

# introdução à recuperação de informação

**13.5 Seleção de recursos**

***Seleção de recursos*** é o processo de seleção de um subconjunto dos termos que ocorrem em um conjunto de treinamento e usando apenas este subconjunto como recursos na classificação de texto.

A seleção de recursos serve a dois propósitos principais.

1. Primeiro, torna o treinamento e a aplicação de um classificador mais eficientes ao diminuir o tamanho do vocabulário efetivo. Isso é de particular importância para classificadores que, ao contrário do NB (Naive Bayes), são caros de treinar.
2. Em segundo lugar, a seleção de recursos muitas vezes aumenta a classificação de acurácia eliminando recursos de ruído. Um recurso de ruído é aquele que, quando adicionado para a representação do documento, aumenta o erro de classificação em novos dados. Suponha um termo raro, digamos aracnocêntrico, não tem informações sobre uma classe, digamos China, mas todas as instâncias de aracnocêntrico aconteceu de ocorrer em documentos da China no conjunto de treinamento. Então, o método de aprendizagem pode produzir um classificador que atribui erroneamente documentos de teste contendo aracnocêntrico para China. Tal generalização incorreta de uma propriedade acidental do treinamento, isso é um **overfitting**.

Podemos ver a seleção de recursos como um método para substituir um classificador complexo (usando todos os recursos) por um mais simples (usando um subconjunto dos recursos). Pode parecer contra-intuitivo no início que um classificador aparentemente mais fraco seja vantajoso na classificação de texto estatístico, mas ao discutir a compensação de viés-variância, veremos que os modelos mais fracos são geralmente preferíveis quando os dados de treinamento limitados estão disponíveis.

O algoritmo básico de seleção de recursos é mostrado na Figura 13,6. Para uma classe *c,* nós calculamos uma medida de utilidade *A (t, c)* para cada termo do vocabulário e selecione o *k (número de termos que se deseja)* termos que têm os valores mais altos de *A (t, c).* Todos os outros termos são descartados e não são usados na classificação.

Vamos apresentar três diferentes medidas de utilidade nesta seção:

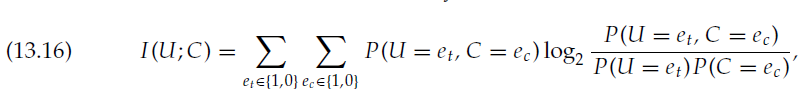
1. informações mútuas, *A (t, c) = I (Ut;* *Cc);*
2. a *χ*2teste, *A (t, c) = X*2 (*t, c);*
3. e frequência, *A (t, c) = N (t, c).*

Dos dois modelos NB, o **modelo Bernoulli** é particularmente sensível aos recursos de ruído. Um classificador Bernoulli NB requer alguma forma de seleção de recursos ou então sua precisão será baixa.

Esta seção aborda principalmente a seleção de recursos para tarefas de classificação de duas classes, como *China* contra *não-China.*

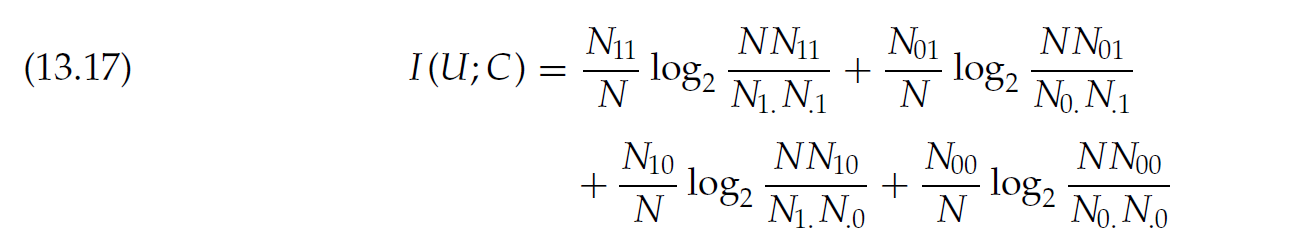
***13.5.1 Informação mútua***

Um método comum de seleção de recursos é calcular A (t, c) como o esperado **informação mútua (MI)** do prazo t e classe c.5 MI mede quanta informação a presença / ausência de um termo contribui para tornar correto classificação decidida em c.



Onde *U* é uma variável aleatória que assume valores et = 1 (o documento contém o termo t) e et = 0 (o documento não contém t), C é uma variável aleatória que assume valores ec = 1 (o documento está na classe c) e ec = 0 (o documento não está na classe c). Nós escrevemos Ut e Cc se não estiver claro a partir do contexto qual termo t e classe c estamos nos referindo.

Para MLEs das probabilidades, Equação (13,16) é equivalente a Equação (13,17): onde o N*s* são contagens de documentos que têm os valores de et e ec isso são indicados pelos dois subscritos. Por exemplo, N10 é o número de documentos que contêm t (et = 1) e não estão em c (ec = 0). N1. = N10 + N11 é o número de documentos que contêm t (et = 1) e contamos documentos independente da filiação à classe (ec ∈ {0, 1}). N = N00 + N01 + N10 + N11 é o número total de documentos. Um exemplo de um dos MLE estima que transforma a equação (13,16) na Equação (13,17) é P (U = 1, C = 1) = N11 /N.



**Exemplo 13.3:** Considere a classe *aves* e o termo export em ReutersRCV1. As contagens do número de documentos com as quatro possíveis combinações de indicador valores são os seguintes:

|  |  |  |  |
| --- | --- | --- | --- |
|  | *ec = e*aves =1 | *ec = e*aves =0 |  |
| *et = e*export =1 |  |  |  |
| *N*11 =49 | *N*10 =141 |  |
| *et = e*export =0 |  |  |  |
| *N*01 =27.652 | *N*00 =774.106 |  |
|  |  |  |  |

5 Tome cuidado para não confundir **informações mútuas** esperadas com *informações mútuas pontuais,* que é definido como log *N*11 /*E*11 Onde *N*11 e *E*11 são definidos como na Equação (13,17) As duas medidas têm propriedades diferentes.

Selecionar *k* termos *t*1,. . . ,*tk* para uma determinada classe, usamos a seleção de recursos algoritmo na figura 13,6: Calculamos a medida de utilidade como *A(t, c) = I (Ut,* *Cc)* e selecione o *k* termos com os maiores valores.

A **informação mútua** mede a quantidade de informação - no sentido teórico da informação - que um termo contém sobre a classe. Se a distribuição de um termo for a mesma na classe e na coleção como um todo, então *I(U; C) =* 0. O MI atinge seu valor máximo se o termo for um indicador perfeito de pertencimento à classe, ou seja, se o termo estiver presente em um documento se e somente se o documento estiver na classe.

Figura 13,7 mostra termos com altas pontuações de informações mútuas para as seis classes na Figura 13,1.6 Os termos selecionados (por exemplo, london, uk, british for the class UK*)* são de utilidade óbvia para tomar decisões de classificação para suasrespectivas classes. No final da lista para uk encontramos termos como peripherals

e tonight (não mostrado na figura) que claramente não são úteis para decidir se o documento está na classe. Como você pode esperar, manter os termos informativos e eliminar os não informativos tende a reduzir o ruído e melhorar a precisão do classificador.

Esse aumento de precisão pode ser observado na Figura 13,8, que mostra *F*1em função do tamanho do vocabulário após a seleção do recurso para Reuters-RCV1.7

Comparando *F*1 em 132.776 **recursos (features / variáveis)** (correspondendo à seleção de todos os recursos) e em 10-100 recursos, vemos que a seleção de recursos MI aumenta *F*1 por cerca de 0,1 para o **modelo multinomial** e mais de 0,2 para o **modelo de Bernoulli**.

Para o **modelo Bernoulli**, *F*1 atinge o pico mais cedo, em dez recursos selecionados. Nesse ponto, o modelo de Bernoulli é melhor do que o modelo multinomial. Ao basear uma decisão de classificação em apenas alguns recursos, é mais robusto considerar apenas a ocorrência binária.

Para o **modelo multinomial** (seleção de recursos MI), o pico ocorre mais tarde, em 100 recursos, e sua eficácia se recupera um pouco no final, quando usamos todos os recursos. A razão é que o multinomial leva o número de ocorrências é levado em consideração na estimativa e classificação dos parâmetros e, portanto, explora melhor um número maior de recursos do que o modelo de Bernoulli. **Independentemente das diferenças entre os dois métodos, o uso de um subconjunto cuidadosamente selecionado dos recursos resulta em melhor eficácia do que o uso de todos os recursos.**

6 - As pontuações dos recursos foram calculadas nos primeiros 100.000 documentos, exceto para *aves,* uma classe rara, para a qual 800.000 documentos foram usados. Omitimos números e outras palavras especiais das dez listas principais.

7 - Treinamos os classificadores nos primeiros 100.000 documentos e calculamos *F*1 nos próximos 100.000. Os gráficos são médias de cinco classes.

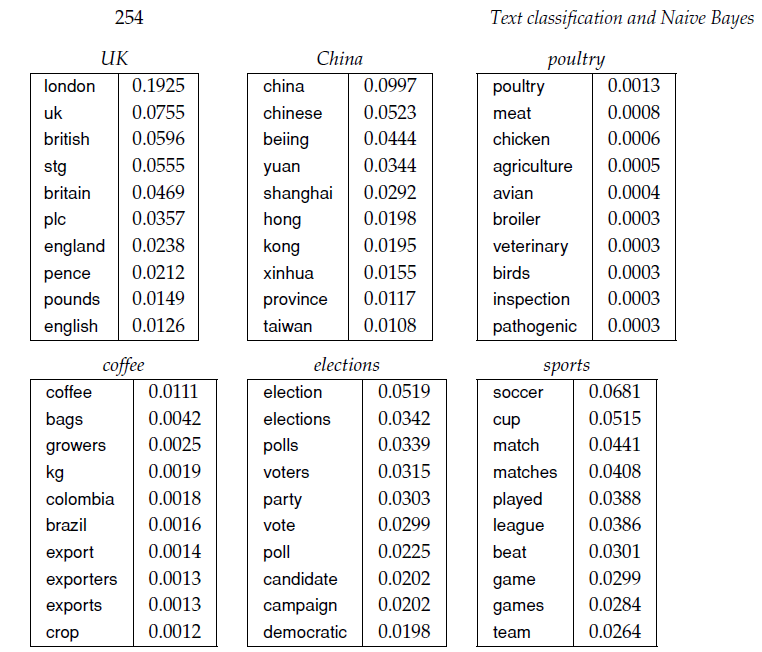


Figura 13.7 **Recursos (features/variáveis)** com altas pontuações de informações mútuas para **seis classes Reuters-RCV1**.

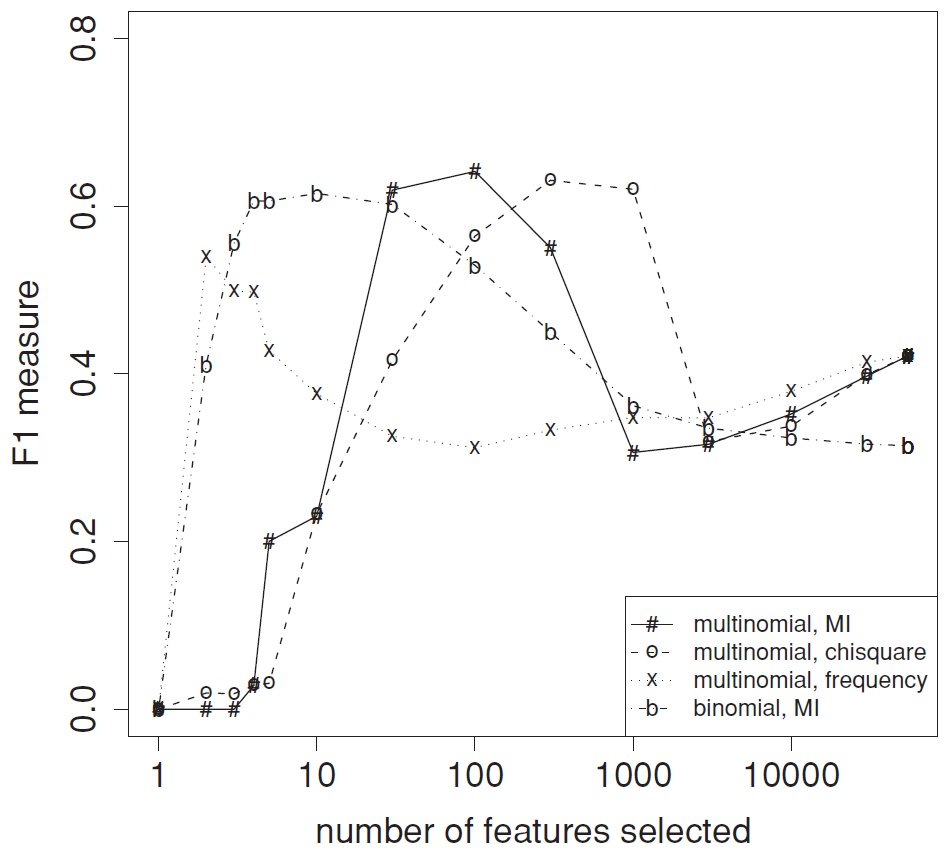
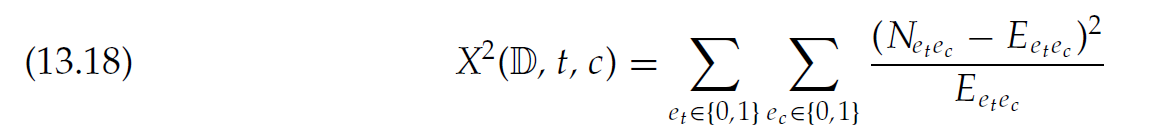


Figura 13.8 Efeito do tamanho do conjunto de recursos na precisão dos modelos multinomiais e de Bernoulli.

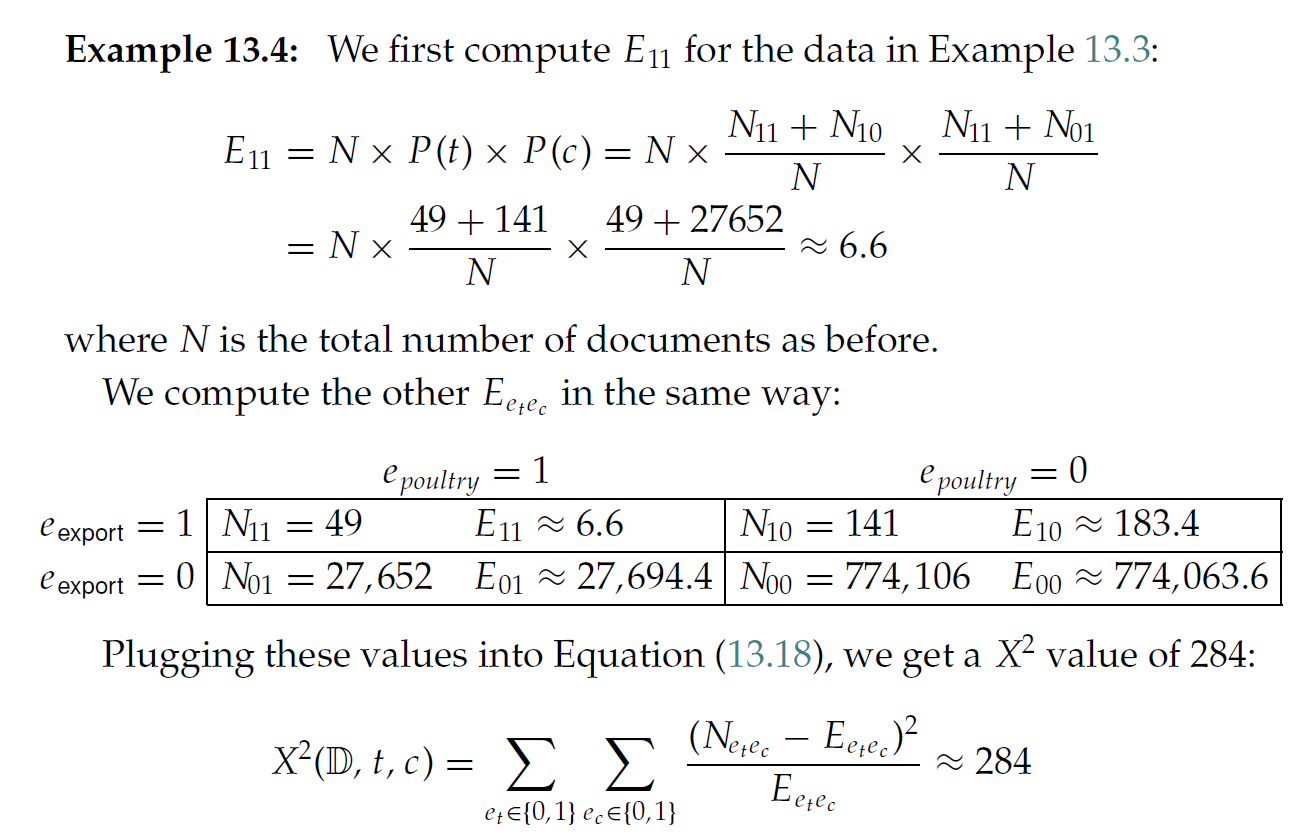
***13.5.2 χ*2 *Seleção de recursos***

Outro método popular de **seleção de recursos é χ2** Nas estatísticas, o χ2 teste é aplicado para testar a independência de dois eventos, onde dois eventos A e B são definidos como **independente** se estar P(AB) = P(A) P(B) ou equivalente, P(A | B) = P(A) e P(B|A) = P(B). Na seleção de recursos, os dois eventos são ocorrências do prazo e ocorrência na seguinte quantidade:



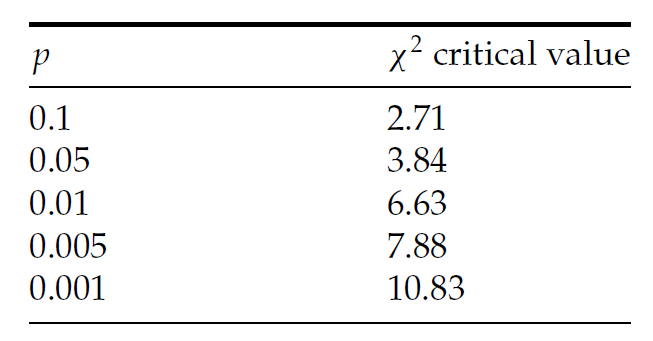
Onde et e ec são definidos como na Equação (13,16) N é o observado frequência no D e E a esperado frequência.

Por exemplo, E11 é a frequência esperada de t e c ocorrendo juntos em um documento assumindo que o termo e a classe são independente.



X2 é uma medida de quantas contagens esperadas E e contagens observadas N desviar um do outro. Um alto valor de X2 indica que a hipótese de independência, que implica que as contagens esperadas e observadas são semelhantes, está incorreta. Em nosso exemplo, X2 ≈ 284> 10,83. Com base na tabela 13,6, podemos rejeitar a hipótese de que aves e export são independentes com apenas 0,001 de chance de estar errado.8 Equivalentemente, dizemos que o resultado X2 ≈ 284> 10,83 é estatisticamente significativo no nível 0,001. Se os dois eventos forem dependentes, a ocorrência do termo torna a ocorrência da classe mais provável (ou menos provável), portanto, deve ser útil como um recurso. Este é o fundamento lógico de χ2 seleção de recursos.

**Tabela 13.6** Valores críticos do χ2 distribuição com um grau de liberdade. Por exemplo, se os dois eventos são independentes, então P(X2>6,63) <0,01. Então para X2> 6,63 a suposição de independência pode ser rejeitada com 99% de confiança.



**Avaliando *χ*2 como um método de seleção de recursos**

Do ponto de vista estatístico, *χ*2 a seleção de recursos é problemática. Para um teste com um grau de liberdade, a chamada correção de Yates deve ser usada, o que torna mais difícil alcançar significância estatística. **Além disso, sempre que um teste estatístico é usado várias vezes, a probabilidade de obter pelo menos um erro aumenta.** **Se 1.000 hipóteses forem rejeitadas, cada uma com probabilidade de erro de 0,05, então 0,05× 1000 = 50 chamadas do teste estarão erradas em média. No entanto, na classificação de texto, raramente importa se alguns termos adicionais são adicionados ao conjunto de recursos ou removidos dele.** Em vez disso, a importância *relativa* dos recursos é importante. Enquanto *χ*2 a seleção de recursos apenas classifica os recursos com relação à sua utilidade e não é usada para fazer afirmações sobre a dependência estatística ou independência de variáveis, não precisamos nos preocupar excessivamente com o fato de que ela não segue estritamente a teoria estatística.

8 Podemos fazer essa inferência porque, se os dois eventos são independentes, então *X*2 ∼ *χ*2, Onde *χ*2é o *χ*2distribuição.

***13.5.3 Seleção de recursos com base em frequência***

Ou seja, selecionando os termos que são mais comuns na classe. A frequência pode ser definida como a frequência do documento (o número de documentos na classe c que contém o termo t) ou como frequência de coleta (o número de tokens de t que ocorrem em documentos em c). **A frequência do documento é mais apropriada para o modelo de Bernoulli, a frequência de coleta para o modelo multinomial.**

A seleção de recursos com base em frequência seleciona alguns termos frequentes que não têm informações específicas sobre a classe, por exemplo, os dias da semana (Segunda-feira, terça-feira, . . . ), que são frequentes em todas as classes em textos de notícias. Quando muitos milhares de recursos são selecionados, a seleção de recursos com base na frequência geralmente funciona bem. Assim, se uma precisão um pouco abaixo do ideal for aceitável, a seleção de recursos com base em frequência pode ser uma boa alternativa para métodos mais complexos. No entanto, Figura13,8 é um caso em que a seleção de recursos com base em frequência tem um desempenho muito pior do que MI e *χ*2 e não deve ser usado.

***13.5.4 Seleção de recursos para vários classificadores***

Em um sistema operacional com um grande número de classificadores, é desejável selecionar um único conjunto de recursos em vez de um diferente para cada classificador. Uma maneira de fazer isso é calcular o *X*2 estatística para um *n* × 2 tabela onde as colunas são ocorrência e não ocorrência do termo e cada linha corresponde a uma das classes. Podemos então selecionar o *k* termos com o maior *X*2estatística como antes.

Mais comumente, as estatísticas de seleção de recursos são calculadas primeiro separadamente para cada classe na tarefa de classificação de duas classes *c* contra *c* e então combinados. Um método de combinação calcula uma única figura de mérito para cada característica, por exemplo, calculando a média dos valores *A(t, c)* para recurso *t,* e então seleciona o *k* recursos com maiores valores de mérito. Outro método de combinação freqüentemente usado seleciona o topo *k / n* recursos para cada um *n* classificadores e, em seguida, combina esses *n* conjuntos em um conjunto de recursos global.

**A precisão da classificação muitas vezes diminui ao selecionar *k* características comuns para um sistema com *n* classificadores em oposição a *n* diferentes conjuntos de tamanho *k.* Mas mesmo que isso aconteça, o ganho em eficiência devido a uma representação de documento comum pode compensar a perda de precisão.**

***13.5.5 Comparação de métodos de seleção de recursos***

Informações mútuas e *χ*2 representam métodos de seleção de recursos bastante diferentes. A independência de termo *t* e classe *c* às vezes pode ser rejeitado com alta confiança, mesmo se *t* carrega poucas informações sobre a associação de um documento em *c.* Isso é particularmente verdadeiro para termos raros. Se um termo ocorre uma vez em uma grande coleção e essa ocorrência está na classe *aves*, então isso é estatisticamente significativo. Mas uma única ocorrência não é muito informativa de acordo com a definição de informação teórica da informação. Porque seu critério é significativo, *χ*2 seleciona mais termos raros (que geralmente são indicadores menos confiáveis) do que **informações mútuas**. Mas o critério de seleção de informações mútuas também não seleciona necessariamente os termos que maximizam a precisão da classificação.

Apesar das diferenças entre os dois métodos, a precisão da classificação dos conjuntos de recursos selecionados com *χ*2 e MI não parece diferir sistematicamente. Na maioria dos problemas de classificação de texto, existem alguns indicadores fortes e muitos indicadores fracos. Desde que todos os indicadores fortes e um grande número de indicadores fracos sejam selecionados, espera-se que a precisão seja boa. Ambos os métodos fazem isso.

Figura 13,8 compara MI e *χ*2 seleção de recursos para o modelo multinomial. A eficácia máxima é virtualmente a mesma para os dois métodos. *χ*2 atinge esse pico mais tarde, com 300 recursos, provavelmente porque os raros, mas altamente significativos recursos que ele seleciona inicialmente não cobrem todos os documentos da classe. No entanto, os recursos selecionados posteriormente (na faixa de 100–300) são de melhor qualidade do que aqueles selecionados por MI.

**Todos os três métodos - MI, *χ*2 e com base na frequência** - são ***ambicioso* métodos**. Eles podem selecionar recursos que não contribuem com informações incrementais sobre seleção recursos selecionados anteriormente. Na figura13,7, kong é selecionado como o sétimo termo, embora seja altamente correlacionado com o previamente selecionado hong e, portanto, redundante. **Embora tal redundância possa afetar negativamente a precisão, os métodos não gananciosos raramente são usados na classificação do texto devido a seu custo computacional.**