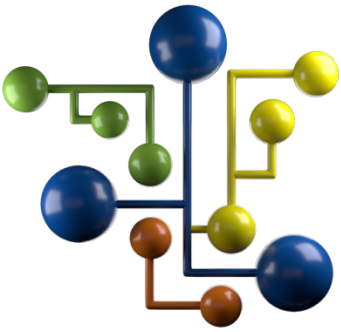

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Classificação de Texto

Vamos agora considerar em profundidade a tarefa de classificação de texto, também conhecida como categorização: dado algum tipo de texto, decidimos qual conjunto predefinido de classes pertence a ele. A identificação da linguagem e a classificação do gênero são exemplos de classificação de texto, como também é a análise do sentimento (classificação de um filme ou revisão de produto como positivo ou negativo) e a descoberta de spam (classificação de uma mensagem de e mail como spam ou não spam).

Uma vez que "não spam" é estranho, os pesquisadores cunharam o termo ham para não spam. Podemos tratar a descoberta de spam como um problema em aprendizagem supervisionada. Um conjunto de treinamento está disponível: os exemplos positivos (spam) estão na minha pasta de spam, os exemplos negativos (ham) estão na minha caixa de entrada. Veja esta extração:

Spam: Venda por atacado de óculos modernos –57% hoje. Relógios de grife por preço baixo…
Spam: Você pode comprar ViagraFr $ 1,85 Todos os medicamentos a preços imbatíveis!…
Spam: PODEMOS TRATAR DE QUALQUER COISA QUE VOCÊ TENHA APENAS CONFIE EM NÓS …
Spam: Come.ce a ganhar* o salário que vo,cê m-erece o'btendo referên'cias apropriada,s!

Ham: O significado prático de largura da hiperárvore identificando mais…
Ham: Resumo: Vamos motivar o problema de aglomeração de identidade social:…
Ham: É bom ver você meu amigo. Ei Pedro, Foi bom saber de você. …
Ham: PDS implica convexidade do problema de otimização resultante (Kernel Ridge …

A partir dessa extração, podemos começar a ter uma ideia de quais seriam as boas características para incluir no modelo de aprendizagem supervisionada n-gramas de palavras e como "preço baixo" e "Você pode comprar" parecem indicadores de spam (embora tenham probabilidade diferente de zero de ser ham também). Atributos de caracteres também parecem importantes: é mais provável que o spam esteja todo em letras maiúsculas e tenha pontuações incorporadas às palavras. Aparentemente, os spammers pensaram que a palavra bigrama "você merece" seria muito indicativa de spam e então escreveram "vo,cê m-erece". Um modelo de caractere deveria descobrir isso. Ou poderíamos criar um modelo de n-grama de caracteres completo de spam e ham ou poderíamos criar as características manualmente como "número de sinais de pontuação incorporados às palavras".

Observe que temos duas formas complementares de falar sobre classificação. Na abordagem de modelagem de linguagem, definimos um modelo de linguagem de n-grama para P(Mensagem | spam) pelo treinamento em uma a pasta de spam, e um modelo P(Message | ham) pelo treinamento em uma caixa de entrada. Então, podemos classificar uma nova mensagem com a aplicação da regra de Bayes:

$$\underset{c \in \{spam, ham\}}{\operatorname{argmax}} P(c \mid mensagem) = \underset{c \in \{spam, ham\}}{\operatorname{argmax}} P(mensagem \mid c)\, P(c)$$

onde P(c) é estimado apenas pela contagem do número total de mensagens de spam e ham. Essa abordagem funciona bem para a descoberta de spam, tal como aconteceu para a identificação do idioma. Nesta fórmula, P(c) é estimado apenas pela contagem do número total de mensagens de spam e ham.

Na abordagem de aprendizagem de máquina representamos a mensagem como um conjunto de pares de característica/valor e aplicamos um algoritmo de classificação h à característica de vetor X. Podemos tornar as abordagens de modelagem de linguagem e de aprendizagem de máquina compatíveis pensando em n-gramas como características. É mais fácil visualizar com um modelo unigrama. As características são as palavras no vocabulário: “a,” “aardvark”…, e os valores são o número de vezes que cada palavra aparece na mensagem. Isso torna o vetor de característica grande e esparso. Se existirem 100 mil palavras no modelo de linguagem, o vetor de características terá comprimento 100.000, mas para uma mensagem de e-mail curta quase todas as características terão resultado zero.

Essa representação do unigrama tem sido chamada de modelo de saco de palavras ou bag of words. Você pode pensar sobre o modelo como colocar as palavras de treinamento do corpus em um saco e em seguida selecionar as palavras uma de cada vez. A noção de ordem das palavras é perdida; um modelo de unigrama oferece a mesma probabilidade para qualquer permutação de um texto. Modelos de n-grama de ordem superior mantêm alguma noção local da ordem das palavras.

Com bigramas e trigramas, o número de características é elevado ao quadrado ou ao

cubo e podemos acrescentar outras características não n-grama: o tempo em que a mensagem foi enviada, se uma URL ou uma imagem faz parte da mensagem, um número de identificação do remetente da mensagem, o número do remetente das mensagens ham e spam anteriores, e assim por diante. A escolha das características é a parte mais importante na criação de um bom detector de spam — mais importante do que a escolha do algoritmo para processamento das características. Em parte, isso é porque existe grande quantidade de dados de treinamento; por isso, se pudermos propor uma característica, os dados poderão determinar com precisão se é boa ou não. É necessário atualizar constantemente as características de atualização porque a detecção de spam é uma tarefa adversarial; os spammers modificam seus spams em resposta às alterações no detector de spam.

Pode ser caro executar algoritmos em um vetor de características muito grandes, por isso muitas vezes um processo de seleção de características ou Feature Selection é usado para manter apenas as características que melhor distinguem entre spam e ham. Por exemplo, o bigrama “of the” é frequente em inglês e pode ser igualmente frequente em spam e ham, então não há sentido em contá-lo. Muitas vezes, os cem melhores ou mais característicos fazem um bom trabalho de distinção entre classes.

Uma vez escolhido um conjunto de características, podemos aplicar qualquer uma das técnicas de aprendizagem supervisionada que vimos; as mais populares para a categorização de texto incluem kvizinhos mais próximos, máquinas de vetores de suporte, árvores de decisão, Naive Bayes e regressão logística. Todos eles têm sido aplicados para descoberta de spam, geralmente com precisão na faixa de 98-99%. Com um conjunto de atributos projetado cuidadosamente, a precisão pode ultrapassar 99,9%.