



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP.HCM

Khoa Công nghệ thông tin

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

BÀI TẬP 03

Phân lớp
Classification


Giảng viên hướng dẫn: Nguyễn Ngọc Thảo

Người thực hiện: Nhóm 37

Hà Tiến Đạt – 18424023

Vũ Mạnh Hùng -18424029

TP.HCM – 09/2019



Mục lục

I) Nội dung thực hiện báo cáo viết (10 điểm)	3
1) Bộ lọc StringToWordVector chuyển chuỗi ký tự thành nhiều thuộc tính số (@attribute). Bạn đếm được bao nhiêu thuộc tính số trong bảng classifier output?	4
2) Thuộc tính class (tức là “ý kiến” của mỗi tweet) có bị ảnh hưởng bởi bộ lọc không?	4
3) Ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Tóm lại, bạn nhận thấy bộ phân lớp đã thực thi như thế nào? Bạn có hài lòng với kết quả phân lớp này không? Tại sao?.....	5
4) Nhấn StringToWordVector để hiển thị cửa sổ chứa nhiều tùy chọn. Các tùy chọn này là tham số ảnh hưởng đến hành vi của bộ lọc và do đó cũng ảnh hưởng đến bộ phân lớp về mặt tổng thể. Nhấn More và đọc mô tả của các tham số. Sau khi đã đọc hiểu mọi tham số, bạn hãy tập trung vào tham số minTermFreq. Hiệu chỉnh giá trị của tham số này. Đầu tiên đặt giá trị bằng 5. Chạy lại bộ phân lớp, phân tích kết quả đầu ra, và ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bạn nhận thấy bộ phân lớp đã thực thi như thế nào?	7
5) Tiếp đó đặt giá trị của tham số minTermFreq bằng 10. Chạy lại bộ phân lớp, phân tích kết quả đầu ra, và ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bạn nhận thấy bộ phân lớp đã thực thi như thế nào?	8
6) Bạn có thể giải thích chức năng của tham số minTermFreq thông qua cách thức mà tham số này tác động đến hiệu quả phân lớp?	10
7) Phục hồi giá trị của tham số minTermFreq về 1. Tải tập tin hư từ (stopword) về máy tính từ địa chỉ sau, http://stp.lingfil.uu.se/~santini/sais/2016/stopwords_eng.txt . Thiết lập tham số useStoplist thành True và chỉ định các tập tin stopwords_eng.txt vào trường stopwords. Đọc kỹ nội dung của bảng classifier output. Bạn đếm được bao nhiêu thuộc tính trong bảng classifier output?	11
8) Ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bộ phân lớp hoạt động như thế nào so với kết quả thực thi trong những câu hỏi trước?.....	12
9) Bạn sẽ làm thế nào để tăng sức ảnh hưởng của danh sách hư từ lên việc phân lớp? Hãy đưa ra một vài kiến nghị (ví dụ thêm nhiều từ trong tweet vào tập tin danh sách hư từ, hoặc giảm số từ trong tập tin, loại bỏ/thêm vào/xử lý phủ định, v.v.)	13
10) Bạn được tùy chọn một tham số từ danh sách tham số của bộ lọc, ngoài những tham số bạn đã trải nghiệm trong các câu hỏi bên trên. Mô tả tham số và giải thích lý do bạn chọn tham số này. Ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bộ phân lớp hoạt động như thế nào với cấu hình tham số mà bạn đã chọn? So sánh với các lượt chạy trước đó.....	15
II) Nội dung thực hiện cài đặt (10 điểm)	17
1) (1.0đ) Chương trình nhận dữ liệu đầu vào là tập tin *.csv có cấu trúc như sau.....	17
2) (3.0đ) Chương trình phát sinh dữ liệu đầu ra là tập tin model.txt chứa thông tin tương tự như trong phần văn bản của cửa sổ Classifier output (tab Classify – WEKA), bao gồm.....	18

- 3) (1.0đ) Chương trình thực thi giải thuật ID3 và đánh giá giải thuật bằng phương pháp n-folds cross validation với cú pháp tham số dòng lệnh như sau 18
- 4) (5.0đ) Tùy chọn 3 tập dữ liệu có quy mô nhỏ (~100 mẫu), trung bình (~500 mẫu), và lớn (~1000 mẫu). Chạy chương trình cài đặt với các tập dữ liệu đã chọn và đối chiếu kết quả phát sinh được với kết quả của WEKA ID3 trên cùng bộ tham số 19
- III) Nguồn tham khảo: 19

I) Nội dung thực hiện báo cáo viết (10 điểm)

Dữ liệu thực nghiệm

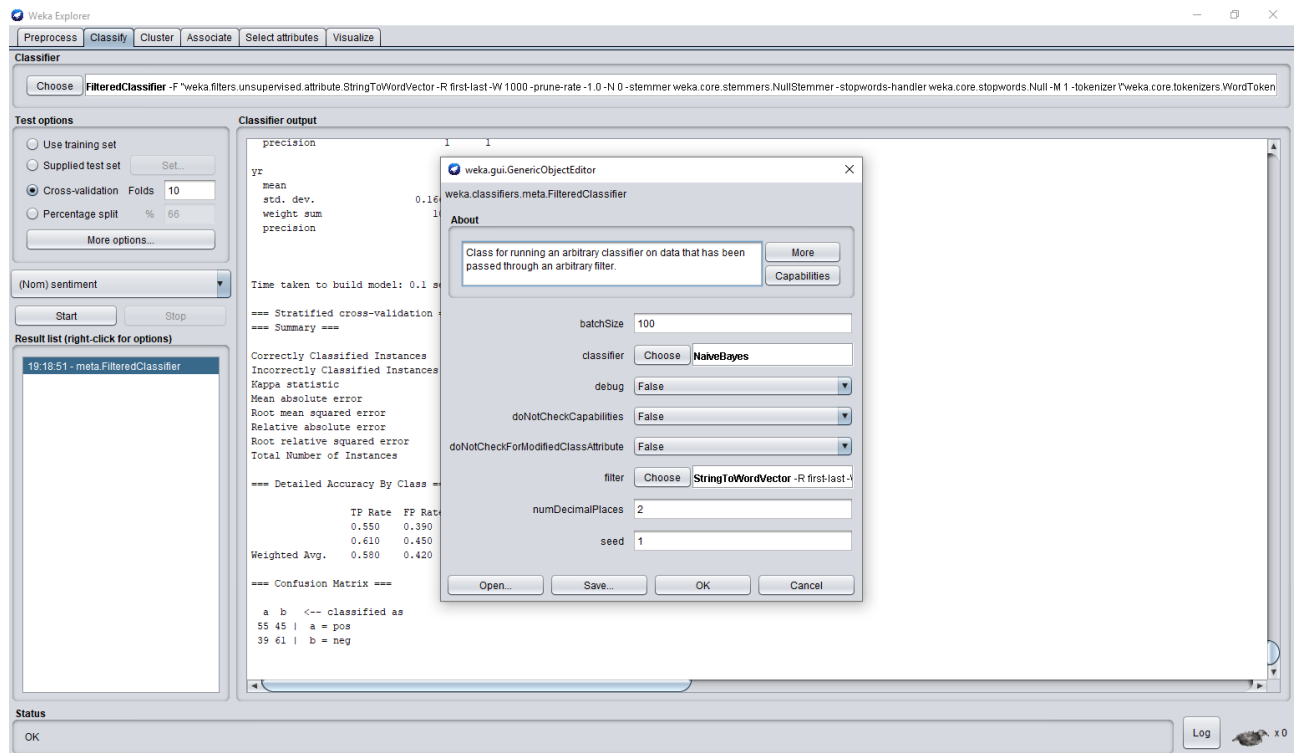
Sinh viên làm việc trên tập dữ liệu gồm các bài viết đăng trên Twitter (còn gọi là tweet). Tập dữ liệu này được trích từ nguồn “ngữ liệu Edinburgh Twitter” trong công trình khoa học của Petrovic và cộng sự (2010). Tweets được sử dụng rộng rãi trong bài toán phân tích ý kiến (sentiment analysis), và nhà kiến tạo ngữ liệu đã đưa ra lý do như sau: “Dịch vụ blog mini Twitter đã và đang trở thành công cụ phổ biến để thể hiện ý kiến, loan truyền tin tức hay đơn giản là liên lạc với bạn bè. Người ta thường bình luận về những sự kiện trong thời gian thực, với hàng trăm bài viết nhỏ (tweets) được đăng mỗi giây cho những sự kiện quan trọng.” (Petrovic và cộng sự, 2010). Bài thực hành này chỉ thao tác trên một tập dữ liệu con nhỏ hơn của toàn bộ ngữ liệu, vốn được tạo ra nhằm mục đích học tập. Tập dữ liệu nhỏ ở định dạng ARFF, bao gồm 100 tweets dương (positive) và 100 tweet âm (negative).

Tập dữ liệu gồm hai thuộc tính, `tweet_body` (kiểu dữ liệu string, chứa nội dung văn bản của mỗi tweet) và `class` (kiểu dữ liệu nominal, mang một trong hai giá trị, dương (positive, pos) hoặc âm (negative, neg)).

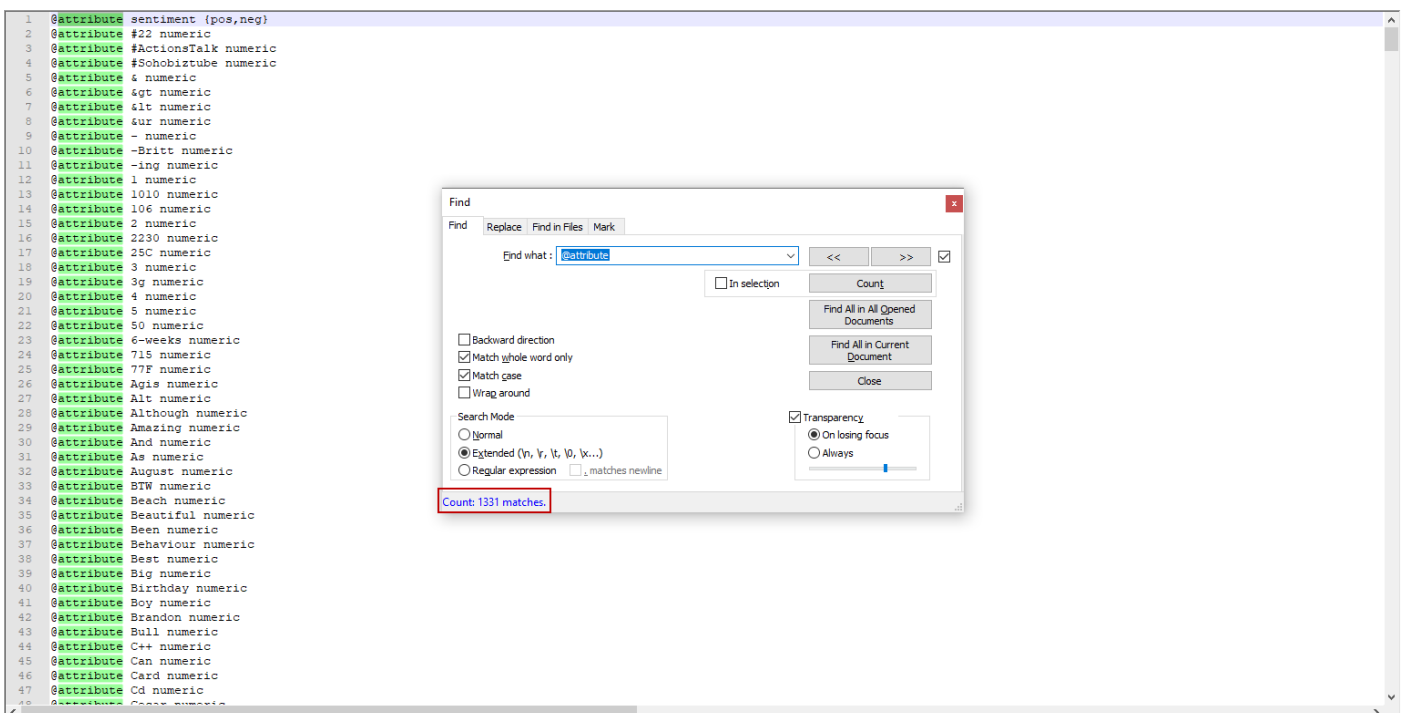
Khởi động Weka và chọn giao diện Explorer. Mở tập dữ liệu thực nghiệm trong Weka (Preprocess → Open File). Di chuyển đến tab Classify. Nhấn nút Choose và tiếp đó chọn Meta → FilteredClassifier. Nhấn lên tên của FilteredClassifier để hiển thị cửa sổ tham số.

Trong cửa sổ này, bạn có thể chọn bộ phân lớp và bộ lọc rút trích đặc trưng tương ứng. Chọn bộ phân lớp naïve Bayes (Classifier → bayes → NaiveBayes) và bộ lọc StringToWordVector (filter → unsupervised → StringToWordVector). StringToWordVector có chức năng chuyển đổi chuỗi ký tự (tức là nội dung của tweet) thành vector từ khóa.

Nhấn OK và tiếp đó nhấn Start.



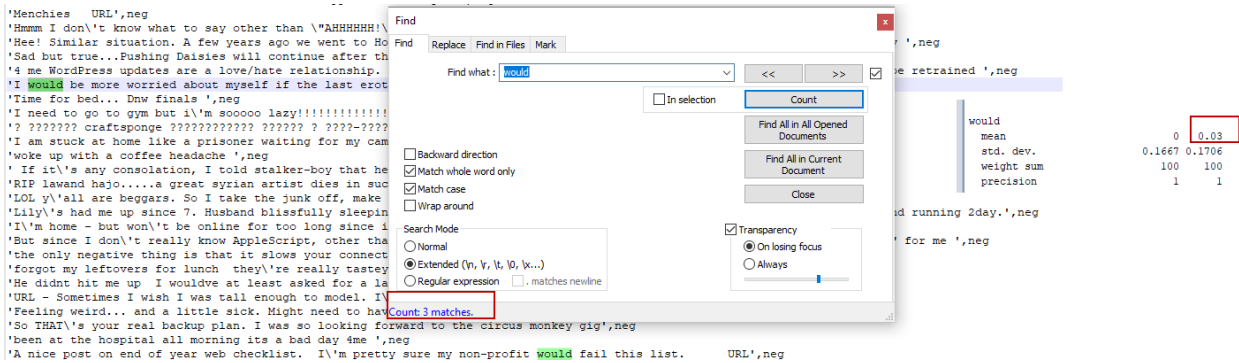
- 1) Bộ lọc StringToWordVector chuyển chuỗi ký tự thành nhiều thuộc tính số (@attribute). Bạn đếm được bao nhiêu thuộc tính số trong bảng classifier output?



Số thuộc tính trong bảng classifier output là 1131 thuộc tính. Trong đó có **1130** thuộc tính số

- 2) Thuộc tính class (tức là “ý kiến” của mỗi tweet) có bị ảnh hưởng bởi bộ lọc không?

Thuộc tính class không bị ảnh hưởng bởi bộ lọc



- 3) Ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Tóm lại, bạn nhận thấy bộ phân lớp đã thực thi như thế nào? Bạn có hài lòng với kết quả phân lớp này không? Tại sao?

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      116          58      %
Incorrectly Classified Instances    84           42      %
Kappa statistic                    0.16
Mean absolute error                 0.4348
Root mean squared error             0.5692
Relative absolute error             86.9544 %
Root relative squared error         113.8342 %
Total Number of Instances          200

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.550	0.390	0.585	0.550	0.567	0.160	0.593	0.587	pos
	0.610	0.450	0.575	0.610	0.592	0.160	0.593	0.579	neg
Weighted Avg.	0.580	0.420	0.580	0.580	0.580	0.160	0.593	0.583	

```

=== Confusion Matrix ===
 a b  <-- classified as
55 45 | a = pos
39 61 | b = neg

```

- Confusion matrix:

		Predicted Class		
		pos	neg	
Actual Class	pos	TP = 55	FN = 45	P =100
	neg	FP = 39	TN =61	N = 100
		P' = 94	N' = 106	All = 200

- Accuracy:

$$= \frac{TP + TN}{All} = \frac{55 + 61}{200} = 0.58$$

- TP-Rate:

- Theo pos: 0.55
- Theo neg: 0.61
- Trung bình: 0.58

Tài liệu

- FP-Rate:
 - Theo pos: 0.39
 - Theo neg: 0.45
 - Trung bình: 0.42
- Precision:
 - Theo pos: 0.585
 - Theo neg: 0.575
 - Trung bình: 0.58
- Recall:
 - Theo pos: 0.55
 - Theo neg: 0.61
 - Trung bình: 0.58
- F-Measure:
 - Theo pos: 0.567
 - Theo neg: 0.592
 - Trung bình: 0.58
- Nhận xét:

Bộ phân lớp thực thi tương đối tốt khi kết quả phân lớp có 116 giá trị phân lớp đúng chiếm 58%. Tỷ lệ mẫu mà bộ phân lớp gán nhãn pos thực sự là pos (độ chính xác) là 0.585. Tỷ lệ mẫu pos mà bộ phân lớp đã gán nhãn được (độ toàn vẹn) là 0.55.

Tóm lại khi xét trên mẫu dữ liệu là kiểu text (văn bản) muốn phân lớp thì phải tiền xử lý thông qua bộ lọc. Khi đó kết quả phân lớp sẽ bị ảnh hưởng bởi kết quả của bộ lọc. Nên với các độ đo đều trên mức trung bình thì bộ phân lớp thực thi tốt.

File kết quả đính kèm: **Result_minTermFreq_1.txt**

- 4) Nhấn StringToWordVector để hiển thị cửa sổ chứa nhiều tùy chọn. Các tùy chọn này là tham số ảnh hưởng đến hành vi của bộ lọc và do đó cũng ảnh hưởng đến bộ phân lớp về mặt tổng thể. Nhấn More và đọc mô tả của các tham số. Sau khi đã đọc hiểu mọi tham số, bạn hãy tập trung vào tham số minTermFreq. Hiệu chỉnh giá trị của tham số này. Đầu tiên đặt giá trị bằng 5. Chạy lại bộ phân lớp, phân tích kết quả đầu ra, và ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bạn nhận thấy bộ phân lớp đã thực thi như thế nào?

```
Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      105          52.5  %
Incorrectly Classified Instances    95          47.5  %
Kappa statistic                    0.05
Mean absolute error                 0.4609
Root mean squared error             0.5731
Relative absolute error             92.1732 %
Root relative squared error         114.6284 %
Total Number of Instances          200

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.550	0.500	0.524	0.550	0.537	0.050	0.562	0.557	pos
	0.500	0.450	0.526	0.500	0.513	0.050	0.562	0.541	neg
Weighted Avg.	0.525	0.475	0.525	0.525	0.525	0.050	0.562	0.549	

```

=== Confusion Matrix ===
  a  b  <-- classified as
55 45 |  a = pos
50 50 |  b = neg

```

- Confusion matrix:

		Predicted Class		
		pos	neg	
Actual Class	pos	TP = 55	FN = 45	P = 100
	neg	FP = 50	TN = 50	N = 100
		P' = 105	N' = 95	All = 200

- Accuracy:

$$= \frac{TP + TN}{All} = \frac{55 + 61}{200} = 0.58$$

- TP-Rate:

- Theo pos: 0.55
- Theo neg: 0.5
- Trung bình: 0.525

- FP-Rate:

- Theo pos: 0.5
- Theo neg: 0.45

- Trung bình: 0.475
- Precision:
 - Theo pos: 0.524
 - Theo neg: 0.526
 - Trung bình: 0.525
- Recall:
 - Theo pos: 0.55
 - Theo neg: 0.5
 - Trung bình: 0.525
- F-Measure:
 - Theo pos: 0.537
 - Theo neg: 0.513
 - Trung bình: 0.525
- Nhận xét:

Khi minTermFreq = 5 thì kết quả phân lớp thay đổi theo có 105 giá trị phân lớp đúng chiếm 52.5%. So sánh các độ đo với minTermFreq = 1 thì giảm

Như vậy minTermFreq tác động đến hiệu quả phân lớp và có thể tỉ lệ với hiệu quả phân lớp. Tăng minTermFreq thì hiệu quả phân lớp có thể giảm đi và ngược lại

File kết quả đính kèm: **Result_minTermFreq_5.txt**

- 5) Tiếp đó đặt giá trị của tham số minTermFreq bằng 10. Chạy lại bộ phân lớp, phân tích kết quả đầu ra, và ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bạn nhận thấy bộ phân lớp đã thực thi như thế nào?

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      109          54.5 %
Incorrectly Classified Instances    91           45.5 %
Kappa statistic                    0.09
Mean absolute error                 0.4737
Root mean squared error             0.5324
Relative absolute error             94.7486 %
Root relative squared error         106.47 %
Total Number of Instances          200

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.620	0.530	0.539	0.620	0.577	0.091	0.571	0.585	pos
	0.470	0.380	0.553	0.470	0.508	0.091	0.571	0.530	neg
Weighted Avg.	0.545	0.455	0.546	0.545	0.542	0.091	0.571	0.558	

```

=== Confusion Matrix ===

 a b  <-- classified as
62 38 | a = pos
53 47 | b = neg

```

- Confusion matrix:

		Predicted Class		
		pos	neg	
Actual Class	pos	TP = 62	FN = 38	P =100
	neg	FP = 53	TN =47	N = 100
		P' = 115	N' = 85	All = 200

- Accuracy:

$$= \frac{TP + TN}{All} = \frac{62 + 47}{200} = 0.545$$

- TP-Rate:

- Theo pos: 0.62
- Theo neg: 0.47
- Trung bình: 0.545

- FP-Rate:

- Theo pos: 0.53
- Theo neg: 0.38
- Trung bình: 0.455

- Precision:

- Theo pos: 0.539
- Theo neg: 0.553
- Trung bình: 0.546

- Recall:

- Theo pos: 0.62
- Theo neg: 0.47
- Trung bình: 0.545

- F-Measure:

- Theo pos: 0.577
- Theo neg: 0.508
- Trung bình: 0.542

- Nhận xét:

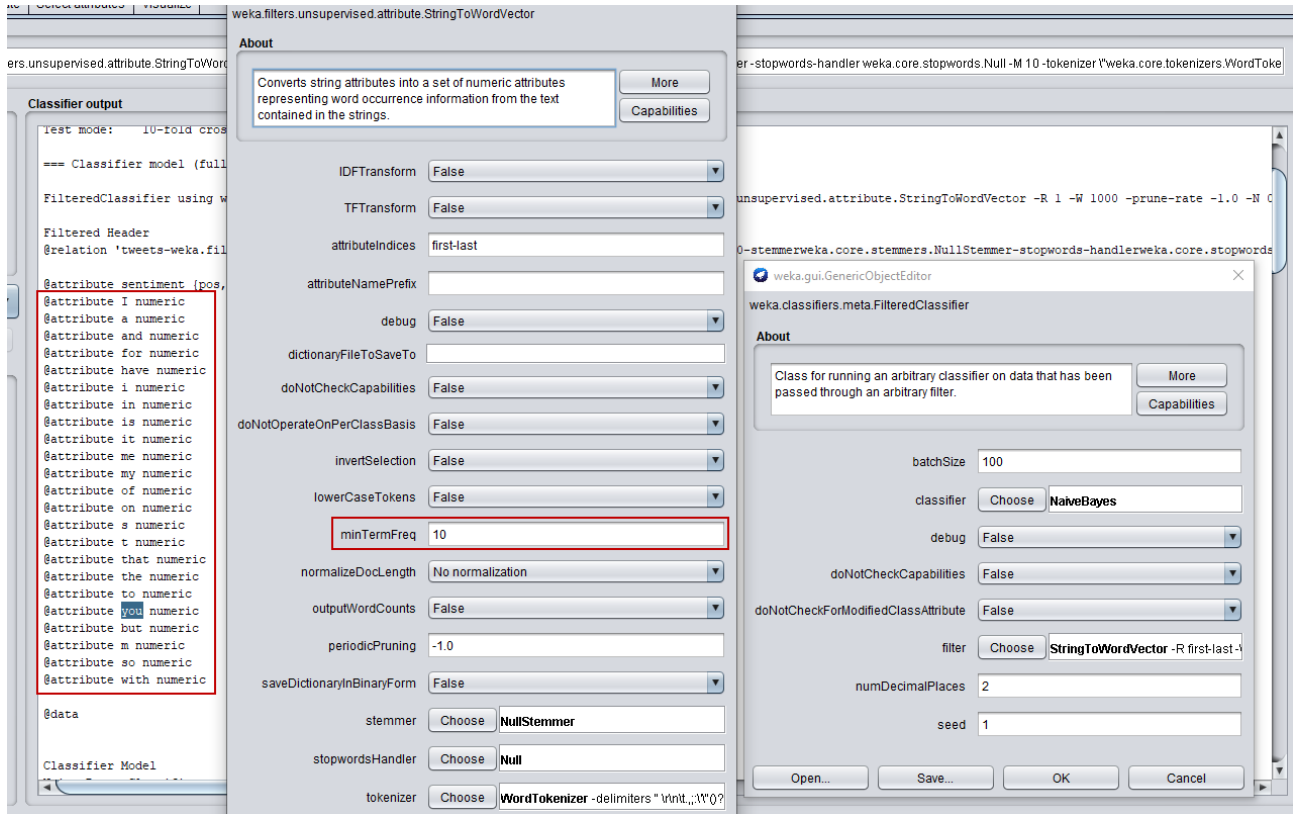
Khi minTermFreq = 10 thì kết quả phân lớp thay đổi theo có 109 giá trị phân lớp đúng chiếm 54.5%. So sánh các độ đo với minTermFreq = 5 thì tăng nhưng vẫn nhỏ hơn so với minTermFreq = 1

Như vậy minTermFreq không thật sự tỷ lệ với hiệu quả phân lớp. Tăng minTermFreq thì hiệu quả phân lớp chưa chắc đã giảm đi và ngược lại

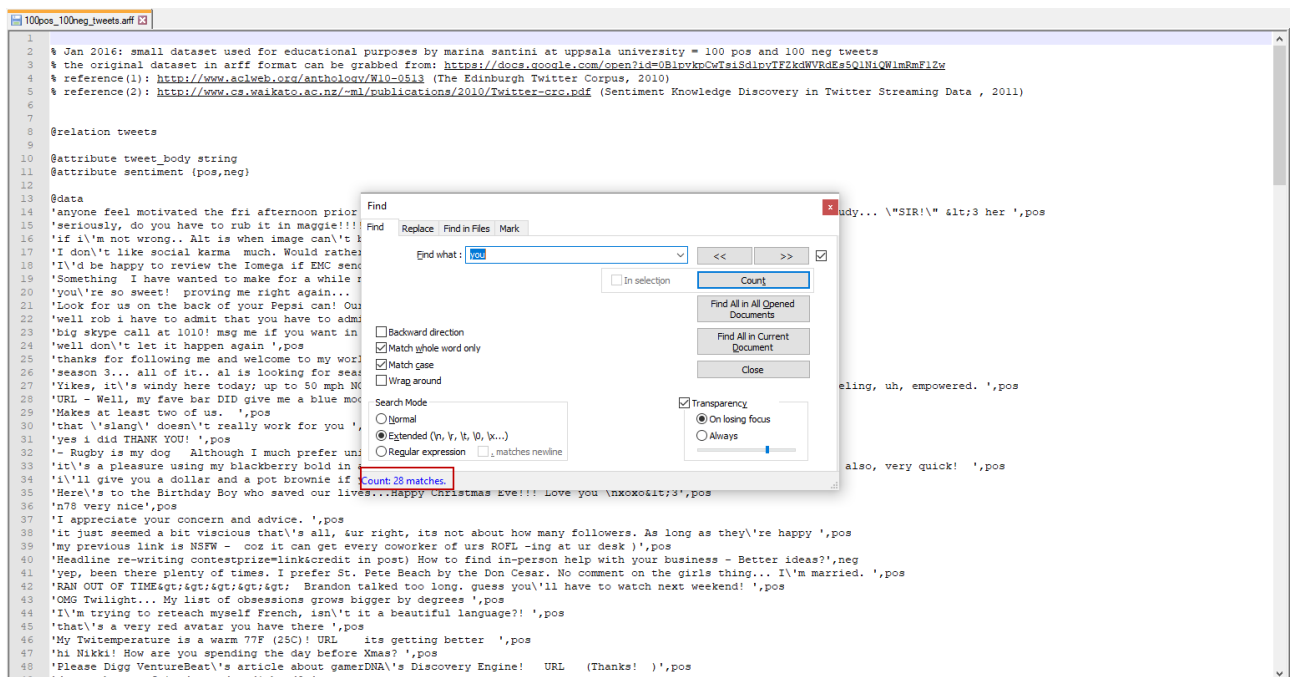
File kết quả đính kèm: **Result_minTermFreq_10.txt**

6) Bạn có thể giải thích chức năng của tham số minTermFreq thông qua cách thức mà tham số này tác động đến hiệu quả phân lớp?

- Tham số minTermFreq là The minimum term frequency – **Tần số giới hạn tối thiểu**. Được thực thi trên cơ sở của mỗi lớp. Cụ thể khi thiết lập tham số minTermFreq = 10 cho filter StringToWordVector thì các từ được xem xét phải xuất hiện ít nhất 10 lần.



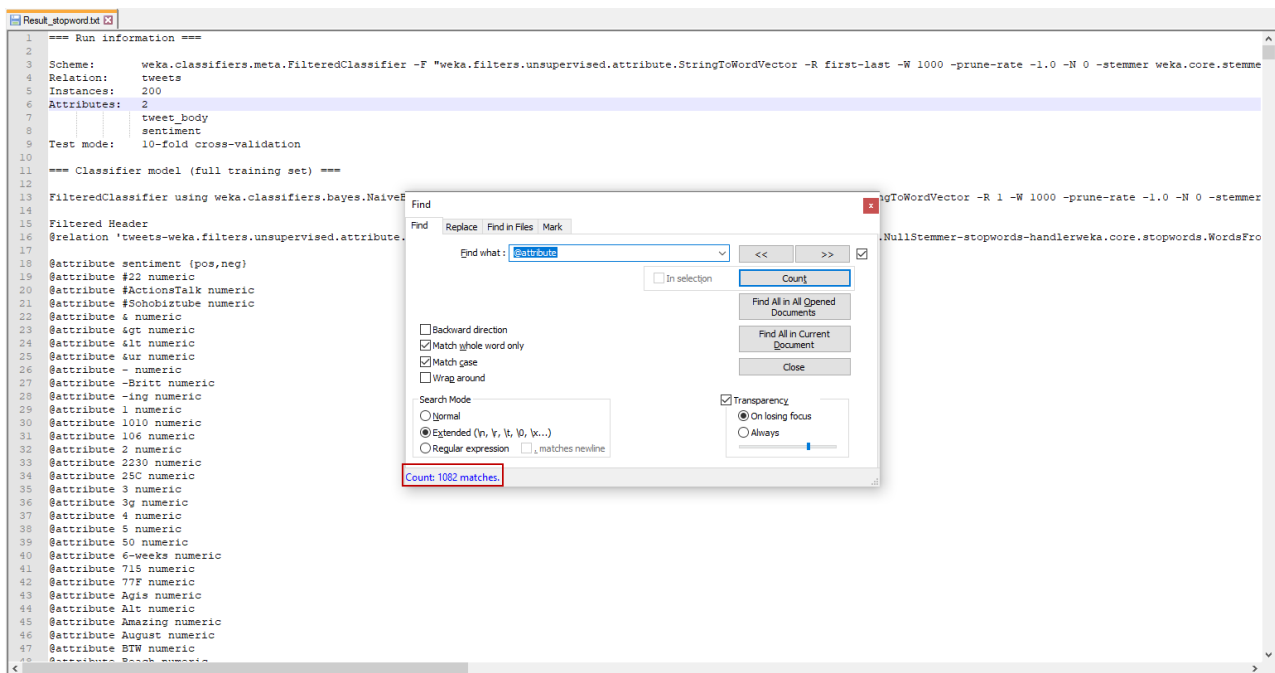
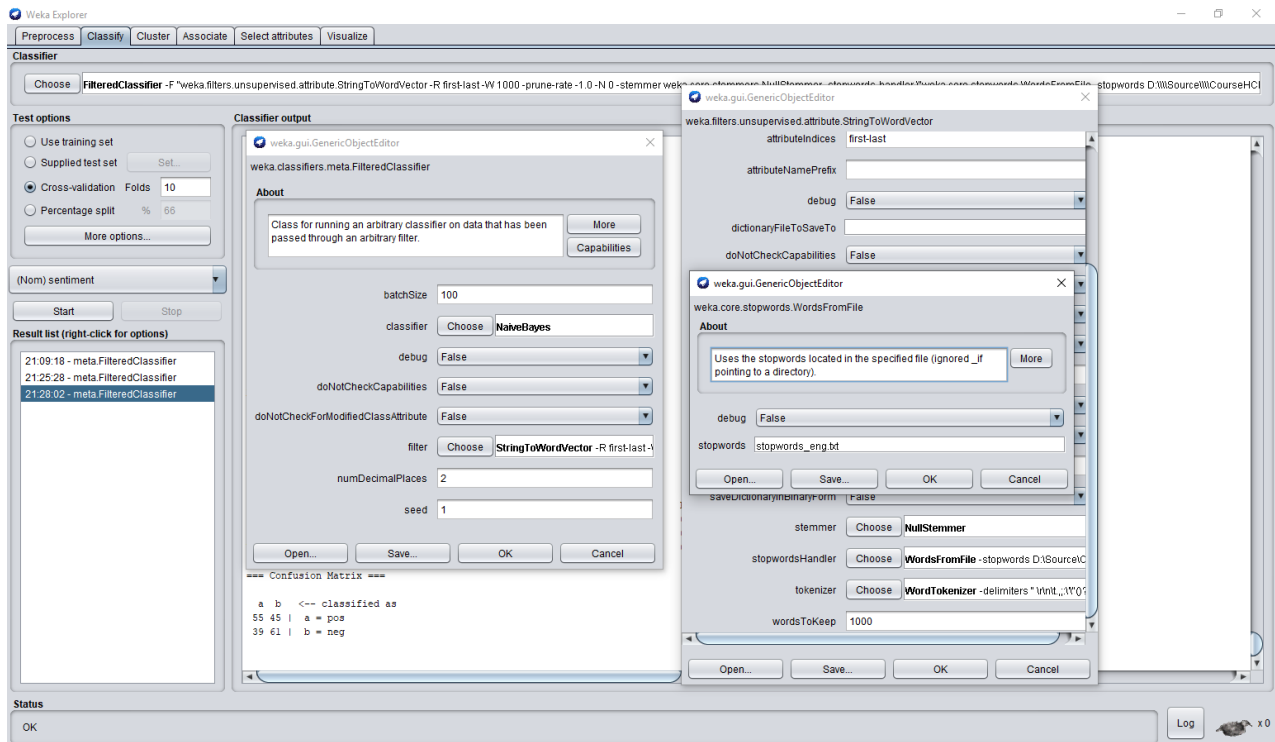
Ví dụ từ "you" xuất hiện 28 lần được xem xét



- Trong trường hợp này việc dựa trên tần số xuất hiện của các từ trong văn bản để filter rồi từ đó tiến hành phân lớp không đem lại hiệu quả hơn. Từ xuất hiện nhiều chưa chắc đã thể

hiện đúng phân lớp. Còn phải phụ thuộc vào ý nghĩa của từ. Tác động của các từ không có ý nghĩa hay mang ý nghĩa chung chung không cụ thể nhưng lại xuất hiện nhiều (stopword) .

- 7) Phục hồi giá trị của tham số minTermFreq về 1. Tải tập tin hư từ (stopword) về máy tính từ địa chỉ sau, http://stp.lingfil.uu.se/~santinim/sais/2016/stopwords_eng.txt. Thiết lập tham số useStoplist thành True và chỉ định các tập tin stopwords_eng.txt vào trường stopwords. Đọc kỹ nội dung của bảng classifier output. Bạn đếm được bao nhiêu thuộc tính trong bảng classifier output?



Số thuộc tính trong bảng classifier output là **1082** thuộc tính. Trong đó có **1081** thuộc tính số

- 8) Ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bộ phân lớp hoạt động như thế nào so với kết quả thực thi trong những câu hỏi trước?

```
Time taken to build model: 0.08 seconds
```

```
=== Stratified cross-validation ===
=== Summary ===
```

```
Correctly Classified Instances      107          53.5  %
Incorrectly Classified Instances    93           46.5  %
Kappa statistic                    0.07
Mean absolute error                0.4773
Root mean squared error            0.5461
Relative absolute error            95.4535 %
Root relative squared error        109.2126 %
Total Number of Instances         200
```

```
=== Detailed Accuracy By Class ===
```

```

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0.580    0.510    0.532    0.580    0.555      0.070    0.546    0.538    pos
      0.490    0.420    0.538    0.490    0.513      0.070    0.546    0.549    neg
Weighted Avg.  0.535    0.465    0.535    0.535    0.534      0.070    0.546    0.543

```

```
=== Confusion Matrix ===
```

```

 a  b  <-- classified as
58 42 |  a = pos
51 49 |  b = neg

```

- Confusion matrix:

		Predicted Class		
		pos	neg	
Actual Class	pos	TP = 58	FN = 42	P = 100
	neg	FP = 51	TN = 49	N = 100
		P' = 109	N' = 91	All = 200

- Accuracy:

$$= \frac{TP + TN}{All} = \frac{58 + 49}{200} = 0.535$$

- TP-Rate:

- Theo pos: 0.58
- Theo neg: 0.49
- Trung bình: 0.535

- FP-Rate:

- Theo pos: 0.51
- Theo neg: 0.42
- Trung bình: 0.465

- Precision:

Tài liệu

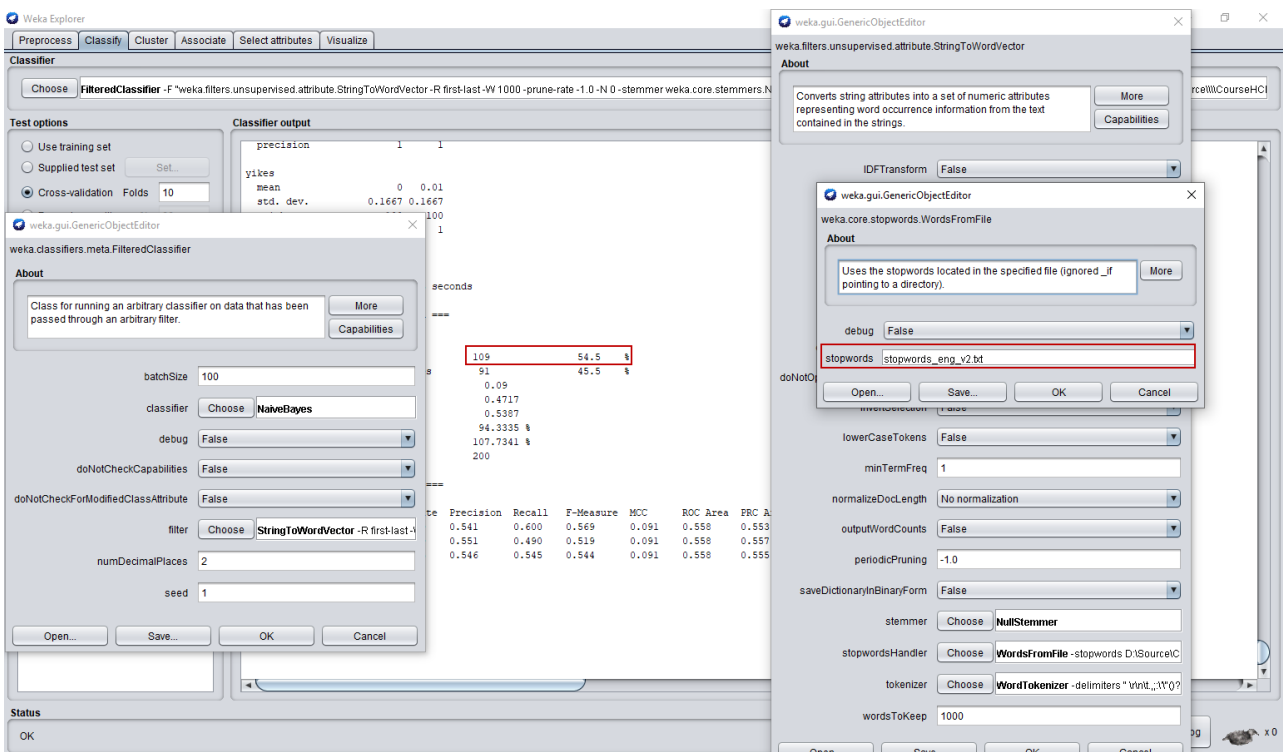
- Theo pos: 0.32
- Theo neg: 0.38
- Trung bình: 0.535
- Recall:
 - Theo pos: 0.58
 - Theo neg: 0.49
 - Trung bình: 0.535
- F-Measure:
 - Theo pos: 0.555
 - Theo neg: 0.513
 - Trung bình: 0.534
- Nhận xét:

So với minTermFreq = 5 thì thi tốt hơn nhưng không bằng khi minTermFreq = 1 hay minTermFreq = 10

File kết quả đính kèm: **Result_stopword.txt**

9) Bạn sẽ làm thế nào để tăng sức ảnh hưởng của danh sách hư từ lên việc phân lớp? Hãy đưa ra một vài kiến nghị (ví dụ thêm nhiều từ trong tweet vào tập tin danh sách hư từ, hoặc giảm số từ trong tập tin, loại bỏ/thêm vào/xử lý phủ định, v.v.)

- Thêm nhiều từ trong tweet vào tập tin danh sách hư từ:



Tài liệu

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	109	54.5	%
Incorrectly Classified Instances	91	45.5	%
Kappa statistic	0.09		
Mean absolute error	0.4717		
Root mean squared error	0.5387		
Relative absolute error	94.3335 %		
Root relative squared error	107.7341 %		
Total Number of Instances	200		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.600	0.510	0.541	0.600	0.569	0.091	0.558	0.553	pos
	0.490	0.400	0.551	0.490	0.519	0.091	0.558	0.557	neg
Weighted Avg.	0.545	0.455	0.546	0.545	0.544	0.091	0.558	0.555	

=== Confusion Matrix ===

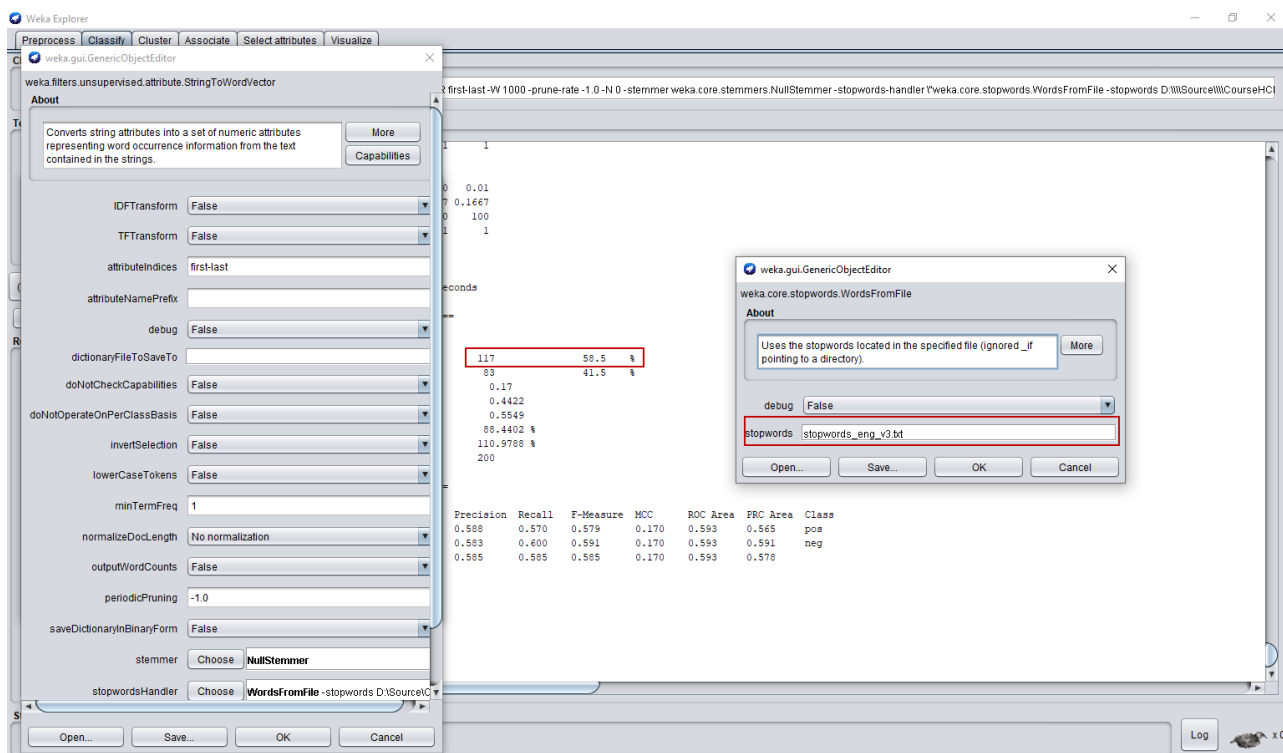
```

a b  <-- classified as
60 40 | a = pos
51 49 | b = neg

```

File kết quả đính kèm: **Result_stopword_v2.txt**

- Giảm số từ trong tập tin hư từ:



```
Time taken to build model: 0.04 seconds
```

```
=== Stratified cross-validation ===
```

```
=== Summary ===
```

Correctly Classified Instances	117	58.5	%
Incorrectly Classified Instances	83	41.5	%
Kappa statistic	0.17		
Mean absolute error	0.4422		
Root mean squared error	0.5549		
Relative absolute error	88.4402	%	
Root relative squared error	110.9788	%	
Total Number of Instances	200		

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.570	0.400	0.588	0.570	0.579	0.170	0.593	0.565	pos
	0.600	0.430	0.583	0.600	0.591	0.170	0.593	0.591	neg
Weighted Avg.	0.585	0.415	0.585	0.585	0.585	0.170	0.593	0.578	

```
=== Confusion Matrix ===
```

```

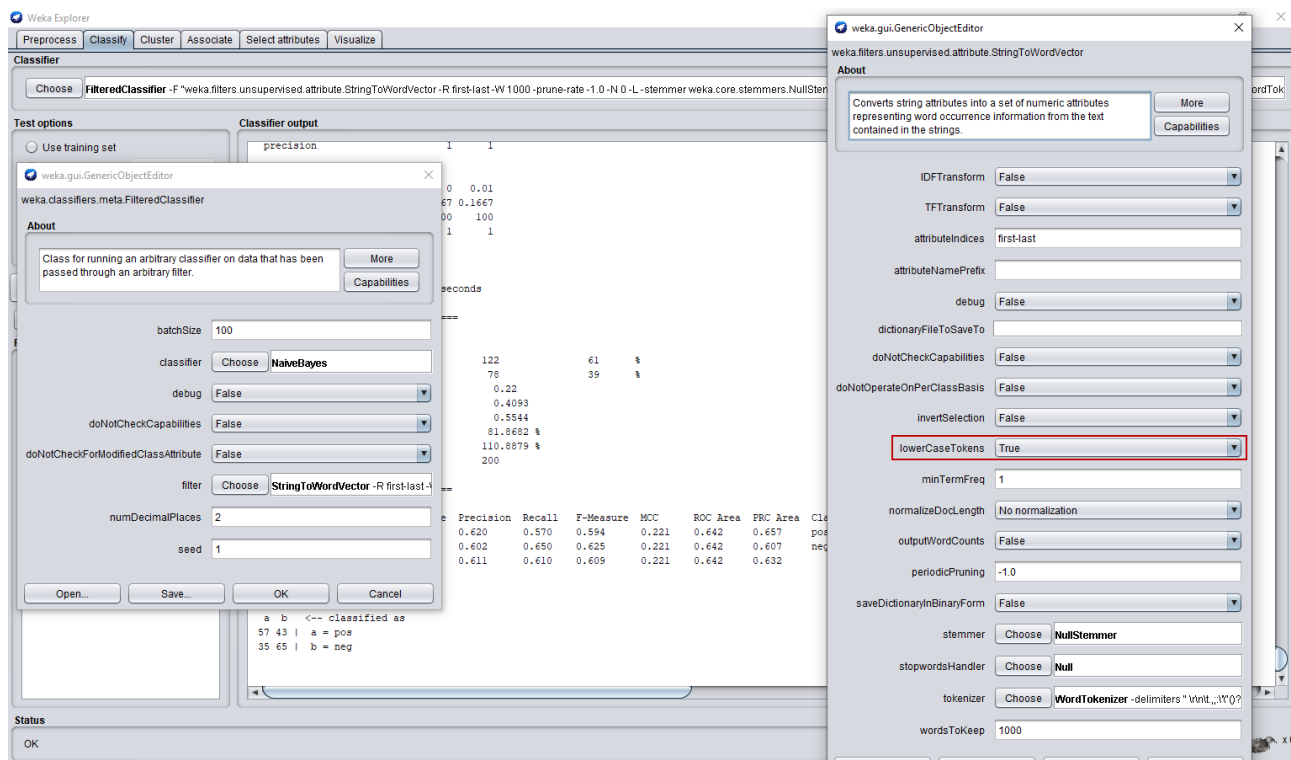
a  b  <-- classified as
57 43 | a = pos
40 60 | b = neg
```

File kết quả đính kèm: **Result_stopword_v3.txt**

- 10) Bạn được tùy chọn một tham số từ danh sách tham số của bộ lọc, ngoài những tham số bạn đã trải nghiệm trong các câu hỏi bên trên. Mô tả tham số và giải thích lý do bạn chọn tham số này. Ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bộ phân lớp hoạt động như thế nào với cấu hình tham số mà bạn đã chọn? So sánh với các lượt chạy trước đó.

Chọn lowerCaseTokens = True để chuyển tất cả các ký tự thành thường. Do trong các câu tweet chữ Hoa và chữ Thường không khác nhau mấy về ý nghĩa. Ngoài ra do các câu tweet do người dùng viết sẽ không kiểm soát chính tả, ngữ pháp nên việc thêm hoa thường lùm tùm có thể xảy ra. Chữ Hoa thường xuất hiện đầu câu và cũng không mang mấy ý nghĩa cho phân lớp. Vì vậy để tăng hiệu quả phân lớp thì nên chuyển tất cả về ký tự thường.

Tài liệu



Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
 === Summary ===

Correctly Classified Instances	122	61	%
Incorrectly Classified Instances	78	39	%
Kappa statistic	0.22		
Mean absolute error	0.4093		
Root mean squared error	0.5544		
Relative absolute error	81.8682 %		
Root relative squared error	110.8879 %		
Total Number of Instances	200		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.570	0.350	0.620	0.570	0.594	0.221	0.642	0.657	pos
	0.650	0.430	0.602	0.650	0.625	0.221	0.642	0.607	neg
Weighted Avg.	0.610	0.390	0.611	0.610	0.609	0.221	0.642	0.632	

=== Confusion Matrix ===

```
a b  <-- classified as
57 43 | a = pos
35 65 | b = neg
```

- Confusion matrix:

		Predicted Class		
		pos	neg	
Actual Class	pos	TP = 57	FN = 43	P = 100
	neg	FP = 35	TN = 65	N = 100
		P' = 92	N' = 108	All = 200

- Accuracy:

$$= \frac{TP + TN}{All} = \frac{57 + 65}{200} = 0.61$$

- TP-Rate:
 - Theo pos: 0.57
 - Theo neg: 0.65
 - Trung bình: 0.61
- FP-Rate:
 - Theo pos: 0.35
 - Theo neg: 0.43
 - Trung bình: 0.39
- Precision:
 - Theo pos: 0.62
 - Theo neg: 0.602
 - Trung bình: 0.611
- Recall:
 - Theo pos: 0.57
 - Theo neg: 0.65
 - Trung bình: 0.61
- F-Measure:
 - Theo pos: 0.594
 - Theo neg: 0.625
 - Trung bình: 0.609
- Nhận xét:

Bộ phân lớp thực thi tốt hơn so với các lượt chạy trước kết quả phân lớp có 122 giá trị phân lớp đúng chiếm 61%. Cao nhất trong các kết quả chạy trước đó

File kết quả đính kèm: **Result_lowerCaseTokens.txt**

II) Nội dung thực hiện cài đặt (10 điểm)

Cài đặt chương trình đọc vào một tập dữ liệu bất kỳ có định dạng *.csv, xây dựng mô hình phân lớp bằng giải thuật ID3 và đánh giá giải thuật bằng phương pháp cross validation, rồi xuất ra tập tin kết quả.

1) (1.0đ) Chương trình nhận dữ liệu đầu vào là tập tin *.csv có cấu trúc như sau

- Giả sử tập dữ liệu có N thuộc tính rời rạc (thuộc tính phân lớp nằm cuối cùng) và M mẫu tương ứng với các thuộc tính này. Dữ liệu được tổ chức thành bảng có M+1 dòng và N cột.

Tài liệu

- Dòng đầu tiên chứa tên của N thuộc tính, phân cách nhau bằng dấu phẩy (","), Tên thuộc tính không có khoảng trắng và ký tự đặc biệt.
- M dòng tiếp theo, mỗi dòng gồm N giá trị, phân cách nhau bằng dấu phẩy (","), Tên giá trị thuộc tính không có khoảng trắng và ký tự đặc biệt.

2) (3.0đ) Chương trình phát sinh dữ liệu đầu ra là tập tin model.txt chứa thông tin tương tự như trong phần văn bản của cửa sổ Classifier output (tab Classify – WEKA), bao gồm

- Mô hình cây quyết định ID3 rút ra từ toàn bộ tập dữ liệu (full training set). Lưu ý, cây này có thể khác với cây thu được trong mỗi lần chạy cross validation.
- Tiêu chí chọn thuộc tính tốt nhất (Entropy, Information Gain, Information Gain Ratio, và Gini Index).
- Số lượng mẫu phân lớp đúng/sai và tỉ lệ tương ứng
- Các giá trị TP Rate, FP Rate, Precision, Recall và F-Measure cho mỗi phân lớp.

3) (1.0đ) Chương trình thực thi giải thuật ID3 và đánh giá giải thuật bằng phương pháp n-folds cross validation với cú pháp tham số dòng lệnh như sau

<ID nhóm> <input> <output> <folds> <best_att>

- <ID nhóm>: tên của tập tin thực thi chương trình là ID của nhóm.
- <input>: tập tin dữ liệu đầu vào có định dạng *.csv
- <output FI>: tập tin đầu ra model.txt
- <folds>: số lượng fold chỉ định cho phương pháp cross validation.
- <best_att>: chiến lược chọn thuộc tính tốt nhất, 0: Entropy, 1: Information Gain, 2: Information Gain Ratio, và 3: Gini Index.

```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.17763.678]
(c) 2018 Microsoft Corporation. All rights reserved.

D:\Source\CourseHCMUS\BigData\KTDLUD\Lab03\Source>python 37_Lab03_Classification.py input_play.csv model.txt 10 0
----- PHÂN LỚP ID3 -----
----- Chọn thuộc tính lần 1 -----
- Tính Entropy trung bình:
outlook: AE = 0.694
temperature: AE = 0.911
humidity: AE = 0.788
windy: AE = 0.892
```

Mới tính được entropy cho thuộc tính phân lớp thứ nhất

Tài liệu

- 4) (5.0đ) Tùy chọn 3 tập dữ liệu có quy mô nhỏ (~100 mẫu), trung bình (~500 mẫu), và lớn (~1000 mẫu). Chạy chương trình cài đặt với các tập dữ liệu đã chọn và đối chiếu kết quả phát sinh được với kết quả của WEKA ID3 trên cùng bộ tham số

III) Nguồn tham khảo:

<http://weka.sourceforge.net/doc.stable/weka/filters/unsupervised/attribute/StringToWordVector.html>