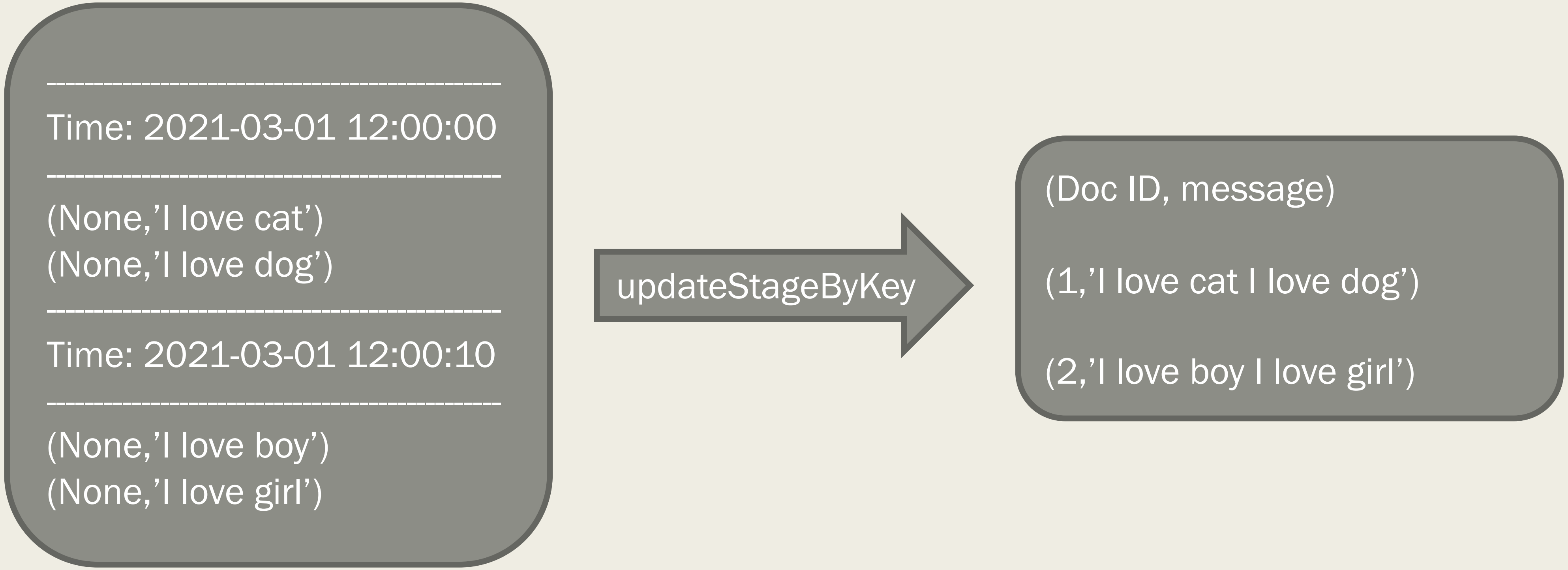
TDIDF
SPARK STREAMIMG

Nothing’s gonna stop us now

Receive message from topic



```
-------------------------------------------
Time: 2021-03-01 12:00:00
-------------------------------------------
(None,’I love cat’)
(None,’I love dog’)
-------------------------------------------
Time: 2021-03-01 12:00:10
-------------------------------------------
(None,’I love boy’)
(None,’I love girl’)
```

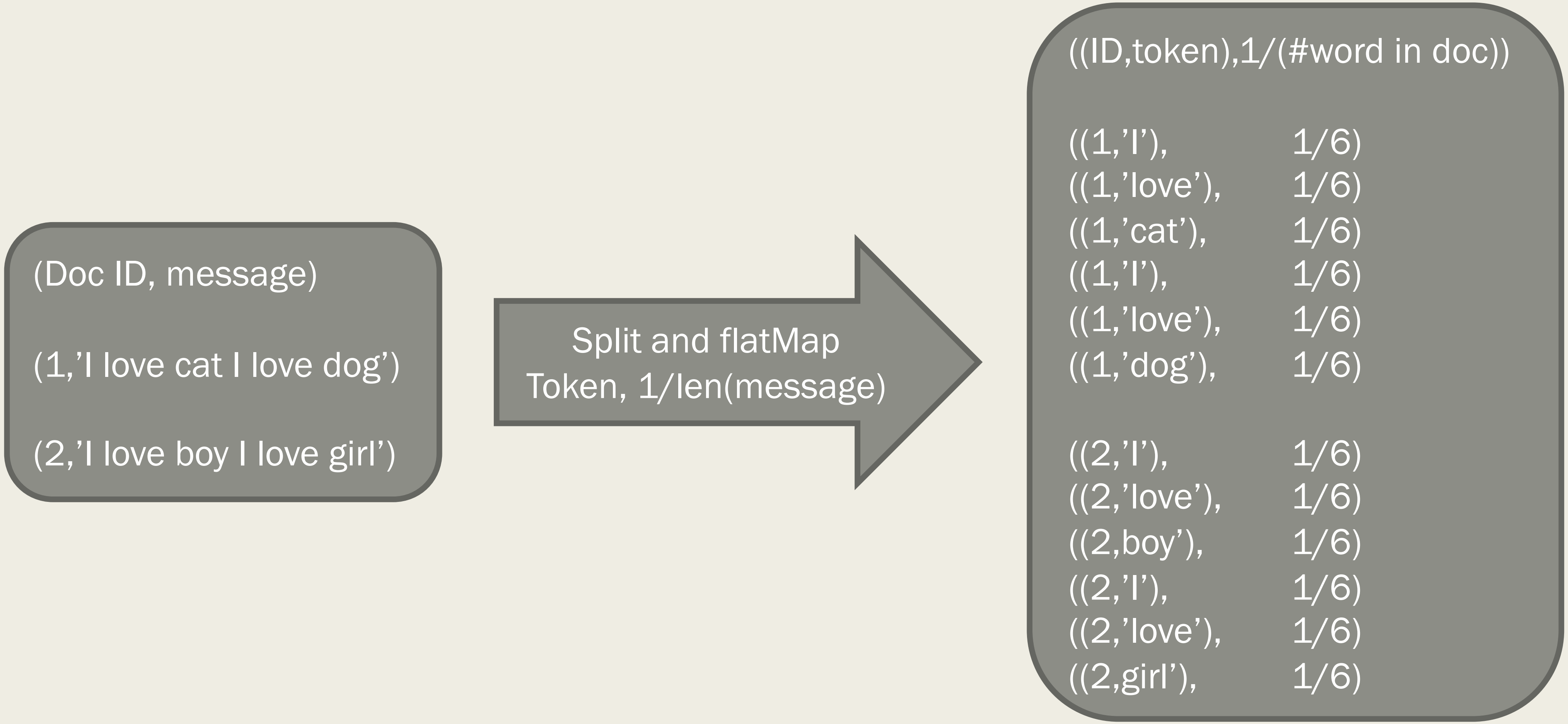
updateStageByKey

(Doc ID, message)

(1,’I love cat I love dog’)

(2,’I love boy I love girl’)

Compute TF



(Doc ID, message)

(1,’I love cat I love dog’)

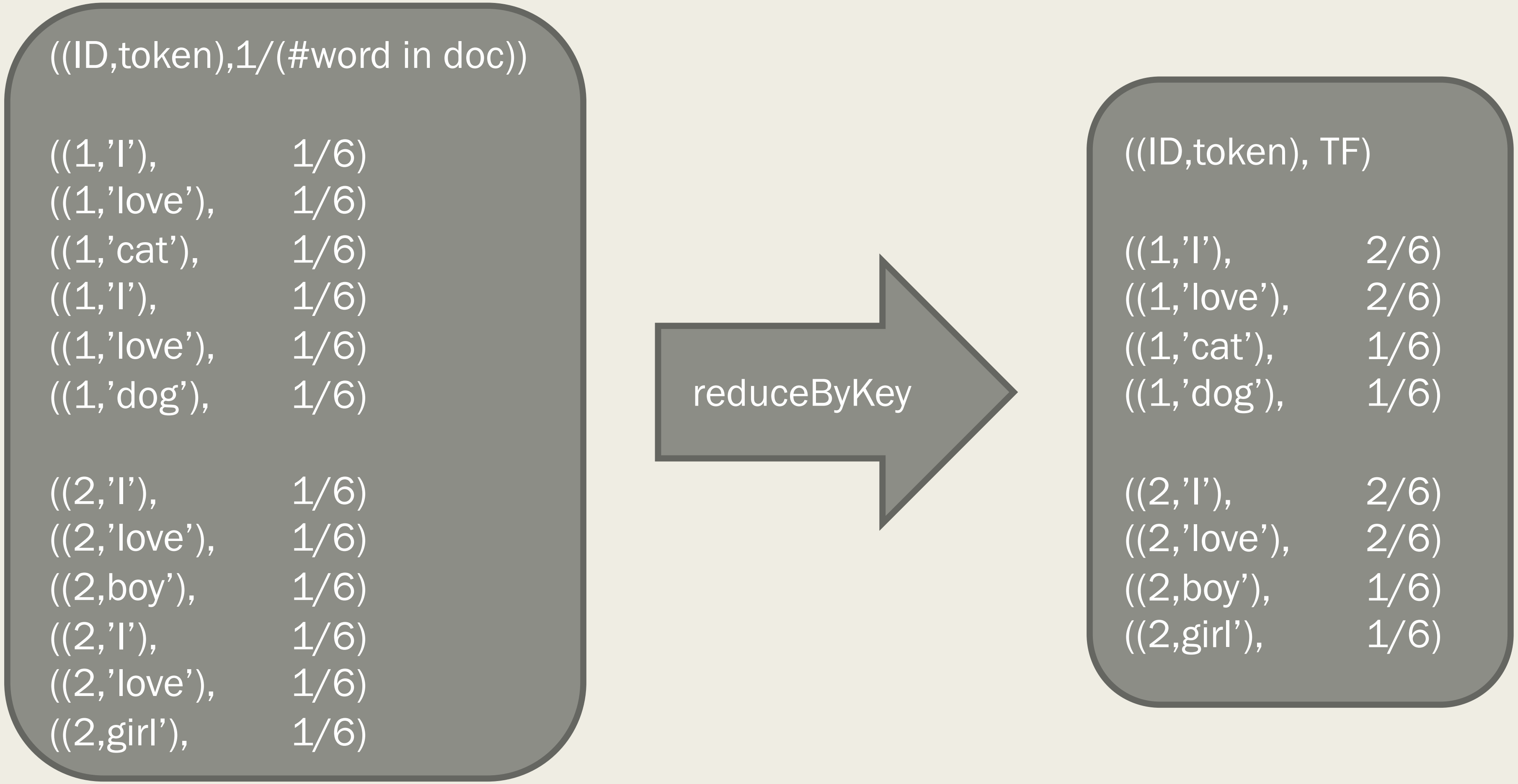
(2,’I love boy I love girl’)

Split and flatMap
Token, 1/len(message)

((ID,token),1/(#word in doc))

```
((1,’I’),          1/6)
((1,’love’),       1/6)
((1,’cat’),        1/6)
((1,’I’),          1/6)
((1,’love’),       1/6)
((1,’dog’),        1/6)

((2,’I’),          1/6)
((2,’love’),       1/6)
((2,boy’),         1/6)
((2,’I’),          1/6)
((2,’love’),       1/6)
((2,girl’),        1/6)
```

Compute TF



((ID,token),1/(#word in doc))

((1,’I’), 1/6)
((1,’love’), 1/6)
((1,’cat’), 1/6)
((1,’I’), 1/6)
((1,’love’), 1/6)
((1,’dog’), 1/6)

((2,’I’), 1/6)
((2,’love’), 1/6)
((2,boy’), 1/6)
((2,’I’), 1/6)
((2,’love’), 1/6)
((2,girl’), 1/6)

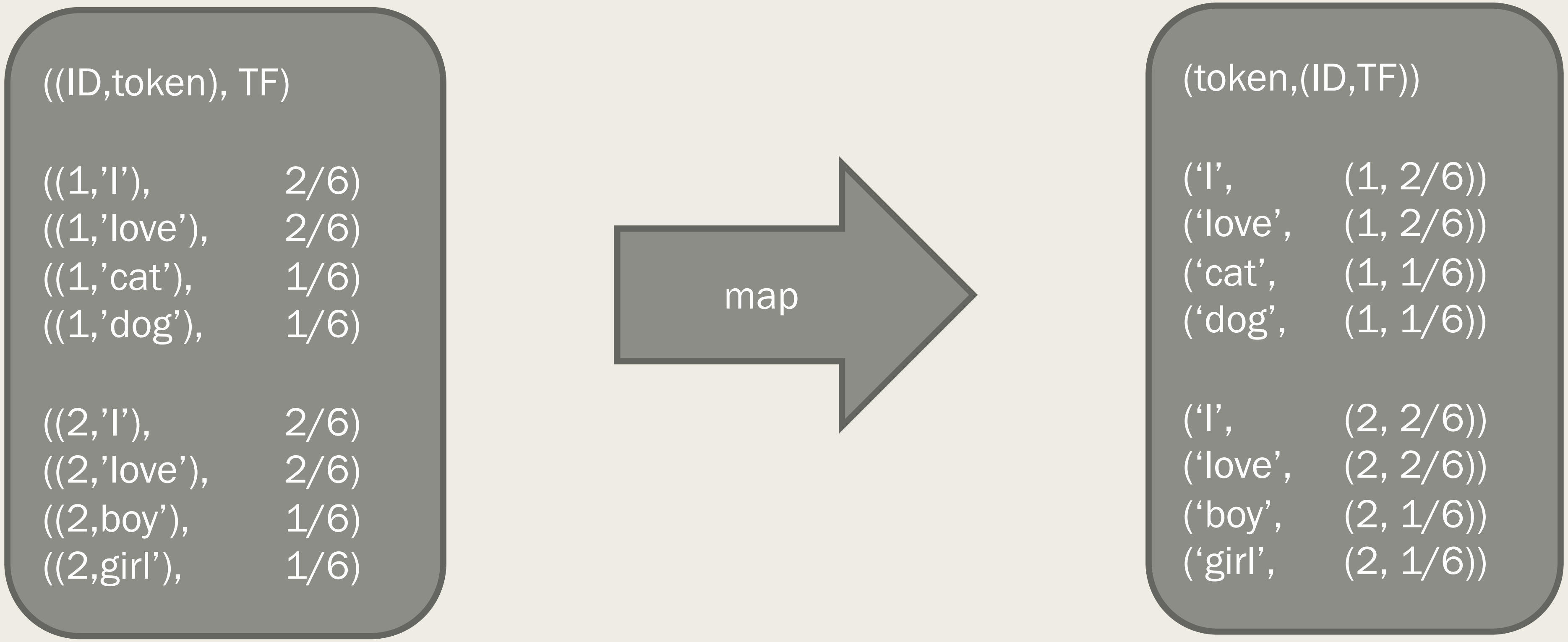
reduceByKey

((ID,token), TF)

((1,’I’), 2/6)
((1,’love’), 2/6)
((1,’cat’), 1/6)
((1,’dog’), 1/6)

((2,’I’), 2/6)
((2,’love’), 2/6)
((2,boy’), 1/6)
((2,girl’), 1/6)

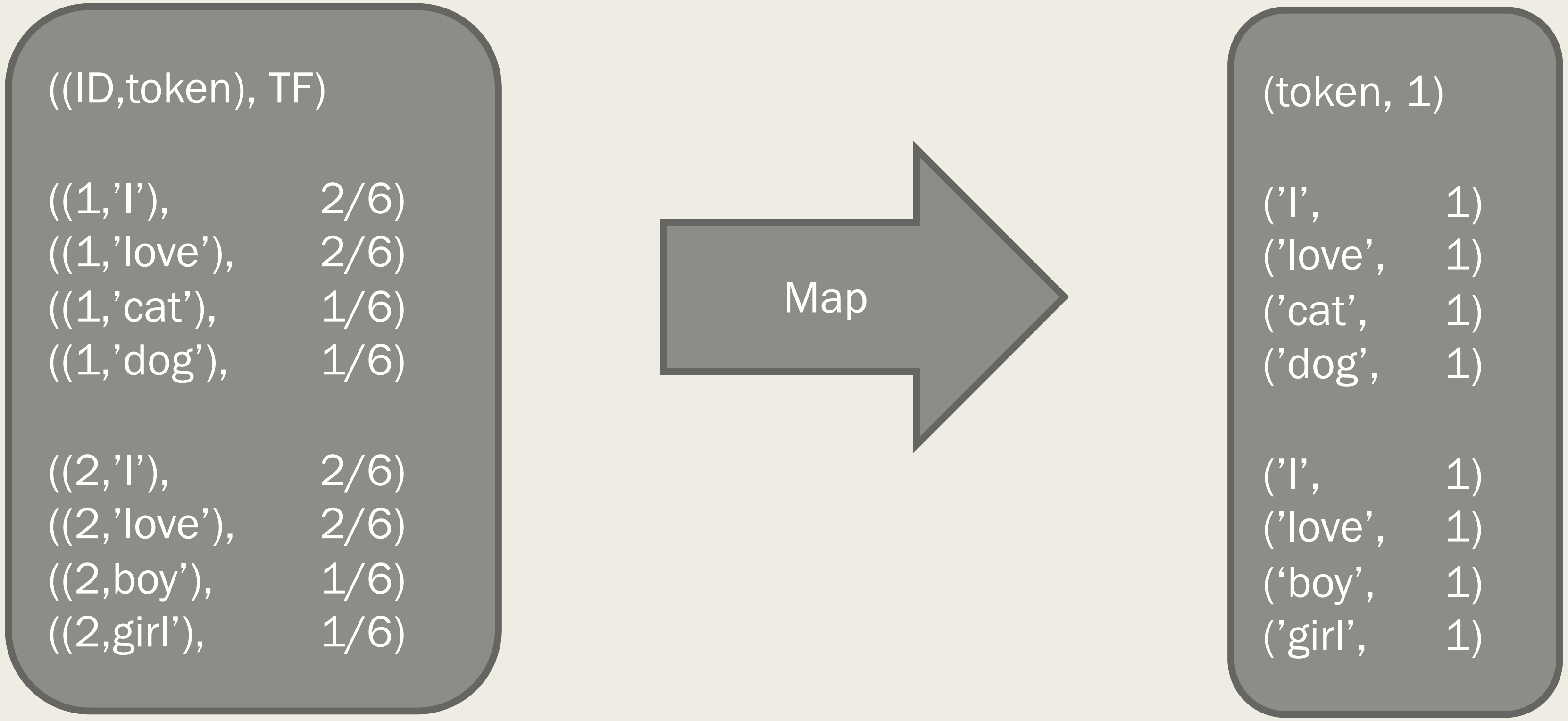
Compute TF



```
((ID,token), TF)

((1,’I’),        2/6)
((1,’love’),     2/6)
((1,’cat’),      1/6)
((1,’dog’),      1/6)

((2,’I’),        2/6)
((2,’love’),     2/6)
((2,boy’),       1/6)
((2,girl’),      1/6)
```

```
(token,(ID,TF))

(‘I’,       (1, 2/6))
(‘love’,    (1, 2/6))
(‘cat’,     (1, 1/6))
(‘dog’,     (1, 1/6))

(‘I’,       (2, 2/6))
(‘love’,    (2, 2/6))
(‘boy’,     (2, 1/6))
(‘girl’,    (2, 1/6))
```

Compute IDF



((ID,token), TF)

((1,’I’), 2/6)
((1,’love’), 2/6)
((1,’cat’), 1/6)
((1,’dog’), 1/6)

((2,’I’), 2/6)
((2,’love’), 2/6)
((2,boy’), 1/6)
((2,girl’), 1/6)

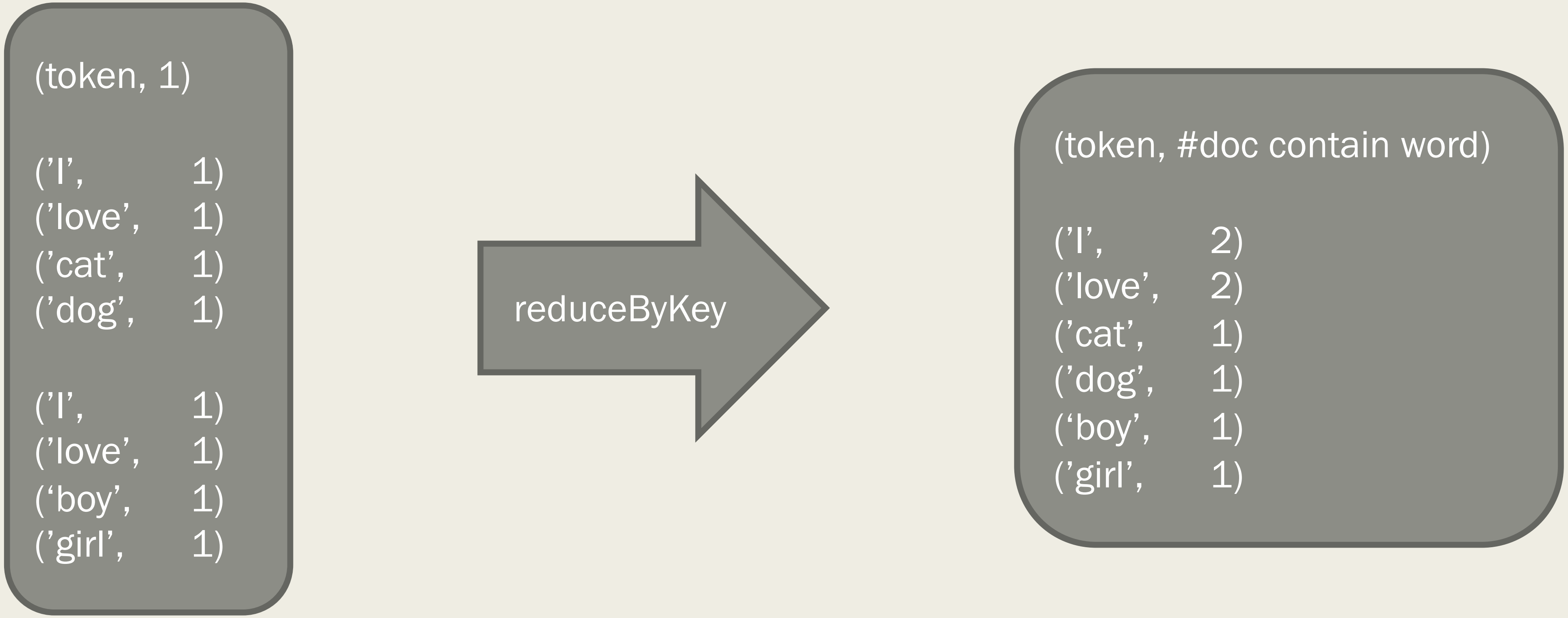
Map

(token, 1)

(’I’, 1)
(’love’, 1)
(’cat’, 1)
(’dog’, 1)

(’I’, 1)
(’love’, 1)
(‘boy’, 1)
(’girl’, 1)

Compute IDF



(token, 1)

(’I’, 1)
(’love’, 1)
(’cat’, 1)
(’dog’, 1)

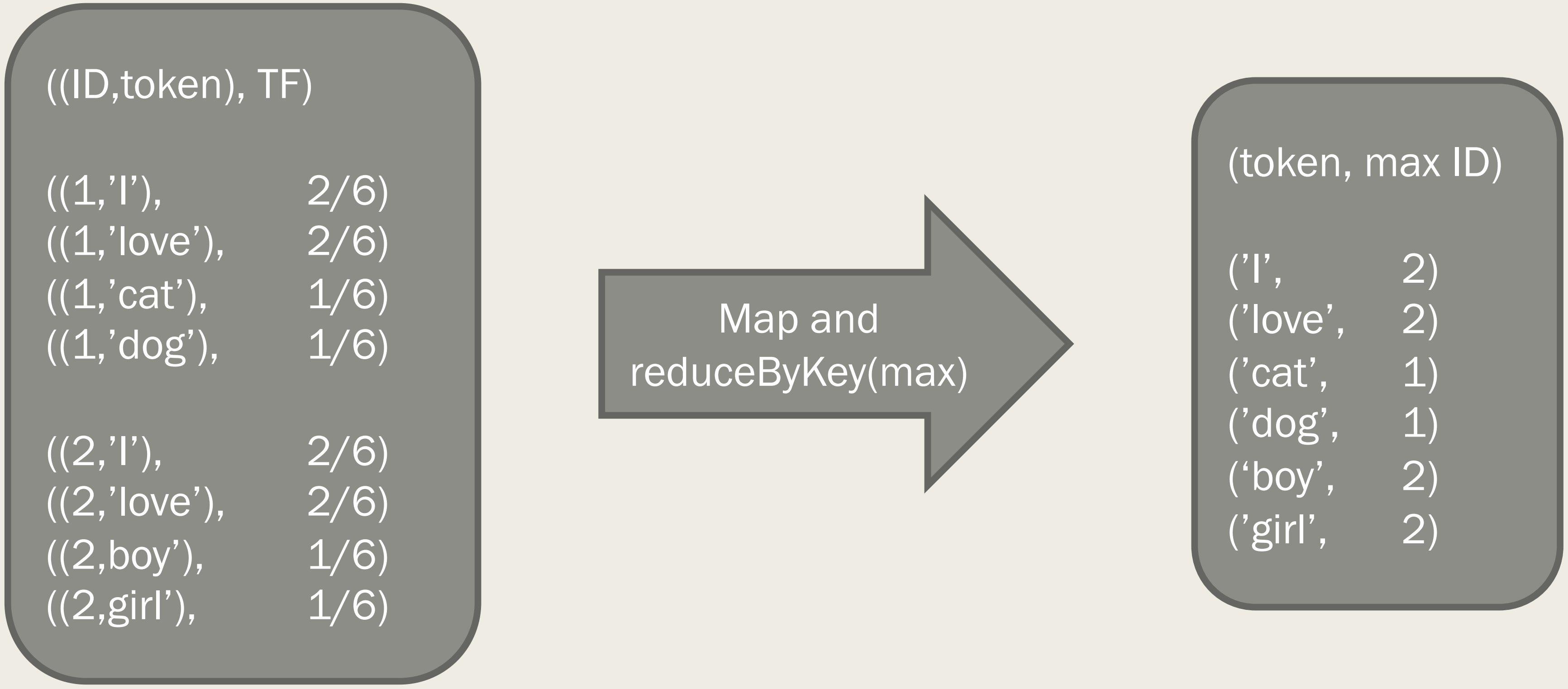
(’I’, 1)
(’love’, 1)
(‘boy’, 1)
(’girl’, 1)

reduceByKey

(token, #doc contain word)

(’I’, 2)
(’love’, 2)
(’cat’, 1)
(’dog’, 1)
(‘boy’, 1)
(’girl’, 1)

Compute IDF



((ID,token), TF)

((1,’I’), 2/6)
((1,’love’), 2/6)
((1,’cat’), 1/6)
((1,’dog’), 1/6)

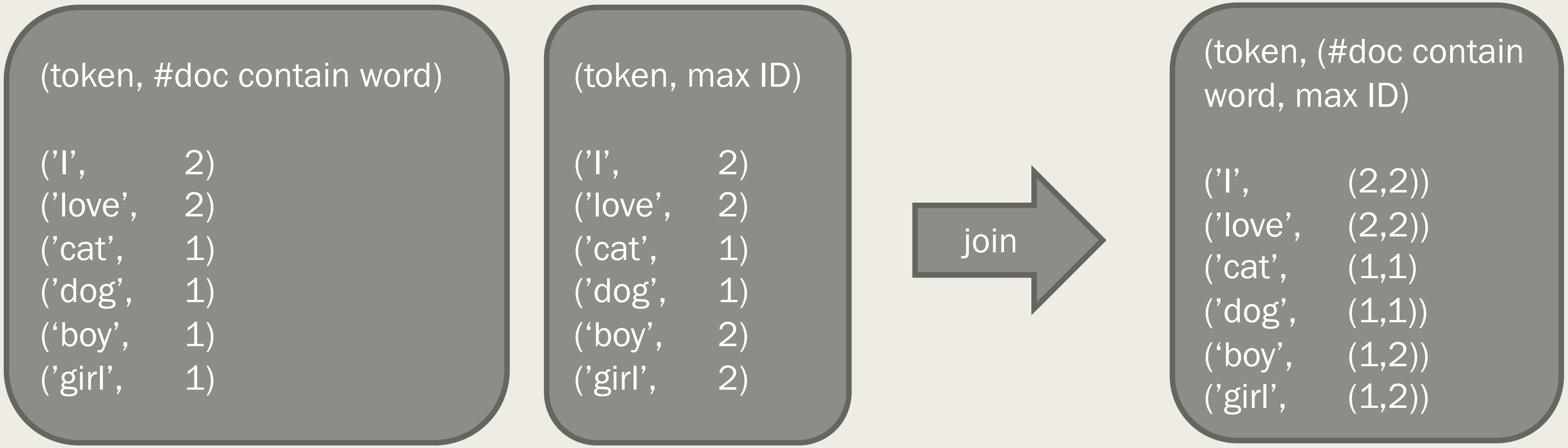
((2,’I’), 2/6)
((2,’love’), 2/6)
((2,boy’), 1/6)
((2,girl’), 1/6)

Map and reduceByKey(max)

(token, max ID)

(’I’, 2)
(’love’, 2)
(’cat’, 1)
(’dog’, 1)
(‘boy’, 2)
(’girl’, 2)

Compute IDF



(token, #doc contain word)

(’I’, 2)
(’love’, 2)
(’cat’, 1)
(’dog’, 1)
(‘boy’, 1)
(’girl’, 1)

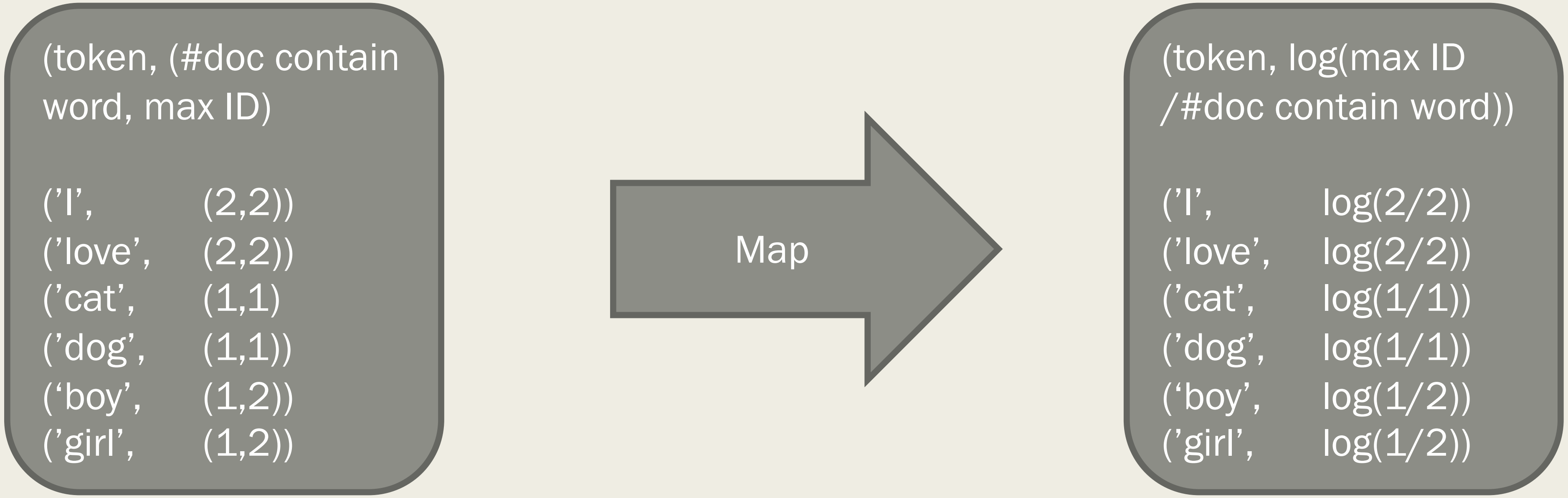
(token, max ID)

(’I’, 2)
(’love’, 2)
(’cat’, 1)
(’dog’, 1)
(‘boy’, 2)
(’girl’, 2)

(token, (#doc contain word, max ID)

(’I’, (2,2))
(’love’, (2,2))
(’cat’, (1,1)
(’dog’, (1,1))
(‘boy’, (1,2))
(’girl’, (1,2))

Compute IDF



(token, (#doc contain word, max ID)

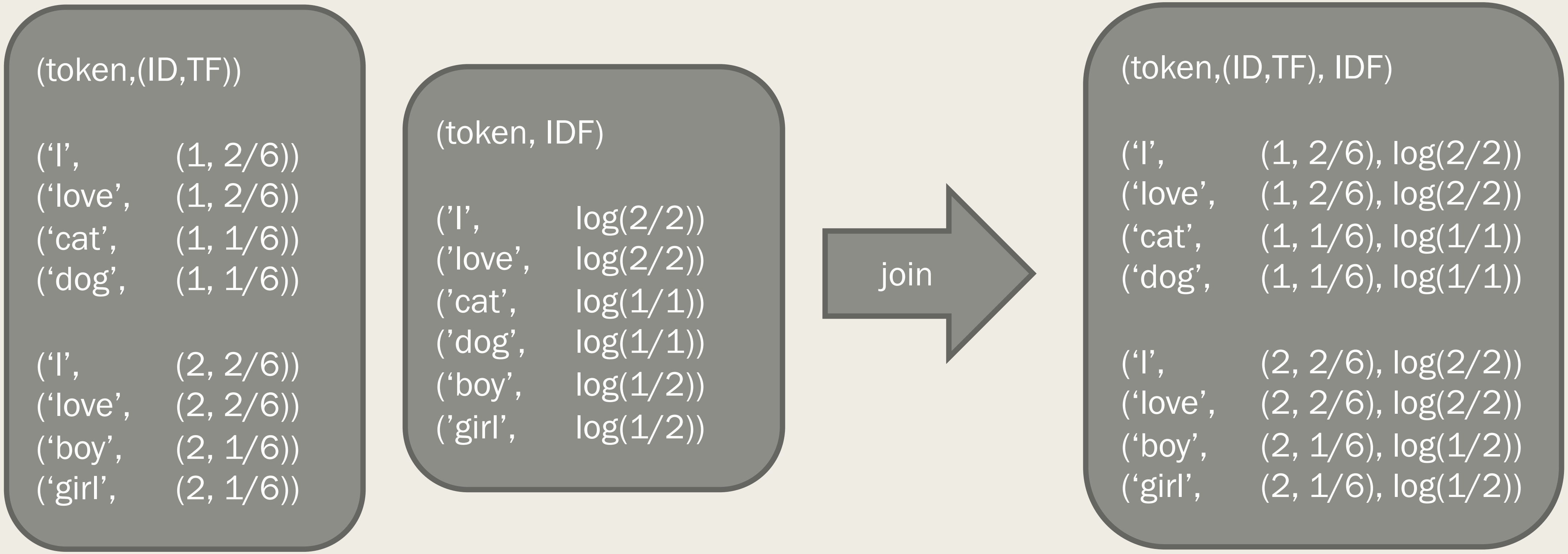
(’I’,	(2,2))
(’love’,	(2,2))
(’cat’,	(1,1)
(’dog’,	(1,1))
(‘boy’,	(1,2))
(’girl’,	(1,2))

Map

(token, log(max ID /#doc contain word))

(’I’,	log(2/2))
(’love’,	log(2/2))
(’cat’,	log(1/1))
(’dog’,	log(1/1))
(‘boy’,	log(1/2))
(’girl’,	log(1/2))

Compute TF-IDF

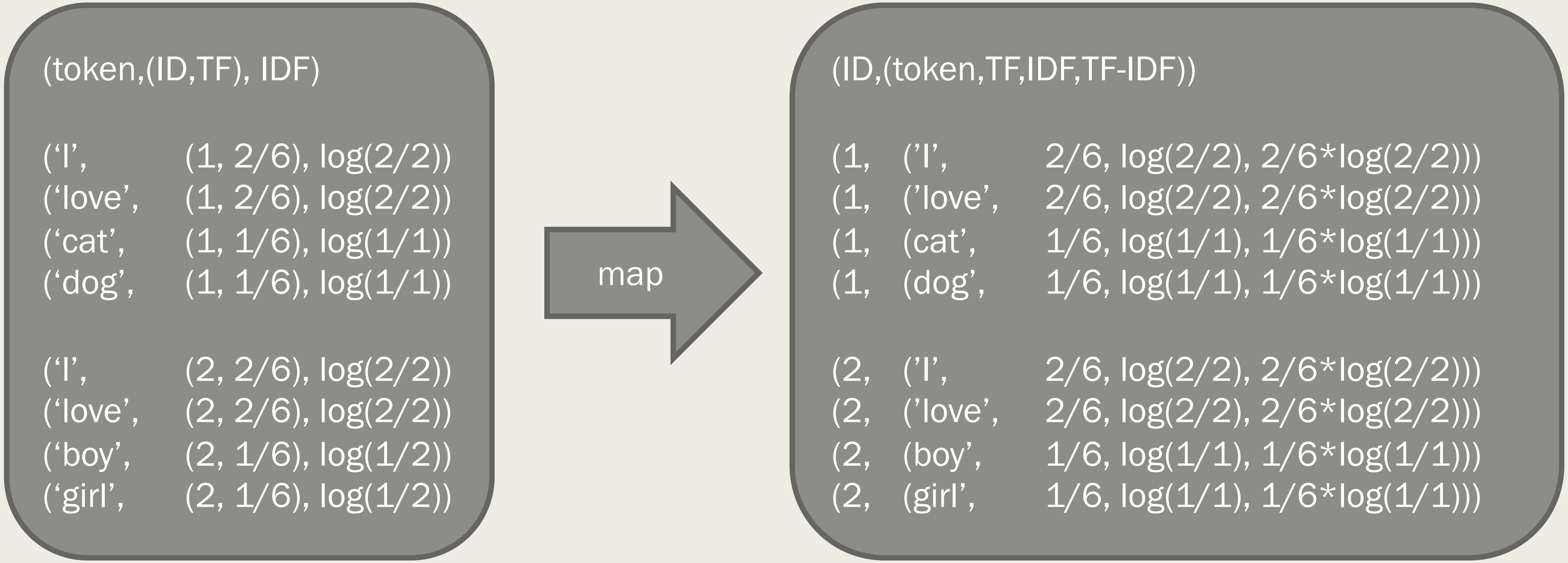


(token,(ID,TF))	
(‘I’,	(1, 2/6))
(‘love’,	(1, 2/6))
(‘cat’,	(1, 1/6))
(‘dog’,	(1, 1/6))
(‘I’,	(2, 2/6))
(‘love’,	(2, 2/6))
(‘boy’,	(2, 1/6))
(‘girl’,	(2, 1/6))

(token, IDF)	
(’I’,	log(2/2))
(’love’,	log(2/2))
(’cat’,	log(1/1))
(’dog’,	log(1/1))
(‘boy’,	log(1/2))
(’girl’,	log(1/2))

(token,(ID,TF), IDF)	
(‘I’,	(1, 2/6), log(2/2))
(‘love’,	(1, 2/6), log(2/2))
(‘cat’,	(1, 1/6), log(1/1))
(‘dog’,	(1, 1/6), log(1/1))
(‘I’,	(2, 2/6), log(2/2))
(‘love’,	(2, 2/6), log(2/2))
(‘boy’,	(2, 1/6), log(1/2))
(‘girl’,	(2, 1/6), log(1/2))

Compute TF-IDF



(token,(ID,TF), IDF)

(‘I’, (1, 2/6), log(2/2))
(‘love’, (1, 2/6), log(2/2))
(‘cat’, (1, 1/6), log(1/1))
(‘dog’, (1, 1/6), log(1/1))

(‘I’, (2, 2/6), log(2/2))
(‘love’, (2, 2/6), log(2/2))
(‘boy’, (2, 1/6), log(1/2))
(‘girl’, (2, 1/6), log(1/2))

(ID,(token,TF,IDF,TF-IDF))

(1, (’I’, 2/6, log(2/2), 2/6*log(2/2)))
(1, (’love’, 2/6, log(2/2), 2/6*log(2/2)))
(1, (cat’, 1/6, log(1/1), 1/6*log(1/1)))
(1, (dog’, 1/6, log(1/1), 1/6*log(1/1)))

(2, (’I’, 2/6, log(2/2), 2/6*log(2/2)))
(2, (’love’, 2/6, log(2/2), 2/6*log(2/2)))
(2, (boy’, 1/6, log(1/1), 1/6*log(1/1)))
(2, (girl’, 1/6, log(1/1), 1/6*log(1/1)))