

N-gram Worksheet

Corpus:

The dog ran.
The dog ate.
The dog ran the zoo.
The cat ate the dog!
The cat ran the zoo.

(Simple) Tokenized Corpus:

['the', 'dog', 'ran', '.', 'the',
'dog', 'ate', '.', 'the', 'dog',
'ran', 'the', 'zoo', '.', 'the',
'cat', 'ate', 'the', 'dog', '!',
'the', 'cat', 'ran', 'the',
'zoo', '.']

Vocabulary:

['!',
'.',
'ate',
'cat',
'dog',
'ran',
'the',
'zoo']

Compute the unigram probabilities given the tokenized corpus above

x_t	!	.	ate	cat	dog	ran	the	zoo
$p(x_t)$	1/26	4/26	2/26	2/26	3/26	3/26	8/26	2/26

Compute the bigram probabilities given the tokenized corpus above

$p(x_{t+1} x_t)$	x_{t+1}	x_t
0/1	!	!
0/1	.	!
0/1	ate	!
0/1	cat	!
0/1	dog	!
0/1	ran	!
1/1	the	!
0/1	zoo	!

$p(x_{t+1} x_t)$	x_{t+1}	x_t
0/3	!	.
0/3	.	.
0/3	ate	.
0/3	cat	.
0/3	dog	.
0/3	ran	.
3/3	the	.
0/3	zoo	.

$p(x_{t+1} x_t)$	x_{t+1}	x_t
0/2	!	ate
1/2	.	ate
0/2	ate	ate
0/2	cat	ate
0/2	dog	ate
0/2	ran	ate
1/2	the	ate
0/2	zoo	ate

$p(x_{t+1} x_t)$	x_{t+1}	x_t
0/2	!	cat
0/2	.	cat
1/2	ate	cat
0/2	cat	cat
0/2	dog	cat
1/2	ran	cat
0/2	the	cat
0/2	zoo	cat

$p(x_{t+1} x_t)$	x_{t+1}	x_t
0/3	!	dog
0/3	.	dog
0/3	ate	dog
0/3	cat	dog
0/3	dog	dog
0/3	ran	dog
0/3	the	dog
0/3	zoo	dog

$p(x_{t+1} x_t)$	x_{t+1}	x_t
0/3	!	ran
1/3	.	ran
0/3	ate	ran
0/3	cat	ran
0/3	dog	ran
0/3	ran	ran
2/3	the	ran
0/3	zoo	ran

$p(x_{t+1} x_t)$	x_{t+1}	x_t
0/8	!	the
0/8	.	the
0/8	ate	the
2/8	cat	the
4/8	dog	the
0/8	ran	the
0/8	the	the
2/8	zoo	the

$p(x_{t+1} x_t)$	x_{t+1}	x_t
0/2	!	zoo
2/2	.	zoo
0/2	ate	zoo
0/2	cat	zoo
0/2	dog	zoo
0/2	ran	zoo
0/2	the	zoo
0/2	zoo	zoo