PDF English N-gram Analysis

EN-PDF-v1 | Book pp. 282-294 | Title, text, captions and notes

Unigram | N = 5378 | Unique = 1547

Combination	Count	Proportion
the	454	8.44%
of	280	5.21%
and	183	3.40%
a	133	2.47%
in	128	2.38%
to	66	1.23%
that	60	1.12%
as	58	1.08%
library	54	1.00%
for	46	0.86%
future	45	0.84%
is	45	0.84%
it	42	0.78%
by	40	0.74%
an	39	0.73%

Bigram | N = 5153 | Unique = 4015

Combination	Count	Proportion
of the	98	1.90%
in the	37	0.72%
the library	28	0.54%
the project	22	0.43%
and the	20	0.39%
future library	18	0.35%
the future	18	0.35%
for the	15	0.29%
as a	14	0.27%
in a	14	0.27%
on the	14	0.27%
the digital	14	0.27%
to the	12	0.23%
space of	11	0.21%
with the	10	0.19%

Top 15 shown. Complete frequency tables are included in the data package.

PDF English Trigram Analysis

EN-PDF-v1 | Trigrams do not cross sentence or distinct text-block boundaries.

Trigram | N = 4936 | Unique = 4649

Combination	Count	Proportion
of the library	15	0.30%
of the project	8	0.16%
space of the	8	0.16%
it doesn't exist	7	0.14%
of the book	7	0.14%
of the digital	6	0.12%
the library and	5	0.10%
the silence of	5	0.10%
a hundred years	4	0.08%
a kind of	4	0.08%
aesthetic of bookishness	4	0.08%
if it doesn't	4	0.08%
in a hundred	4	0.08%
in which the	4	0.08%
of the future	4	0.08%

Scope: supplied excerpt only, not the complete edited volume.

Accepted source: clear book pages 282-294; overlapping final right half excluded.

Printed page numbers and rotated running headers are excluded.

Native-font ligatures and line-wrap splits were checked and corrected.

Short sentence segments are handled with max(length - n + 1, 0).

Original Chinese statistics remain unchanged: 724 / 699 / 674 windows.

PDF English Conditional Probability

English-trained reference | EOS included in context denominators | Unsmoothed

First-order: P(B | A)

Condition	Next (count/context)	Probability
of	the (98/280)	35.00%
in	the (37/128)	28.91%
the	library (28/454)	6.17%
the	project (22/454)	4.85%
and	the (20/183)	10.93%
future	library (18/45)	40.00%
the	future (18/454)	3.96%
for	the (15/46)	32.61%
as	a (14/58)	24.14%
in	a (14/128)	10.94%
on	the (14/33)	42.42%
the	digital (14/454)	3.08%

Second-order: P(C | A,B)

Condition	Next (count/context)	Probability
of the	library (15/98)	15.31%
of the	project (8/98)	8.16%
space of	the (8/11)	72.73%
it doesn't	exist (7/7)	100.00%
of the	book (7/98)	7.14%
of the	digital (6/98)	6.12%
the library	and (5/28)	17.86%
the silence	of (5/5)	100.00%
a hundred	years (4/4)	100.00%
a kind	of (4/4)	100.00%
accessed april	<EOS> (4/7)	57.14%
aesthetic of	bookishness (4/4)	100.00%

C2 remains Chinese-trained. These English probabilities are a separate reference.