Exercício – Inova Talentos

README

A proposta deste exercício é apoiar a avaliação do seu perfil para atuar no projeto Inova Talentos. A área de pesquisa trabalha em vários projetos técnicos, incluindo pesquisas em análise e mineração de dados.

Na pesquisa, além do desejo de resolver o problema, necessitamos de pessoas que gostem de entender os conceitos e principalmente de explicar de forma escrita e oral como foi a abordagem e como se chegou a uma determinada conclusão.

Um problema fictício, mas similar aos problemas que estamos atuando na nossa Empresa:

A quantidade de spams – mensagens não solicitadas – que recebemos diariamente não para de crescer, inclusive nas mensagens SMS. O conceito de "spam" é diverso: anúncios de produtos / web sites, esquemas para ganhar dinheiro rápido, correntes, pornografia... etc.

>> Input

A base de dados SMS_Senior.csv contém vários exemplos de mensagens comuns (4827) e mensagens spams (747). As mensagens foram submetidas a uma etapa de mineração de texto, com o objetivo de identificar as palavras mais frequentes na base de dados. Seguem informações dos atributos da base de dados:

- 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 mensangem (Word_Count);
- 1 coluna contendo a data de recebimento da mensagem (Date);
- 1 coluna que identifica se a mensagem é spam ou não (IsSpam).

A **primeira etapa** do seu trabalho consiste em extrair estatísticas da 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).

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 ou software para extrair as informações das duas etapas do trabalho. Lembre-se de explicar o método de classificação utilizado, como a etapa de treinamento e classificação foram executadas e quais resultados foram encontrados.

Descreva o trabalho realizado em um artigo com uma ou duas páginas no modelo abaixo. Lembre-se de apontar as estatísticas extraídas e de explicar o método de classificação utilizado, como a etapa de treinamento e classificação foram executadas e quais resultados foram encontrados.

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.

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.