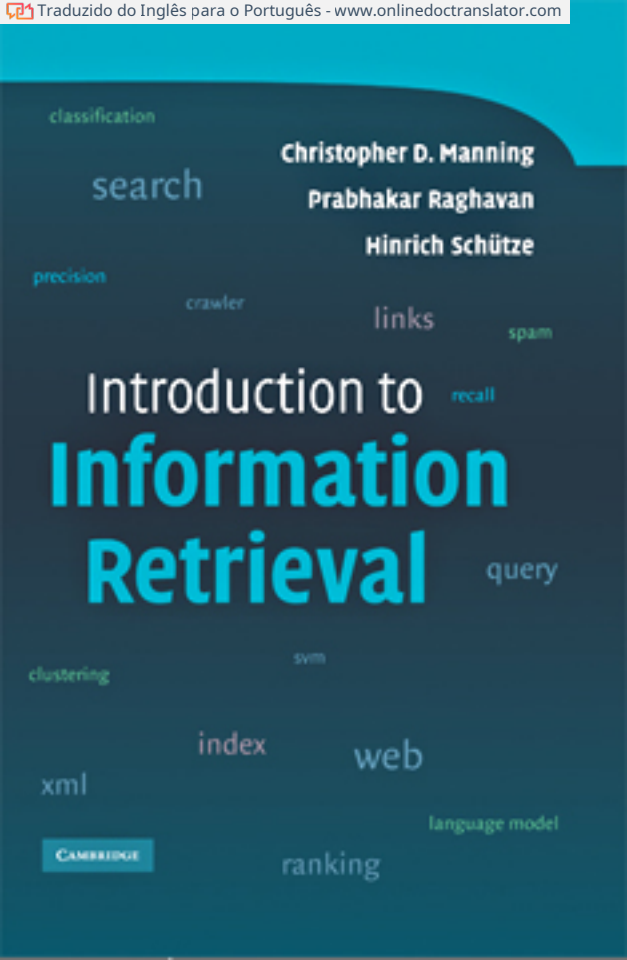

CAMBRIDGE www.cambridge.org/9780521865715

```
SelectFeatures (D, c, k) V ←
1  ExtractVocabulary (D)
2  eu ← []
3  para cada t ∈ V
4  Faz A (t, c) ← ComputeFeatureUtility (D, t, c)
5      Acrescentar(EU, ⟨A (t, c), t⟩ )
6  Retorna FeaturesWithLargestValues (L, k)
```

Figura 13.6 Algoritmo de seleção de recurso básico para selecionar o *k* Melhores características.

Exercício 13.4 Mesa 13,3 dá Bernoulli e estimativas multinomiais para o palavra a. Explique a diferença.

13.5 Seleção de recursos

recurso seleção

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

A seleção de recursos serve a dois propósitos principais. Primeiro, torna o treinamento e a aplicação de um classificador mais e fi cientes ao diminuir o tamanho do vocabulário efetivo. Isso é de particular importância para classificadores que, ao contrário do NB, são caros de treinar. Em segundo lugar, a seleção de recursos muitas vezes aumenta a classificação de classificação

recurso de ruído

curacia, eliminando recursos de ruído. UMA*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, digamosaracnocêntrico, não tem informações sobre uma classe, digamos *China,* mas todas as instâncias de aracnocêntrico aconteceu de ocorrer em *China* documentos em nosso conjunto de treinamento. Então, o método de aprendizagem pode produzir um classificador que atribui erroneamente documentos de teste contendoaracnocêntrico para *China.* Tal generalização incorreta de uma propriedade acidental do treinamento

superdimensionado

conjunto é chamado *superajuste.*

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 na Seção14,6 (página 284), 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 um dado classe *c,* nós calculamos uma medida de utilidade *A (t, c)* para cada termo do vocabulário e selecione o *k* 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: informações mútuas, $A(t, c) = I(U_t; C_c)$; a χ^2 teste, $A(t, c) = X^2(t, c)$; 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.* Seção 13.5.5 brevemente discute otimizações para sistemas com mais de duas classes.

13.5.1 Informação mútua

mútuo em formação

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 o correto classificação dezembrision o∑n c. Formalmente:

$$I(U; C) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t) P(C = e_c)} \quad (13,16)$$

Onde *você* é uma variável aleatória que assume valores $e_t = 1$ (o documento contém o termo *t)* e $e_t = 0$ (o documento não contém *t),* conforme definido na página 246, e C é uma variável aleatória que assume valores $e_c = 1$ (o documento está em aula *c)* e $e_c = 0$ (o documento não está em aula *c).* Nós escrevemos *você*$_t$ e C_c 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):

$$I(U; C) = \frac{N_{11}}{N} \log_2 \frac{N N_{11}}{N_{1.} N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{N N_{01}}{N_{0.} N_{.1}} + \frac{N_{10}}{N} \log_2 \frac{N N_{10}}{N_{1.} N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{N N_{00}}{N_{0.} N_{.0}} \quad (13,17)$$

onde o Ns são contagens de documentos que têm os valores de e_t e e_c isso são indicados pelos dois subscritos. Por exemplo, N_{10} é o número de documentos que contêm t ($e_t = 1$) e não estão em c ($e_c = 0$). $N_{1.} = N_{10} + N_{11}$ é o número de documentos que contêm t ($e_t = 1$) e contamos documentos independente da filiação à classe ($e_c \in \{0, 1\}$). $N = N_{00} + N_{01} + N_{10} + N_{11}$ é o número total de documentos. Um exemplo de um dos MLE estima que transformEquation (13,16) na Equação (13,17) é $P(U = 1, C = 1) = N_{11}/N$.

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

	$e_c = e_{aves} = 1$	$e_c = e_{aves} = 0$
$e_t = e_{export} = 1$	$N_{11} = 49$	$N_{10} = 141$
$e_t = e_{export} = 0$	$N_{01} = 27.652$	$N_{00} = 774.106$

[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. Veja a seção13,7.

Depois de conectar esses valores na Equação (13,17) Nós temos:

$$
\begin{aligned}
I(U;C) = {} & \frac{49}{801.948} \log_2 \frac{801.948 \cdot 49}{(49 + 27.652)(49 + 141)} \\
& + \frac{141}{801.948} \log_2 \frac{801.948 \cdot 141}{(141 + 774.106)(49 + 141)} \\
& + \frac{27.652}{801.948} \log_2 \frac{801.948 \cdot 27.652}{(49 + 27.652)(27.652 + 774.106)} \\
& + \frac{774.106}{801.948} \log_2 \frac{801.948 \cdot 774.106}{(141 + 774.106)(27.652 + 774.106)} \\
\approx {} & 0{,}000105
\end{aligned}
$$

Selecionar k termos $t_1, \ldots, t_k$ para uma determinada classe, usamos a seleção de recursos algoritmo na figura 13,6: Calculamos a medida de utilidade como $A(t, c) = I(U_t, C_c)$ 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, Londres, Reino Unido, Reino Unido para a aula *REINO UNIDO)* são de utilidade óbvia para tomar decisões de classificação para suas respectivas classes. No final da lista para *Reino Unido* encontramos termos como periféricos e esta noite (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_1 em função do tamanho do vocabulário após a seleção do recurso para Reuters-RCV1.[7] Comparando F_1 em 132.776 recursos (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 um 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

[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.

Reino Unido

Londres	0,1925
Reino Unido	0,0755
britânico	0,0596
stg	0,0555
Grã-Bretanha	0,0469
plc	0,0357
Inglaterra	0,0238
centavos	0,0212
libras	0,0149
inglês	0,0126

China

China	0,0997
chinês	0,0523
beiij ng	0,0444
yuan	0,0344
Xangai	0,0292
hong	0,0198
kong	0,0195
xinhua	0,0155
província	0,0117
Taiwan	0,0108

aves

aves	0,0013
eu no	0,0008
frango	0,0006
agricultura	0,0005
ave	0,0004
frango	0,0003
veterinário	0,0003
pássaros	0,0003
inspeção	0,0003
patogênico	0,0003

café

café	0,0111
bolsas	0,0042
cultivadores	0,0025
kg	0,0019
Colômbia	0,0018
Brasil	0,0016
exportar	0,0014
exportadores	0,0013
exportações	0,0013
cortar	0,0012

eleições

eleição	0,0519
eleições	0,0342
enquetes	0,0339
eleitores	0,0315
Festa	0,0303
voto	0,0299
votação	0,0225
candidato	0,0202
campanha	0,0202
democrático	0,0198

Esportes

futebol	0,0681
xícara	0,0515
partida	0,0441
fósforos	0,0408
reproduziu	0,0388
liga	0,0386
bater	0,0301
jogos	0,0299
jogos	0,0284
equipe	0,0264

Figura 13.7 Recursos com altas pontuações de informações mútuas para seis classes Reuters-RCV1.

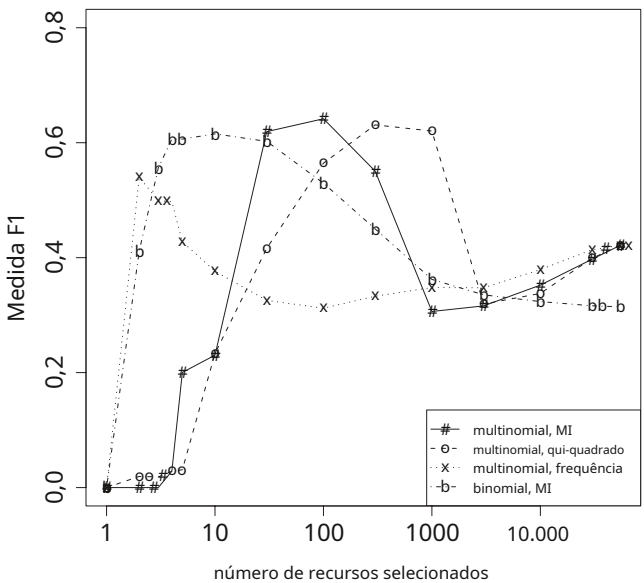

Figura 13.8 Efeito do tamanho do conjunto de recursos na precisão dos modelos multinomiais e de Bernoulli.

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.

13.5.2 χ^2 Seleção de recursos

χ^2 recurso seleção independência

Outro método popular de seleção de recursos é χ^2 Nas estatísticas, o χ^2 teste é aplicado para testar a independência de dois eventos, onde dois eventos A e B são definidos como *independente* E se $P(AB) = P(A)\,P(B)$ ou equivalente, $P(A \mid B) = P(A)$ e $P(B \mid A) = P(B)$. Na seleção de recursos, os dois eventos são ocorrências do prazo e ocorrência a da classe. Em seguida, classificamos os termos em relação a seguinte quantidade:

$$X^2(D, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \tag{13,18}$$

Onde e_t e e_c são definidos como na Equação (13,16) N é o *observado* frequência no D e E a *esperado* frequência. Por exemplo, E_{11} é a frequência esperada de t e c ocorrendo juntos em um documento assumindo que o termo e a classe são independente.

Exemplo 13.4: Nós primeiro calculamos E_{11} para os dados em Exemplo 13,3:

$$E_{11} = N \times P(t) \times P(c) = N \times \frac{N_{11} + N_{10}}{N} \times \frac{N_{11} + N_{01}}{N}$$

$$= N \times \frac{49 + 141}{N} \times \frac{49 + 27{,}652}{N} \approx 6{,}6$$

Onde N é o número total de documentos como antes.

Nós calculamos o outro $E_{e_t e_c}$ do mesmo jeito:

	$e_{aves} = 1$		$e_{aves} = 0$	
$e_{export} = 1$	$N_{11} = 49$	$E_{11} \approx 6{,}6$	$N_{10} = 141$	$E_{10} \approx 183{,}4$
$e_{export} = 0$	$N_{01} = 27.652$	$E_{01} \approx 27.694{,}4$	$N_{00} = 774.106$	$E_{00} \approx 774.063{,}6$

Conectando esses valores na Equação (13,18), temos um X^2 valor de 284:

$$X^2(D, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \approx 284$$

X^2 é uma medida de quanto conta esperada E e contagens observadas N desviar um do outro. Um alto valor de X^2 indica que a hipótese de independência, que implica que as contagens esperadas e observadas são semelhantes, está incorreta. Em nosso exemplo, $X^2 \approx 284 > 10{,}83$. Com base na tabela 13,6, podemos rejeitar a hipótese de que *aves* e exportar são independentes com

Tabela 13.6 Valores críticos do χ^2 distribuição com um grau de liberdade. Por exemplo, se os dois eventos são independentes, então $P(X^2 > 6{,}63) < 0{,}01$. Então para $X^2 > 6{,}63$ a suposição de independência pode ser rejeitada com 99% de con
confiança.

p	χ^2 valor crítico
0,1	2,71
0,05	3,84
0,01	6,63
0,005	7,88
0,001	10,83

estatístico signi fi cância

apenas 0,001 de chance de estar errado.[8] Equivalentemente, dizemos que o resultado $X^2 \approx 284 > 10{,}83$ 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 χ^2 seleção de recursos.

Uma maneira aritmeticamente mais simples de calcular X^2 é o seguinte:

$$X^2(D, t, c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})} \tag{13,19}$$

Isso é equivalente à Equação (13,18) (Exercício 13,14)

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 (ver Seção13,7), 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, o*relativo* A importância dos recursos é importante. Enquantoχ^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 $X^2 \sim \chi^2$, Onde χ^2 é o χ^2 distribuição. Veja, por exemplo,Arroz (2006)

13.5.3 Seleção de recursos com base em frequência

Um terceiro método de seleção de recursos é *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 $n \times 2$ 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^2 estatística como antes.

Mais comumente, as estatísticas de seleção de recursos são calculadas primeiro separadamente para cada classe na tarefa de classificação de duas classes $\overline{c}$ 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 classi fi ers 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 prazo 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 no *aves* classe, 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.

ambicioso recurso seleção

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 recursos selecionados anteriormente. Na figura 13,7, 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 (consulte a Seção 13,7 para referências) raramente são usados na classificação do texto devido a seu custo computacional.

? **Exercício 13.5** Considere o termos mais primeiros 100.000 seguintes frequências para a classe *café* para quatro documentos da Reuters-RCV1:

prazo	N_{00}	N_{01}	N_{10}	N_{11}
Brasil	98.012	102	1835	51
conselho	96.322	133	3525	20
produtores	98.524	119	1118	34
assado	99.824	143	23	10

Selecione dois desses quatro termos com base em (i) χ^2, (ii) informação mútua, (iii) frequência.

13.6 Avaliação da classificação do texto

Historicamente, a coleção clássica Reuters-21578 foi a principal referência para avaliação de classificação de texto. Esta é uma coleção de 21.578 artigos de notícias, originalmente coletados e rotulados por Carnegie Group, Inc. e Reuters,