Harshavardhan Reddy Mallannagari

11603389

NLP - HW2

k=0 [redacted] alphabets are [A-Z, a-z]

Initial	Corpus
2	B,a,k,e,s,-
2	b,e,t,t,y,-
2	L,O,U,-
2	b,O,U,g,h,t,-
2	S,O,m,e,-
2	b,i,t,t,e,r,-
2	b,u,t,t,e,r,-
1	b,u,t-
1	i,t,-
1	m,a,d,e,-
2	h,e,r-
2	b,a,t,t,e,r,-
1	S,q-
2	b,e,t,t,e,r,-
1	t,o,-
1	m,a,k,e-

Vocabulary	
[redacted]-	26
a|A	6
b|B	15
d|D	1
e|E	20
g|G	2
h|H	4
i|I	3
k|K	3
l|L	2
m|M	4
o|O	8
r|R	12
s|S	3
t|T	25
u|U	7

Iteration - 0

Initial values

k = 1

occurences - 12

Iteration 1

combine e,r → er

2	B, a, k, er, -
2	b, e, t, t, y, -
2	d, o, u, -
2	b, o, u, g, h, t, -
2	s, o, m, e, -
2	b, i, t, t, er, -
2	b, u, t, t, er, -
1	b, u, t, -
1	i, t, -
1	m, a, d, e, -
2	h, er, -
2	b, a, t, t, er, -
1	s, o, -
2	b, e, t, t, er, -
1	t, o, -
1	m, a, k, e, -

Vocabulary

-	26
a	6
b	15
d	1
e	8
g	2
h	4
i	3
k	3
l	2
m	4
o	8
r	0
s	3
t	25
u	7
er	12

k = 2

occurrences (12)

er, - → er-

Combine [redacted]

Iteration - 2

Count	Corpus
2	b, a, k, er-
2	b, e, t, t, y, -
2	d, o, u, [redacted] -
2	b, o, u, g, h, t, -
2	s, o, m, e, -
2	b, i, t, t, er-
2	b, u, t, t, er-
1	b, u, t, -
1	i, t, -
1	m, a d, e, -
2	h, er-
2	b a, t, t, er-
1	s, o, -
2	b, e, t, t, er-
1	to, -
1	m, a, k, e, -

Vocabulary

-	14
a	6
b	15
d	1
e	[redacted]8
g	2
h	4
i	3
k	3
l	2
m	4
o	8
s	3
t	[redacted]25
u	7
er-	12
er	0

k=3

occurrences (10)

Iteration-3

combine t,t → tt

count	corpus
2	b, a, k, er_
2	b, e, tt, y, _
2	d, o, u, _
2	b, o, u, g, h, t, _
2	s, o, m, e, _
2	b, i, tt, er_
2	b, u, tt, er_
1	b, u, t, _
1	i, t, _
1	m, a, d, e
2	[illegible] h, er_
2	b, a, tt, er_
1	s, o, _
2	b, e, tt, er_
1	t, o, _
1	m, a, k, e

vocabulary.

_	14
a	6
b	15
d	1
e	8
g	2
h	4
i	3
k	3
l	2
m	4
o	8
s	3
t	5
u	7
er_	12
tt	10

$k = 4$

occurrences (8)

Iteration - 1

Combine tt, er_ → tter_

Count	Corpus
2	b, a, k, er_
2	b, e, tt, y, _
2	d, o, u, _
2	b, au, g, h, t, _
2	s, o, m, e_
2	b, i, tter_
2	b, u, tter_
1	b, u, t, _
1	i, t, _
1	m, a, d, e, _
2	h, er_
2	b, a, tter_
1	s, o, _
2	b, e, tter_
1	t, o, _
1	m, a, k, e, _

Vocabulary

_	14
a	6
b	15
d	1
e	8
g	2
h	4
i	3
k	3
l	2
m	4
o	8
s	3
t	5
u	7
er_	4
tt	2
tter_	8

K=5 occurences (4)

Iteration 5

combine b,e → be

Count	Corpus
2	b,a, k, er_
2	~~be~~ be, tt, y, -
2	d, o, u, -
2	b, o, u, g, h, t, -
2	s, o, m, e, -
2	b, i, tter_
2	b, u, tter_
1	b, u, t, -
1	i, t, -
1	m, a, d, e, -
2	h, er_ -
2	b, a, tter_
1	s, o, -
2	be, tter_ -
1	t, o, -
1	m, a, k, e, -

Vocabulary

-	14
a	6
b	11
d	1
e	4
g	2
h	4
i	3
k	3
l	2
m	4
o	8
s	3
t	5
u	7
er_	4
tt	2
tter_	8
be	4

K=6

occurences (4)

Iteration - 6

combine o,u → ou

Count	Corpus
2	b,a,k,er_
2	be,tt,y,_
2	d, ou,_
2	b, ou,g,h,t,_
2	s,o,m,e,_
2	b,i, tter_
2	b,u,tter_
1	b,u,t,_
1	i,t,_
1	m,a,d,e,_
2	h, er_
2	b,a, tter_
1	sp,_
2	be, tter_
1	t,o,_
1	m,a,k,e,_

Vocabulary

_	14
a	6
b	11
d	1
e	4
g	2
h	4
i	3
k	3
d	2
m	4
o	4
s	3
t	5
u	3
er_	4
tt	2
tter_	8
be	4
ou	4

k=7 occurences (4)

Iteration - 7

Combine b,a → ba

Count	Corpus
2	ba, k, ers-
2	be, tt, y,-
2	d, ou, -
2	b, ou, g, h, t, -
2	s, o, m, e, -
2	b, i, tter-
2	b, u, tter-
1	b, u, t,-
1	i, t, -
1	m, a, d, e,-
2	h, ier-
2	ba, tter-
1	s, o, -
2	be, tter-
1	t, o, -
1	m, a, k, e, -

Vocabulary - 14

a	2
b	7
d	1
e	4
g	2
h	4
i	3
k	3
l	2
m	4
o	4
s	3
t	5
u	3
ers-	4
tt	2
tter-	8
be	4
ou	4
ba	4

Considering a/A → a ie, [A-Z, a-z] ∈ Alphabets

case insensitive.

Iteration - 0

All the Initial values that are present in the given text.

Iteration - 1

In this iteration, combining e, r → er

count - 12 [(baker,2), (biter,2) (butter,2) (better,2) (batter,2) (her,2)]

Iteration - 2

In this iteration combining. e, r, - → er -

count - 12

Iteration - 3

In the third iteration, we are combining t, t → tt

which uses 20 t's but the total count is 10

betty - 2

biter - 2

better - 2

butter - 2

batter - 2

Iteration 4
In this iteration, we are considering the vocabulary tt, er- → tter- which has a total count of 8

Iteration 5
Here in this combining b,e
count -4 (betty, 2) (better, 2)

Iteration 6
For the 6th iteration considers o,u vocabulary.
combine o,u → ou
count -4 (lou, 2) (bought, 2)

Iteration 7
For the 7th iteration considering b,a vocabulary
for combining b,a → ba
count - 4
(baker -2
batter -4