This is the assignment for week 8's supervision.

An old exam question

A spy would like you to write a computer program that recognises, given a small number of consecutive characters from the middle of a computer file, whether the file is an English-language document. Assuming that the two alternative hypotheses are that the file is an English-language document ($\mathcal{H}_E$), or that it is a random string of characters drawn from the same alphabet ($\mathcal{H}_R$), describe how you would solve this problem.

Estimate how many characters your method would need in order to work reasonably well.

Further questions

Maybe you would enjoy writing a program that implements your method?

How would your answers differ if instead the task were to distinguish

(a) English from German?

(b) English from Hsilgne (backwards English)?

[When I say German, let's assume unaccented German – German with all the umlauts stripped out.]

(Some facts about English and German are supplied below.)

LETTER FREQUENCIES OF ENGLISH AND GERMAN

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
English (**e**)	.07	.01	.03	.03	.10	.02	.02	.05	.06	.001	.006	.03	.02	.06
German (**g**)	.06	.02	.03	.04	.15	.01	.03	.04	.07	.002	.01	.03	.02	.08

	O	P	Q	R	S	T	U	V	W	X	Y	Z	–
English (**e**)	.06	.02	.0009	.05	.05	.08	.02	.008	.02	.002	.01	.0008	.17
German (**g**)	.02	.007	.0002	.06	.06	.05	.04	.006	.02	.0003	.0003	.01	.14

The entropies of these two distributions are $H(\mathbf{e}) = 4.1\,\text{bits}$; $H(\mathbf{g}) = 4.1\,\text{bits}$; and the relative entropies between them are $D_{\text{KL}}(\mathbf{e}||\mathbf{g}) = 0.16\,\text{bits}$ and $D_{\text{KL}}(\mathbf{g}||\mathbf{e}) = 0.12\,\text{bits}$. The relative entropies between the uniform distribution $\mathbf{u}$ and the English distribution $\mathbf{e}$ are $D_{\text{KL}}(\mathbf{e}||\mathbf{u}) \simeq 0.6\,\text{bits}$ and $D_{\text{KL}}(\mathbf{u}||\mathbf{e}) \simeq 1\,\text{bits}$.

How well calibrated are your estimates of uncertainty?

According to BBC News, the Earth was almost put on impact alert by some astronomers who, on 13 January 2004, reckoned a 30m object, later designated 2004 AS1, had a one-in-four chance of hitting the planet within 36 hours. 2004 AS1 in fact passed the Earth at a distance of about 12 million km – 32 times the Earth-Moon distance, posing no danger to us whatsoever. Clearly the correct assessment of the probability of earth being hit should have been more like one in 10^7.

This sheet is intended to give insight into the accuracy of our own probability estimates.

Give a 94% confidence interval for the following quantities. Give the tightest* interval you can, while remaining 94% sure that the true value is in the interval.

	Quantity	Lower bound	Guess	Upper bound	Ratio	Score
1	Mass of textbook (g)					
2	Population of Britain					
3	Population of Turkey					
4	Population of Luxembourg					
5	Number of British MEPs					
6	Starting pay of University Lecturer					
7	Parliamentary salary of MP					
8	Council tax, South Cambs. (£/house)					
9	Fraction of central government expenditure that goes to 'Defence'					
10	UK prison population (as fraction of whole)					
11	Number of USA nuclear warheads					
12	Distance to sun (miles)					
13	Mean radius of earth (km)					
14	Speed of light ($\mathrm{m\,s^{-1}}$)					
15	Density of Gold ($\mathrm{g\,cm^{-3}}$)					
16	The ratio $\dfrac{\text{Density of Uranium}^{238}}{\text{Density of Gold}}$					

$\star$ My definition of 'tightest' is that an interval (l, u) gets tighter if the ratio u/l is reduced. If you want, you can score how tight your intervals are using $\log_2(\log_2 u/l)$ – the smaller, the better.

Ratio	$\log_2(\log_2 \text{Ratio})$
1.01	-6
1.02	-5
1.04	-4
1.09	-3
1.19	-2
1.41	-1
2	0
4	1
16	2
250	3
100,000	4