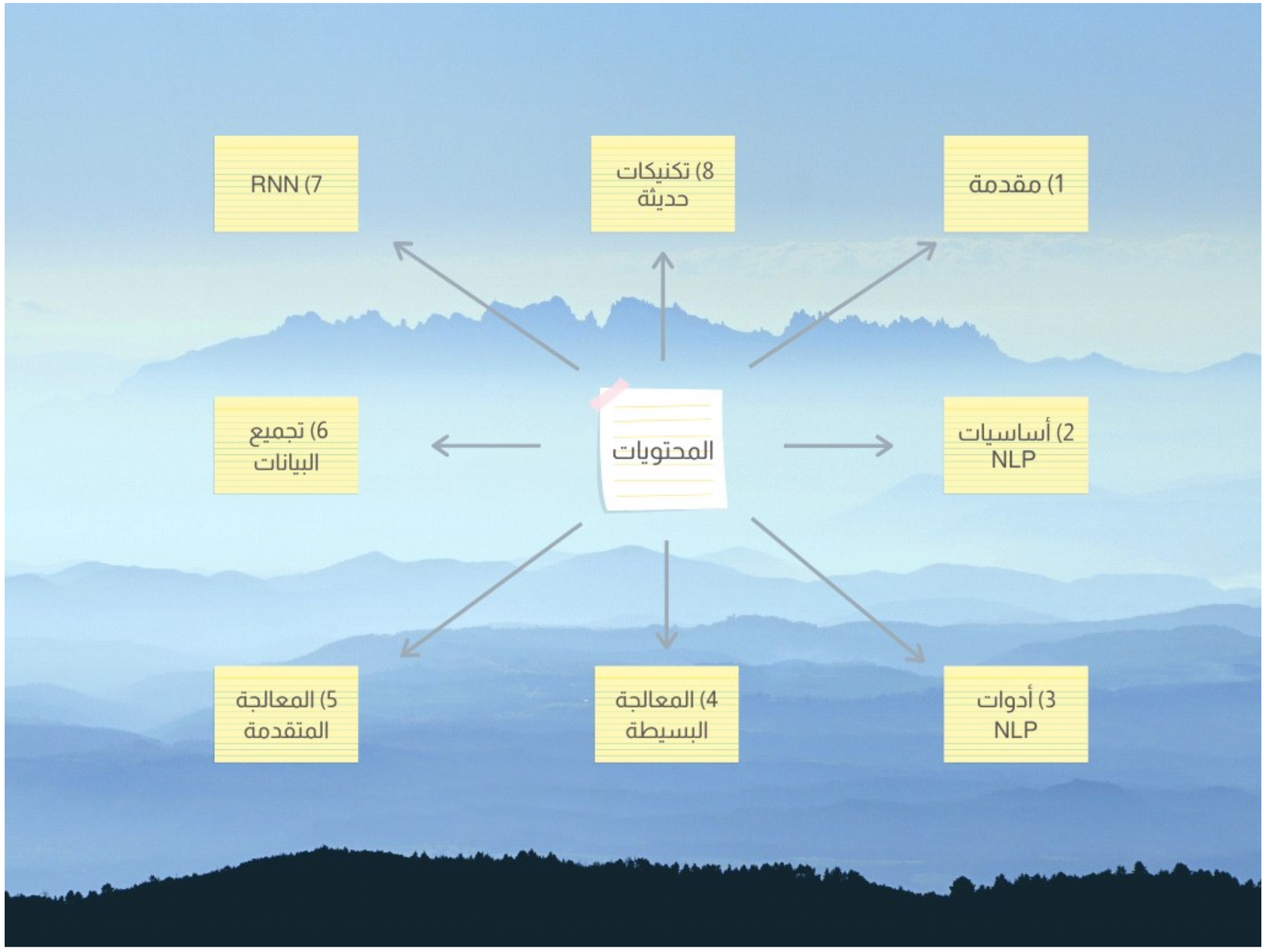
# Natural Language Processing المعالجة اللغوية الطبيعية



# المحتويات

## القسم الخامس : المعالجة المتقدمة للنصوص

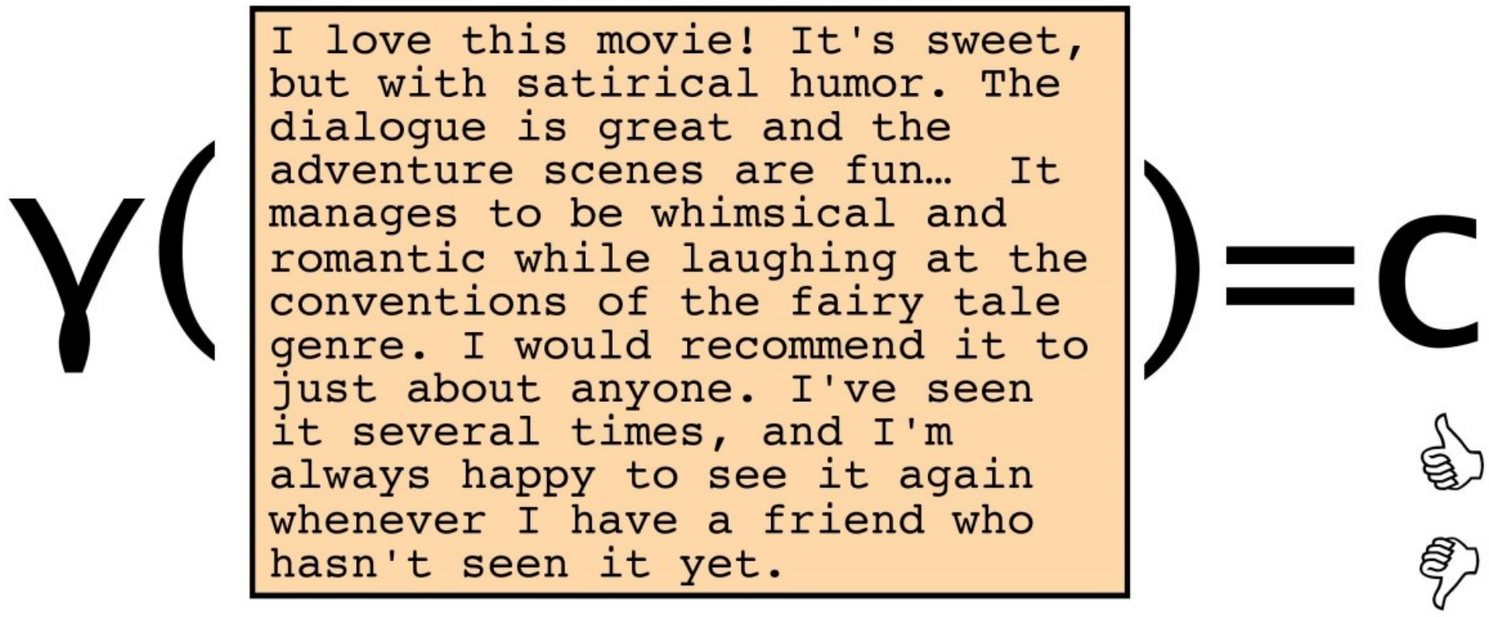
---

## الجزء الحادي عشر : Naïve Bayes

==========

تعتمد فكرة Naïve Bayes علي أساس فكرة bag of words

فلو كان لدينا تعليقا علي احد الافلام و كان فيه هذا النص , ونريد تحديد هل هو تعليق سلبي ام ايجابي . .



فنلحظ أن به بعض الكلمات المميزة التي يجب أن نتوقف عليها , و التي سيكون لها التأثير في اخيار نوع النص

x **love** xxxxxxxxxxxxxxxx **sweet**
xxxxxxx **satirical** xxxxxxxxxx
xxxxxxxxxxx **great** xxxxxxx
xxxxxxxxxxxxxxxxxxx **fun** xxxx
xxxxxxxxxxxxx **whimsical** xxxx
**romantic** xxxx **laughing**
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx **recommend** xxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xx **several** xxxxxxxxxxxxxxxxx
xxxxx **happy** xxxxxxxxx **again**
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxx

و من هنا يمكن أن نقوم باختيار هذه الكلمات و معرفة عدد كل منها , ويكون هذا هو احد العناصر المعتمد عليها في التصنيف

| | |
|---|---|
| great | 2 |
| love | 2 |
| recommend | 1 |
| laugh | 1 |
| happy | 1 |
| ... | ... |

*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*

ما هي الفكرة الرياضية للـ Naïve Bayes

تقوم علي فكرة ان احتمالية وجود صنف (كلاس) معين , بناء علي النص المعطي , ستساوي احتمالية وجود النص , بناء علي الصنف المعطي , مضروب في احتمالية وجود هذا الصنف , مقسوم علي احتمالية وجود النص

For a document $d$ and a class $c$

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

و الطرف الأيسر : احتمالية وجود صنف معين بناء علي النص المعطي , مثل ان يكون التعليق سلبي او ايجابي , وهو الشئ المطلوب

و احتمالية وجود النص , بناء علي الصنف المعطي : معناه مدي تواجد هذه الكلمات في هذا النوع من الكلام , فلو كان لدينا نوع معروف ( كلام ايجابي ) فما مدي احتمالية تواجد كلمات معينة ( Impressive) فيها

و احتمالية وجود هذا الصنف : اي مدي تواجد و انتشار هذا النوع من النصوص

احتمالية وجود النص : متعلقة بمدي انتشار و شيوع الكلمات المستخدمة في النص

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Dropping the denominator

و سيتم تكرار هذه المعادلة في عدد من الاصناف لنفس النص , لاختيار ذات الاحتمالية الاكبر , و يمكن اختصار القانون ليكون البسط فقط , إذ أن المقام هو نفسه , و سيكون ثابت , وبالتالي اختيار القيمة الأعلى تعني البسط الاعلي

*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*

و فكرة حساب احتمالية النص , بناء علي الصنف المعطي , يساوي احتمالية الكلمات الموجودة في النص , بناء علي الصنف

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

و يتم هذا عبر عمل ضرب للاحتمالات مع بعضها البعض باستخدام باي , لذا تسمي multinomial لأنها يتم ضربها في بعضها البعض

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*

و يتم حساب احتمالية الصنف , بناء علي عدد النصوص التي بها هذا الصنف , مقسومة علي كل عدد النصوص , فلو كان لدينا الف جملة من النوع الطبي و العدد الكلي للجمل هو 5 الالف , بتكون النسبة 0.2

اما احتمالية وجود كلمة في هذا الصنف , فهي عدد تكرار هذه الكلمة في الجمل ذات هذا النوع من الصنف , مقسومة علي العدد الكلي لجميع الكلمات في هذا النوع من الصنف

First attempt: maximum likelihood estimates

- simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

و لكن اذا قمنا بعمل اختبار لكلمة معينة , وكانت الكلمة غير موجودة في عينة التدريب , مثلا كلمة fantastic لم تكن لسبب ما موجودة في عينة التدريب في الجمل الايجابية , فالمشكلة ان الاحتمالية النهائية لها ستكون بصفر , وهذا غير مقبول

What if we have seen no training documents with the word ***fantastic*** and classified in the topic **positive** (***thumbs-up)***?

$$\hat{P}(\text{"fantastic"} \mid \text{positive}) = \frac{count(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

Zero probabilities cannot be conditioned away, no matter the other evidence!

$$c_{MAP} = \text{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

لذا فإننا نقوم بتطبيق قانون لابلاس , بإضافة رقم 1 في البسط و المقام , ولاحظ ان رقم 1 في المقام سيتم ترجمته الي عدد ال vocabulary لان السمشن في المقام هي لجميع الكلمات الموجودة في الصنف

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left( count(w, c) + 1 \right)}$$

$$= \frac{count(w_i, c) + 1}{\left( \sum_{w \in V} count(w, c) \right) + |V|}$$

بالتالي اذا صادفنا كلمة ما غير موجودة , فلن يكون رقمها صفر و لكن قليل

*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*

,و ماذا عن الكلمة غير الموجودة ؟ ؟

يتم وضع مكان لـ Wu و هنا تختفي قيمة count(w,c) و تكون بصفر لانها غير موجودة , و يكون 1 علي المقام

$$\hat{P}(w_u \mid c) = \frac{count(w_u, c) + 1}{\left(\sum_{w \in V} count(w, c)\right) + |V + 1|}$$

$$= \frac{1}{\left(\sum_{w \in V} count(w, c)\right) + |V + 1|}$$

و تكون هنا الخطوات مجمعة

From training corpus, extract *Vocabulary*

Calculate $P(c_j)$ terms

- For each $c_j$ in $C$ do

  $docs_j \leftarrow$ all docs with class $= c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calculate $P(w_k \mid c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word $w_k$ in *Vocabulary*

    $n_k \leftarrow$ # of occurrences of $w_k$ in $Text_j$

$$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*

و نلحظ أن هناك تشابها بين فكرة الـ naïve bayes و نموذج اللغة الذي تحدثنا عنه من قبل

فنموذج اللغة و بالأخص نوع unigram , هو يبحث عن مدي تواجد كلمة معينة في العينة , ليكون لها نسبة معينة في تواجدها , كذلك يقوم NB

فلو كان لدينا جملة I love this fun film , وقمنا بحساب P(w/c) للنوعين : تعليق ايجابي او سلبي

فسنجد ان احتمالية تواجد كل كلمة فيهم في العينة السلبية و العينة الايجابية بارقام مختلفة , ثم حساب احتمالية الجملة كلها بضرب احتمالية الكلمات معا , كما نري هنا

| Model pos | |
|---|---|
| 0.1 | I |
| 0.1 | love |
| 0.01 | this |
| 0.05 | fun |
| 0.1 | film |

| Model neg | |
|---|---|
| 0.2 | I |
| 0.001 | love |
| 0.01 | this |
| 0.005 | fun |
| 0.1 | film |

| I | love | this | fun | film |
|---|---|---|---|---|
| 0.1 | 0.1 | 0.01 | 0.05 | 0.1 |
| 0.2 | 0.001 | 0.01 | 0.005 | 0.1 |

P(s|pos) > P(s|neg)

- Assigning each word: P(word | c)
- Assigning each sentence: P(s|c)=Π P(word|c)

Class *pos*

| | |
|---|---|
| 0.1 | I |
| 0.1 | love |
| 0.01 | this |
| 0.05 | fun |
| 0.1 | film |

| I | love | this | fun | film |
|---|---|---|---|---|
| 0.1 | 0.1 | .05 | 0.01 | 0.1 |

P(s | pos) = 0.0000005

و من هنا يمكن تصنيف هذه الجملة انها ايجابية لان احتماليتها اكبر

مع ملاحظة ان نسبة الكلمات العامة مثل ( I , this , film ) متشابهة نسبيا في العينة الايجابية و السلبية , بينما كلمات ايجابية مثل ( love , fun ) ذات نسبة تواجد اعلي في العينة الايجابية عن السلبية , مما جعل تصنيفها ايجابيا , و العكس صحيح

مع العلم ان NB يمكن أن يقوم بالتركيز علي تواجد كلمات او فيتشرز معينة مثل

*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*

و هنا مثال عملي علي تطبيق القواعد السابقة

فلو كان لدينا ثلاث جمل و التي تنتمي للفئة "الصين" و جملة تنتمي لفئة "اليابان" , وهناك جملة اخيرة نريد تحديد الفئة الخاصة بها test data

| | Doc | Words | Class |
|---|---|---|---|
| Training | 1 | Chinese Beijing Chinese | c |
| | 2 | Chinese Chinese Shanghai | c |
| | 3 | Chinese Macao | c |
| | 4 | Tokyo Japan Chinese | j |
| Test | 5 | Chinese Chinese Chinese Tokyo Japan | ? |

فعلينا حساب كلا من P( c ) لكل فئة و P ( w/c) لكل كلمة مرتين , فتكون P( c ) للصين هي ¾ و لليابان هي ¼

بينما تكون اول كلمة من اول جملة Chinese تكون احتماليتها هكذا P(Chinese/c) اي احتمالية وجودها في فئة الصين يساوي البسط/ المقام

البسط يساوي عدد مرات تكرار هذه الكلمة في الجمل في فئة الصين + 1 الخاص بـ لابلاس و هذا يساوي 6

و المقام هو عدد جميع الكلمات (8) مضاف اليه اعداد 1 بعدد مرات الكلمات ( مع حذف التكرار ) و هي الـ ( voabulary) , و ستساوي 6

و يتم تكرار الامر مع جميع الكلمات الواردة في جملة الاختبار (و ليس كل الكلمات في الجمل كلها ) في الفئتين , ثم نقوم بحساب احتمالية ان تكون جملة الاختبار في الفئة الاولي , و ستساوي P( c ) , مضروبة في P ( w/c) لكل كلمة

ونقارن بين الفئتين , مع ملاحظة انه يفترض قسمتها علي احتمالية الجملة , لكن هي ثابتة فيتم حذفها

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

| | Doc | Words | Class |
|---|---|---|---|
| Training | 1 | Chinese Beijing Chinese | c |
| | 2 | Chinese Chinese Shanghai | c |
| | 3 | Chinese Macao | c |
| | 4 | Tokyo Japan Chinese | j |
| Test | 5 | Chinese Chinese Chinese Tokyo Japan | ? |

**Priors:**

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

**Conditional Probabilities:**

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

**Choosing a class:**

$P(c|d5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$

$P(j|d5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$

*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*

نتكلم الآن عن تقييم الموديل بعد تدريبه

و من البيانات المشهورة هي بيانات رويترز , والتي تحتوي علي التفاصيل التالية :

- Most (over)used data set, 21,578 docs (each 90 types, 200 toknens)
- 9603 training, 3299 test articles (ModApte/Lewis split)
- 118 categories
  - An article can be in more than one category
  - Learn 118 binary category distinctions
- Average document (with at least one category) has 1.24 classes
- Only about 10 out of 118 categories are large

Common categories
(#train, #test)

- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)
- Trade (369,119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)

57

وهذه نوعية منها

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

58

&#3;</BODY></TEXT></REUTERS>

و يتم عمل مصفوفة التشتت في الفئات المختلفة هكذا :

- For each pair of classes $<c_1,c_2>$ how many documents from $c_1$ were incorrectly assigned to $c_2$?
  - $c_{3,2}$: 90 wheat documents incorrectly assigned to poultry

| Docs in test set | Assigned UK | Assigned poultry | Assigned wheat | Assigned coffee | Assigned interest | Assigned trade |
|---|---|---|---|---|---|---|
| True UK | 95 | 1 | 13 | 0 | 1 | 0 |
| True poultry | 0 | 1 | 0 | 0 | 0 | 0 |
| True wheat | 10 | 90 | 0 | 1 | 0 | 0 |
| True coffee | 0 | 0 | 0 | 34 | 3 | 7 |
| True interest | - | 1 | 2 | 13 | 26 | 5 |
| True trade | 0 | 0 | 2 | 14 | 5 | 10 |

و من ثم يتم حساب precision , recall & accuracy لكل فئة من الفئات

**Recall**:
Fraction of docs in class *i* classified correctly: $\frac{c_{ii}}{\sum_j c_{ij}}$

**Precision**:
Fraction of docs assigned class *i* that are actually about class *i*: $\frac{c_{ii}}{\sum_j c_{ji}}$

**Accuracy**: (1 - error rate)
Fraction of docs classified correctly: $\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$

مع العلم ان هناك فارق بين micro& macro average للفئات المختلفة , وهي سواء للـ precision , recall or accuracy

فال macro معتمدة علي حساب الدقة لكل فئة علي حدة , ثم عمل mean لها , بينما ال micro معتمدة علي جمع النتائج معا في CM كبيرة , وحساب الدقة لها

Class 1

| | Truth: yes | Truth: no |
|---|---|---|
| Classifier: yes | 10 | 10 |
| Classifier: no | 10 | 970 |

Class 2

| | Truth: yes | Truth: no |
|---|---|---|
| Classifier: yes | 90 | 10 |
| Classifier: no | 10 | 890 |

Micro Ave. Table

| | Truth: yes | Truth: no |
|---|---|---|
| Classifier: yes | 100 | 20 |
| Classifier: no | 20 | 1860 |

- Macroaveraged precision: (0.5 + 0.9)/2 = 0.7
- Microaveraged precision: 100/120 = .83
- Microaveraged score is dominated by score on common classes

كما اننا نقوم بتقطيع البيانات الي : تدريب , تطوير , اختبار , تجنبا لمشكلة ال OF

و غالبا ما يتم تكرار تقسيم التدريب و التطوير علي اجزاء مختلفة من الداتا للتأكد من دقتها

*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*