1. Title: SPAM E-mail Database

2. Sources:
   (a) Creators: Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt
        Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304
   (b) Donor: George Forman (gforman at nospam hpl.hp.com) 650-857-7835
   (c) Generated: June-July 1999

3. Past Usage:
   (a) Hewlett-Packard Internal-only Technical Report. External forthcoming.
   (b) Determine whether a given email is spam or not.
   (c) ~7% misclassification error.
       False positives (marking good mail as spam) are very undesirable.
       If we insist on zero false positives in the training/testing set,
       20-25% of the spam passed through the filter.

4. Relevant Information:
        The "spam" concept is diverse: advertisements for products/web
        sites, make money fast schemes, chain letters, pornography...
        Our collection of spam e-mails came from our postmaster and
        individuals who had filed spam. Our collection of non-spam
        e-mails came from filed work and personal e-mails, and hence
        the word 'george' and the area code '650' are indicators of
        non-spam. These are useful when constructing a personalized
        spam filter. One would either have to blind such non-spam
        indicators or get a very wide collection of non-spam to
        generate a general purpose spam filter.

        For background on spam:
        Cranor, Lorrie F., LaMacchia, Brian A. Spam!
        Communications of the ACM, 41(8):74-83, 1998.

5. Number of Instances: 4601 (1813 Spam = 39.4%)

6. Number of Attributes: 58 (57 continuous, 1 nominal class label)

7. Attribute Information:
The last column of 'spambase.data' denotes whether the e-mail was
considered spam (1) or not (0), i.e. unsolicited commercial e-mail.
Most of the attributes indicate whether a particular word or
character was frequently occuring in the e-mail. The run-length
attributes (55-57) measure the length of sequences of consecutive
capital letters. For the statistical measures of each attribute,
see the end of this file. Here are the definitions of the attributes:

48 continuous real [0,100] attributes of type word_freq_WORD
= percentage of words in the e-mail that match WORD,
i.e. 100 * (number of times the WORD appears in the e-mail) /
total number of words in e-mail. A "word" in this case is any
string of alphanumeric characters bounded by non-alphanumeric

characters or end-of-string.

6 continuous real [0,100] attributes of type char_freq_CHAR
= percentage of characters in the e-mail that match CHAR,
i.e. 100 * (number of CHAR occurences) / total characters in e-mail

1 continuous real [1,...] attribute of type capital_run_length_average
= average length of uninterrupted sequences of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_longest
= length of longest uninterrupted sequence of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_total
= sum of length of uninterrupted sequences of capital letters
= total number of capital letters in the e-mail

1 nominal {0,1} class attribute of type spam
= denotes whether the e-mail was considered spam (1) or not (0),
i.e. unsolicited commercial e-mail.

8. Missing Attribute Values: None

9. Class Distribution:

Spam	1813	(39.4%)
Non-Spam	2788	(60.6%)

Attribute Statistics:

	Min:	Max:	Average:	Std.Dev:	Coeff.Var_%:
1	0	4.54	0.10455	0.30536	292
2	0	14.28	0.21301	1.2906	606
3	0	5.1	0.28066	0.50414	180
4	0	42.81	0.065425	1.3952	2130
5	0	10	0.31222	0.67251	215
6	0	5.88	0.095901	0.27382	286
7	0	7.27	0.11421	0.39144	343
8	0	11.11	0.10529	0.40107	381
9	0	5.26	0.090067	0.27862	309
10	0	18.18	0.23941	0.64476	269
11	0	2.61	0.059824	0.20154	337
12	0	9.67	0.5417	0.8617	159
13	0	5.55	0.09393	0.30104	320
14	0	10	0.058626	0.33518	572
15	0	4.41	0.049205	0.25884	526
16	0	20	0.24885	0.82579	332
17	0	7.14	0.14259	0.44406	311
18	0	9.09	0.18474	0.53112	287
19	0	18.75	1.6621	1.7755	107
20	0	18.18	0.085577	0.50977	596

21	0	11.11	0.80976	1.2008	148
22	0	17.1	0.1212	1.0258	846
23	0	5.45	0.10165	0.35029	345
24	0	12.5	0.094269	0.44264	470
25	0	20.83	0.5495	1.6713	304
26	0	16.66	0.26538	0.88696	334
27	0	33.33	0.7673	3.3673	439
28	0	9.09	0.12484	0.53858	431
29	0	14.28	0.098915	0.59333	600
30	0	5.88	0.10285	0.45668	444
31	0	12.5	0.064753	0.40339	623
32	0	4.76	0.047048	0.32856	698
33	0	18.18	0.097229	0.55591	572
34	0	4.76	0.047835	0.32945	689
35	0	20	0.10541	0.53226	505
36	0	7.69	0.097477	0.40262	413
37	0	6.89	0.13695	0.42345	309
38	0	8.33	0.013201	0.22065	1670
39	0	11.11	0.078629	0.43467	553
40	0	4.76	0.064834	0.34992	540
41	0	7.14	0.043667	0.3612	827
42	0	14.28	0.13234	0.76682	579
43	0	3.57	0.046099	0.22381	486
44	0	20	0.079196	0.62198	785
45	0	21.42	0.30122	1.0117	336
46	0	22.05	0.17982	0.91112	507
47	0	2.17	0.0054445	0.076274	1400
48	0	10	0.031869	0.28573	897
49	0	4.385	0.038575	0.24347	631
50	0	9.752	0.13903	0.27036	194
51	0	4.081	0.016976	0.10939	644
52	0	32.478	0.26907	0.81567	303
53	0	6.003	0.075811	0.24588	324
54	0	19.829	0.044238	0.42934	971
55	1	1102.5	5.1915	31.729	611
56	1	9989	52.173	194.89	374
57	1	15841	283.29	606.35	214
58	0	1	0.39404	0.4887	124

This file: 'spambase.DOCUMENTATION' at the UCI Machine Learning Repository
http://www.ics.uci.edu/~mlearn/MLRepository.html