Atividade 03 - Projeto e Correlatos

Documento de Projeto

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 - Perguntas de pesquisa:

1. Como o sentimento dos *reviews* se comportam quando estão - ou não - recomendando o jogo analisado? Tem alguma relação com a quantidade de horas jogadas?
2. Quais tópicos são abordados nos *reviews* dos jogos mais populares? Há alguma evolução deles ao longo do tempo?
3. Os tópicos dos *reviews* possuem alguma relação com a quantidade de horas jogadas?
4. Jogos com características similares possuem intersecção de tópicos abordados?
5. Qual é a relação entre a aceitação do jogo (medida pela porcentagem de recomendação dos *reviews*) e a quantidade de horas jogadas? Jogos com mais recomendações tendem a ser jogados por mais tempo?

4 - Hipóteses:

1. Como os *reviews* se comportam quando estão - ou não - recomendando o jogo analisado? Tem alguma relação com a quantidade de horas jogadas?
    a. Os *reviews* que recomendam o jogo possuem um sentimento mais positivo em comparação com os que não recomendam;
    b. Jogadores que investem mais horas no jogo tendem a ter um sentimento mais positivo sobre o jogo, e vice-versa;
    c. Um *review* com sentimento positivo e com muitas horas jogadas está recomendando o jogo;
2. Qual é a relação entre a aceitação do jogo (medida pela porcentagem de recomendação dos *reviews*) e a quantidade de horas jogadas? Jogos com mais recomendações tendem a ser jogados por mais tempo?
    a. A qualidade percebida do jogo (medida pela porcentagem de recomendação dos *reviews*) pode influenciar a motivação dos jogadores para continuar jogando e, portanto, uma maior média de quantidade de horas jogadas;
3. Quais tópicos são abordados nos *reviews* dos jogos mais populares? Há alguma evolução deles ao longo do tempo?
    a. Os tópicos mais abordados nos reviews dos jogos populares são jogabilidade, gráficos, e história;
    b. A evolução dos tópicos ao longo do tempo está relacionada a novas atualizações nos jogos;
4. Os tópicos dos *reviews* possuem alguma relação com a quantidade de horas jogadas?
    a. Jogadores que investem mais horas em um jogo tendem a abordar tópicos relacionados à jogabilidade e progressão no jogo;
    b. Jogadores que jogam por menos tempo podem abordar tópicos mais gerais, como *bugs*;
5. Jogos com características similares possuem intersecção de tópicos abordados?
    a. Jogos com características similares tendem a ter uma intersecção significativa nos tópicos abordados em seus reviews;
    b. A intersecção de tópicos em reviews de jogos com características similares pode ser explicada pelas similaridades nos elementos de jogabilidade e design;
    c. Jogos que apresentam inovações ou novas abordagens em seus gêneros podem ter tópicos específicos abordados em seus *reviews*.

5 - Dados e modelos:

1. **Dados e modelos:** Descrever os dados e modelos que serão usados para responder as perguntas e/ou testar as hipóteses.

Para responder às perguntas e testar as hipóteses levantadas, serão utilizados dados de reviews de jogos coletados da plataforma Steam, disponível no Kaggle[1]. 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. Os dados citados já foram majoritariamente limpos e preparados para as próximas etapas.

Aplicaremos um modelo de análise de sentimentos da biblioteca *TextBlob* para classificar o sentimento geral de cada *review* numa escala entre menos 1 e -1, onde um é positivo, menos um é negativo e zero representa neutro, e utilizaremos modelos estatísticos para testar algumas das nossas hipóteses, como regressão logística para analisar a influência da quantidade de horas jogadas e o sentimento do review na recomendação ou não de um jogo, e regressão linear para ver o efeito da quantidade de horas de jogo no sentimento do *review* realizado, e o efeito da aceitação do jogo (medida pela porcentagem de recomendação dos *reviews)* na quantidade de horas jogadas.

Para analisar os tópicos mais abordados nos reviews, utilizaremos técnicas de modelagem de tópicos para cada jogo isoladamente, como o *Latent Dirichlet Allocation* (LDA, probabilístico), e BERTopics (semântico). É necessário realizar a análise separadamente para cada jogo para prevenir que jogos com mais comentários influenciem nos tópicos dos demais jogos, e isso permitirá agrupar os reviews em tópicos comuns e identificar as palavras-chave mais relevantes para cada tópico nos passos seguintes.

Aliado a poder detectar a similaridade de tópicos entre jogos, executaremos análises de frequência para identificar os tokens mais importantes dos tópicos abordados nos reviews, e análises de clusterização (como *K-means*), onde representaremos os tokens de cada tópico de cada jogo no espaço vetorial por meio de *Word Embeddings,* o que permitirá identificar padrões e intersecções entre os tópicos abordados nos reviews de jogos com características similares. Para validar quais tópicos entre jogos diferentes com características similares são de fato parecidos ou não, realizaremos uma análise de variância (ANOVA), por exemplo. Com tudo isso feito, é possível escolher partes dos dados para testar diferentes hipóteses, como agrupar por tempo de jogo, temporalmente e jogos com características similares.

Por fim, utilizaremos gráficos e visualizações de dados para comunicar nossos resultados de forma clara e acessível, como gráficos e tabelas, para apresentar os resultados e *insights* obtidos a partir da análise dos dados.

[1] https://www.kaggle.com/datasets/luthfim/steam-reviews-dataset

6 - Cronograma:

1. **Limpeza e filtro dos dados**: uma parte dos dados ainda não está completamente tratado, então a filtragem e seleção dos *reviews* e jogos que serão trabalhados é a primeira coisa a ser feita;
2. **Classificação:** é necessária uma classificação manual dos jogos em categoria/gênero, separação temporal dos *reviews* dos jogos mais populares, e separação de grupos que jogam muito e jogam pouco (por jogo);
3. **Regressão:** faremos regressões e correlações para verificar os itens das hipóteses 1 e 2;
4. **Identificação e processamento tópicos:** realizaremos a análise de tópicos para cada jogo e os procedimentos citados respectivamente na seção 5.
    a. Identificar os *tokens* mais relevantes de cada tópico;
    b. Transformar os *tokens* de cada tópico de cada jogo em uma representação vetorial;
    c. Realizar análises de clusterização nos tokens para buscar similaridade dos tópicos conforme hipótese 5;
5. **Tópicos temporais:** com todos os tópicos calculados e processados, faremos a interpretação deles por tempo para responder o item na hipótese 3-b;
6. **Tópicos separados por tempo de jogo:** com os tópicos calculados com a separação dos dados entre um grupo que jogou muito e outro grupo que jogou pouco, faremos a interpretação do resultado para responder os itens na hipótese 4;
7. **Similaridade:** usando a análise de variância, poderemos validar a similaridade entre os categorias diferentes identificadas nas etapas anteriores;
8. **Visualização**: após realizar todos os procedimentos, utilizaremos gráficos e visualizações de dados distintas para compartilhar nossas descobertas.