Análise Exploratória de Dados de Comentários e Publicações no Instagram

Objetivo

Realizar uma análise exploratória dos dados de comentários e publicações do Instagram de usuários, utilizando Python para análise de dados e conceitos de estatística com foco na fixação dos conceitos e prática. O escopo da atividade é aberto, e espera-se que, como futuros bons analistas de dados, vocês extrapolem as opções sugeridas e reportem as análises em uma breve apresentação no formato PDF. Note que essa atividade **não** é avaliativa. Portanto, a realização da mesma é opcional, mas **fortemente** recomendada.

Base de Dados

Os dados foram coletados de perfis públicos no Instagram e incluem contextos políticos e não políticos. Os dados cobrem o período de 2018-09-01 até 2019-11-10. Portanto, textos inapropriados podem estar presentes. Além disso, a publicação destes dados em repositórios públicos não está autorizada, já que contêm nomes de pessoas e outras informações que podem ser consideradas sensíveis. Os dados podem ser baixados **clicando aqui**. A base é muito grande, mas vocês podem focar em um período de interesse. Por exemplo, por ser uma base política, você pode querer explocar um período próximo ao primeiro e ao segundo turno das eleições da época ou analisar apenas alguns aspectos de forma temporal.

Descrição da Atividade

1. Pré-processamento dos Dados

- Importar e carregar os dados utilizando bibliotecas como Pandas.
- Limpar os dados, tratando valores ausentes e duplicados.
- Converter colunas de datas para o formato `datetime` para facilitar a análise temporal.

2. Análise Exploratória

Distribuição de Frequência e Estatísticas Descritivas

- Aplicar técnicas estatísticas para resumir os dados e identificar padrões, tendências e anomalias. Calcular medidas de tendência central (média, mediana, moda) e de dispersão (variância, desvio padrão, amplitude), etc.
- Identificar e visualizar a distribuição das publicações e dos comentários ao longo do tempo.
- Utilizar histogramas, box plots e CDFs para visualizar a distribuição dos dados.
- Calcular a covariância e o coeficiente de correlação para investigar possíveis associações entre variáveis.

Análise Temporal

- Criar séries temporais para analisar tendências e padrões sazonais nos dados de publicações e comentários. Explorar distribuições, plots e métricas discutidas na aula.
- Identificar períodos de alta e baixa atividade.
- Utilizar gráficos de linha para visualizar essas tendências.

Análise dos Usuários

- Identificar os usuários mais ativos em termos de publicações e comentários.
- Analisar a popularidade dos conteúdos publicados (curtidas, comentários).
- Utilizar gráficos de barras e scatter plots para visualizar a atividade dos usuários.
- Analisar as distribuições de popularidades dos usuários, das postagens, etc.

Visualização de Dados

- Utilizar bibliotecas de visualização como Matplotlib e Seaborn para criar gráficos (barras, linhas, dispersão, calor, etc) para representar a distribuição, composição, tendência e associação dos dados.

Entrega

- Entregar uma breve apresentação com as principais análises e visualizações.
- Interpretar os resultados encontrados e destacar resultados interessantes obtidos a partir dos dados.

Considerações Finais

O sucesso desta atividade dependerá da sua capacidade de explorar os dados de maneira criativa e eficiente (já que os dados são grandes), utilizando as ferramentas e técnicas aprendidas nas aulas.