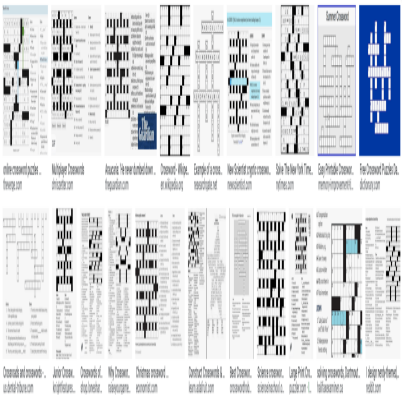


How to make up words

Description

ETAOINS are the 7 most commonly used letters in the English language. See post: Counting letters. Perhaps we could communicate with just those seven letters.

https://unscramblex.com is a website that provides all anagrams of up to 15 characters. There are 178 anagrams of ETAOINS. If you want include words where letters occur together (EE, TT, etc), or words in which letters appear more than once then you need to include letters in the initial character string more than once. EETTAAOOIINNSS gives 690 words, ranging from *on* to *assentation*. Thatâ??s as good a corpus as any to derive a table of probabilities suitable for analysis (or at least illustration) of a Markov chain.

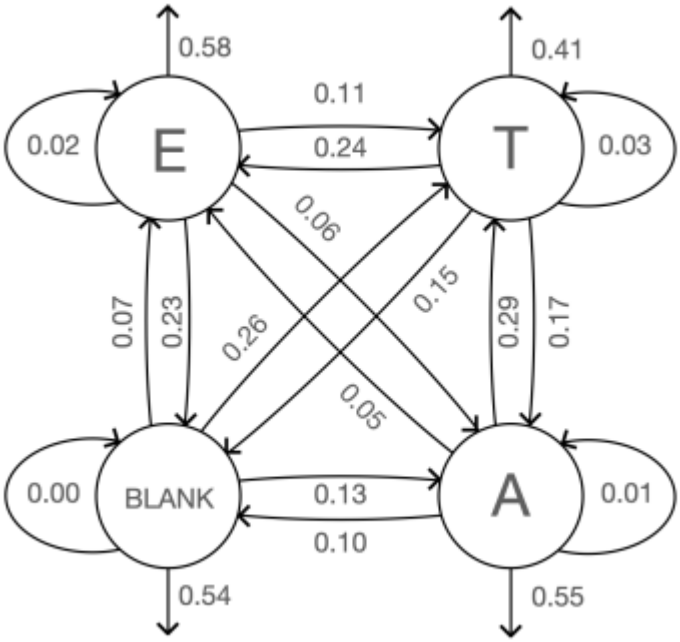
The letter E followed by the letter E occurs in ten words in the corpus, E followed by T occurs 55 times. E at the end of a word occurs 110 times. I calculated these numbers by feeding the corpus of 690 words into a spreadsheet. I included a blank space to signal the start or the end of a word

	E	T	A	O	I	N	S	BLANK	Total
E	10	55	31	16	15	104	146	110	487
T	149	19	107	88	82	3	71	95	614
A	21	111	4	4	20	112	71	39	382
O	8	45	18	28	13	123	43	19	297
I	34	35	24	29	0	99	42	19	282
N	107	70	75	60	40	40	84	69	545
S	104	110	46	36	48	10	25	277	656
BLANK	47	165	79	39	37	89	172	0	628

From that data, the spreadsheet can calculate probabilities. The probability that an E is followed by another E is 0.02, that itâ??s followed by a T is 0.11, or that it comes at the end of a word is 0.23.

	E	T	A	O	I	N	S	BLANK
E	0.02	0.11	0.06	0.03	0.03	0.21	0.30	0.23
T	0.24	0.03	0.17	0.14	0.13	0.00	0.12	0.15
A	0.05	0.29	0.01	0.01	0.05	0.29	0.19	0.10
O	0.03	0.15	0.06	0.09	0.04	0.41	0.14	0.06
I	0.12	0.12	0.09	0.10	0.00	0.35	0.15	0.07
N	0.20	0.13	0.14	0.11	0.07	0.07	0.15	0.13
S	0.16	0.17	0.07	0.05	0.07	0.02	0.04	0.42

A Markov network graph of the above tabulated data would have 8 nodes connected by 64 directed arrows. Thatâ??s too cumbersome to draw. Hereâ??s the graph for just four nodes E, T, A and a blank space. Lines exiting a node have to add up to a probability of 1.0.



The matrix and network would be much larger for all the letters of the alphabet. Unscramblex.com is useful for generating words in Scrabble and crossword puzzles. The corpus of words I used is pretty high-brow. A corpus of words used in everyday speech and writing would show a different matrix. With further refinements, the matrix provides a signature of someoneâ??s writing style and can be used to identify style and authorship of the kind used in psychometric profiling, e.g. the Cambridge University

psychometric profiling tool.

The method also provides a step on the way to decrypting a message in a simple substitution cipher. Knowing what letters are likely to follow in sequence reduces the search space for the original plain text message.

With greater sophistication, the Markov model can also generate words and sentences. My crude implementation here generates existing and new words like TE NESTIN AS TESATESA NOO and SESESES, but also AAAAAAAAA!

Bibliography

- Hayes, Brian. 2013. First Links in the Markov Chain. *American Scientist*, (102) 2.
- Hayes, Brian. 2013. First Links in the Markov Chain. *American Scientist*. Available online: https://www.americanscientist.org/article/first-links-in-the-markov-chain (accessed 8 December 2021).
- Lee, Dar-Shyang. 2002. Substitution Deciphering Based on HMMs with Applications to Compressed Document Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (24) 12, 1661-1666.
- Vobbilisetty, Rohit, Fabio Di Troia, Richard M. Low, Corrado Aaron Visaggio, and Mark Stamp. 2017. Classic cryptanalysis using hidden Markov models. *Cryptologia*, (41) 1, 1-28.

Category

1. Artificial Intelligence

Tags

1. cryptography

Date Created
December 11, 2021
Author
rcoyne99