Análise de base de dados de e-mails

Caroline Salvador
caroline.farias.salvador@gmail.com

Introdução

SPAM é uma mensagem eletrônica não solicitada que contém conteúdo de marketing digital e pode ser encaminhada para vários destinatários através de e-mails, WhatsApp, blogs e rede sociais. Levando em consideração o crescente número de SPAM encaminhado diariamente, este artigo apresenta a análise de uma base de dados que contém informações de e-mails. O desenvolvimento da análise foi feito utilizando a linguagem de programação Python em conjunto com a biblioteca Pandas que permite a manipulação e análise de dados.

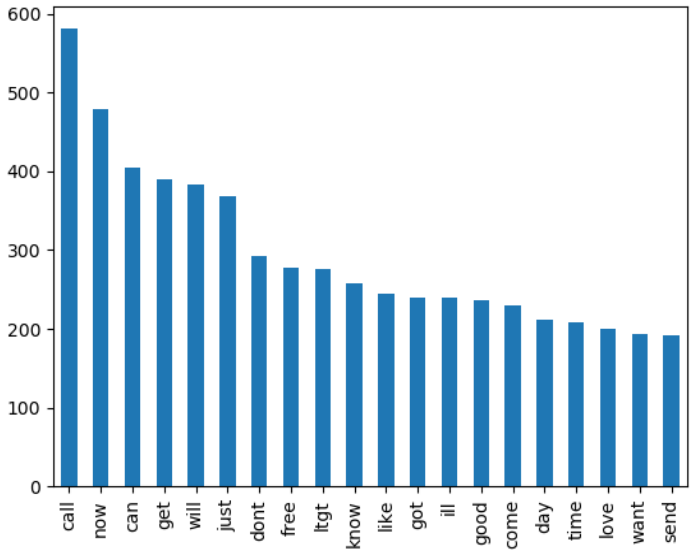
Resultados

O conjunto de dados utilizado para análise foi extraído do arquivo sms_senior.csv. Este arquivo contém várias mensagens comuns (4827 mensagens) e mensagens SPAM (747 mensagens) que foram submetidas a uma etapa de mineração de texto para identificar a frequência das palavras no corpo dos e-mails.
Os seguintes atributos constituem o arquivo:

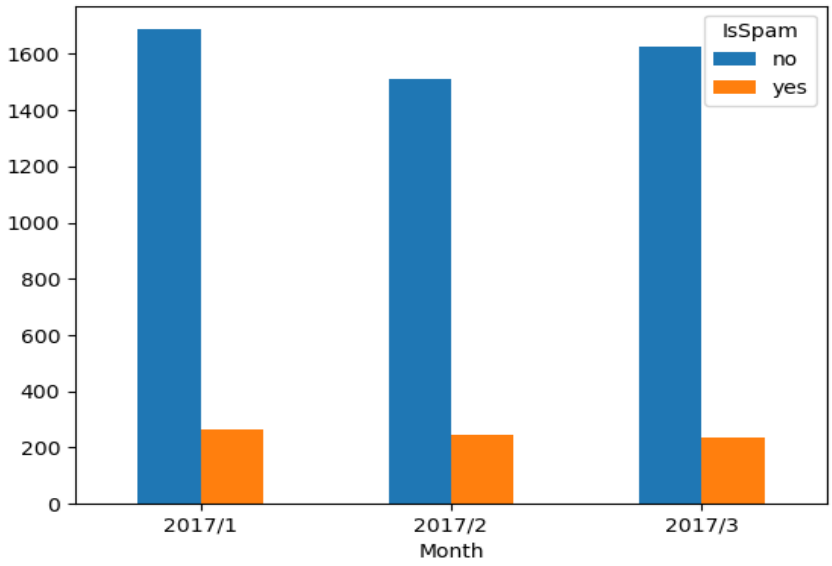
- Coluna contendo a mensagem original (*Full_Text*);
- 149 colunas com valores inteiros que indicam a frequência de uma determinada palavra na mensagem (*"got"... "wan"*);
- 1 coluna contendo a quantidade de palavras frequentes na mensagem (*Common_Words_Count*);
- 1 coluna contendo a quantidade total de palavras da mensagem (*Word_Count*);
- 1 coluna contendo a data de recebimento da mensagem (*Date*);
- 1 coluna que identifica se a mensagem é spam ou não (*IsSpam*).

Análise

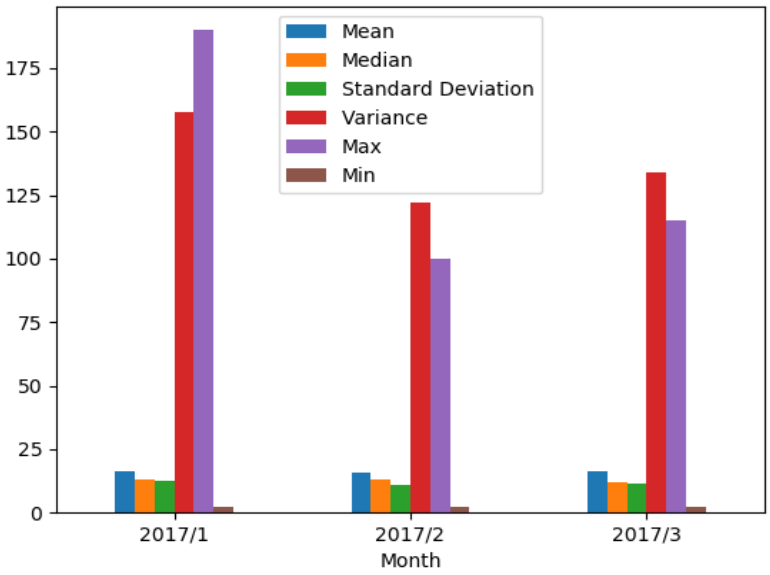
1. As palavras mais frequentes em toda a base de dados: O gráfico abaixo foi gerado utilizando o subconjunto das 20 primeiras palavras mais frequentes do conjunto (*"got"... "wan"*):



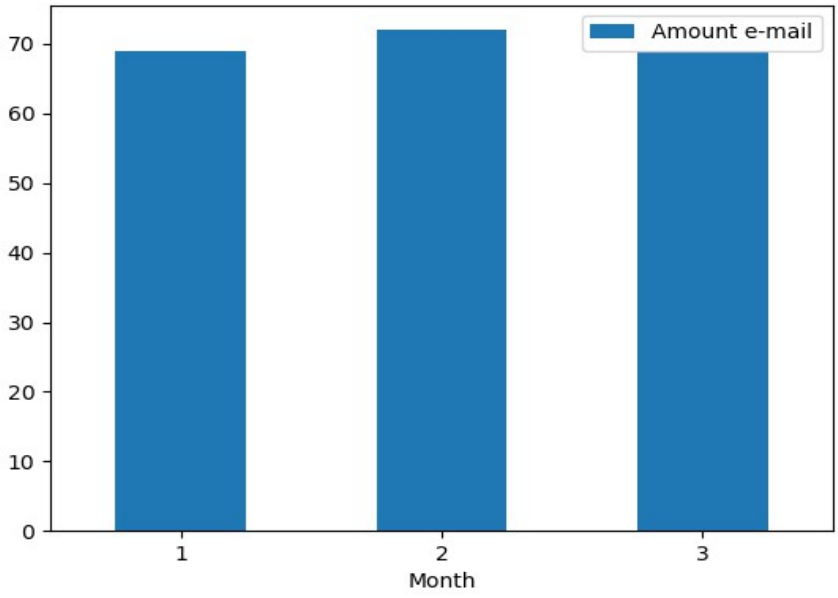
2. As quantidades de mensagens comuns e SPAMs para cada mês: O gráfico abaixo foi gerado utilizando o agrupamento mensal e categorização (*IsSpam*) dos e-mails como mostra a legenda . Neste gráfico é possível observar que a variação mensal de SPAM é pequena.



3. Calcular medidas máximo, mínimo, média, mediana, desvio padrão e variância da quantidade total de palavras (*Word_Count*) para cada mês. O gráfico abaixo apresenta técnicas de estatística descritiva como: medida de tendência central (média, mediana), medida de variabilidade (desvio padrão, variância) e os valores máximos e mínimos.



4. Exibir o dia de cada mês que possui a maior sequência de mensagens comuns (não SPAM). O gráfico abaixo foi gerado usando agrupamento de mês/dia e filtro para buscar somente e-mails comuns.



Conclusão

Os resultados obtidos apresentam pouca variação entre quantidade de SPAM e e-mails comuns recebidos dentro do intervalo de tempo de um mês. Também foi possível observar que e a contagem de palavras não gera informações relevantes e a palavra que mais aparece nos e-mails e a "*call*", tanto para SPAM como e-mail comum.
O cenário mais relevante é o que apresenta as quantidades de e-mails comuns e SPAM para cada mês, permitindo verificar o aumento de recebimento de SPAM mensalmente.

Referências

Wikipédia. **Estatística descritiva**. 2019. Disponível em: <https://pt.wikipedia.org/wiki/Estat%C3%ADstica_descritiva> Acesso em: 20 fev. 2020.

GALVÃO, Felipe. **Estatística Descritiva com Python**. 2016. Disponível em: <http://felipegalvao.com.br/blog/2016/03/31/estatistica-descritiva-com-python/>. Acesso em: 20 fev. 2020.

ALECRIM, Emerson. **O que é SPAM e como evitá-lo?**. 2019. Disponível em: <https://www.infowester.com/spam.php>. Acesso em: 20 fev. 2020.