

BayesStopList

In Bayes.pm, there's a list of common words to skip during tokenization, aka the "stop-list". Use something like Regexp::Trie to generate an optimized regexp to catch them all. ala:

```
perl -MRegexp::Trie -nle 'BEGIN{$r=Regexp::Trie->new;} $r->add($_); END {print $r->regexp,"\n"}'
```

Original list:

```
able
all
already
and
any
are
because
both
can
come
each
email
even
few
first
for
from
give
has
have
http
information
into
it's
just
know
like
long
look
made
mail
mailing
mailto
make
many
more
most
much
need
not
now
number
off
one
only
out
own
people
place
right
same
see
such
that
the
this
through
time
```

using
web
where
why
with
without
work
world
year
years
you
you're
your

If it's any use, a listing of the top 500 most common English words is available at: <http://www.world-english.org/english500.htm>

(words with length < 3 removed)

able
about
above
act
add
after
again
against
age
ago
air
all
also
always
among
and
animal
answer
any
appear
are
area
ask
back
base
beauty
bed
been
before
began
begin
behind
best
better
between
big
bird
black
blue
boat
body
book
both
box
boy
bring
brought
build
busy

but
call
came
can
car
care
carry
cause
center
certain
change
check
children
city
class
clear
close
cold
color
come
common
complete
contain
correct
could
country
course
cover
cross
cry
cut
dark
day
decide
deep
develop
did
differ
direct
does
dog
don't
done
door
down
draw
drive
dry
during
each
early
earth
ease
east
eat
end
enough
even
ever
every
example
eye
face
fact
fall
family
far
farm
fast
father
feel

feet
few
field
figure
fill
final
find
fine
fire
first
fish
five
fly
follow
food
foot
for
force
form
found
four
free
friend
from
front
full
game
gave
get
girl
give
gold
good
got
govern
great
green
ground
group
grow
had
half
hand
happen
hard
has
have
head
hear
heard
heat
help
her
here
high
him
his
hold
home
horse
hot
hot
hour
house
how
hundred
idea
inch
interest
island
just

keep
kind
king
knew
know
land
language
large
last
late
laugh
lay
lead
learn
leave
left
less
let
letter
life
light
like
line
list
listen
little
live
long
look
lot
love
low
machine
made
main
make
man
many
map
mark
may
mean
measure
men
might
mile
mind
minute
miss
money
moon
more
morning
most
mother
mountain
move
much
music
must
name
near
need
never
new
next
night
north
note
nothing
notice

noun
now
number
numeral
object
off
often
old
once
one
only
open
order
other
our
out
over
own
page
paper
part
pass
pattern
people
perhaps
person
picture
piece
place
plain
plan
plane
plant
play
point
port
pose
possible
pound
power
press
problem
produce
product
pull
put
question
quick
rain
ran
reach
read
ready
real
record
red
remember
rest
right
river
road
rock
room
round
rule
run
said
same
saw
say
school

science
sea
second
see
seem
self
sentence
serve
set
several
shape
she
ship
short
should
show
side
simple
since
sing
sit
six
size
sleep
slow
small
snow
some
song
soon
sound
south
special
spell
stand
star
start
state
stay
step
still
stood
stop
story
street
strong
study
such
sun
sure
surface
table
tail
take
talk
teach
tell
ten
test
than
that
the
their
them
then
there
these
they
thing
think
this

those
though
thought
thousand
three
through
time
together
told
too
took
top
toward
town
travel
tree
true
try
turn
two
under
unit
until
use
usual
very
voice
vowel
wait
walk
want
war
warm
was
watch
water
way
week
weight
well
went
were
west
what
wheel
when
where
which
while
white
who
whole
why
will
wind
with
wonder
wood
word
work
world
would
write
year
yes
yet
you
young
your

History

(this part by jm, who selected the original set)

The original selection of words in the stop-list was based on words that scored around 0.4 to 0.6 in Bayes score after learning a "typical" spam/nospam corpus, with a very large number of hits – in other words, they were both (a) very common and (b) likely to always be ignored by the Bayes code anyway, since they were always going to fall within the \$MIN_PROB_STRENGTH range. They weren't chosen just as "common" words.

However, it's arguable that the stop-list makes an assumption that everyone speaks English – in some non-English-language countries, a nonspam corpus may contain no English terms while the spam corpus is mostly-English, in which case those stop-words would actually make good spam signs.

Hence, I don't think it's a good idea to increase the stop-list with additional "common" english words.