**Mixing Letters**

**TEXTEN1.txt**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Probability of mixing letters inside a word | | Min | Max | Avg |
| 0% | Entropy | 5.2874 | 5.2874 | 5.2874 |
| Perplexity | 39.0546 | 39.0546 | 39.0546 |
| 10% | Entropy | 4.7252 | 4.7377 | 4.7308 |
| Perplexity | 26.4494 | 26.6810 | 26.5539 |
| 5% | Entropy | 5.0530 | 5.0614 | 5.0568 |
| Perplexity | 33.1982 | 33.3907 | 33.2856 |
| 1% | Entropy | 5.2487 | 5.2521 | 5.25 |
| Perplexity | 38.0201 | 38.1101 | 38.0536 |
| 0.1% | Entropy | 5.2828 | 5.2847 | 5.2839 |
| Perplexity | 38.9305 | 38.9819 | 38.9590 |
| 0.01% | Entropy | 5.2867 | 5.2872 | 5.2870 |
| Perplexity | 39.0353 | 39.0493 | 39.0429 |

**TEXTCZ1.txt**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Probability of mixing letters inside a word | | Min | Max | Avg |
| 0% | Entropy | 4.7478 | 4.7478 | 4.7478 |
| Perplexity | 26.8681 | 26.8681 | 26.8681 |
| 10% | Entropy | 3.9947 | 4.0121 | 4.0054 |
| Perplexity | 15.9413 | 16.1343 | 16.0599 |
| 5% | Entropy | 4.3314 | 4.3399 | 4.3360 |
| Perplexity | 20.1323 | 20.2505 | 20.1968 |
| 1% | Entropy | 4.6553 | 4.6591 | 4.6578 |
| Perplexity | 25.1988 | 25.2655 | 25.2421 |
| 0.1% | Entropy | 4.7379 | 4.7398 | 4.7387 |
| Perplexity | 26.6839 | 26.7190 | 26.6994 |
| 0.01% | Entropy | 4.7467 | 4.7472 | 4.7469 |
| Perplexity | 26.8465 | 26.8564 | 26.8518 |

**Observations**

We can easily see that the entropy decreases when the probability of changing a letter increase.

When you change a letter in a word, the probability that the same word inside the file will have the same change, in the same position, with the same new letter is particularly low (we will call this event )

We can actually compute this probability. We can pick a word in our file. We assign .  
Given the number of letters in our alphabet, the probability of changing a letter inside the , the number of words in the file and the number of occurences of in the file.

Let us consider the file TEXTEN1.txt.

Empirically, we find ( contains not only letters uppercase/lowercase but also numbers and special symbols).

is the highest occurency of a word inside the file (we consider the probability in the best case), which is related to the character .

and

The probability .

We can see, therefore, that the probability of such a change repeats itself somewhere else in the file is nearly 0 (and this is one of the best cases!).

As this probability is nearly zero, then we can easily note that this change will produce a lot of pairs (w1, w2) that will happen only once in the file.

We can then assume that for many strings.

As we evaluate through the following formula,

We can see that having many will bring to 0 many terms of this sum as .

In conclusion, we can say that the entropy dicreases by incrementing the probability because of the unlikelyhood of having some couples of tokens repeated.

For sure, the result would have been way more different by adding some smoothing tecnique to the data available.

**Mixing words**

**TEXTEN1.txt**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Probability of mixing words inside the file | | Min | Max | Avg |
| 0% | Entropy | 5.2874 | 5.2874 | 5.2874 |
| Perplexity | 39.0546 | 39.0546 | 39.0546 |
| 10% | Entropy | 5.9960 | 6.0069 | 6.0005 |
| Perplexity | 63.8267 | 64.3097 | 64.025 |
| 5% | Entropy | 5.7093 | 5.7226 | 5.7164 |
| Perplexity | 52.3205 | 52.8053 | 52.5808 |
| 1% | Entropy | 5.3889 | 5.4003 | 5.3932 |
| Perplexity | 41.9022 | 42.2335 | 42.0257 |
| 0.1% | Entropy | 5.2980 | 5.3003 | 5.2987 |
| Perplexity | 39.3408 | 39.4055 | 39.3637 |
| 0.01% | Entropy | 5.2882 | 5.2892 | 5.2887 |
| Perplexity | 39.0765 | 39.1033 | 39.0912 |

**TEXTCZ1.txt**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Probability of mixing words inside the file | | Min | Max | Avg |
| 0% | Entropy | 4.7478 | 4.7478 | 4.7478 |
| Perplexity | 26.8681 | 26.8681 | 26.8681 |
| 10% | Entropy | 5.0750 | 5.0847 | 5.0805 |
| Perplexity | 33.7085 | 33.9358 | 33.8368 |
| 5% | Entropy | 4.9480 | 4.9513 | 4.9497 |
| Perplexity | 30.8662 | 30.9382 | 30.9029 |
| 1% | Entropy | 4.7929 | 4.7995 | 4.7964 |
| Perplexity | 27.7207 | 27.8483 | 27.7877 |
| 0.1% | Entropy | 4.7526 | 4.7537 | 4.7532 |
| Perplexity | 26.9571 | 26.9771 | 26.9682 |
| 0.01% | Entropy | 4.7481 | 4.7485 | 4.7483 |
| Perplexity | 26.8726 | 26.8814 | 26.8772 |

**Observations**

In this case, instead, we can observe that the entropy rises by incrementing the probability of changing the position of a word inside the text.

This may be related to the fact that by switching the position of words inside the text, the structure of the words is maintained, so we don’t have the variable related to the choice of the letter and its position.

This may result in more pairs that have an occurrence greater than one, that results in . This way, we are actually adding up more terms together (and getting therefore a larger entropy).