A Stop List for General Text

Christopher Fox

AT&T Bell Laboratories
Lincroft, New Jersey 07738

1. Introduction

A *stop list*, or *negative dictionary* is a device used in automatic indexing to filter out words that would make poor index terms[1]. Traditionally[2] stop lists are supposed to have included only the most frequently occurring words. In practice, however, stop lists have tended to include infrequently occurring words, and have not included many frequently occurring words. Infrequently occurring words seem to have been included because stop list compilers have not, for whatever reason, consulted empirical studies of word frequencies. Frequently occurring words seem to have been left out for the same reason, and also because many of them might still be important as index terms.

This paper reports an exercise in generating a stop list for general text based on the Brown corpus[3] of 1,014,000 words drawn from a broad range of literature in English. We start with a list of tokens occurring more than 300 times in the Brown corpus. From this list of 278 words, 32 are culled on the grounds that they are too important as potential index terms. Twenty-six words are then added to the list in the belief that they may occur very frequently in certain kinds of literature. Finally, 149 words are added to the list because the finite state machine based filter in which this list is intended to be used is able to filter them at almost no cost. The final product is a list of 421 stop words that should be maximally efficient and effective in filtering the most frequently occurring and semantically neutral words in general literature in English.

2. The Most Frequently Occurring Words in English

In compiling our list of the most frequently occurring words in English, several arbitrary decisions were made. First, we had to choose a cut-off point for the list. We wanted a large list (more than a few dozen words) to keep many useless words out of our indexes, but it soon becomes obvious to a stop list compiler that there are thousands of words in English that might go into a large stop list. Browsing the data from the Brown corpus, it is also obvious that many words, including words important as index terms, occur at the rate of one or two hundred per million in English. With these facts in mind, we settled on a cut-off of an occurrence of 300 in the Brown corpus, estimating that this would produce a list of about 250 words, almost none of which would be potentially good index terms.

We also had to decide how to count words. The data in the book from which we derived our list is based on counts of word lemmas, which are based on parts of speech. Thus, for example, "go," "went," and "gone," are counted as the same word because they are different versions of the verb "to go," but "keep" is counted as two words, once as the verb "to keep," and once as the noun "keep." Since stop list processing (and automatic indexing) is based purely on spelling, this way of counting is not appropriate for our purposes. Instead we tallied occurrences of particular spellings of words, including the roots of contractions, but not of hyphenated forms. Thus our frequency ranking, though based on the data in the book, is not identical to the rankings found in the book. It is also likely that our counts and rankings are incorrect, since this work was done by hand, not by computer (we decided it was easier to spend a few hours with the book than many hours getting the tape from Brown, reading it, writing a program, and so on). Nevertheless, the counts are close, and the rankings probably very close, to the actual values; certainly they are close enough to generate a good stop list, which is our only concern.

Appendix A lists the tokens in English that occurred more than 300 times in the Brown corpus. There are 278 words, listed in order of occurrence along with their the frequency counts.

3. Culling Important Terms

As it turns out, many words of potentially great importance as index terms occur frequently. Hence the next step in forming our list was to examine it and pull out terms that we thought were too important as

potential index terms to be filtered from our indexes. Choice of these words was again an arbitrary decision, although we suspect that our choices will not be controversial. Thirty-two words were culled from the list in this way; they are listed in alphabetical order in Appendix B, along with their frequency of occurrence and their rank in the original list.

4. Adding A Bit More Fluff

In examining our list of the most frequently occurring words in the Brown corpus, we were surprised that many words traditionally appearing in stop lists did not make the cut. For example, "above" occurs only 296 times, "sure" only 265 times, and "whether" only 286 times. On the assumption that many of these traditional stop words would occur more frequently in certain kinds of writing, we decided to add some of them that barely missed the cut. The cut-off point that we arbitrarily chose for this part of the exercise was an occurrence of 200. We found 26 traditional stop words meeting this criterion, and added them to the list. These extra fluff words are listed in alphabetic order in Appendix C, along with their frequency of occurrence. At this point the stop list included 272 words.

5. Adding Words for Free

The stop list filter in the system we are building uses a minimum state deterministic finite automaton (minimal DFA) to tokenize an input stream and recognize stop words[4]. This sort of recognizer can often be adjusted to recognize larger sets of words virtually for free, that is, with no cost in memory or processing speed. For example, if the minimal DFA already recognizes words beginning with the letter "a," then often it can recognize the same set of words, along with the letter "a" itself, simply by making one of its states a final state. Since a little extra word stopping for free saves further processing down the line, for a net increase in efficiency, we decided to add words to the stop list likely to be recognizable for free, or almost for free.

We added words according to the following criteria:

- Add all letters for which a word already in the list starts with the letter.
- Add any traditional stop word occurring at least 100 times that differs from a word already in list only in a single letter.
- Add words with the same prefix ending in "body," "one," "thing," or "where," provided at least one of these words, or the prefix, already appears in the list. Also add the prefix, if it is a word. For example, since "nothing" occurs in the list, the words "noone," "nobody," and "nowhere" were added. The prefix "no" already appears in the list.
- Add words with the same prefix ending in "ed," "ing," or "s," provided at least one of these words, or the prefix, already appears in the list. Also add the prefix, it it is a word. For example, since "asked" appears in the list, the words "ask," "asking," and "asks" were added as well.
- Add words with the same prefix ending in "er" or "est," provided at least one of these words, or the prefix, already appears in the list. Also add the prefix, it it is a word.
- If a word in the list can take the suffix "ly," then add the suffixed word. Thus "clearly" was added to the list to join "clear."
- If a word in the list can take the suffix "self," then add the suffixed word.
- If a word in the list can take the suffix "s", then add the suffixed word.
- Add proper prefixes of words appearing in the list.

If the same word could be supplemented with more than one sort of ending, then a choice was made between the alternatives so as to stop the most occurrences. For example, either "clearly" or all three of "cleared," "clearing," and "clears" could be added to the list; the former word occurs 128 times, and the latter three a total of 40 times, so "clearly" was added instead of the others.

Altogether, this list has 149 words in it. These words are listed in alphabetical order in Appendix D. With these words, the stop list is brought to its final size of 421 words. A minimal DFA recognizer for this list will contain 317 states and 552 arcs, which is remarkably small considering that the list contains 421 words and 2450 characters.

6. Conclusion

Selecting a stop word list is more difficult than it appears, especially if its members are chosen based on empirical data about word usage in English. The stop list that we have generated can serve as the basis for stop lists for specialized data bases, or as a list for general English literature.

Appendix A—Most Frequently Occurring Words

The first number is the rank, the second the raw frequency of occurrence.

1	69975	the	2	36432	of
3	28872	and	4	26190	to
5	23073	a	6	21337	in
7	10790	that	8	10066	is
9	9806	was	10	9500	he
11	9495	for	12	8730	it
13	7289	with	14	7254	as
15	6976	not	16	6891	his
17	6742	on	18	6361	be
19	5377	at	20	5246	by
21	5149	i	22	5145	this
23	5133	had	24	4383	but
25	4372	are	26	4371	from
27	4204	or	28	3925	have
29	3727	an	30	3668	you
31	3621	they	32	3560	which
33	3298	one	34	3283	were
35	3062	would	36	3032	her
37	3002	all	38	2857	she
39	2844	there	40	2798	will
41	2668	their	42	2628	we
43	2572	him	44	2470	been
45	2430	has	46	2380	who
47	2333	when	48	2202	more
49	2199	if	50	2143	no
51	2082	out	52	1985	so
53	1961	what	54	1961	said
55	1890	up	56	1854	its
57	1816	about	58	1790	than
59	1784	into	60	1774	them
61	1768	can	62	1748	only
63	1701	other	64	1635	new
65	1618	some	66	1598	could
67	1595	time	68	1573	these
69	1414	two	70	1398	may
71	1377	then	72	1361	first
73	1350	do	74	1348	any
75	1314	now	76	1306	my
77	1303	such	78	1294	like
79	1281	man	80	1235	over
81	1233	our	82	1173	me
83	1169	even	84	1159	most
85	1110	made	86	1070	also
87	1070	after	88	1043	did
89	1027	many	90	1018	before
91	1013	must	92	971	through
93	957	years	94	946	where
95	937	much	96	936	back
97	912	your	98	910	way
99	898	down	100	897	well
101	888	should	102	883	because

103	878	each	104	872	just
105	868	people	106	850	those
107	842	state	108	834	too
109	833	mr	110	825	world
111	823	how	112	797	very
113	794	make	114	789	good
115	782	still	116	772	see
117	772	own	118	770	men
119	763	work	120	761	here
121	753	long	122	742	get
123	731	both	124	730	between
125	707	under	126	699	year
127	697	never	128	690	another
129	686	same	130	681	being
131	680	while	132	678	life
133	676	last	134	674	know
135	672	might	136	668	us
137	642	day	138	639	off
139	628	since	140	627	against
141	622	come	142	618	came
143	615	great	144	614	right
145	613	three	146	613	states
147	613	go	148	610	take
149	601	few	150	596	himself
151	591	use	152	588	during
153	583	without	154	580	again
155	570	place	156	567	around
157	561	old	158	552	however
159	542	small	160	536	mrs
161	530	home	162	517	thought
163	508	went	164	499	part
165	499	once	166	499	high
167	496	school	168	495	upon
169	492	every	170	490	say
171	481	united	172	474	used
173	472	number	174	467	war
175	467	does	176	462	until
177	458	away	178	456	always
179	450	water	180	449	something
181	446	fact	182	444	public
183	439	though	184	438	put
185	434	enough	186	433	think
187	433	almost	188	430	head
189	426	took	190	426	far
191	424	night	192	421	hand
193	419	system	194	415	set
195	414	general	196	412	nothing
197	410	better	198	405	point
199	404	why	200	400	house
201	399	end	202	397	later
203	397	find	204	397	eyes
205	397	asked	206	395	next
207	395	going	208	394	program
209	394	knew	210	387	give
211	386	toward	212	385	white

213	385	room	214	382	group
215	381	social	216	381	side
217	378	young	218	378	several
219	377	present	220	377	let
221	376	order	222	376	national
223	375	second	224	375	given
225	374	possible	226	373	rather
227	373	light	228	371	per
229	370	face	230	369	often
231	369	important	232	369	god
233	369	among	234	368	things
235	367	early	236	361	large
237	360	need	238	360	big
239	359	within	240	359	become
241	359	business	242	358	case
243	357	felt	244	355	along
245	353	best	246	348	four
247	344	ever	248	343	power
249	343	least	250	338	saw
251	338	got	252	337	less
253	334	mind	254	333	thing
255	328	want	256	326	today
257	325	others	258	324	interest
259	323	although	260	320	turned
261	318	open	262	318	members
263	318	area	264	314	family
265	314	done	266	313	problem
267	313	certain	268	312	kind
269	312	door	270	312	began
271	312	different	272	311	thus
273	311	sense	274	311	seemed
275	311	help	276	309	whole
277	307	perhaps	278	304	itself

Appendix B—Frequently Occurring but Important Words

The first number is the raw frequency of occurrence, the second number is the rank in the list in Appendix A.

business	359	243	day	642	139
door	312	271	eyes	397	206
family	314	266	god	369	234
hand	421	194	head	430	190
help	311	277	home	530	163
house	400	202	life	678	134
light	373	229	mind	334	255
national	376	224	night	424	193
own	772	119	people	868	107
power	343	250	program	394	210
public	444	184	school	496	169
sense	311	275	set	415	196
social	381	217	system	419	195
time	1595	69	united	481	173
war	467	176	water	450	181
white	385	214	world	825	112

Appendix C—Extra Fluff Words

The number is the raw frequency of occurrence.

above	296	across	282
already	274	behind	258
cannot	258	clear	219
either	285	full	230
further	218	gave	285
having	279	keep	260
known	245	making	231
necessary	222	quite	281
really	275	shall	268
show	289	sure	265
taken	281	therefore	205
together	268	whether	286
whose	251	yet	283

Appendix D—Free or Nearly Free Words

alone	anybody	anyone
anything	anywhere	areas
ask	asking	asks
b	backed	backing
backs	becomes	became
beings	c	cases
certainly	clearly	d
differ	differently	downed
downing	downs	e
ended	ending	ends
evenly	everybody	everyone
everything	everywhere	f
faces	facts	finds
fully	furthered	furthering
furthers	g	generally
gets	gives	goods
greater	greatest	grouped
grouping	groups	h
herself	higher	highest
interested	interesting	interests
j	k	keeps
knows	l	largely
latest	lets	likely
longer	longest	m
member	mostly	myself
n	needed	needing
needs	newer	newest
non	nobody	noone
nowhere	numbers	o
older	oldest	opened
opening	opens	ordered
ordering	orders	p
parted	parting	parts
places	pointed	pointing
points	presented	presenting
presents	problems	puts
q	r	rooms
s	says	seconds
sees	seem	seeming
seems	showed	showing
shows	sides	smaller
smallest	somebody	someone
somewhere	t	thinks
thoughts	turn	turning
turns	u	uses
v	w	wanted
wanting	wants	ways
wells	worked	working
works	y	younger
youngest	yours	

Appendix E—The Final List

The first number is the raw frequency of occurrence. The second number, if there is one, is the rank of the word in the original list of the 278 most frequently occurring words. The final annotation explains why extra words have been added to the list.

a	23073	5	
about	1816	59	
above	296	-	extra fluff
across	282	-	extra fluff
after	1070	89	
again	580	156	
against	627	142	
all	3002	37	
almost	433	189	
alone	195	-	cheap with "along"
along	355	246	
already	274	-	extra fluff
also	1070	88	
although	323	261	
always	456	180	
among	369	235	
an	3727	29	
and	28872	3	
another	690	130	
any	1348	76	
anybody	45	-	(body,one,thing,where) endings
anyone	146	-	(body,one,thing,where) endings
anything	280	-	(body,one,thing,where) endings
anywhere	39	-	(body,one,thing,where) endings
are	4372	25	
area	318	265	
areas	236	-	cheap with "area"
around	567	158	
as	7254	14	
ask	128	-	(ed,ing,s) endings
asked	397	207	
asking	67	-	(ed,ing,s) endings
asks	18	-	(ed,ing,s) endings
at	5377	19	
away	458	179	
b	140	-	free
back	936	98	
backed	24	-	(ed,ing,s) endings
backing	8	-	(ed,ing,s) endings
backs	15	-	(ed,ing,s) endings
be	6361	18	
because	883	104	
become	359	242	
becomes	104	-	cheap with "become"
became	246	-	cheap with "become"
been	2470	46	
before	1018	92	
began	312	272	
behind	258	-	extra fluff

being	681	132	
beings	36	-	cheap with "being"
best	353	247	
better	410	199	
between	730	126	
big	360	240	
both	731	125	
but	4383	24	
by	5246	20	
c	150	-	free
came	618	144	
can	1768	63	
cannot	258	-	extra fluff
case	358	244	
cases	148	-	cheap with "case"
certain	313	269	
certainly	143	-	(ly) ending
clear	219	-	extra fluff
clearly	128	-	(ly) ending
come	622	143	
could	1598	68	
d	120	-	free
did	1043	90	
differ	18	-	free with "different"
different	312	273	
differently	16	-	(ly) ending
do	1350	75	
does	467	177	
done	314	267	
down	898	101	
downed	5	-	(ed,ing,s) endings
downing	1	-	(ed,ing,s) endings
downs	5	-	(ed,ing,s) endings
during	588	154	
e	110	-	free
each	878	105	
early	367	237	
either	285	-	extra fluff
end	399	203	
ended	49	-	(ed,ing,s) endings
ending	27	-	(ed,ing,s) endings
ends	95	-	(ed,ing,s) endings
enough	434	187	
even	1169	85	
evenly	4	-	(ly) ending
ever	344	249	
every	492	171	
everybody	76	-	(body,one,thing,where) endings
everyone	98	-	(body,one,thing,where) endings
everything	188	-	(body,one,thing,where) endings
everywhere	47	-	(body,one,thing,where) endings
f	70	-	free
face	370	231	
faces	72	-	cheap with "face"
fact	446	183	

facts	87	-	cheap with "fact"
far	426	192	
felt	357	245	
few	601	151	
find	397	205	
finds	59	-	cheap with "find"
first	1361	74	
for	9495	11	
four	348	248	
from	4371	26	
full	230	-	extra fluff
fully	80	-	(ly) ending
further	218	-	extra fluff
furthered	3	-	(ed,ing,s) endings
furthering	2	-	(ed,ing,s) endings
furthers	0	-	(ed,ing,s) endings
g	55	-	free
gave	285	-	extra fluff
general	414	197	
generally	132	-	(ly) ending
get	742	124	
gets	66	-	cheap with "get"
give	387	212	
given	375	226	
gives	114	-	cheap with "give"
go	613	149	
going	395	209	
good	789	116	
goods	57	-	cheap with "good"
got	338	253	
great	615	145	
greater	188	-	(er,est) endings
greatest	88	-	(er,est) endings
group	382	216	
grouped	5	-	(ed,ing,s) endings
grouping	4	-	(ed,ing,s) endings
groups	125	-	(ed,ing,s) endings
h	80	-	free
had	5133	23	
has	2430	47	
have	3925	28	
having	279	-	extra fluff
he	9500	10	
her	3032	36	
herself	125	-	cheap with "her"
here	761	122	
high	499	168	
higher	161	-	(er,est) endings
highest	4	-	(er,est) endings
him	2572	45	
himself	596	152	
his	6891	16	
how	823	113	
however	552	160	
i	5149	21	

if	2199	51	
important	369	233	
in	21337	6	
interest	324	260	
interested	101	-	(ed,ing,s) endings
interesting	82	-	(ed,ing,s) endings
interests	83	-	(ed,ing,s) endings
into	1784	61	
is	10066	8	
it	8730	12	
its	1854	58	
itself	304	280	
j	130	-	free
just	872	106	
k	30	-	free
keep	260	-	extra fluff
keeps	21	-	cheap with "keep"
kind	312	270	
knew	394	211	
know	674	136	
known	245	-	extra fluff
knows	99	-	cheap with "know"
l	60	-	free
large	361	238	
largely	68	-	(ly) ending
last	676	135	
later	397	204	
latest	35	-	(er,est) endings
least	343	251	
less	337	254	
let	377	222	
lets	5	-	cheap with "let"
like	1294	80	
likely	151	-	(ly) ending
long	753	123	
longer	193	-	(er,est) endings
longest	6	-	(er,est) endings
m	70	-	free
made	1110	87	
make	794	115	
making	231	-	extra fluff
man	1281	81	
many	1027	91	
may	1398	72	
me	1173	84	
member	137	-	free
members	318	264	
men	770	120	
might	672	137	
more	2202	50	
most	1159	86	
mostly	44	-	(ly) ending
mr	833	111	
mrs	536	162	
much	937	97	

must	1013	93	
my	1306	78	
myself	129	-	cheap with "my"
n	35	-	free
necessary	222	-	extra fluff
need	360	239	
needed	187	-	(ed,ing,s) endings
needing	5	-	(ed,ing,s) endings
needs	59	-	(ed,ing,s) endings
never	697	129	
new	1635	66	
newer	20	-	(er,est) endings
newest	15	-	(er,est) endings
next	395	208	
no	2143	52	
non	146	-	cheap with "not"
not	6976	15	
nobody	79	-	(body,one,thing,where) endings
noone	0	-	(body,one,thing,where) endings
nothing	412	198	
now	1314	77	
nowhere	30	-	(body,one,thing,where) endings
number	472	175	
numbers	125	-	cheap with "number"
o	35	-	free
of	36432	2	
off	639	140	
often	369	232	
old	561	159	
older	93	-	(er,est) endings
oldest	14	-	(er,est) endings
on	6742	17	
once	499	167	
one	3298	33	
only	1748	64	
open	318	263	
opened	131	-	(ed,ing,s) endings
opening	83	-	(ed,ing,s) endings
opens	16	-	(ed,ing,s) endings
or	4204	27	
order	376	223	
ordered	69	-	(ed,ing,s) endings
ordering	13	-	(ed,ing,s) endings
orders	58	-	(ed,ing,s) endings
other	1701	65	
others	325	259	
our	1233	83	
out	2082	53	
over	1235	82	
p	120	-	free
part	499	166	
parted	5	-	(ed,ing,s) endings
parting	3	-	(ed,ing,s) endings
parts	113	-	(ed,ing,s) endings
per	371	230	

perhaps	307	279	
place	570	157	
places	9	-	cheap with "place"
point	405	200	
pointed	72	-	(ed,ing,s) endings
pointing	26	-	(ed,ing,s) endings
points	143	-	(ed,ing,s) endings
possible	374	227	
present	377	221	
presented	66	-	(ed,ing,s) endings
presenting	10	-	(ed,ing,s) endings
presents	33	-	(ed,ing,s) endings
problem	313	268	
problems	240	-	cheap with "problem"
put	438	186	
puts	20	-	cheap with "put"
q	9	-	free
quite	281	-	extra fluff
r	80	-	free
rather	373	228	
really	275	-	extra fluff
right	614	146	
room	385	215	
rooms	55	-	cheap with "room"
s	130	-	free
said	1961	56	
same	686	131	
saw	338	252	
say	490	172	
says	200	-	cheap with "say"
second	375	225	
seconds	27	-	cheap with "second"
see	772	118	
sees	36	-	cheap with "see"
seem	228	-	(ed,ing,s) endings
seemed	311	276	
seeming	10	-	(ed,ing,s) endings
seems	259	-	(ed,ing,s) endings
several	378	220	
shall	268	-	extra fluff
she	2857	38	
should	888	103	
show	289	-	extra fluff
showed	141	-	(ed,ing,s) endings
showing	63	-	(ed,ing,s) endings
shows	94	-	(ed,ing,s) endings
side	381	218	
sides	98	-	cheap with "side"
since	628	141	
small	542	161	
smaller	78	-	(er,est) endings
smallest	13	-	(er,est) endings
so	1985	54	
some	1618	67	
somebody	65	-	(body,one,thing,where) endings

someone	100	-	(body,one,thing,where) endings
something	449	182	
somewhere	60	-	(body,one,thing,where) endings
state	842	109	
states	613	148	
still	782	117	
such	1303	79	
sure	265	-	extra fluff
t	65	-	free
take	610	150	
taken	281	-	extra fluff
than	1790	60	
that	10790	7	
the	69975	1	
their	2668	42	
them	1774	62	
then	1377	73	
there	2844	39	
therefore	205	-	extra fluff
these	1573	70	
they	3621	31	
thing	333	256	
things	368	236	
think	433	188	
thinks	23	-	cheap with "think"
this	5145	22	
those	850	108	
though	439	185	
thought	517	164	
thoughts	54	-	cheap with "thought"
three	613	147	
through	971	94	
thus	311	274	
to	26190	4	
today	326	258	
together	268	-	extra fluff
too	834	110	
took	426	191	
toward	386	213	
turn	233	-	(ed,ing,s) endings
turned	320	262	
turning	75	-	(ed,ing,s) endings
turns	38	-	(ed,ing,s) endings
two	1414	71	
u	220	-	free
under	707	127	
until	462	178	
up	1890	57	
upon	495	170	
us	668	138	
use	591	153	
uses	57	-	cheap with "use"
used	474	174	
v	19	-	free
very	797	114	

w	90	-	free
want	328	257	
wanted	226	-	(ed,ing,s) endings
wanting	16	-	(ed,ing,s) endings
wants	72	-	(ed,ing,s) endings
was	9806	9	
way	910	100	
ways	128	-	cheap with "way"
we	2628	44	
well	897	102	
wells	6	-	cheap with "well"
went	508	165	
were	3283	34	
what	1961	55	
when	2333	49	
where	946	96	
whether	286	-	extra fluff
which	3560	32	
while	680	133	
who	2380	48	
whole	309	278	
whose	251	-	extra fluff
why	404	201	
will	2798	40	
with	7289	13	
within	359	241	
without	583	155	
work	763	121	
worked	128	-	(ed,ing,s) endings
working	151	-	(ed,ing,s) endings
works	130	-	(ed,ing,s) endings
would	3062	35	
y	20	-	free
year	699	128	
years	957	95	
yet	283	-	extra fluff
you	3668	30	
young	378	219	
younger	41	-	(er,est) endings
youngest	14	-	(er,est) endings
your	912	99	
yours	25	-	cheap with "your"

REFERENCES

1. van Rijsbergen, C.J., *Information Retrieval*, Butterworths, 1975.

2. Luhn, H.P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal of Research and Development* 1(4), October, 1957.

3. Francis, W. Nelson, and Henry Kucera, *Frequency Analysis of English Usage*, Houghton Mifflin, 1982.

4. Aho, Alfred, Ravi Sethi, and Jeffrey Ullman, *Compilers: Principles, Techniques, and Tools*, Addison-Wesley, 1986.