CSE 4062 - Semester 2021

Intro to Data Science and Analytics

Delivery #2: Exploring your data

Group 2 Members:

Aleyna BOZACI 150319630 - aleynabozacii@gmail.com

Bekir ÖZKAN 150319557 - bekirozkan9698@gmail.com

Merve YAYIN 150116051 - yayinm8@gmail.com

Metehan ERTAN 150117051 - metehan.ertan@hotmail.com

Mustafa USCA 150416054 - uscamustafa64@gmail.com

1- Dataset

1 - Deceptive Opinion Spam Corpus v1.4s (shortly Turk). We will be using negative reviews for testing.

- 400 truthful positive reviews from TripAdvisor
- 400 deceptive positive reviews from Mechanical Turk
- 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp
- 400 deceptive negative reviews from Mechanical Turk

2 - We extracted a smaller dataset from YELP Dataset as it was too large. We are going to use 800 truthful and 800 deceptive reviews for the learning part and 200 reviews for the testing part.

Deceptive Opinion Spam Corpus v1.4s has 800 rows for learning and 800 rows for testing. For the Yelp dataset there are 1600 rows for learning and 200 rows for learning. Target attribute is truthful or deceptive as boolean.

2 - Statistics

2.1- Yelp Train

Number of documents: 1600

Number of words: 148339

Number of characters: 769776

Min word count: 2

Average word count: 92.711875

Max word count: 903

Min character count: 11

Average character count: 481.11

Max character count: 4809

2.2- Yelp Test

Number of documents: 200

Number of words: 21059

Number of characters: 108822

Min word count: 1

Average word count: 105.295

Max word count: 432

Min character count: 4

Average character count: 544.11

Max character count: 2240

2.3- Turk Train

Number of documents: 800

Number of words: 145407

Number of characters: 747588

Min word count: 32

Average word count: 181.75875

Max word count: 797

Min character count: 174

Average character count: 934.485

Max character count: 4048

2.4- Turk Test

Number of documents: 800

Number of words: 97604

Number of characters: 513557

Min word count: 25

Average word count: 122.005

Max word count: 438

Min character count: 147

Average character count: 641.94625

Max character count: 2389

2 - Bag of Words

2.1- Yelp Train

Size of bag of words is 10137 , without stop words size of bag of words is 8326:

With stop words		Without stop words	
Word	Count	Word	Count
the	6440	food	1095
and	5126	place	861
a	3787	good	815
I	3660	great	723

to	2893	service	565
was	2844	restaurant	464
of	2127	like	455
is	1952	pizza	425
it	1822	one	421
for	1689	really	405

2.2- Yelp Test

Size of bag of words is 3310 , without stop words size of bag of words is 2701:

With stop words		Without stop words	
Word	Count	Word	Count
the	928	food	130
and	765	place	126
a	569	good	108
I	535	french	101
was	450	great	97
to	392	brunch	92

of	285	back	89
it	276	like	84
for	251	delicious	69
is	250	chicken	67

2.3- Turk Train

Size of bag of words is 8697 , without stop words size of bag of words is 7480 :

With stop words		Without stop words	
Word	Count	Word	Count
the	7770	room	1850
to	4300	hotel	1755
and	4090	stay	649
I	4033	chicago	636
a	3602	would	572
was	3464	service	470
in	2311	one	464
of	1915	get	384

room	1799	desk	380
for	1701	night	370

2.4- Turk Test

Size of bag of words is 6467 , without stop words size of bag of words is 5437 :

With stop words		Without stop words	
Word	Count	Word	Count
the	5046	hotel	1582
and	3718	room	970
a	2746	chicago	891
to	2407	great	664
was	2360	stay	655
I	2343	staff	473
in	1634	stayed	359
The	1628	rooms	355
of	1376	would	351
hotel	1287	service	349