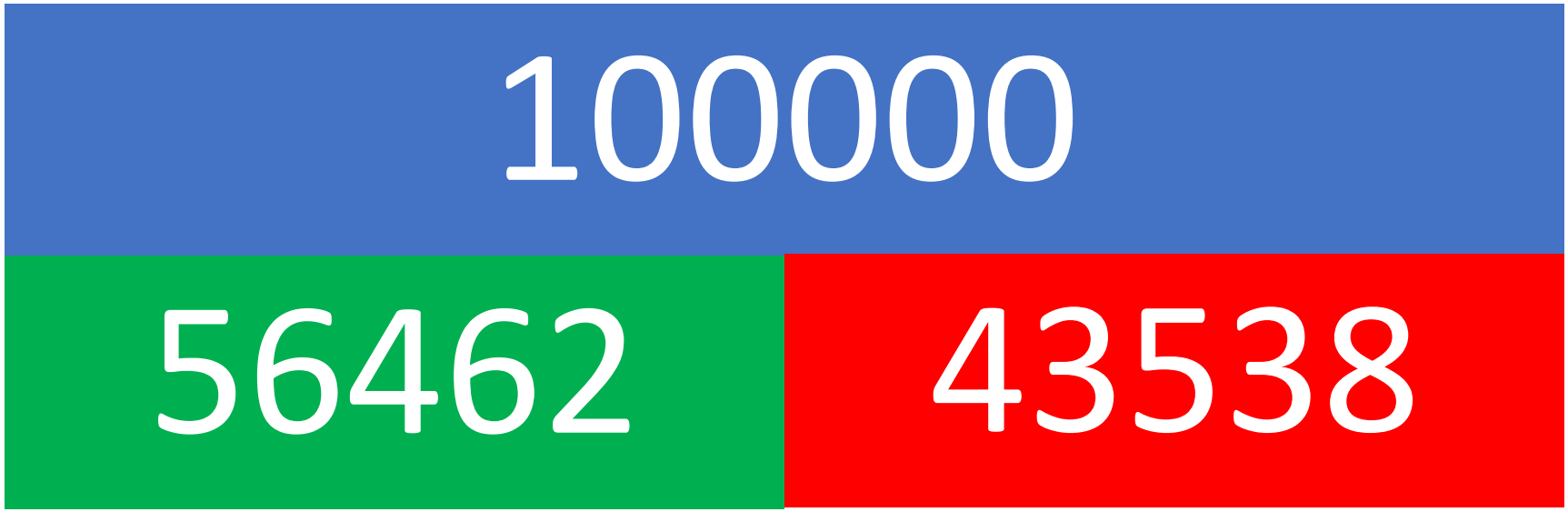
Twitter Sentiment Analysis

สถิติทั่วไปของชุดข้อมูล



จำนวน Token ใน Dataset

$$\frac{1610572}{111608} = 14.43$$

การเตรียมข้อมูล: ตัดคำ

@soultravelers3 http://bit.ly/H9Nqe 15 TravelTips 4 student safety in roughplaces

nltk.word_tokenize:

['@', 'soultravelers3', 'http', ':', '//bit.ly/H9Nqe', '15', 'TravelTips', '4', 'student', 'safety', 'in', 'roughplaces']

nltk.tokenize.TweetTokenizer:

['@soultravelers3', 'http://bit.ly/H9Nqe', '15’, 'TravelTips', '4', 'student', 'safety', 'in', 'roughplaces']

การเตรียมข้อมูล: นำลิงก์ออกจากข้อความ

['@soultravelers3',
'http://bit.ly/H9Nqe', '15’, 'TravelTips',
'4', 'student', 'safety', 'in', 'roughplaces']

['@soultravelers3', '15’, 'TravelTips', '4',
'student', 'safety', 'in', 'roughplaces']

การเตรียมข้อมูล: mention ออกจากข้อความ

['@soultravelers3', '15’, 'TravelTips', '4', 'student', 'safety', 'in', 'roughplaces']

['15’, 'TravelTips', '4', 'student', 'safety', 'in', 'roughplaces']

การเตรียมข้อมูล: ตัวอักษรเบิ้ล

SSSWEETTTTTTT

SWEET

การเตรียมข้อมูล: ตัวอักษรเบิ้ล

SSSWEETTTTTTT

[(s,3),(w,1),(e,2),(t:7)]

[(s,2),(w,1),(e,2),(t:2)]

ssweett	ssweet	sswett	sswet
sweett	sweet	swett	swet

การเตรียมข้อมูล: ภาษา SMS

Gr8 = Great

BRB = Be Right Back

การเตรียมข้อมูล: hashtag

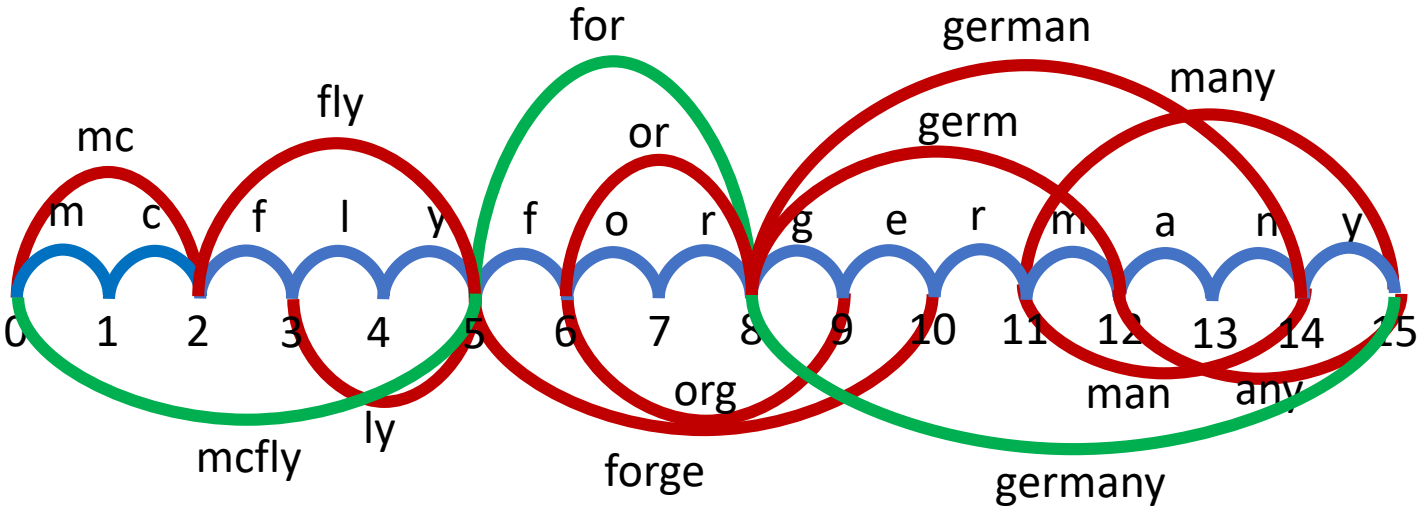
Label	ข้อความ
0	#IMISSCATH #IMISSCATH #IMISSCATH #IMISSCATH #IMISSCATH #IMISSCATH #IMISSCATH #IMISSCATH #IMISSCATH #IMISSCATH #IMISSCATH
1	#mcflyforgermany #mcflyforgermany #mcflyforgermany #mcflyforgermany #mcflyforgermany #mcflyforgermany #mcflyforgermany #mcflyforgermany

การเตรียมข้อมูล: ตัดคำใน hashtag

nltk.word_tokenize([‘mcflyforgermany’]) = [‘mcflyforgermany’]

การเตรียมข้อมูล: ตัดคำใน hashtag

mcflyforgermany



การแบ่งชุดข้อมูล

การเลือก Feature

Bag of word model

['15’, 'TravelTips', '4', 'student', 'safety','in', 'roughplaces']

{

15:1,

TravelTips:1,

4:1,

student:1,

safety:1,

in:1,

roughplaces:1

}

Classification Technique

Naive Bayes classifier

$$p(D \mid C) = \prod_i p(w_i \mid C)$$

food	happy	label
1	1	True
0	1	True
1	0	False

ถ้า input ได้ว่า food = 0 และ happy = 1

p(food=0,happy=1|true) = p(food=0|true)*p(happy=1|true)
= (1/2)*(1) = 0.5

p(food=0,happy=1|false) = p(food=0|false) * p(happy=1|false)
= (1/2)*(0) = 0

จะเห็นว่า p(food=0,happy=1|true) > p(food=0,happy=1|false)

จึงได้ว่า label ของข้อมูล food = 0 และ happy = 1 คือ false

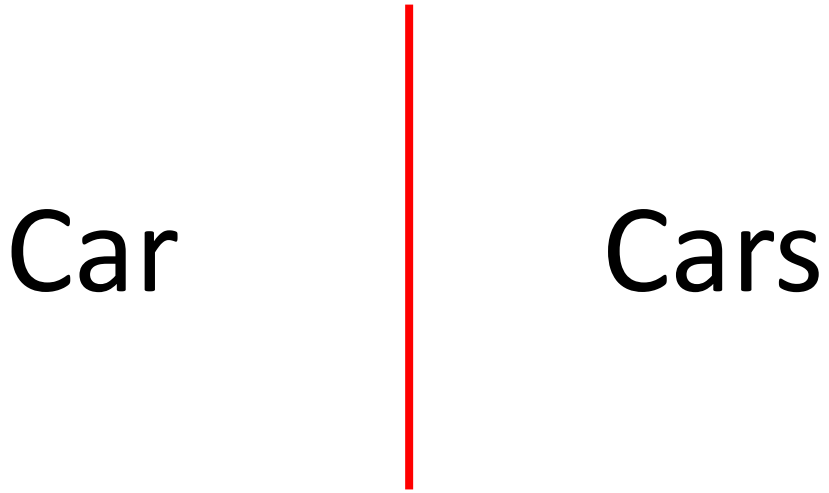
ความถูกต้อง

Bag of word model

อ้างอิง / ทดสอบ	1	0
1	36.4%	20.6%
0	6.5%	36.4%

Precious : 0.6386
Recall : 0.8485
F1 Score : 0.7287
Accuracy : 0.7284

วิเคราะห์ข้อผิดพลาด: Bag of word model



การเลือก Feature

Bag of word model + Lemmariztion

WordNetLemmatizer

Cars  Car

ความถูกต้อง

Bag of word model + Lemmariztion

อ้างอิง / ทดสอบ	1	0
1	36.5%	20.5%
0	6.4%	36.5%

Precious : 0.6404
Recall : 0.8508
F1 Score : 0.7307
Accuracy : 0.7307

วิเคราะห์ข้อผิดพลาด

Bag of word model + Lemmariztion

Fair | Fairly

การเลือก Feature

Bag of word model + Stemming

SnowballStemmer

Fairly  Fair

ความถูกต้อง

Bag of word model + Stemming

อ้างอิง / ทดสอบ	1	0
1	36.2%	20.8%
0	6.3%	36.6%

Precious : 0.6351
Recall : 0.8518
F1 Score : 0.7276
Accuracy : 0.7289

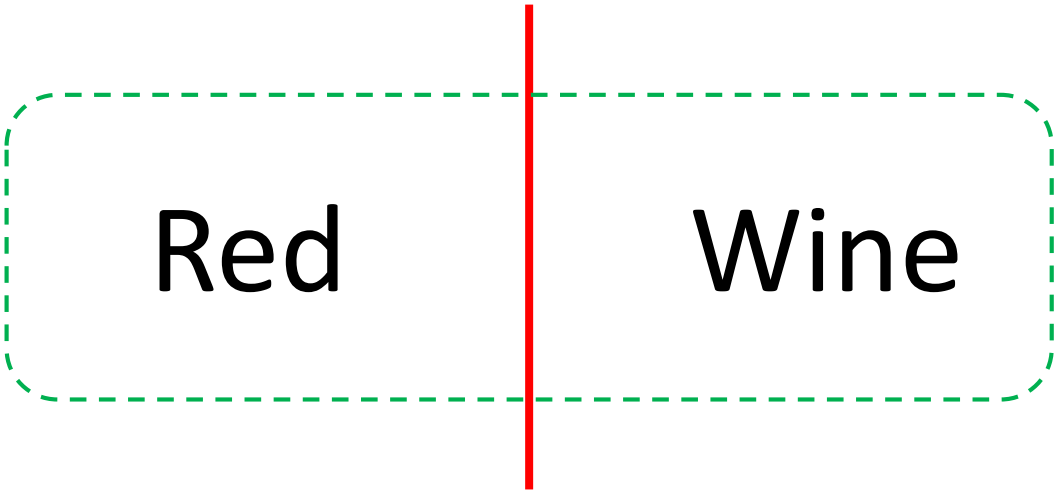
วิเคราะห์ข้อผิดพลาด

Bag of word model + Stemming

การทำ Lemmatization มีค่า Accuracy มากกว่า

การเลือก Feature

Bag of word model + Lemmarization + Bigrams



ความถูกต้อง

Bag of word model + Lemmarization + Bigrams

อ้างอิง / ทดสอบ	1	0
1	40.1%	17%
0	8.2%	34.8%

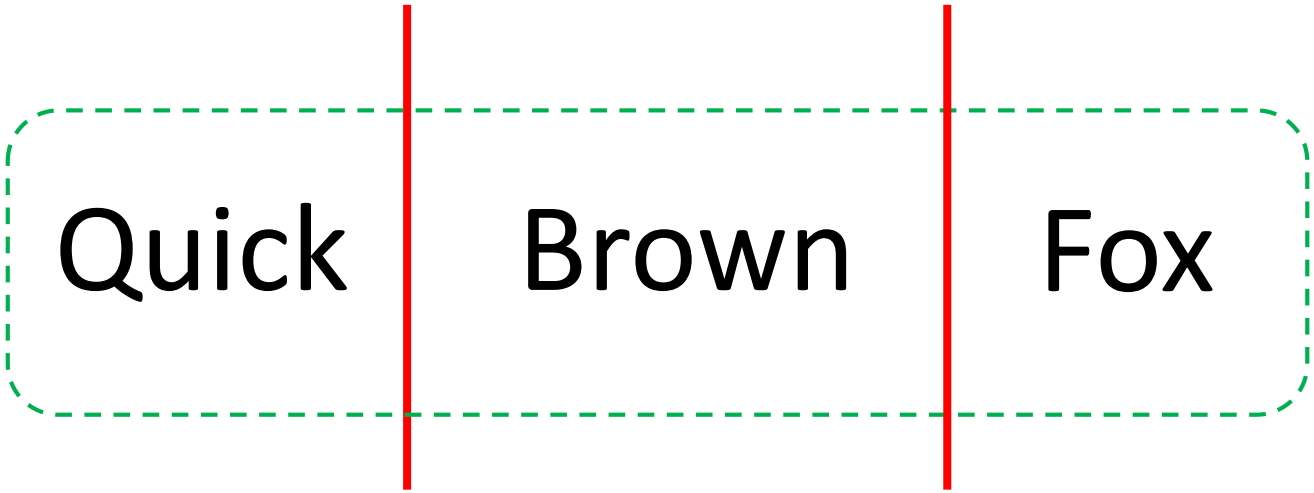
Precious : 0.7023
Recall : 0.8302
F1 Score : 0.7609
Accuracy : 0.7486

วิเคราะห์ข้อผิดพลาด

Bag of word model + Lemmarization + Bigrams

การเลือก Feature

Bag of word model + Lemmarization + Trigram



ความถูกต้อง

Bag of word model + Lemmarization + Trigram

อ้างอิง / ทดสอบ	1	0
1	41.3%	15.8%
0	14%	29%

Precious : 0.7233
Recall : 0.7468
F1 Score : 0.7349
Accuracy : 0.7028

วิเคราะห์ข้อผิดพลาด

Bag of word model + Lemmarization + Trigrams

แบบใช้ Bigrams มีค่า Accuracy มากกว่า

การเลือก Feature

Bag of word model + Synsets

Automobile Motor

car.n.01

การเลือก Feature

Bag of word model + Synsets

I motorbike to my home

motorbike.v.01

I was ride a motorbike

motorbike.n.01

ความถูกต้อง

Bag of word model + Synsets

อ้างอิง / ทดสอบ	1	0
1	37.6%	19.5%
0	11.1%	31.9%

Precious : 0.6585
Recall : 0.7721
F1 Score : 0.7108
Accuracy : 0.6943

วิเคราะห์ข้อผิดพลาด

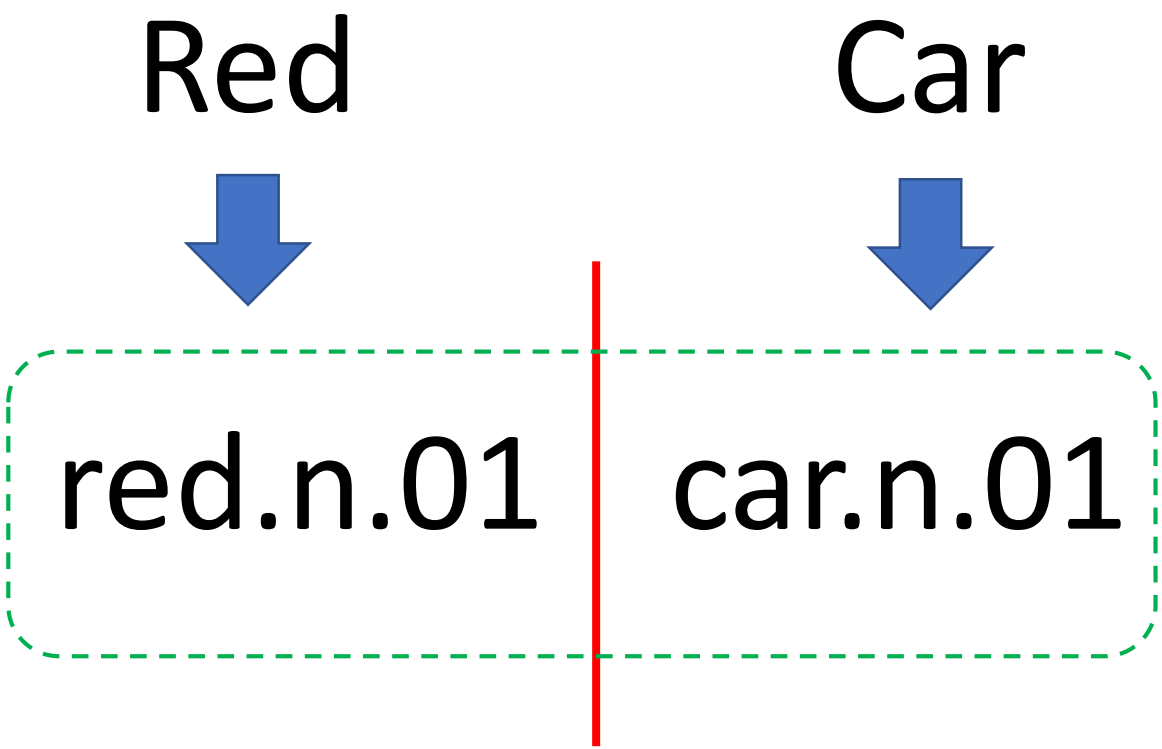
Bag of word model + Synsets

list(set([‘red.n.01’,‘car.n.01’]]))

[‘car.n.01’, ‘red.n.01’]

การเลือก **Feature**

Bag of word model + Synsets +Bigrams



ความถูกต้อง

Bag of word model + Synsets + Bigrams

อ้างอิง / ทดสอบ	1	0
1	37.6%	19.5%
0	11.1%	31.9%

Precious : 0.7040
Recall : 0.7053
F1 Score : 0.7046
Accuracy : 0.6637

วิเคราะห์ข้อผิดพลาด

Bag of word model + Synsets +Bigrams

แบบไม่ใช้ Bigrams มีค่า Accuracy มากกว่า

สรุป

Bag of word model + Lemmarization + Bigrams

ความถูกต้อง

ทดสอบกับข้อมูล Test Set ด้วยวิธีการ

Bag of word model + Lemmarization + Bigrams

อ้างอิง / ทดสอบ	1	0
1	37.6%	19.5%
0	11.1%	31.9%

Precious : 0.6853
Recall : 0.8083
F1 Score : 0.7398
Accuracy : 0.7325