2 Classification naïve bayésienne pour détecter les pourriels

La classification naïve bayésienne est un algorithme d'apprentissage supervisé qui fonctionne avec les probabilités. Un problème particulier qui peut être traité avec cet algorithme est la classification de courriels. On peut tenter d'estimer la probabilité qu'un courriel soit en fait un pourriel en prenant en compte les mots particuliers qu'il contient, l'idée étant que certains mots auront tendance à être plus souvent utilisés dans des pourriels.

La probabilité qui nous intéresse est en fait une probabilité conditionnelle, celle du fait qu'un courriel particulier soit ou non un pourriel, étant donné les mots particuliers qui le composent. Un courriel sera jugé un pourriel si :

$$\text{Prob}(\text{pourriel} \mid \text{mots}) > \text{Prob}(\text{ok} \mid \text{mots})$$

2.1 Entraînement du modèle

Voyons comment on peut calculer ces probabilités en entraînant un modèle sur une série de courriels particuliers.

Copiez tout d'abord ces données dans les colonnes A et B d'un feuillet Excel:

Bonjour, on se voit demain?	ok
Voici le compte-rendu de la réunion	ok
N'oublie pas d'apporter le rapport	ok
Invitation à dîner ce week-end	ok
Merci pour ta réponse rapide	ok
Le document est attaché à ce courriel	ok
C'est noté, à bientôt!	ok
As-tu vu les dernières nouvelles?	ok
Gagnez un iPhone gratuit maintenant!	pourriel
Cliquez ici pour réclamer votre prix $$$	pourriel

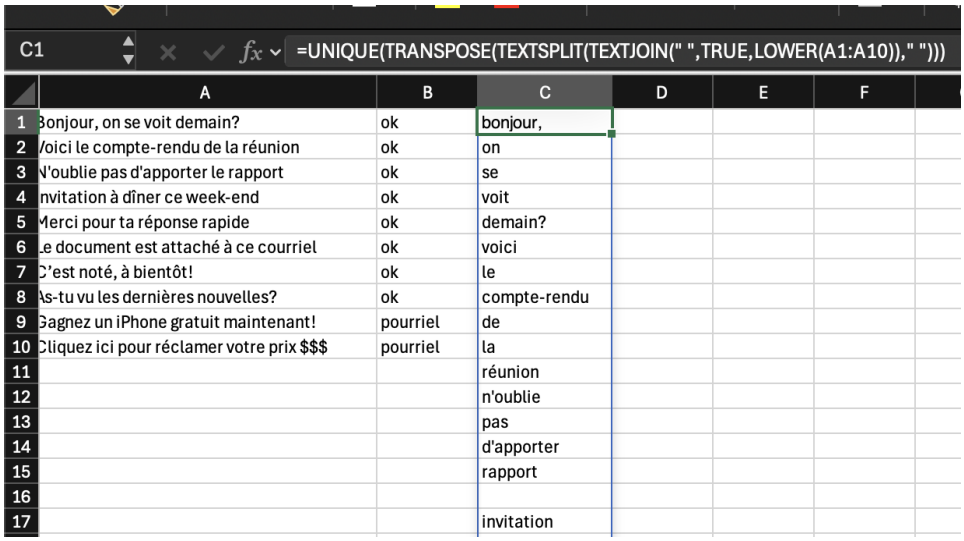
Calculons tout d'abord dans la colonne C la probabilité à priori qu'un courriel quelconque soit "ok" ou non (sans prendre en considérations les mots donc, pour le moment):

```
=COUNTIF(B1:B10, UNIQUE(B1:B10)) / COUNTA(B1:B10)
```

Ces probabilités à priori nous serviront plus loin. Définissez ensuite la colonne D avec cette formule :

```
=UNIQUE(TRANSPOSE(TEXTSPLIT(TEXTJOIN(" ",TRUE,LOWER(A1:A10))," ")))
```

La colonne D devrait maintenant contenir le vocabulaire des courriels:



	A	B	C	D	E	F
1	Bonjour, on se voit demain?	ok	bonjour,			
2	Voici le compte-rendu de la réunion	ok	on			
3	N'oublie pas d'apporter le rapport	ok	se			
4	Invitation à dîner ce week-end	ok	voit			
5	Merci pour ta réponse rapide	ok	demain?			
6	Le document est attaché à ce courriel	ok	voici			
7	C'est noté, à bientôt!	ok	le			
8	As-tu vu les dernières nouvelles?	ok	compte-rendu			
9	Gagnez un iPhone gratuit maintenant!	pourriel	de			
10	Cliquez ici pour réclamer votre prix $$$	pourriel	la			
11			réunion			
12			n'oublie			
13			pas			
14			d'apporter			
15			rapport			
16						
17			invitation			

La colonne E devrait correspondre au nombre de fois où les mots de la colonne D apparaissent dans les courriels de la catégorie "ok" :

```
=SUMPRODUCT((B$1:B$10="ok") *
       ISNUMBER(SEARCH(" " & D1 & " ", " " & LOWER(A$1:A$10) & " ")))
```

et de manière similaire pour la colonne F et les mots de la catégorie "pourriel":

```
=SUMPRODUCT((B$1:B$10="pourriel") *
       ISNUMBER(SEARCH(" " & D1 & " ", " " & LOWER(A$1:A$10) & " ")))
```

Notez que les colonnes E et F doivent avoir le même nombre d'éléments que la colonne D (il faut donc utiliser la fonction de remplissage automatique, pour laquelle le plus simple est de soit glisser (drag) la première cellule vers le bas, ou encore de double-cliquer sur le petit "+" noir qui apparait en bas à droite de la première cellule, une fois la formule exécutée).

À partir de ces fréquences de mots pour chaque classe ("ok" ou "pourriel"), on peut maintenant calculer la probabilité conditionnelle de chaque mot du vocabulaire, étant donné la classe. Donc la colonne G correspond à la probabilité des mots étant donné un courriel "ok":

```
=(E1 + 1) / (SUM(E:E) + COUNTA(D:D))
```

et de manière similaire la colonne H est la probabilité des mots quand on sait qu'on a affaire à un pourriel:

```
=(F1 + 1) / (SUM(F:F) + COUNTA(D:D))
```

2.2 Utilisation du modèle (inférence)

Nous allons maintenant utiliser le modèle pour déterminer si un nouveau courriel est un pourriel ou non. Dans la colonne I entrez le courriel à tester:

```
Salut voici la facture de votre iphone
```

Faites l'extraction des mots du courriel dans la colonne J:

```
=TRANSPOSE(TEXTSPLIT(LOWER(I1), " "))
```

Nous avons maintenant besoin de la probabilité des mots de ce courriel de test dans l'hypothèse où ça serait un courriel "ok", donc pour la colonne K:

```
=IFERROR(XLOOKUP(FILTER(J:J, J:J<>""), D:D, G:G), 1e-10)
```

et de manière similaire pour la probabilité des mots du courriel dans l'hypothèse où il s'agit d'un pourriel, donc pour la colonne L:

```
=IFERROR(XLOOKUP(FILTER(J:J, J:J<>""), D:D, H:H), 1e-10)
```

La probabilité que le courriel soit ok est la colonne M:

```
=PRODUCT(K:K) * C1
```

Et la probabilité qu'il soit un pourriel est la colonne N:

```
=PRODUCT(L:L) * C2
```

Notre classification finale sera dans la colonne O:

```
=IF(M1 > N1, "ok", "pourriel")
```