Segundo Exercício-Programa
Spamminator

Norton Trevisan Roman

16 de maio de 2019

1 Tarefa

Sua tarefa é desenvolver um filtro de spam: um programa que receba como entrada um texto e, como saída, diga se é um spam ou não.

Para isso, você seguirá a seguinte metodologia:

1. Ao longo de uma semana, colete todos os e-mails que chegarem a você (inclusive os marcados como spam por sua ferramenta de leitura de e-mails)
2. Classifique-os manualmente como "spam" ou "não spam"
3. Separe 20% dos e-mails, aleatoriamente, para testes (esse será seu **conjunto de testes**). Os 80% restantes formarão seu **conjunto de treino**.
4. Construa seu Spammitator, usando o modelo Bayesiano Ingênuo, conforme visto em sala (você pode criar variações deste)
5. Treine seu Spamminator no conjunto de treino
6. Rode seu Spamminator no conjunto de teste (os 20% separados). Qual a taxa de acerto?
   (a) Compare, para esses mesmos e-mails de teste, a taxa de acerto de seu Spamminator com a de seu software de leitura de e-mail (ex: gmail). Quem acertou mais? De quanto foi a diferença? (para saber a taxa de acertos de seu software de e-mail use a classificação manual feita por você)
7. Crie uma lista de *stop words* – palavras comuns que estão presentes em todos os textos (veja abaixo como)
8. Remova as stop words de seus conjuntos de treino e de teste
9. Treine seu Spamminator no conjunto de treino modificado
10. Rode seu Spamminator no conjunto de teste modificado. Qual a taxa de acerto?
    (a) Compare a taxa de acerto com o acerto sem a remoção das stop-words

(b) Compare a taxa de acerto de seu Spamminator à da classificação feita (e medida anteriormente) em seu software de leitura de e-mail (ex: gmail). Quem acertou mais? De quanto foi a diferença? Melhorou em relação ao valor anterior?

Ao calcular a probabilidade de ocorrência de uma palavra, não esqueça de usar algum método de suavização (*smoothing*).

1.1 Lista de Stop-Words

Para criar a lista de stop-words, você pode verificar, em todos os e-mails de treino, quais as palavras mais comuns. Para tal, construa um gráfico mostrando a frequência de cada palavra nesse conjunto, definindo então um ponto de corte (*threshold*), a partir do qual você ignorará as palavras mais frequentes.

A ideia por trás desse procedimento é remover palavras muito comuns, como "o", "e", "um" etc, que muito provavelmente aparecem tanto em spams quanto em não spams, reduzindo assim o ruído nos dados.

Você é livre para seguir outra metodologia de criação, desde que a justifique, mesmo que subjetivamente. Você deverá também justificar sua escolha para o ponto de corte.

Uma metodologia alternativa seria verificar a frequência das palavras no conjunto de treino, mas separadas entre spam e não spam. Uma stop-word seria, então, uma palavra com frequência maior que um determinado limiar tanto em spams quanto em não spams (não necessariamente o mesmo limiar para ambos os conjuntos), evitando, assim, que palavras frequentes em apenas um dos conjuntos – e portanto significativas – sejam incluídas erroneamente nas stop words.

1.2 Representação do Texto

Para essa tarefa, o texto será representado como um conjunto de suas palavras (modelo *bag of words*), sem levar em conta sua ordem, mas mantendo sua multiplicidade (ou seja, repetições da mesma palavra). Cabe a você decidir se usa apenas as palavras, ou também inclui outros marcadores, como pontuação, por exemplo.

Lembre de tomar o cuidado de normalizar a representação das palavras, ou seja, de deixá-las num formato comum (ou as deixe todas em maiúscula, ou em minúscula). A partir das palavras do texto você poderá então construir tabelas de frequência, usadas em seu classificador.

1.3 Entrada

A entrada será composta por um texto, em formato txt "puro", armazenado em um arquivo.

1.4 Saída

A saída de seu programa é a classificação do texto de entrada como "spam" ou "não spam".

1.5 Material a Ser Entregue

Você deve entregar, via e-disciplinas, apenas o relatório produzido nesse experimento. Por isso é de suma importância que seu relatório detalhe os algoritmos e valores utilizados, bem como qualquer decisão de projeto sua, apresentando de forma clara os experimentos e conclusões. Também é importante que seu relatório contenha uma introdução ao problema, e esteja apresentável.

Não esqueça de incluir, em seu relatório, os seguintes pontos (essa não é uma lista exaustiva):

- O modelo matemático utilizado (e respectivo algoritmo de implementação), detalhando e justificando suas decisões (como tipo de suavização etc);
- A lista de stop-words criada, bem como seu gráfico de frequência (ainda que simplificado, apontando intervalos de palavras). Detalhe como essa lista foi construída, justificando sua escolha para o ponto de corte;
- A apresentação e discussão dos resultados das comparações feitas entre seu Spamminator e o filtro de spams de seu leitor de e-mails, bem como entre as versões com e sem o uso de stop-words.

O trabalho pode ser feito em grupos de até 4 pessoas. Assim, apenas um integrante do grupo deve fazer a submissão. A submissão deverá ser feita via um arquivo zip (ou rar). O nome do arquivo deverá ser o número USP do aluno que fez a entrega (por exemplo, 1234567.zip). Este arquivo deverá conter tão somente o arquivo pdf de seu relatório. Formatos de entrega fora desses padrões acarretarão desconto em nota.

Qualquer tentativa de fraude ou cola implicará em nota zero para todos os envolvidos. Guarde uma cópia do trabalho entregue.

2 Avaliação

Para avaliação serão observados os pontos constantes da Seção 1, bem como o formato de entrega, definido na Seção 1.5

É de sua responsabilidade verificar:

- Se o material entregue está de acordo com as especificações
- Se a entrega realmente ocorreu (ou seja, se o upload foi feito corretamente). Então faça o upload, baixe e teste o que baixou.

Falhas nos itens acima não serão toleradas.