Exercício Pesquisador 2020

README

A proposta deste exercício, é apoiar a avaliação do seu perfil para atuar como pesquisador. A área de Pesquisa Aplicada tem como o objetivo realizar trabalhos de desenvolvimento nas mais diversas áreas e documentar todo seu processo e as conclusões. Em geral, documentar é a última coisa que um desenvolvedor deseja fazer, mas na nossa equipe, além do desejo de resolver o problema, necessitamos de pessoas que gostem de entender os conceitos e explica-los de forma escrita e/ou oral.

O PROBLEMA

A quantidade de spams (mensagens não solicitadas) que recebemos diariamente, não para de crescer. Os tipos de spam são diversos: anúncios de produtos / web sites, esquemas para ganhar dinheiro rápido, correntes, pornografia e etc.

Input

O arquivo `sms_senior.csv` contém vários exemplos de mensagens comuns (4827 unidades) e mensagens spams (747 unidades). As mensagens foram submetidas a uma etapa de mineração de texto, com o objetivo de identificar as palavras mais frequentes na base de dados. Segue as informações dos atributos do arquivo:

- 1 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*).

Primeira Etapa

A primeira etapa do seu trabalho consiste em extrair estatísticas desta base de dados:

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

2. Exibir gráfico com as quantidades de mensagens comuns e spams para cada mês;
3. 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;
4. Exibir o dia de cada mês que possui a maior sequência de mensagens comuns (não spam).

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.

Output

Você pode utilizar qualquer linguagem de programação e ferramentas de software para extrair as informações das duas etapas do trabalho. Por fim, descreva o trabalho realizado em um artigo com uma ou duas páginas no modelo anexo. Lembre-se de apontar as estatísticas extraídas e de explicar o método de classificação utilizado, como a etapa de treinamento e classificação foram executadas, além dos resultados que foram encontrados.

O modelo parece grande, mas você pode ser bem objetivo. Os códigos fontes ou arquivos utilizados no trabalho deverão ser postados no github, onde o README deve explicar como proceder para executar sua solução.

ANEXOS

Titulo

Nome do Autor
email@senior.com.br

Introdução

Descrever o problema de forma simples e dedicar um ou dois parágrafos para descrever o que existe nesse sentido - semelhante aos trabalhos correlatos da faculdade, mas de forma bem objetiva para mostrar que você "olhou para fora" antes de fazer essa avaliação para não reinventar a roda (OK, talvez aqui nesse exercício podemos reinventar a roda um pouco...).

Metodologia

Basicamente você deve descrever aqui como fez para realizar a pesquisa. Ou seja, se alguém quiser repetir ela, poderá fazer isso olhando esse item.

Resultados

Mostrar os tempos medidos em tabelas ou gráficos.

Conclusão

Descrever o que você conclui do experimento com resultados obtidos. Existe mesmo vantagem em usar um ou outro no cenário que você montou?

Referências

Colocar somente o que você usou no trabalho. Sempre que algo aparece aqui, é porque você usou no seu texto.