Análise de Mensagens para Identificação de Spam

Roberto Luiz Debarba

roberto.debarba@senior.com.br

Introdução

Atualmente, a quantidade de mensagens não solicitadas (spam) que recebemos diariamente continua crescendo. Os tipos de mensagens são diversos e englobam anúncios de produtos, correntes, conteúdo ilícito, entre outros. Segundo Awad e Elseuofi (2011) estatísticas recentes apontam que 40% dos e-mails são spam, gerando 15,4 bilhões de mensagens por dia, totalizando um custo de US$355 milhões de dólares por ano.

> Filtragem automática de e-mails aparenta ser o método mais efetivo de contenção de spams no momento e uma competição apertada entre geradores de spam e métodos de filtragem está ocorrendo. Geradores de spam começaram a usar diversos métodos complicados parar superar os métodos de filtragem como o uso aleatório de endereços de remetente e/ou a adição de caracteres aleatório no início ou fim da linha de assunto da mensagem. Engenharia do conhecimento e aprendizado de máquina são duas abordagens comuns usadas na filtragem de e-mails. Na abordagem da engenharia do conhecimento um conjunto de regras devem ser especificadas de acordo com quais e-mails são considerados spam ou comuns. A abordagem de aprendizado de máquina é mais eficiente que a abordagem de engenharia do conhecimento; ela não requer a especificação de nenhuma regra. Ao invés disso, é usado um conjunto de exemplos de treinamento, sendo esses um conjunto de mensagens de e-mail pré-classificadas. (AWAD; ELSEUOFI, 2011, p. 173, tradução nossa).

Diante do exposto, este trabalho apresenta técnicas de análise de mensagens para categorização de spam. Também é apresentado o estudo e implementação do algoritmo de classificação Naive Bayes para identificação automática dessas mensagens.

Metodologia

Buscando entender o cenário de mensagens atual com o objetivo de desenvolver um método de classificação automática de spam, realiza-se uma análise e implementação sobre um conjunto de dados contendo 4827 mensagens comuns (não spam) e 747 mensagens spam. O conjunto está estruturado no formato *.csv* e contém os seguintes dados: (i) mensagem original; (ii) frequência de determinada palavra; (iii) quantidade de palavras frequentes; (iv) quantidade total de palavras; (v) data de recebimento; (vi) identificador de spam.

A primeira etapa consiste na extração de informações a partir do conjunto de dados. Para a realização da tarefa, foi implementado um algoritmo na linguagem de programação Python executado sobre a plataforma Jupyter Notebook, que permitiu a execução e visualização dos dados de forma rápida durante o desenvolvimento da rotina. Também foram utilizadas as bibliotecas Pandas, para manipulação do conjunto de dados, Matplotlib, para impressão dos gráficos e Scikit-learn, para execução do modelo de classificação das mensagens.

O gráfico com as palavras mais frequentes em todo o conjunto é gerado através da soma de cada em palavra em todas as mensagens seguida da ordenação pela quantidade, gerando uma classificação de frequência de uso. O gráfico com a quantidade de mensagens comuns e spam para cada mês é obtido através do agrupamento das mensagens por mês e da agregação da quantidade através de soma dos itens agrupados.

O cálculo de máximo, mínimo, média, mediana, desvio padrão e variância usa o agrupamento das mensagens por mês e agrega as mensagens de cada grupo com cada operação listada. Por fim, para exibir o dia de cada mês que possui a maior sequência de mensagens comuns, é realizado um agrupamento por dia de recebimento das mensagens aplicando uma agregação que identifica qual a maior sequência de mensagens comuns. É aplicado novamente sobre o resultado um agrupamento por mês com a agregação mantendo apenas o dia com a maior quantidade de mensagens.

A segunda etapa consiste no desenvolvimento de um método capaz de classificar automaticamente as mensagens comuns e spam. Para execução foi utilizado um algoritmo de aprendizado de máquina com aprendizado supervisionado, chamado método de classificação de Naive Bayes, através da biblioteca Scikit-learn junto da linguagem de programação Python.

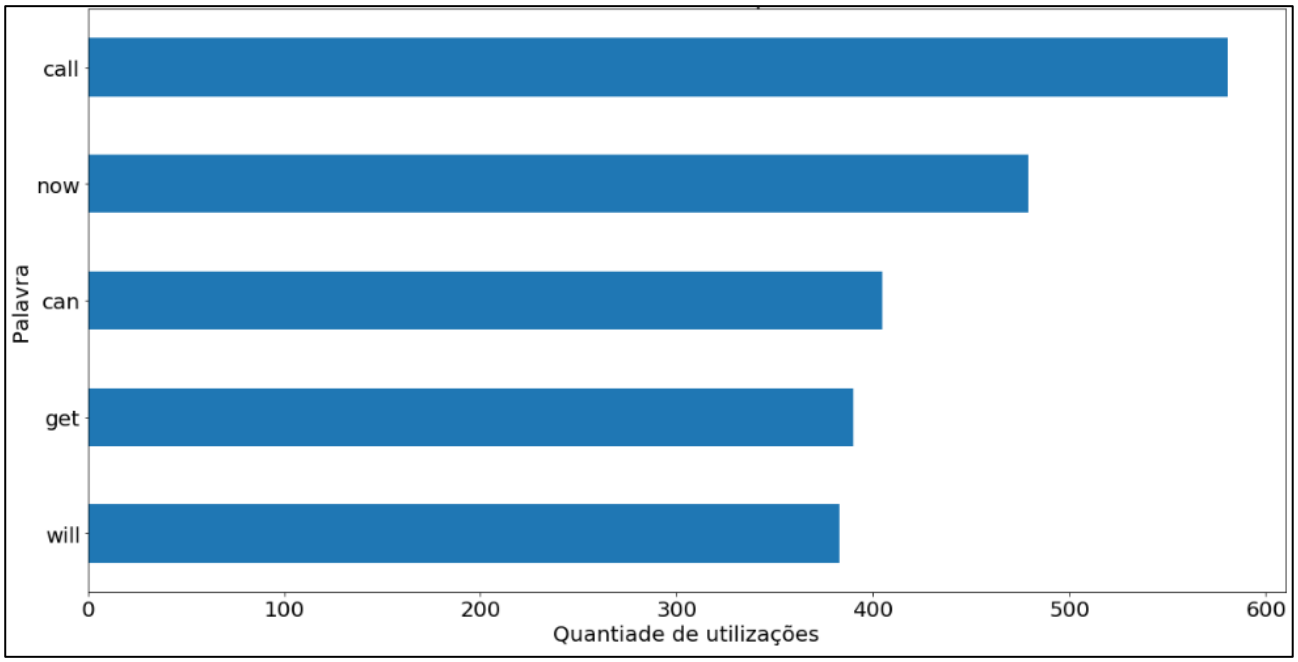
> Na tarefa de filtragem de e-mails algumas características podem ser um conjunto de palavravas ou a linha do assunto. A tarefa de classificação de e-mails normalmente é dividida em várias sub tarefas. A primeira, coleta de dados e representação, é a mais complicada. A segunda, seleção e redução das características dos e-mails. Por fim, a fase de classificação dos e-mails é onde o processo encontra o relacionamento real entre o treinamento e o conjunto de testes. (AWAD; ELSEUOFI, 2011, p. 174, tradução nossa).

Com apenas duas categorias necessárias, mensagens comuns e spam, Awad e Elseuofi (2011) descrevem que o algoritmo de Naive Bayes pode ser usado para classificação, onde a frequência das palavras possui o maior papel. Se algumas palavras ocorrem mais em spam do que em mensagens comuns, então o e-mail sendo analisado provavelmente é spam. A técnica de classificação Naive Bayes tem se tornado um método muito popular em softwares de filtragem de e-mail. Os dados selecionados para treinamento e os resultados alcançados são descritos na seção Resultados.

Resultados

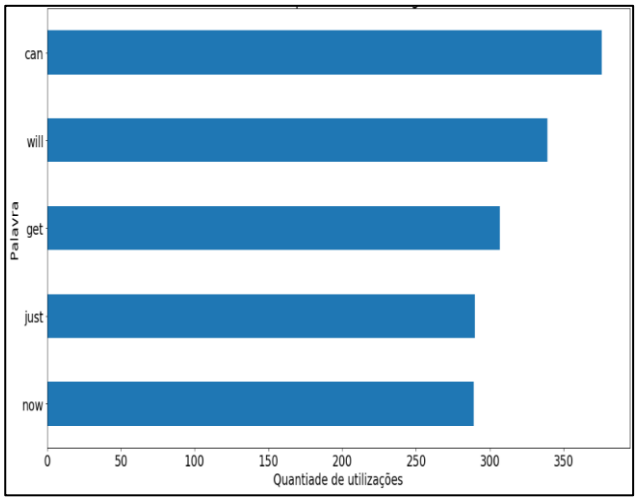
A Figura 1 apresenta, em ordem decrescente, as 5 palavras mais utilizadas nas mensagens. A Figura 2 e Figura 3 apresentam as 5 palavras mais frequentes categorizadas por comuns e spam, respectivamente.

Figura 1 – Palavras mais frequentes



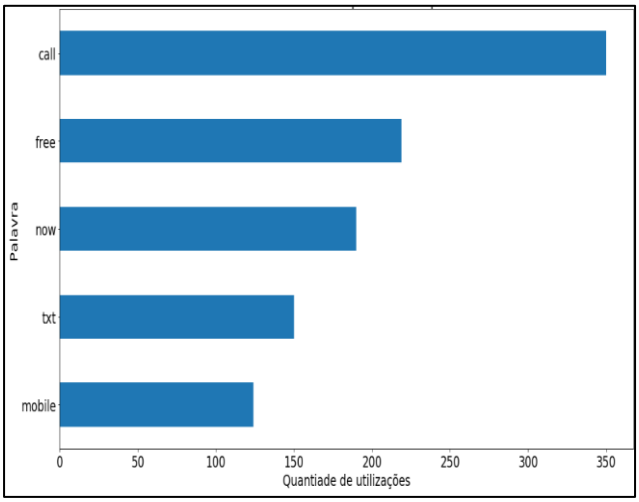
Fonte: elaborado pelo autor.

Figura 2 – Palavras comuns mais frequentes



Fonte: elaborado pelo autor.

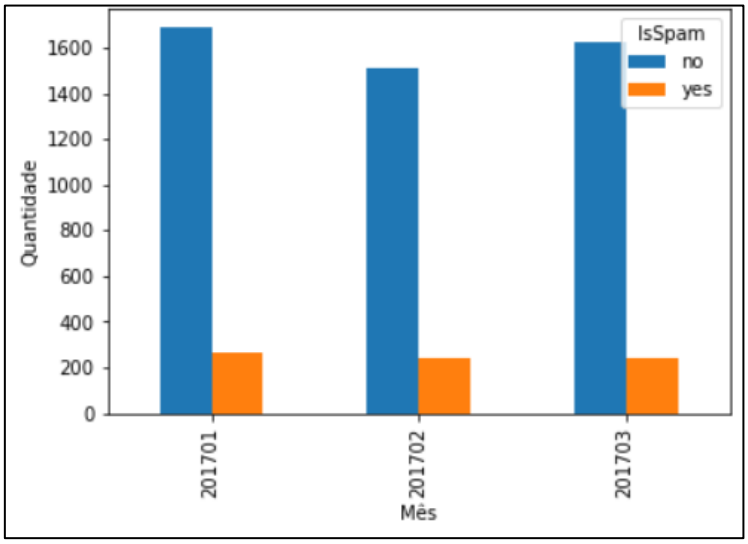
Figura 3 - Palavras spam mais frequentes



Fonte: elaborado pelo autor.

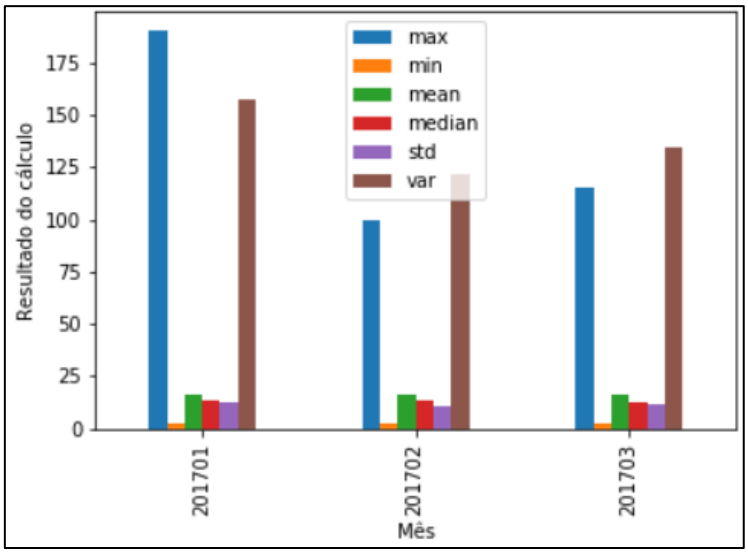
A Figura 4 apresenta a quantidade de mensagens por mês classificadas como comuns ou spam. Em complemento, a Figura 5 mostra o resultado do cálculo de máximo, mínimo, média, mediana, desvio padrão e variância sobre as mensagens agrupadas mensalmente.

Figura 4 – Quantidade de mensagens por mês



Fonte: elaborado pelo autor.

Figura 5 – Resultado do cálculo de mensagens por mês



Fonte: elaborado pelo autor.

O Quadro 1 apresenta o dia com a maior quantidade, mensalmente, de mensagens comuns em sequência, acompanhadas pela quantidade.

Quadro 1 – Dias com a maior sequência de mensagens comuns

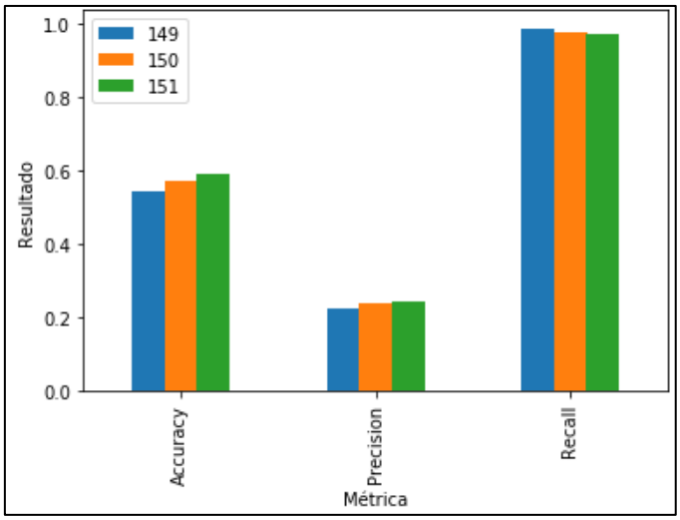
Mês	Dia	Quantidade
01/2017	26/01/2017	31
02/2017	04/02/2017	39
03/2017	31/03/2017	46

Fonte: elaborado pelo autor.

A Figura 6 mostra o resultado do método de classificação de Naive Bayes. São apresentados três modelos testados com o resultado de três métricas de avaliação: (i) *accuracy*; (ii) *precision*; (iii) *recall*. No primeiro modelo foram utilizadas as frequências que determinadas palavras aparecem na mensagem. No segundo, foram utilizados os dados do primeiro mais a quantidade de palavras frequentes. O terceiro utilizou os dados primeiro e segundo modelo com adição da quantidade total de palavras. O aumento da quantidade de dados para treinamento (*features*) está diretamente relacionado com o

aumento da *accuracy*. Foram realizados testes incluindo também as mensagens originais e suas datas, ambos com valores únicos no conjunto, e o valor da *accuracy* se aproximou de 100%, mostrando um possível problema de *overfitting*.

Figura 6 – Resultado por quantidade de *features*



Fonte: elaborado pelo autor.

Conclusão

Através do desenvolvimento desse trabalho, em conjunto com a aplicação das técnicas discutidas, é possível identificar, primeiramente, baixa sazonalidade na quantidade e distribuição das mensagens comuns e spam, tornando o mês um parâmetro de pouco valor para classificação das mensagens. Em contra partida, a frequência das palavras nas mensagens possui grande variação entre as categorias comum e spam, mostrando que sua presença nos dados de treinamento de classificação possui, possivelmente, grande impacto para o aumento da precisão.

A execução do método de classificação Naive Bayes apresentou baixo valor de *accuracy* para o problema de classificação de mensagens. Ao utilizar a solução em uma ferramenta de filtragem de mensagens, por exemplo, não é possível excluir as mensagens identificadas como spam. O baixo resultado de acerto iria remover com uma frequência muito alta mensagens comuns. Uma aplicação possível para o modelo e resultado atual seria a adição de alertas nas mensagens, delegando ao usuário a decisão final sobre a categorização.

O método de classificação utilizado, apesar de indicado por Awad e Elseuofi (2011), necessita ser comparado com o resultado de outros algoritmos, como K-Nearest Neighbour (KNN). O resultado da comparação pode resultar na escolha de outro método, possibilitando a aplicação em filtros de mensagens sem intervenção humana ou, em contra partida, concluir que o método Naive Bayes é o mais indicado e sua aplicação deve ser usada conscientemente considerando seu resultado pouco preciso. Alterações nos dados do modelo de treinamento também podem aumentar, possivelmente, o valor de *accuracy* alcançado.

Referências

AWAD, W.a.; ELSEUOFI, S.m.. Machine Learning Methods for Spam E-Mail Classification. **International Journal Of Computer Science And Information Technology**. Porto Fuad, p. 173-184. 28 fev. 2011. Disponível em: <https://www.researchgate.net/publication/50211017_Machine_Learning_Methods_for_Spam_E-Mail_Classification>. Acesso em: 21 fev. 2020.