

Desafio Senior Labs Challenge

O objetivo desse documento é apresentar soluções para o desafio da Senior Labs Challenge.

Nesse desafio foi proposto criar um artigo detalhando qual a metodologia aplicada e desenvolver soluções através de ferramentas de software para extrair dados estatísticos e relatórios gráficos da base de dados disponibilizada.

1. Introdução

Nas últimas décadas, o e-mail se tornou uma das principais formas de comunicação no meio corporativo, e cada vez é mais comum que as pessoas sejam inundadas por uma grande quantidade de mensagens.

Infelizmente, muitas dessas mensagens são spams, são enviadas com o objetivo de promover produtos ou serviços de maneira invasiva e não solicitada.

Além de ser desagradável o recebimento de mensagens indesejáveis, os spams também podem representar uma ameaça à segurança dos usuários, já que muitos desses e-mails podem conter links maliciosos que podem infectar o computador com vírus ou roubar informações pessoais.

No desafio proposto, foi disponibilizada uma base de dados no formato CSV contendo vários exemplos de mensagens comuns e de spams.

Utilizando essas informações, o objetivo é extrair dados estatísticos, relatórios gráficos e por fim propor um modelo que seja possível classificar de forma automática se o e-mail é uma mensagem comum ou spam.

2. Problema

Extrair informações estáticas da base de dados disponibilizada, através de qualquer software ou linguagem de programação.

3. Objetivo

Realizar a criação de um artigo evidenciando os dados extraídos, consolidados em relatórios e apresentar metodologia aplicada para obter os resultados esperados.

4. Solução Proposta

O software utilizado para extrair as informações solicitadas é o Power BI da Microsoft,

Através dele foi possível importar a base de dados e gerar os relatórios solicitados por meio dos recursos e tecnologia da ferramenta.

6. Primeira Etapa

Primeira etapa do desafio consiste em extrair estatísticas desta base de dados

6.1 Exibir gráfico as palavras

Exibir gráfico as palavras mais frequentes em toda a base de dados (Ex.: gráfico de barras, nuvem de palavras, etc).

6.1.1 Solução implementada

Através de um modelo gráfico “Nuvem de Palavras”, foi gerado uma representação visual da frequência de palavras em um texto ou conjunto de dados.

As palavras extraídas desse gráfico são referentes ao campo “Full_Text”, sem filtros, os dados foram exportados considerando todas as mensagens de e-mails.



6.2 Quantidades de mensagens comuns e spams

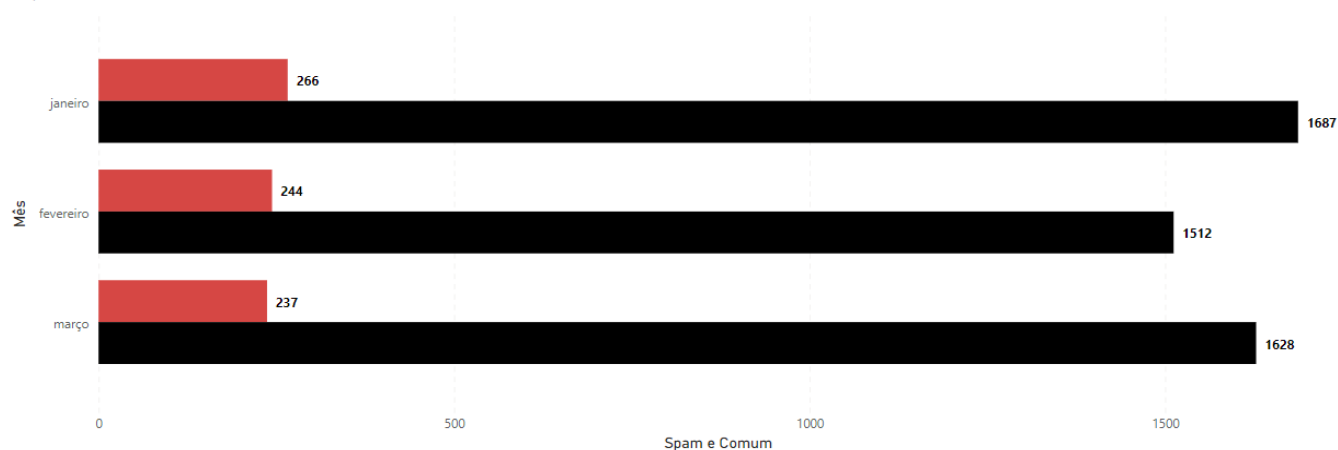
Exibir gráfico com as quantidades de mensagens comuns e spams para cada mês.

6.2 Solução implementada

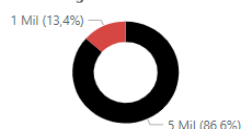
Através de um modelo gráfico “Gráfico de barras”, é exibido a quantidade de mensagens comuns e spam por mês, uma matriz com os totalizadores e um “Gráfico em Pizza” com os respectivos percentuais.

Mensagens por Mês

● Spam ● Comum



Percentual de Msg's



Mensagens
● no
● yes

Mês	Spam	Comum	Total
janeiro	266	1687	1953
fevereiro	244	1512	1756
março	237	1628	1865
Total	747	4827	5574

6.3 Calcular indicadores do campo "Word_Count" por Mês

Calcular o máximo, o mínimo, a média, a mediana, o desvio padrão e a variância da quantidade total de palavras (Word_Count) para cada mês;

6.3.1 Solução implementada

Abaixo os resultados para cada grupo de informações, a consolidação dos dados foi executada através do agrupamento dos dados por mês além das respectivas fórmulas para cálculo.

Mês	Máximo de Palavras
janeiro	190
março	115
fevereiro	100
Total	190

Mês	Média de Palavras
janeiro	16,34
fevereiro	16,03
março	16,29
Total	16,22

Mês	Desvio padrão de Palavras
janeiro	12,55
fevereiro	11,04
março	11,57
Total	11,77

Mês	Mínimo de Palavras
janeiro	2
fevereiro	2
março	2
Total	2

Mês	Mediana de Palavras
janeiro	13
fevereiro	13
março	12
Total	13

Mês	Varição de Palavras
janeiro	157,60
fevereiro	121,87
março	133,94
Total	138,44

6.4 Sequência de mensagens comuns por Mês

Exibir o dia de cada mês que possui a maior sequência de mensagens comuns (não spam).

6.4.1 Solução implementada

Dashboard exibindo informações dos dias de cada mês com maior sequência de mensagens comuns enviadas. Para chegar nesse resultado foi necessário criar um novo modelo de dados agrupando as informações por mês e filtrando apenas e-mails que não são spams.

Mês	Emails por dia
☒ janeiro	69
domingo, 1 de janeiro de 2017	69
☒ março	69
quarta-feira, 8 de março de 2017	69
☒ fevereiro	72
segunda-feira, 13 de fevereiro de 2017	72

7 Segunda etapa

A segunda etapa consiste em aplicar um método capaz de classificar automaticamente as mensagens como “comum” e “spam”. Como você considera os resultados encontrados? Justifique.

7.1 Solução implementada

Analisando os dados obtidos através dos relatórios é possível identificar alguns parâmetros para diferenciar e-mails comuns dos spams.

Para devida análise e elaboração do modelo estatístico, foi necessário criar uma nova base de dados para que seja equiparado as quantidades de e-mails comum com spams, na amostragem saneada, ficaram 747 mensagens para cada grupo de mensagens, totalizando 1.494 e-mails.

De posse desse modelo, podemos concluir que existem várias métricas que podem ser baseadas para diferencia-los, sendo elas:

1º parâmetro: a variação de palavras em um e-mail comum, tem em média 30% mais palavras que um Email de spam.

Spam	Variação de Palavras
☒ no	126,49
☒ yes	35,11

2º parâmetro: E-mails comuns, tem em média 55% menos palavras se comparado com um e-mail de spam.

Spam	Soma de Palavras
no	10622
yes	19040
Total	29662

3º parâmetro: E-mails de spams tem até 39 palavras, sendo que mensagens comuns esse numero é por volta de 100 palavras.

Spam	Máximo de Palavras
⊕ no	100
⊕ yes	39

4º parâmetro: E-mails de spams em média tem 25 palavras, sendo que mensagens comuns esse número é por volta de 14 palavras.

Spam	Média de Palavras
⊕ no	14,22
⊕ yes	25,49
Total	19,85

AA

5º parâmetro: Analisando o gráfico de nuvem, fica evidente algumas palavras são muito mais frequentes em e-mails de spam do que mensagens comuns, essas palavras são:

“Urgente”, “Have”, “Guaranteed”, “Want”, “Awarded”, “Award” e “Cash”.

7.2 Conclusão

A união dos parâmetros apresentados aplicados a uma política de AntiSpam é possível classificar qual o tipo de mensagem que está sendo transmitida, com pequena margem de desvio padrão.

O modelo apresentando é uma amostragem com poucos dados analisados, para que se tenha uma assertividade maior, é importante que seja aplicado os mesmos métodos para em um base com maior numero de registros, dessa forma o modelo terá um referencial mais apurado, aumentando as possibilidades de acerto.

8 Referencias

Documentação Power BI: <https://learn.microsoft.com/pt-br/power-bi/consumer/>

Instruções desafio: <https://github.com/SeniorSA/seniorlabs-challenge>

