Assignment 2

# 

1. See P1

$ ./P1.exe ../../Texts/Novels/DostoevskyKaramazov.txt 0.5

the, 15173

and, 11436

to, 9528

he, 8149

i, 7995

of, 7281

a, 6804

you, 6305

that, 6177

it, 5852

in, 5514

was, 4776

his, 4345

s, 3480

for, 3471

him, 3377

but, 3267

at, 3004

with, 2986

not, 2907

had, 2854

is, 2433

on, 2334

me, 2292

all, 2196

as, 2132

t, 2108

have, 1983

her, 1885

be, 1874

she, 1784

what, 1743

so, 1705

my, 1691

one, 1603

from, 1486

there, 1357

they, 1310

alyosha, 1243

this, 1241

are, 1222

by, 1205

no, 1205

will, 1168

if, 1162

been, 1105

would, 1102

up, 1091

your, 1064

only, 1055

were, 1000

said, 995

them, 987

out, 959

an, 943

now, 927

mitya, 917

man, 908

who, 898

do, 866

50.0261 %

60 words are necessary for understanding 50% of the text, representing 0.5% of the language.

$ ./P1.exe ../../Texts/Novels/DrSeuss.txt 0.5

the, 89

and, 66

i, 62

of, 45

a, 41

king, 32

that, 31

he, 23

to, 22

yertle, 19

all, 19

in, 19

turtle, 18

m, 16

turtles, 15

you, 15

one, 15

s, 12

they, 12

my, 12

was, 11

up, 11

them, 11

throne, 10

it, 10

going, 10

his, 10

but, 9

on, 9

see, 9

t, 9

down, 8

here, 8

south, 8

zax, 8

ll, 8

50.5176 %

36 words are necessary for understanding 50% of the text, representing 8.53081% of the language. (Some of these are not real words.)

# 

1. See P2.cpp

./P2.exe …DostoevskyPart1.txt …DostoevskyPart2.txt N 1

For N-Grams of size: 18 --> 100%

For N-Grams of size: 18 --> 100%

For N-Grams of size: 17 --> 99.9994%

repulsion that s what i m afraid of that s what may be too much for me

For N-Grams of size: 16 --> 99.9989%

For N-Grams of size: 15 --> 99.9983%

For N-Grams of size: 14 --> 99.9978%

For N-Grams of size: 13 --> 99.9972%

For N-Grams of size: 12 --> 99.9961%

For N-Grams of size: 11 --> 99.995%

For N-Grams of size: 10 --> 99.9922%

For N-Grams of size: 9 --> 99.9889%

For N-Grams of size: 8 --> 99.9805%

For N-Grams of size: 7 --> 99.9504%

For N-Grams of size: 6 --> 99.837%

For N-Grams of size: 5 --> 99.2807%

For N-Grams of size: 4 --> 96.7292%

For N-Grams of size: 3 --> 87.5173%

For N-Grams of size: 2 --> 68.6534%

For N-Grams of size: 1 --> 33.1106%

No common N-Grams: N > 17

Longest common N-Gram:

* repulsion that s what i m afraid of that s what may be too much for me

./P2.exe …Dickens.txt …KafkaTrial.txt N 1

For N-Grams of size: 8 --> 100%

For N-Grams of size: 7 --> 99.9977%

in the middle of the table and

there is no such thing as a

For N-Grams of size: 6 --> 99.9768%

For N-Grams of size: 5 --> 99.8854%

For N-Grams of size: 4 --> 99.1954%

For N-Grams of size: 3 --> 94.5332%

For N-Grams of size: 2 --> 77.4582%

For N-Grams of size: 1 --> 32.8801%

No common N-Grams: N > 7

Longest common N-Grams:

* In the middle of the table and
* there is no such thing as a

./P2.exe …MarxEngelsManifest.txt …SmithWealthNations.txt N 1

For N-Grams of size: 7 --> 100%

For N-Grams of size: 6 --> 99.9984%

of nature and of reason the

is the same as that of

to keep up the rate of

in order to keep up the

of a man s own labour

from them what they have not

For N-Grams of size: 5 --> 99.987%

For N-Grams of size: 4 --> 99.9199%

For N-Grams of size: 3 --> 99.5337%

For N-Grams of size: 2 --> 97.468%

For N-Grams of size: 1 --> 84.1851%

No common N-Grams: N > 6

Longest common N-Grams:

* of nature and of reason the
* is the same as that of
* to keep up the rate of
* in order to keep up the
* of a man s own labour
* from them what they have not

1. As expected, comparing both the Dostoevsky texts would yield the longest N-Grams, attributable to the fact that Dostoevsky likely used a very similar writing style across the same novel. As the works and their authors began to differ, there is a noticeable decrease in the length of common N-Grams, as well as with the number of common N-Grams. Barring the works of Marx and Smith, they seemed to converge as the length of N-Grams approached 1, likely because much of written text can be largely represented by a small proportion of the English language.
2. See P3.cpp

N = 1: k that had wall parts t but the the be into <END>

N = 2: not and opened the front of the judge who would i didn t understand but she made doors were <END>

N = 3: but now just tell me now <END>

N = 4: at himself at his own na vety in court matters <END>

N = 5: that for the time being questioning and observing the accused are much more important than anything written <END>

N = 6: for his own business <END>

The length of sentences seem to be consistently increasing except for the cases of N=3 and N=6. For N=1, there is almost no meaning that can be derived from the sentence, whereas for N=4 and 5, there is a lot more structure. N=6, for some reason, although it makes sense, is very short.

1. N = 3: it creates capital i <END>

Interestingly, the Manifest case created a very short sentence of only four words, which is very similar to that of the Kafka case.

1. See P4.cpp
2. (In order of the assignment)

-183.035

-200.678

-111.105

-187.372

1. See P6.exe

|  |  |  |
| --- | --- | --- |
| Command | Error Rate (percent) | Confusion Matrix |
| P6 1 0.0000 50 | 7.41036 | 135 3 0 0 0 0  8 347 4 4 2 1  4 0 109 6 6 0  5 3 4 214 11 0  0 0 7 9 244 0  14 2 0 0 0 113 |
| P6 2 0.0000 50 | 40.6375 | 134 3 0 0 0 1  40 325 1 0 0 0  107 2 16 0 0 0  103 0 0 134 0 0  168 3 0 2 87 0  78 2 0 0 0 49 |
| P6 3 0.0000 50 | 84.7012 | 138 0 0 0 0 0  341 25 0 0 0 0  125 0 0 0 0 0  218 0 0 19 0 0  249 1 0 0 10 0  128 1 0 0 0 0 |

|  |  |  |
| --- | --- | --- |
| Command | Error Rate (percent) | Confusion Matrix |
| P6 1 0.0500 50 | 5.9761 | 135 3 0 0 0 0  8 347 4 4 2 1  0 0 113 6 6 0  0 3 4 219 11 0  0 0 7 8 245 0  5 3 0 0 0 121 |
| P6 2 0.0500 50 | 1.03586 | 136 1 0 0 0 1  0 365 0 0 1 0  0 0 123 0 2 0  0 0 0 236 1 0  0 0 1 2 257 0  4 0 0 0 0 125 |
| P6 3 0.0500 50 | 43.8247 | 59 75 0 4 0 0  0 366 0 0 0 0  0 123 0 2 0 0  0 20 0 217 0 0  0 187 0 15 58 0  0 116 0 8 0 5 |

|  |  |  |
| --- | --- | --- |
| Command | Error Rate (percent) | Confusion Matrix |
| P6 3 0.0500 50 | 43.8247 | 59 75 0 4 0 0  0 366 0 0 0 0  0 123 0 2 0 0  0 20 0 217 0 0  0 187 0 15 58 0  0 116 0 8 0 5 |
| P6 3 0.0050 50 | 10.8367 | 122 11 1 2 1 1  0 363 0 1 0 2  0 53 56 13 3 0  0 0 0 237 0 0  0 9 0 10 241 0  3 22 0 4 0 100 |
| P6 3 0.0005 50 | 3.34661 | 127 4 1 0 1 5  0 361 3 0 0 2  0 6 113 4 2 0  0 0 0 237 0 0  0 1 0 2 257 0  4 5 1 1 0 118 |

1. There are a few large trends here: as the length of N-Grams increases, so does the error rate, and as delta increases, so does the error rate.

The former could be because, as the N-Gram length increases, it is less likely to appear in the training text. With that being the case, the program cannot identify what language the text belongs to, and is thus less accurate.

The latter could indicate that higher values of delta are adding a lot of noise, causing the other languages to be overweighed, and thus causing the program to be less accurate. However, that c)’s second configuration shows a case where the delta and N-Gram length is balanced, and thus able to predict the language with great accuracy.



|  |  |  |
| --- | --- | --- |
| Command | Error Rate (percent) | Confusion Matrix |
| P6 2 0.0500 10 | 21.0434 | 538 49 7 16 18 63  53 1448 83 87 113 50  13 35 445 57 64 14  14 33 32 1027 69 12  16 29 44 164 1031 17  86 32 13 17 23 475 |
| P6 2 0.0500 50 | 1.03586 | 136 1 0 0 0 1  0 365 0 0 1 0  0 0 123 0 2 0  0 0 0 236 1 0  0 0 1 2 257 0  4 0 0 0 0 125 |
| P6 2 0.0500 100 | 0.159744 | 69 0 0 0 0 0  0 183 0 0 0 0  0 0 62 0 0 0  0 0 0 118 0 0  0 0 0 0 130 0  1 0 0 0 0 63 |



|  |  |  |
| --- | --- | --- |
| Command | Error Rate (percent) | Confusion Matrix |
| P6 1 0.0000 50 | 17.2443 | 87 4 1 0 0 34  2 326 10 2 0 13  2 0 80 15 18 0  0 4 17 187 11 1  0 0 36 16 200 0  14 4 0 0 0 99 |
| P6 2 0.0000 50 | 4.81826 | 111 1 0 0 0 14  4 347 1 1 0 0  1 6 102 1 4 1  1 0 0 218 1 0  1 1 2 1 247 0  14 2 0 0 0 101 |
| P6 3 0.0000 50 | 75.8242 | 122 2 0 0 0 2  302 51 0 0 0 0  99 5 10 0 1 0  148 2 0 70 0 0  232 1 0 0 19 0  103 0 0 0 0 14 |
| P6 1 0.0500 50 | 17.2443 | 87 4 1 0 0 34  2 326 10 2 0 13  2 0 80 15 18 0  0 4 17 187 11 1  0 0 36 16 200 0  14 4 0 0 0 99 |
| P6 2 0.0500 50 | 2.11327 | 113 1 0 0 0 12  0 351 1 0 0 1  0 0 113 1 1 0  0 0 0 219 1 0  0 0 1 1 250 0  5 0 0 0 0 112 |
| P6 3 0.0500 50 | 1.0989 | 122 0 0 1 0 3  0 351 0 1 1 0  0 0 114 1 0 0  0 0 0 220 0 0  0 0 0 0 252 0  4 0 0 1 1 111 |
| P6 3 0.0500 50 | 1.0989 | 122 0 0 1 0 3  0 351 0 1 1 0  0 0 114 1 0 0  0 0 0 220 0 0  0 0 0 0 252 0  4 0 0 1 1 111 |
| P6 3 0.0050 50 | 1.69062 | 120 1 0 1 0 4  0 352 0 0 1 0  0 3 107 3 2 0  0 0 0 220 0 0  0 0 0 0 252 0  4 0 0 0 1 112 |
| P6 3 0.0005 50 | 2.36686 | 116 3 0 1 0 6  0 353 0 0 0 0  0 4 105 4 2 0  0 0 0 220 0 0  0 0 0 1 251 0  7 0 0 0 0 110 |