópicos que aparentam estarem relacionados a qualidade positivas, como ‘g*reat game’, ‘receive free’* e *‘would recommend game’*, que investigaremos futuramente.
Algo observado em ambos os casos é a presença do tópico ‘*regionlockchina’,* que para estar sendo citado massivamente, pode ter sido uma grande revolta da população ou até um ‘ataque’ de bots, o que também investigaremos para elucidar o que ocorreu.

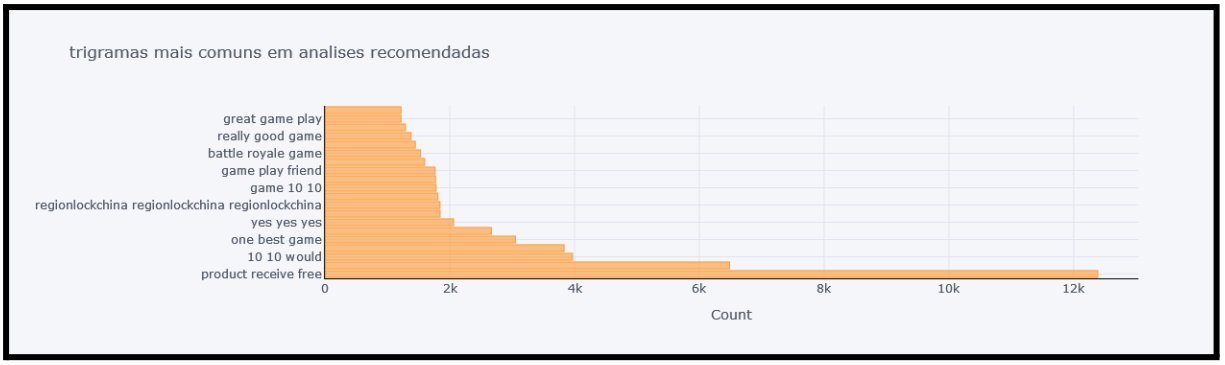
- Bigramas mais comuns de análises que recomendaram o jogo:



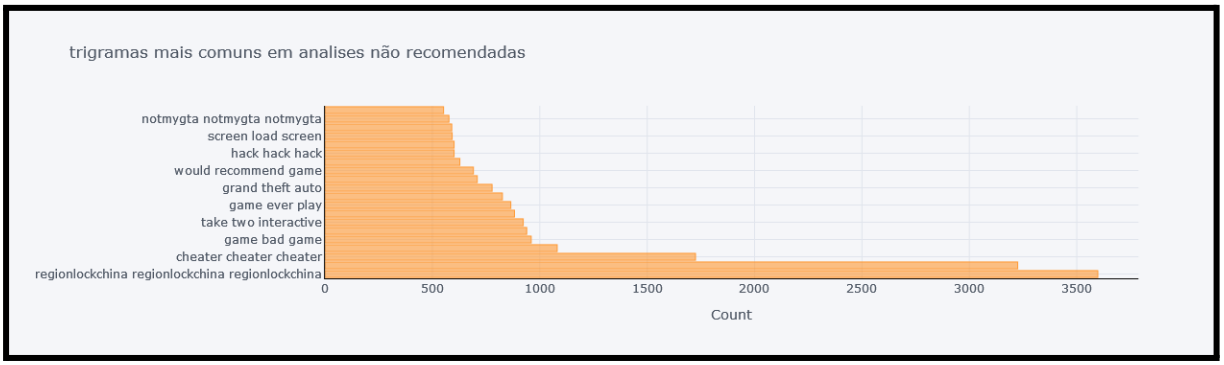
- Bigramas mais comuns de análises que não recomendaram o jogo:



- Trigramas mais comuns de análises que recomendaram o jogo:



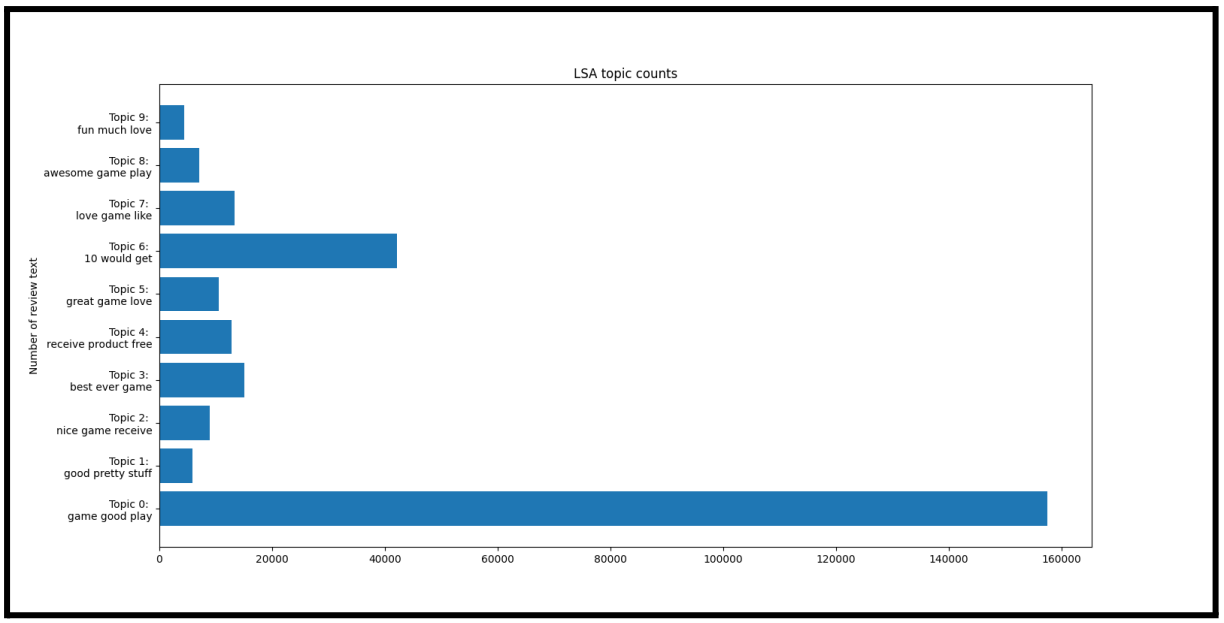
- Trigramas mais comuns de análises que não recomendaram o jogo:



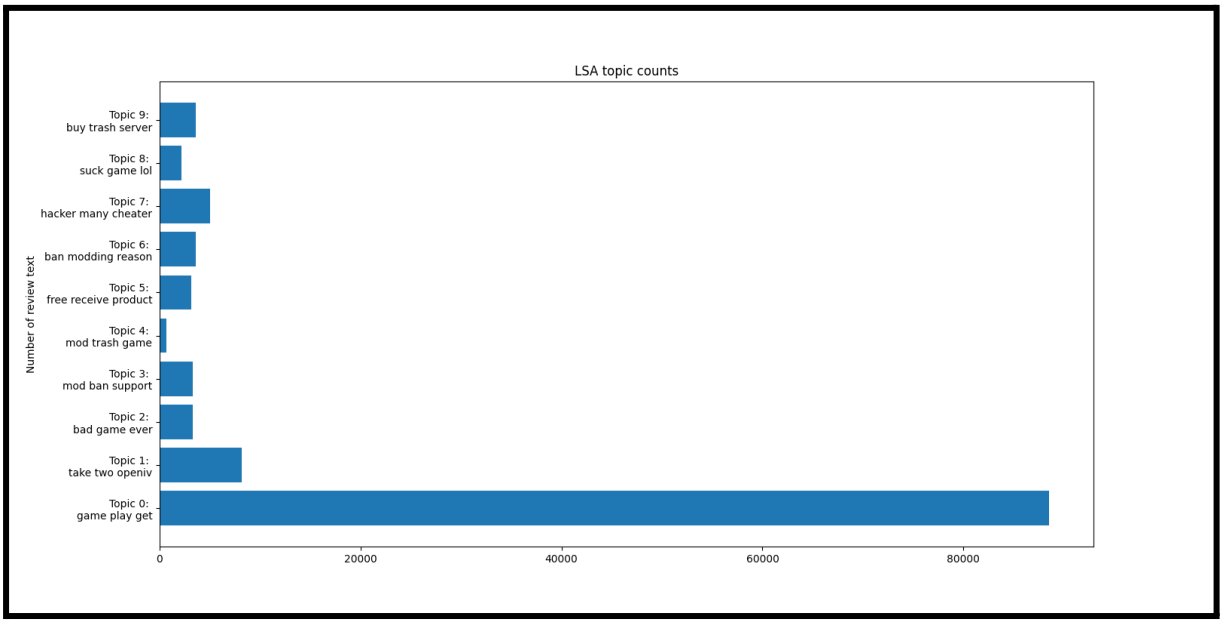
Uma outra forma de analisar tópicos é utilizando algoritmos mais robustos, como Latent Semantic Analysis (LSA) e Latent Dirichlet Allocation (LDA). Devido a limitações de poder computacional e problemas técnicos, utilizamos apenas o LSA, o qual acreditamos ser o suficiente, visto que já nos fornece bons resultados e que corroboram as análises anteriores.

Assim sendo, modelamos os tópicos mais uma vez para os *reviews* que recomendam e os que não, e comparamos os tópicos extraídos a partir do LSA. Entre os recomendados, não surgiu nada novo. Contudo, entre os não recomendados, observamos críticas a servidores dos jogos e em relação aos '*mods'*, que normalmente é referência a '*plugins*' criados pela comunidade que permite alterar aspectos dos jogos, mas pode também referenciar as pessoas que gerenciam o jogo, os moderadores.

- Tópicos mais comuns de análises que recomendaram o jogo:



- Tópicos mais comuns de análises que não recomendaram o jogo:



6 - Perguntas de pesquisa e explorações iniciais:

6.1 - Análise de sentimentos dos *reviews*:

Como se comportam dependendo se está recomendando ou não o jogo em questão, se tem relação com o total de horas jogadas.

Todas as informações necessárias para encontrar se existe ou não alguma relação está disponível na base de dados e nos resultados das análises iniciais.

6.2 - Análise de tópicos:

A: Analisar tópicos pertinentes aos jogos, sobre o que se trata, como evoluem ao longo do tempo.
B: Investigar se os tópicos de cada *review* possuem relação com o tempo de jogo.
C: Investigar se os tópicos dos *reviews* entre jogos com características similares possuem intersecção.

Será possível analisar os tópicos, contudo na questão da evolução ao longo do tempo, talvez não seja possível, uma vez que a base de dados possui informação da data em que a análise foi escrita, porém não são todos os jogos que poderão ser analisados, pois alguns têm poucos dados e analisá-los em períodos específicos diminuiria ainda mais para estes jogos.

Na comparação com o tempo de jogo, é possível separarmos grupos de *reviews* com tempo de jogo parecidos, mas em alguns casos pode ocorrer o mesmo citado anteriormente. E na questão da análise de jogos parecidos, a base de dados não possui categoria dos jogos, então será necessário a classificação manual, que é viável devida a baixa quantidade de títulos distintos.

6.3 - Análise geral:

A: Analisar se a aceitação dos jogos impacta na quantidade de horas jogadas;
B: Analisar se os picos de novos *reviews* possuem relação com novos lançamentos ou atualizações grandes.

A primeira já foi praticamente respondida na análise exploratória, e a segunda será necessário mais algumas informações sobre os jogos, como data de lançamento e datas de grande mudanças, se houver. Mas assim como a classificação das categorias que os jogos se enquadram, é possível realizar a verificação manual devida a baixa quantidade de títulos distintos.

7 - Discussão e próximos passos:

Conseguimos identificar diversos pontos interessantes, e pelo menos uma pergunta inicial já foi respondida. Quanto à completude dos dados, não serão necessárias novas bases, uma vez que temos praticamente todas informações essenciais na base atual, entretanto precisaremos realizar uma categorização manual, como citado anteriormente.

Teremos que melhorar nossos filtros pois algumas análises realizadas com dados já filtrados não foram capazes de identificar palavras corretamente, algumas palavras não foram lematizadas e foram consideradas tokens separados, e reticências e dois pontos (‘..’) apareceram nas palavras mais frequentes, mesmo quando utilizado os dados já limpos, assim precisamos verificar como melhorar.

Exploraremos outras opções para as análises de tópicos: conseguimos utilizar a análise semântica (LSA), todavia outros métodos como LDA e NMF tiveram problemas técnicos, então buscaremos outras formas/bibliotecas para implementar essa tarefa e investigar se há resultados mais relevantes que a análise semântica.

Algumas perguntas iniciais tiveram pequenas alterações para adequação após a análise exploratória dos dados, mas elas permanecem com a mesma ideia, por isso acreditamos que não será necessário a mudança de tema.