Email Classification

Aim

To implement text classification of emails using Naive Bayes classifier.

Spam emails are usually sent to derive cash from the user either directly by tricking the user into purchasing something or indirectly by ticking them into parting with information. At the least, classificatino of spam emails aims to save the user the time and hassle of deleting spam. At best the detection of spam emails allows for the protection of unaware users from the malicious intent of these spam emails.

Data Preprocessing and Feature Selection

The files in the sample set *Lingpspam-mini600* were read individually and line by line.

- Open a file in the set and read it line by line.
- If the line begins with S̈ubject:ässume it is of the subject corpus, otherwise it is of the body corpus.
- All punctuation is replaced with whitespace.
- The line is split by words (e.g. d́on'th́as now become d́on;t́) and added to a counter of the words in the current document
- For each word that belongs to the counter of a single document it is added to the counter for the sub-corpus once.
- The words with the top 200 Document Frequency (words which are appeared in the most number of documents) in each corpus was selected for the features of that corpus.
- The TD-IDF for each feature was calculated.
- TD-IDF values were then normalised using Cosine Normalisation.

information	205	the	549
language	192	of	541
mail	183	and	533
university	179	a	530
time	178	to	530
list	171	in	514
address	165	for	477
english	159	is	459
linguistics	156	on	420
http	156	this	410
send	146	be	410
people	146	or	380
free	144	are	368
make	140	with	366
email	133	that	358

work	128	i	351
number	128	s	348
www	122	have	345
languages	119	by	344
find	118	it	341
fax	116	at	340
order	108	you	338
call	103	from	338
form	101	as	321
research	100	not	320
linguistic	99	if	315
state	99	an	301
subject	98	can	297
world	98	will	270
years	98	all	262
contact	97	one	254
de	96	do	251
money	94	e	247
message	91	any	240
word	91	we	239
ll	89	would	239
phone	88	but	238
check	88	which	233
receive	88	there	230
good	87	your	229
year	86	about	228
interested	86	also	226
day	86	more	224
working	85	no	218
case	85	t	217
include	85	please	217
ve	84	has	211
based	84	like	211
made	83	other	208
home	83	only	206
part	83	information	205
note	83	some	204
mailing	81	so	202
including	81	n	200
type	80	what	199
place	79	new	196
program	79	language	192
give	79	who	191
web	79	out	191
date	78	us	187
special	78	mail	183
line	78	me	182
days	77	my	181
internet	76	university	179
back	76	time	178

american	75	these	177
service	75	our	177
business	74	been	174
full	74	list	171
system	74	now	170
ac	73	they	169
today	73	up	169
interest	72	was	165
remove	72	address	165
questions	72	may	162
john	71	english	159
found	70	their	158
related	70	use	157
usa	69	get	156
linguist	69	linguistics	156
site	69	http	156
text	68	many	155
read	68	here	150
point	68	know	147
ago	67	following	147
week	67	than	146
book	67	how	146
dear	66	people	146
making	66	send	146
cost	66	two	144
simply	65	free	144
question	65	most	143
general	63	edu	143
received	63	m	142
offer	63	just	142
important	62	very	142
data	62	them	140
summary	61	am	140
ca	61	make	140
long	61	does	136

sum	30	re	82
summary	26	in	49
english	24	and	49
language	21	for	34
free	20	sum	30
disc	19	on	29
query	18	of	27
linguistics	15	new	27
sex	13	the	27
comparative	13	summary	26
opposites	12	english	24
words	12	s	23
book	10	a	22

email	10	language	21
call	9	you	21
job	9	free	20
method	9	disc	19
japanese	8	your	19
correction	8	query	18
million	7	to	16
chinese	7	linguistics	15
announcement	7	are	15
syntax	7	that	14
qs	7	sex	13
program	7	comparative	13
slip	6	their	12
lang	6	opposites	12
business	6	own	12
part	6	words	12
armey	6	is	11
dick	6	do	11
internet	6	only	11
mail	6	book	10
german	6	this	10
workshop	6	email	10
speaker	6	call	9
conference	6	at	9
money	6	n	9
list	5	job	9
time	5	method	9
resources	5	q	8
linguist	5	japanese	8
software	5	correction	8
languages	5	million	7
word	5	help	7
native	5	chinese	7
needed	5	announcement	7
jobs	5	syntax	7
grammar	5	qs	7
spanish	5	program	7
research	5	just	7
address	4	or	6
unlimited	4	slip	6
american	4	lang	6
programs	4	non	6
phonetics	4	out	6
books	4	business	6
banning	4	part	6
systems	4	dick	6
request	4	internet	6
fwd	4	armey	6
intuitions	4	t	6
web	4	mail	6

summer	4	german	6
www	4	workshop	6
information	4	we	6
read	4	speaker	6
great	4	conference	6
secrets	4	money	6
pig	4	list	5
change	3	who	5
sites	3	now	5
cd	3	time	5
names	3	please	5
site	3	best	5
people	3	resources	5
teaching	3	linguist	5
released	3	word	5
french	3	want	5
double	3	software	5
dialect	3	jobs	5
make	3	know	5
synthetic	3	languages	5
policy	3	e	5
chomsky	3	native	5
home	3	needed	5
decimal	3	grammar	5
comparison	3	spanish	5
adjectives	3	research	5
fall	3	better	5
line	3	unlimited	4
latin	3	address	4
hey	3	it	4
credit	3	i	4
ary	3	american	4
uniformitarianism	3	us	4
requested	3	programs	4
verbal	3	phonetics	4
debt	3	hi	4
dental	3	he	4

Subject vs Body: Results and Discussion

Corpus: Subject	
	Accuracy [%]
ZeroR	
OneR	
1-NN	
3-NN	
NB	
DT	
MLP	
MyNB	

Corpus: Body	
	Accuracy [%]
ZeroR	
OneR	
1-NN	
3-NN	
NB	
DT	
MLP	
MyNB	

Challenge Results and Discussion

Conclusions

The tokens used in our classifier are formed from single words. Therefore, it will not analyse common consecutive words that are found in spam emails, leading to the failure of our classifier from detecting these emails.

By taking into account permutations of consecutive words, or words that appear within a specified distance of each other, the accuracy of our Bayesian classifier could be increased.

Reflection

Instructions: How to run code