```
(* Adam Beck *)


(* Select long texts in three different languages. Calculate the entropy
 for each and compare them, draw your conclusions *)

(* I will first select an English text *)


(* Sample text from The Adventures of Sherklock Holmes *)
```

```
text =
  "To Sherlock Holmes she is always the woman. I have seldom heard him mention her under
    any other name. In his eyes she eclipses and predominates the whole of her
    sex. It was not that he felt any emotion akin to love for Irene Adler. All
    emotions, and that one particularly, were abhorrent to his cold, precise but
    admirably balanced mind. He was, I take it, the most perfect reasoning and
    observing machine that the world has seen, but as a lover he would have placed
    himself in a false position. He never spoke of the softer passions, save with
    a gibe and a sneer. They were admirable things for the observer--excellent
    for drawing the veil from men's motives and actions. But for the trained
    reasoner to admit such intrusions into his own delicate and finely adjusted
    temperament was to introduce a distracting factor which might throw a doubt
    upon all his mental results. Grit in a sensitive instrument, or a crack in
    one of his own high-power lenses, would not be more disturbing than a strong
    emotion in a nature such as his. And yet there was but one woman to him,
    and that woman was the late Irene Adler, of dubious and questionable memory.

I had seen little of Holmes lately. My marriage had drifted us away from each other. My
    own complete happiness, and the home-centred interests which rise up around the
    man who first finds himself master of his own establishment, were sufficient
    to absorb all my attention, while Holmes, who loathed every form of society
    with his whole Bohemian soul, remained in our lodgings in Baker Street, buried
    among his old books, and alternating from week to week between cocaine and
    ambition, the drowsiness of the drug, and the fierce energy of his own keen
    nature. He was still, as ever, deeply attracted by the study of crime, and
    occupied his immense faculties and extraordinary powers of observation in
    following out those clues, and clearing up those mysteries which had been
    abandoned as hopeless by the official police. From time to time I heard some
    vague account of his doings: of his summons to Odessa in the case of the
    Trepoff murder, of his clearing up of the singular tragedy of the Atkinson
    brothers at Trincomalee, and finally of the mission which he had accomplished
    so delicately and successfully for the reigning family of Holland. Beyond
    these signs of his activity, however, which I merely shared with all the
    readers of the daily press, I knew little of my former friend and companion.

One night--it was on the twentieth of March, 1888--I was returning from a journey to a
    patient (for I had now returned to civil practice), when my way led me
    through Baker Street. As I passed the well-remembered door, which must always
    be associated in my mind with my wooing, and with the dark incidents of
    the Study in Scarlet, I was seized with a keen desire to see Holmes again,
    and to know how he was employing his extraordinary powers. His rooms were
    brilliantly lit, and, even as I looked up, I saw his tall, spare figure pass
    twice in a dark silhouette against the blind. He was pacing the room swiftly,
    eagerly, with his head sunk upon his chest and his hands clasped behind
    him. To me, who knew his every mood and habit, his attitude and manner told
    their own story. He was at work again. He had risen out of his drug-created
    dreams and was hot upon the scent of some new problem. I rang the bell
    and was shown up to the chamber which had formerly been in part my own.";

lowerEnglishText = ToLowerCase[text]; (* Convert text to lowercase *)
```

```
charcode = ToCharacterCode[lowerEnglishText];
(* Get a numerical character code for each letter in the string *)

charentp = Table[{Length[Position[charcode, i]], FromCharacterCode[i]}, {i, 96, 122}]
 (* make a table with the counts for each character *)
```

```
{{0, `}, {218, a}, {43, b}, {69, c}, {118, d}, {335, e}, {66, f}, {41, g}, {170, h},
 {211, i}, {2, j}, {21, k}, {110, l}, {93, m}, {196, n}, {208, o}, {42, p}, {1, q},
 {161, r}, {191, s}, {214, t}, {61, u}, {20, v}, {81, w}, {4, x}, {50, y}, {1, z}}
```

```
charentp[[1]] = {Length[Position[charcode, 32]], "space"}
 (* Also, get the character count for spaces *)
```

```
{611, space}
```

```
TableForm[charentp] (* Put the counts and characters in a table to visualize easier *)
```

```
611    space
218    a
43     b
69     c
118    d
335    e
66     f
41     g
170    h
211    i
2      j
21     k
110    l
93     m
196    n
208    o
42     p
1      q
161    r
191    s
214    t
61     u
20     v
81     w
4      x
50     y
1      z
```

```
sortcharentp = Reverse[Sort[charentp]]
 (* Sort by most common to least common frequencies *)
```

```
{{611, space}, {335, e}, {218, a}, {214, t}, {211, i}, {208, o}, {196, n}, {191, s},
 {170, h}, {161, r}, {118, d}, {110, l}, {93, m}, {81, w}, {69, c}, {66, f}, {61, u},
 {50, y}, {43, b}, {42, p}, {41, g}, {21, k}, {20, v}, {4, x}, {2, j}, {1, z}, {1, q}}
```

```
sumchars = Sum[sortcharentp[[i, 1]], {i, Length[sortcharentp]}]
 (* Get the total character count in the text *)
```

```
3338
```

```
charfreq = N[Table[{sortcharentp[[i, 1]] / sumchars, sortcharentp[[i, 2]]},
   {i, 1, Length[sortcharentp]}]] (* Get the frequency for each character in the text *)
```

```
{{0.183044, space}, {0.100359, e}, {0.0653086, a}, {0.0641102, t},
 {0.0632115, i}, {0.0623128, o}, {0.0587178, n}, {0.0572199, s}, {0.0509287, h},
 {0.0482325, r}, {0.0353505, d}, {0.0329539, l}, {0.027861, m}, {0.024266, w},
 {0.0206711, c}, {0.0197723, f}, {0.0182744, u}, {0.014979, y}, {0.012882, b},
 {0.0125824, p}, {0.0122828, g}, {0.00629119, k}, {0.00599161, v},
 {0.00119832, x}, {0.000599161, j}, {0.000299581, z}, {0.000299581, q}}
```

```
TableForm[charfreq] (* Put this frequency in a table to visualize easier *)
```

```
0.183044       space
0.100359       e
0.0653086      a
0.0641102      t
0.0632115      i
0.0623128      o
0.0587178      n
0.0572199      s
0.0509287      h
0.0482325      r
0.0353505      d
0.0329539      l
0.027861       m
0.024266       w
0.0206711      c
0.0197723      f
0.0182744      u
0.014979       y
0.012882       b
0.0125824      p
0.0122828      g
0.00629119     k
0.00599161     v
0.00119832     x
0.000599161    j
0.000299581    z
0.000299581    q
```

```
entplang = 0;
(* Loop through the entire table, using the entropy formula,
calculate the entropy of the text given our frequencies *)
For[i = 1, i <= 27, i++,
            If[charfreq[[i, 1]] == 0, entplang,
                entplang = entplang - charfreq[[i, 1]] * Log[charfreq[[i, 1]]]
            ]
        ]


(* This represents the average minimum number of bits needed to
 encode a string of symbols,based on the frequency of the symbols. *)
entplang
```

```
2.83089
```

```
(* Take this entropy, multiply it by the size of our text, and that gives the amount
 of bits required to optimally encode the string *)
optimalEnglishEncoding = entplang * sumchars
```

```
9449.52
```

```
(* Next, I will analyze a French text *)
(* Le tour du monde en quatre-vingts jours by Jules Verne
  is the French version of Around the World in Eighty Days *)
frenchText = "
Mais si le rétablissement de la jeune Indienne ne fit pas question
dans l'esprit du brigadier général, celui-ci se montrait moins rassuré
pour l'avenir. Il n'hésita pas à dire à Phileas Fogg que si Mrs.
Aouda restait dans l'Inde, elle retomberait inévitablement entre les
mains de ses bourreaux. Ces énergumènes se tenaient dans toute la
péninsule, et certainement, malgré la police anglaise, ils sauraient
reprendre leur victime, fût-ce à Madras, à Bombay, à Calcutta. Et Sir
Francis Cromarty citait, à l'appui de ce dire, un fait de même nature
qui s'était passé récemment. A son avis, la jeune femme ne serait
véritablement en sûreté qu'après avoir quitté l'Inde.

Phileas Fogg répondit qu'il tiendrait compte de ces observations et
qu'il aviserait.

Vers dix heures, le guide annonçait la station d'Allahabad. Là
reprenait la voie interrompue du chemin de fer, dont les trains
franchissent, en moins d'un jour et d'une nuit, la distance qui sépare
Allahabad de Calcutta.

Phileas Fogg devait donc arriver à temps pour prendre un paquebot qui
ne partait que le lendemain seulement, 25 octobre, à midi, pour
Hong-Kong.

La jeune femme fut déposée dans une chambre de la gare. Passepartout
fut chargé d'aller acheter pour elle divers objets de toilette, robe,
châle, fourrures, etc., ce qu'il trouverait. Son maître lui ouvrait
un crédit illimité.

Passepartout partit aussitôt et courut les rues de la ville.
Allahabad, c'est la cité de Dieu, l'une des plus vénérées de l'Inde,
en raison de ce qu'elle est bâtie au confluent de deux fleuves sacrés,
le Gange et la Jumna, dont les eaux attirent les pèlerins de toute la
péninsule. On sait d'ailleurs que, suivant les légendes du Ramayana,
le Gange prend sa source dans le ciel, d'où, grâce à Brahma, il
descend sur la terre.";

lowerFrenchText = ToLowerCase[frenchText]; (* Convert text to lowercase *)

charcode = ToCharacterCode[lowerFrenchText];
(* Get a numerical character code for each letter in the string *)

charentp = Table[{Length[Position[charcode, i]], FromCharacterCode[i]}, {i, 96, 122}];
(* make a table with the counts for each character *)
```

```
charentp[[1]] = {Length[Position[charcode, 32]], "space"}
 (* Also, get the character count for spaces *)
```

```
{262, space}
```

```
TableForm[charentp] (* Put the counts and characters in a table to visualize easier *)
```

```
262    space
129    a
19     b
40     c
64     d
198    e
16     f
22     g
16     h
107    i
6      j
1      k
87     l
38     m
97     n
54     o
36     p
14     q
98     r
97     s
108    t
78     u
19     v
0      w
4      x
3      y
0      z
```

```
sortcharentp = Reverse[Sort[charentp]]
 (* Sort by most common to least common frequencies *)
```

```
{{262, space}, {198, e}, {129, a}, {108, t}, {107, i}, {98, r}, {97, s}, {97, n},
 {87, l}, {78, u}, {64, d}, {54, o}, {40, c}, {38, m}, {36, p}, {22, g}, {19, v},
 {19, b}, {16, h}, {16, f}, {14, q}, {6, j}, {4, x}, {3, y}, {1, k}, {0, z}, {0, w}}
```

```
sumchars = Sum[sortcharentp[[i, 1]], {i, Length[sortcharentp]}]
 (* Get the total character count in the text *)
```

```
1613
```

```
charfreq = N[Table[{sortcharentp[[i, 1]] / sumchars, sortcharentp[[i, 2]]},
   {i, 1, Length[sortcharentp]}]]
 (* Get the frequency for each character in the text *)
```

```
{{0.16243, space}, {0.122753, e}, {0.0799752, a}, {0.066956, t},
 {0.066336, i}, {0.0607564, r}, {0.0601364, s}, {0.0601364, n}, {0.0539368, l},
 {0.0483571, u}, {0.0396776, d}, {0.033478, o}, {0.0247985, c}, {0.0235586, m},
 {0.0223187, p}, {0.0136392, g}, {0.0117793, v}, {0.0117793, b},
 {0.0099194, h}, {0.0099194, f}, {0.00867948, q}, {0.00371978, j},
 {0.00247985, x}, {0.00185989, y}, {0.000619963, k}, {0., z}, {0., w}}
```

```
TableForm[charfreq] (* Put this frequency in a table to visualize easier *)
```

```
0.16243        space
0.122753       e
0.0799752      a
0.066956       t
0.066336       i
0.0607564      r
0.0601364      s
0.0601364      n
0.0539368      l
0.0483571      u
0.0396776      d
0.033478       o
0.0247985      c
0.0235586      m
0.0223187      p
0.0136392      g
0.0117793      v
0.0117793      b
0.0099194      h
0.0099194      f
0.00867948     q
0.00371978     j
0.00247985     x
0.00185989     y
0.000619963    k
0.             z
0.             w
```

```
entplang = 0;
(* Loop through the entire table, using the entropy formula,
calculate the entropy of the text given our frequencies *)
For[i = 1, i <= 27, i++,
            If[charfreq[[i, 1]] == 0, entplang,
                entplang = entplang - charfreq[[i, 1]] * Log[charfreq[[i, 1]]]
            ]
        ]
(* This represents the average minimum number of bits needed to encode a string
 of symbols, based on the frequency of the symbols. *)
entplang
```

```
2.78247
```

```
(* Take this entropy, multiply it by the size of our text,
and that gives the amount of bits required to optimally encode the string *)
optimalFrenchEncoding = entplang * sumchars
```

```
4488.12
```

```
(* Next, I will analyze an Italian text *)
(* Orlando Furioso, and Italian epic poem by Ludovico Ariosto *)


italianText = "Le donne, i cavallier, l'arme, gli amori,
  le cortesie, l'audaci imprese io canto,
  che furo al tempo che passaro i Mori
```

d'Africa il mare, e in Francia nocquer tanto,
seguendo l'ire e i giovenil furori
d'Agramante lor re, che si diè vanto
di vendicar la morte di Troiano
sopra re Carlo imperator romano. Dirò d'Orlando in un medesmo tratto
cosa non detta in prosa mai, né in rima:
che per amor venne in furore e matto,
d'uom che sì saggio era stimato prima;
se da colei che tal quasi m'ha fatto,
che 'l poco ingegno ad or ad or mi lima,
me ne sarà però tanto concesso,
che mi basti a finir quanto ho promesso. Piacciavi, generosa Erculea prole,
ornamento e splendor del secol nostro,
Ippolito, aggradir questo che vuole
e darvi sol può l'umil servo vostro.
Quel ch'io vi debbo, posso di parole
pagare in parte e d'opera d'inchiostro;
né che poco io vi dia da imputar sono,
che quanto io posso dar, tutto vi dono.Voi sentirete fra i più degni eroi,
che nominar con laude m'apparecchio,
ricordar quel Ruggier, che fu di voi
e de' vostri avi illustri il ceppo vecchio.
L'alto valore e' chiari gesti suoi
vi farò udir, se voi mi date orecchio,
e vostri alti pensier cedino un poco,
sì che tra lor miei versi abbiano loco. Orlando, che gran tempo innamorato
fu de la bella Angelica, e per lei
in India, in Media, in Tartaria lasciato
avea infiniti ed immortal trofei,
in Ponente con essa era tornato,
dove sotto i gran monti Pirenei
con la gente di Francia e de Lamagna
re Carlo era attendato alla campagna,per far al re Marsilio e al re Agramante
battersi ancor del folle ardir la guancia,
d'aver condotto, l'un, d'Africa quante
genti erano atte a portar spada e lancia;
l'altro, d'aver spinta la Spagna inante
a destruzion del bel regno di Francia.
E così Orlando arrivò quivi a punto:
ma tosto si pentì d'esservi giunto:
che vi fu tolta la sua donna poi:
ecco il giudicio uman come spesso erra!
Quella che dagli esperi ai liti eoi
avea difesa con sì lunga guerra,
or tolta gli è fra tanti amici suoi,
senza spada adoprar, ne la sua terra.
Il savio imperator, ch'estinguer volse
un grave incendio, fu che gli la tolse.Nata pochi dì inanzi era una gara
tra il conte Orlando e il suo cugin Rinaldo,
che entrambi avean per la bellezza rara
d'amoroso disio l'animo caldo.

```
  Carlo, che non avea tal lite cara,
  che gli rendea l'aiuto lor men saldo,
  questa donzella, che la causa n'era,
  tolse, e diè in mano al duca di Bavera;in premio promettendola a quel d'essi,
  ch'in quel conflitto, in quella gran giornata,
  degl'infideli più copia uccidessi,
  e di sua man prestasse opra più grata.
  Contrari ai voti poi furo i successi;
  ch'in fuga andò la gente battezzata,
  e con molti altri fu 'l duca prigione,
  e restò abbandonato il padiglione. ";
```

```
lowerItalianText = ToLowerCase[italianText]; (* Convert text to lowercase *)
```

```
charcode = ToCharacterCode[lowerItalianText];
(* Get a numerical character code for each letter in the string *)
```

```
charentp = Table[{Length[Position[charcode, i]], FromCharacterCode[i]}, {i, 96, 122}]
 (* make a table with the counts for each character *)
```

```
{{0, `}, {260, a}, {14, b}, {97, c}, {88, d}, {220, e}, {27, f}, {47, g}, {34, h},
 {215, i}, {0, j}, {0, k}, {121, l}, {56, m}, {146, n}, {204, o}, {62, p}, {14, q},
 {168, r}, {94, s}, {122, t}, {64, u}, {39, v}, {0, w}, {0, x}, {0, y}, {8, z}}
```

```
charentp[[1]] = {Length[Position[charcode, 32]], "space"}
 (* Also, get the character count for spaces *)
```

```
{550, space}
```

```
TableForm[charentp] (* Put this frequency in a table to visualize easier *)
```

```
550    space
260    a
14     b
97     c
88     d
220    e
27     f
47     g
34     h
215    i
0      j
0      k
121    l
56     m
146    n
204    o
62     p
14     q
168    r
94     s
122    t
64     u
39     v
0      w
0      x
0      y
8      z
```

```
sortcharentp = Reverse[Sort[charentp]]
 (* Sort by most common to least common frequencies *)
```

```
{{550, space}, {260, a}, {220, e}, {215, i}, {204, o}, {168, r}, {146, n}, {122, t},
 {121, l}, {97, c}, {94, s}, {88, d}, {64, u}, {62, p}, {56, m}, {47, g}, {39, v},
 {34, h}, {27, f}, {14, q}, {14, b}, {8, z}, {0, y}, {0, x}, {0, w}, {0, k}, {0, j}}
```

```
sumchars = Sum[sortcharentp[[i, 1]], {i, Length[sortcharentp]}]
 (* Get the total character count in the text *)
```

```
2650
```

```
charfreq = N[Table[{sortcharentp[[i, 1]] / sumchars, sortcharentp[[i, 2]]},
   {i, 1, Length[sortcharentp]}]]
 (* Get the frequency for each character in the text *)
```

```
{{0.207547, space}, {0.0981132, a}, {0.0830189, e}, {0.0811321, i}, {0.0769811, o},
 {0.0633962, r}, {0.0550943, n}, {0.0460377, t}, {0.0456604, l}, {0.0366038, c},
 {0.0354717, s}, {0.0332075, d}, {0.0241509, u}, {0.0233962, p}, {0.0211321, m},
 {0.0177358, g}, {0.014717, v}, {0.0128302, h}, {0.0101887, f}, {0.00528302, q},
 {0.00528302, b}, {0.00301887, z}, {0., y}, {0., x}, {0., w}, {0., k}, {0., j}}
```

```
TableForm[charfreq] (* Put this frequency in a table to visualize easier *)
```

```
0.207547      space
0.0981132     a
0.0830189     e
0.0811321     i
0.0769811     o
0.0633962     r
0.0550943     n
0.0460377     t
0.0456604     l
0.0366038     c
0.0354717     s
0.0332075     d
0.0241509     u
0.0233962     p
0.0211321     m
0.0177358     g
0.014717      v
0.0128302     h
0.0101887     f
0.00528302    q
0.00528302    b
0.00301887    z
0.            y
0.            x
0.            w
0.            k
0.            j
```

```
entplang = 0;
(* Loop through the entire table, using the entropy formula,
calculate the entropy of the text given our frequencies *)
For[i = 1, i <= 27, i++,
            If[charfreq[[i, 1]] == 0, entplang,
                entplang = entplang – charfreq[[i, 1]] * Log[charfreq[[i, 1]]]
            ]
        ]
```

```
(* This represents the average minimum number of bits needed to
 encode a string of symbols, based on the frequency of the symbols .*)
entplang
```

```
2.70014
```

```
(* Take this entropy, multiply it by the size of our text, and that gives the amount
 of bits required to optimally encode the string *)
optimalItalianEncoding = entplang * sumchars
```

```
7155.37
```

```
(* Analysis on languages *)
(* English: 2.83089, French: 2.78246, Italian: 2.70014 *)
(* In general, all three of these languages have a very similar entropy. Analyzing even
 longer texts can yeild more accurate results due to the law of large numbers *)

(* Since English is a Germanic language, and Italian and French are Romance languages,
I expected English' entropy to be very different from French and Italian *)

(* The space, letter e, and a,
were all top 3 for these languages for relative frequency,
and were similar frequencies among
 the three languages. Since the top 3 were similar across all three languages,
it's not suprising
 that the entropies of the languages are relatively the same *)

(* From this test, I am concluding that there isn't a significant difference between the
 language's entropies. *)
```