Princeton University
COS 217: Introduction to Programming Systems
Spam Filter Knowledge Module: Example

Example Features File

```
$ cat featuressmall
lottery
dollars
insurance
thousand
million
```

Example Ham Messages File

```
$ cat hamtrainsmall
From lottery other

From other

From thousand million other

From other

From other

From thousand other

From other

From other
```

Example Spam Messages File

```
$ cat spamtrainsmall
From dollars other

From other

From million dollars other

From other

From dollars other

From dollars other

From insurance other

From dollars other
```

Example Message Read from Stdin

```
From million dollars other
```

The "Learn" Algorithm

(1) Command the FeatureFinder object to read features from features file.

Feature	Index
lottery	0
dollars	1
insurance	2
thousand	3
million	4

(2) Command the Parser object to read messages from hamtrainsmall. Use the FeatureFinder object to compute ham feature counts. Do the same for spamtrainsmall.

For ham:

count = 8

Feature	Index	Count
lottery	0	1
dollars	1	0
insurance	2	0
thousand	3	2
million	4	1

For spam:

count = 8

Feature	Index	Count
lottery	0	0
dollars	1	5
insurance	2	1
thousand	3	0
million	4	1

(3) Guess P(ham) and P(spam).

P(ham) = 0.5
P(spam) = 0.5

Note: Must sum to 1.0.

(4) Compute probabilities.

Note: Must avoid 0 probabilities. So assume the existence of two extra messages: one containing all features, and one containing no features.

For ham:

count = 8
P(ham) = 0.5

Feature	Index	Count	$P(f_i \mid ham)$	$P(\sim f_i \mid ham)$
lottery	0	1	(1+1)/(8+2) = .2	.8
dollars	1	0	(0+1)/(8+2) = .1	.9
insurance	2	0	(0+1)/(8+2) = .1	.9
thousand	3	2	(2+1)/(8+2) = .3	.7
million	4	1	(1+1)/(8+2) = .2	.8

For spam:

count = 8
P(spam) = 0.5

Feature	Index	Count	$P(f_i \mid spam)$	$P(\sim f_i \mid spam)$
lottery	0	0	(0+1)/(8+2) = .1	.9
dollars	1	5	(5+1)/(8+2) = .6	.4
insurance	2	1	(1+1)/(8+2) = .2	.8
thousand	3	0	(0+1)/(8+2) = .1	.9
million	4	1	(1+1)/(8+2) = .2	.8

(5) Compute logarithms.

For ham:

count = 8
P(ham) = 0.5
log(P(ham)) = -0.693

Feature	Index	Count	$P(f_i \mid ham)$	$P(\sim f_i \mid ham)$	$\log(P(f_i \mid ham)$	$\log(P(\sim f_i \mid ham))$
lottery	0	1	.2	.8	-1.609	-0.223
dollars	1	0	.1	.9	-2.303	-0.105
insurance	2	0	.1	.9	-2.303	-0.105
thousand	3	2	.3	.7	-1.204	-0.357
million	4	1	.2	.8	-1.609	-0.223

For spam:

count = 8
P(spam) = 0.5
log(P(spam)) = -0.693

Feature	Index	Count	$P(f_i \mid spam)$	$P(\sim f_i \mid spam)$	$\log(P(f_i \mid spam)$	$\log(P(\sim f_i \mid ham))$
lottery	0	0	.1	.9	-2.303	-0.105
dollars	1	5	.6	.4	-0.511	-0.916
insurance	2	1	.2	.8	-1.609	-0.223
thousand	3	0	.1	.9	-2.303	-0.105
million	4	1	.2	.8	-1.609	-0.223

(6) Write knowledge to the knowledge file.

For ham:

count = 8
P(ham) = 0.5
log(P(ham)) = **-0.693**

Feature	Index	Count	$P(f_i \mid ham)$	$P(\sim f_i \mid ham)$	$\log(P(f_i \mid ham)$	$\log(\sim P(f_i \mid ham))$
lottery	0	1	.2	.8	**-1.609**	**-0.223**
dollars	1	0	.1	.9	**-2.303**	**-0.105**
insurance	2	0	.1	.9	**-2.303**	**-0.105**
thousand	3	2	.3	.7	**-1.204**	**-0.357**
million	4	1	.2	.8	**-1.609**	**-0.223**

For spam:

count = 8
P(spam) = 0.5
log(P(spam)) = **-0.693**

Feature	Index	Count	$P(f_i \mid spam)$	$P(\sim f_i \mid spam)$	$\log(P(f_i \mid spam)$	$\log(P(\sim f_i \mid spam))$
lottery	0	0	.1	.9	**-2.303**	**-0.105**
dollars	1	5	.6	.4	**-0.511**	**-0.916**
insurance	2	1	.2	.8	**-1.609**	**-0.223**
thousand	3	0	.1	.9	**-2.303**	**-0.105**
million	4	1	.2	.8	**-1.609**	**-0.223**

The "Filter" Algorithm

(1) Command the FeatureFinder object to read features from features file.

(2) Read knowledge from the knowledge file.

For each message in stdin...

(3) Command the Parser object to read a message. Use the FeatureFinder object to determine whether each feature is present in the message.

Feature	Index	Present(0/1)?
lottery	0	0
dollars	1	1
insurance	2	0
thousand	3	0
million	4	1

(4) Compute log(P(ham)) + log(P(f_1|ham)) + ... + log(P(~f_n|ham)).

Feature	Index	Present(0/1)?	log(P(f_i|ham) or log(P(~f_i|ham)
lottery	0	0	-0.223
dollars	1	1	-2.303
insurance	2	0	-0.105
thousand	3	0	-0.357
million	4	1	-1.609
log(P(ham)) =			-0.693
sum =			-5.290

(5) Compute log(P(spam)) + log(P(f_1|spam)) + ... + log(P(~f_n|spam)).

Feature	Index	Present(0/1)?	log(P(f_i|spam) or log(P(~f_i|spam)
lottery	0	0	-0.105
dollars	1	1	-0.511
insurance	2	0	-0.223
thousand	3	0	-0.105
million	4	1	-1.609
log(P(spam)) =			-0.693
sum =			-3.246

(6) Compute P(spam|f_1,...,~f_n).

P(spam|f_1,...,~f_n)
= 1 / (exp(-5.290 - (-3.246)) + 1)
= 1 / (exp(-5.290 + 3.246) + 1)
= 1 / (exp(-2.044) + 1)
= 1 / (.1295 + 1)
= 0.885

(7) Write P(spam|f_1,...,~f_n) to stdout.

0.885

So the message probably is spam.