Praxis Business School
CELEBRATE YOUR WORTH

Bagging Significant Reviews

CAPSTONE PROJECT

NEXT →

Presentation Highlights

General Discussion Flow

- About Our Project
- Detailed Design Description
- Scope Of Work
- Findings
- Things We Learnt

NEXT →

OUR IDEA IN A MINUTE

NEXT →

PROBLEM STATEMENT AND OBJECTIVE

What we are trying to solve?

Looking at thousand of reviews can be a time-consuming task and choosing the right reviews for the prefect buy isn’t that easy. So, we aim to help customers **choose the most significant reviews** before making an online purchase by extracting information from consumer data.

NEXT →

PROBLEM STATEMENT AND OBJECTIVE

What we are trying to solve?

Looking at thousand of reviews can be a time-consuming task and choosing the right reviews for the prefect buy isn’t that easy. So, we aim to help customers **choose the most significant reviews** before making an online purchase by extracting information from consumer data.

Human beings are emotional creatures and their needs are constantly evolving with time and **knowing their sentiments** about a product undoubtedly gives them an edge over their competitors. So, with the help of our project we'd like to help both buyers and sellers in making the most out of the reviews.

NEXT →

ABOUT DATA EXTRACTION

Which product

Flipkart Perfect Homes Opus Engineered Wood Queen Box Bed from flipkart.com

NEXT →

ABOUT DATA EXTRACTION

Which product	Flipkart Perfect Homes Opus Engineered Wood Queen Box Bed from flipkart.com
What we scraped	Reviews, Ratings, Likes, Dislikes, Date, Consumer Name

NEXT →

ABOUT DATA EXTRACTION

Which product	Flipkart Perfect Homes Opus Engineered Wood Queen Box Bed from flipkart.com
What we scraped	Reviews, Ratings, Likes, Dislikes, Date, Consumer Name
Tools and Libraries Used	Python, Microsoft Excel Selenium, Request, Chrome Web Driver, Datetime etc.
No. of reviews scraped	2506

NEXT →

Detailed Design Description

- Text Cleaning
- Information Extraction
- Noun Extraction and Pattern Mining
- Feature Selection
- Similar Words
- Polarity and Binary Data Preparation
- Calculation of Weights and Clustering

NEXT →

TEXT CLEANING

GRAMMAR CHECK

Using language tool python

NEXT →

TEXT CLEANING

GRAMMAR CHECK

Using language tool python

RECTIFY SLANG

gud, nyc, gr8,

NEXT →

TEXT CLEANING

GRAMMAR CHECK

Using language tool python

CONTRACTOR

I'm --> Im

RECTIFY SLANG

gud, nyc, gr8

SPELLING CORRECTION

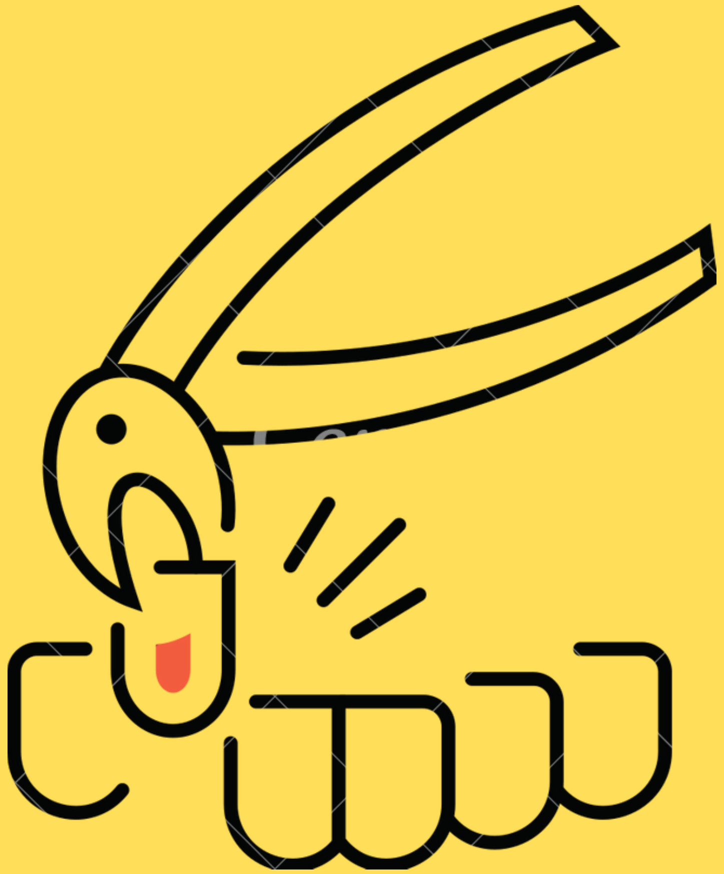
Prodct --> Product

PARSING HTML TAGS

Using html.parser

NEXT →

NOUN EXTRACTION AND PATTERN MINING



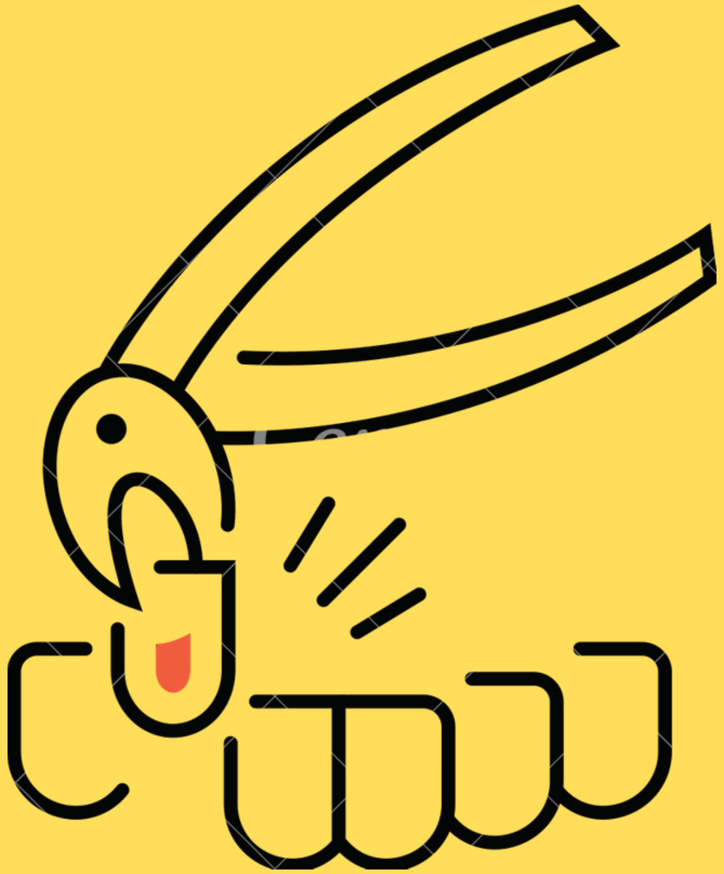
WHY NOUNS?

POS TAGGING

Bag of Nouns

NEXT →

NOUN EXTRACTION AND PATTERN MINING



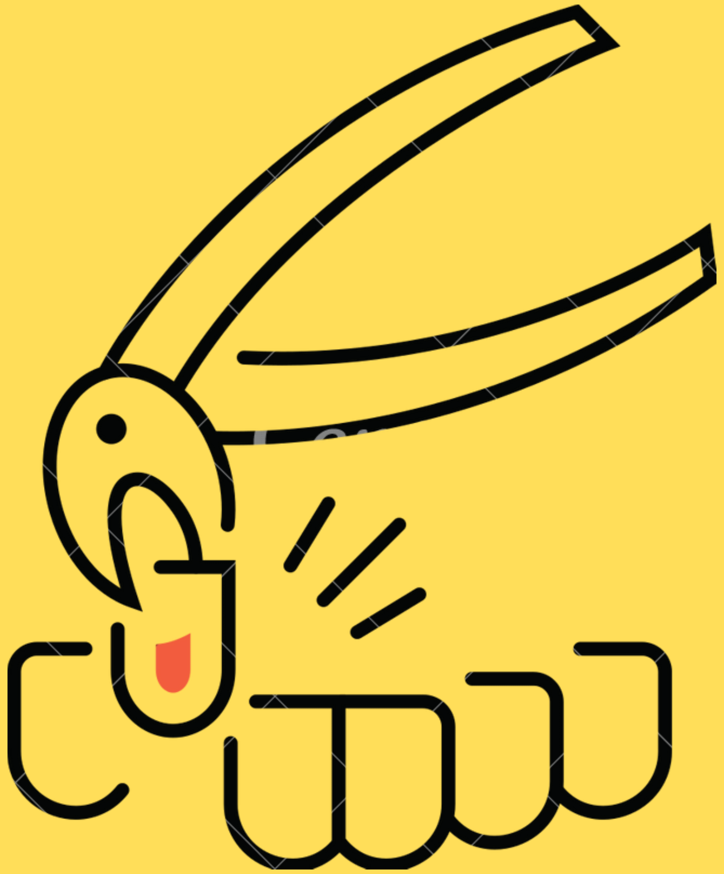
POS TAGGING

Bag of Nouns

APRIORI ALGORITHM VS FP GROWTH ALGORITHM (EXPERIMENT)

NEXT →

NOUN EXTRACTION AND PATTERN MINING



POS TAGGING

Bag of Nouns

APRIORI ALGORITHM VS FP GROWTH ALGORITHM (EXPERIMENT)

↓

FP GROWTH

NEXT →

INFORMATION EXTRACTION

USING NLP, SPACY AND POS TAGGING

PREPOSITION LINKAGE

product within range
installation of bed

ADVERB ADJECTIVE LINKAGE

very good product

ADJECTIVE-NOUN LINKAGE

good installation

ASPECT KEYWORDS

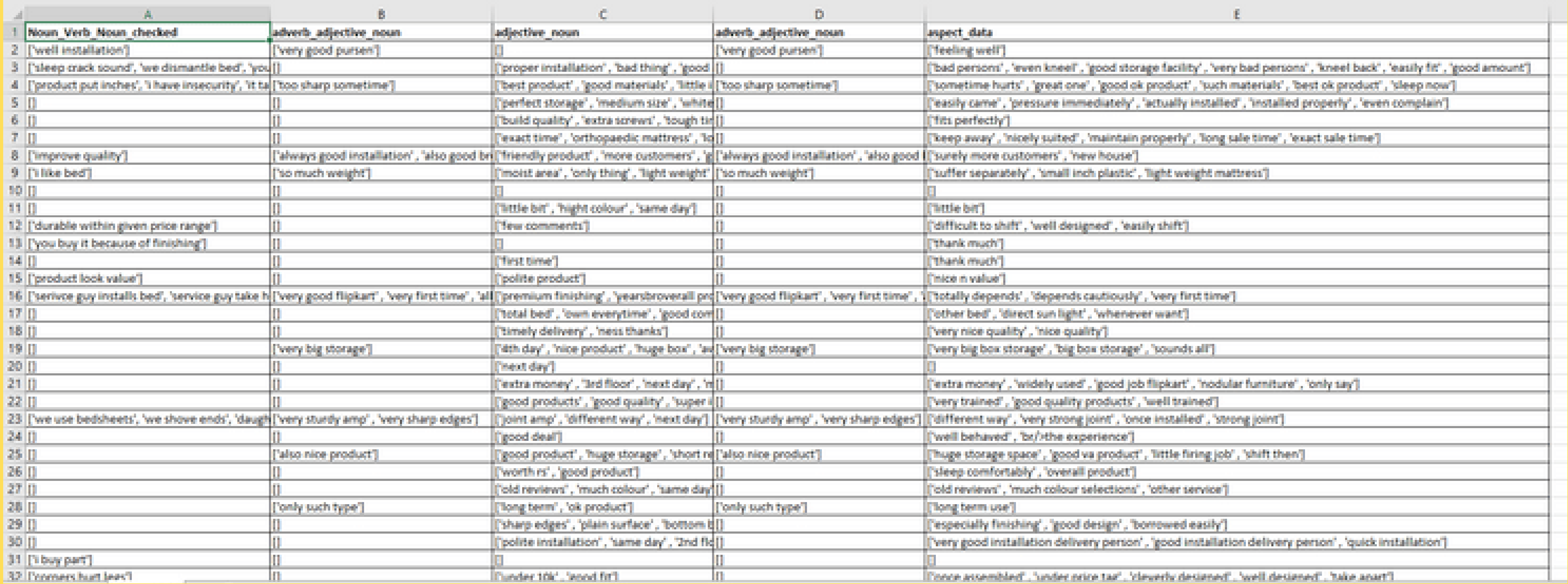
good material
extra screws

NOUN-VERB LINKAGE

improve quality
sharp edges

NEXT →

SNAPSHOT OF INFORMATION EXTRACTION



	A	B	C	D	E
1	**Noun_Verb_Noun_checked**	**adverb_adjective_noun**	**adjective_noun**	**adverb_adjective_noun**	**aspect_data**
2	['well installation']	['very good pursen']	[]	['very good pursen']	['feeling well']
3	['sleep crack sound', 'we dismantle bed', 'you	[]	['proper installation' , 'bad thing' , 'good	[]	['bad persons' , 'even kneel' , 'good storage facility' , 'very bad persons' , 'kneel back' , 'easily fit' , 'good amount']
4	['product put inches', 'i have insecurity', 'it ta	['too sharp sometime']	['best product' , 'good materials' , 'little i	['too sharp sometime']	['sometime hurts' , 'great one' , 'good ok product' , 'such materials' , 'best ok product' , 'sleep now']
5	[]	[]	['perfect storage' , 'medium size' , 'white	[]	['easily came' , 'pressure immediately' , 'actually installed' , 'installed properly' , 'even complain']
6	[]	[]	['build quality' , 'extra screws' , 'tough tir	[]	['fits perfectly']
7	[]	[]	['exact time' , 'orthopaedic mattress' , 'lo	[]	['keep away' , 'nicely suited' , 'maintain properly' , 'long sale time' , 'exact sale time']
8	['improve quality']	['always good installation' , 'also good br	['friendly product' , 'more customers' , 'g	['always good installation' , 'also good	['surely more customers' , 'new house']
9	['i like bed']	['so much weight']	['moist area' , 'only thing' , 'light weight'	['so much weight']	['suffer separately' , 'small inch plastic' , 'light weight mattress']
10	[]	[]	[]	[]	[]
11	[]	[]	['little bit' , 'hight colour' , 'same day']	[]	['little bit']
12	['durable within given price range']	[]	['few comments']	[]	['difficult to shift' , 'well designed' , 'easily shift']
13	['you buy it because of finishing']	[]	[]	[]	['thank much']
14	[]	[]	['first time']	[]	['thank much']
15	['product look value']	[]	['polite product']	[]	['nice n value']
16	['serivce guy installs bed', 'service guy take h	['very good flipkart' , 'very first time' , 'all	['premium finishing' , 'yearsbroverall pro	['very good flipkart' , 'very first time' , 'l	['totally depends' , 'depends cautiously' , 'very first time']
17	[]	[]	['total bed' , 'own everytime' , 'good com	[]	['other bed' , 'direct sun light' , 'whenever want']
18	[]	[]	['timely delivery' , 'ness thanks']	[]	['very nice quality' , 'nice quality']
19	[]	['very big storage']	['4th day' , 'nice product' , 'huge box' , 'av	['very big storage']	['very big box storage' , 'big box storage' , 'sounds all']
20	[]	[]	['next day']	[]	[]
21	[]	[]	['extra money' , '3rd floor' , 'next day' , 'n	[]	['extra money' , 'widely used' , 'good job flipkart' , 'nodular furniture' , 'only say']
22	[]	[]	['good products' , 'good quality' , 'super i	[]	['very trained' , 'good quality products' , 'well trained']
23	['we use bedsheets', 'we shove ends', 'daugh	['very sturdy amp' , 'very sharp edges']	['joint amp' , 'different way' , 'next day']	['very sturdy amp' , 'very sharp edges']	['different way' , 'very strong joint' , 'once installed' , 'strong joint']
24	[]	[]	['good deal']	[]	['well behaved' , 'br/>the experience']
25	[]	['also nice product']	['good product' , 'huge storage' , 'short re	['also nice product']	['huge storage space' , 'good va product' , 'little firing job' , 'shift then']
26	[]	[]	['worth rs' , 'good product']	[]	['sleep comfortably' , 'overall product']
27	[]	[]	['old reviews' , 'much colour' , 'same day'	[]	['old reviews' , 'much colour selections' , 'other service']
28	[]	['only such type']	['long term' , 'ok product']	['only such type']	['long term use']
29	[]	[]	['sharp edges' , 'plain surface' , 'bottom b	[]	['especially finishing' , 'good design' , 'borrowed easily']
30	[]	[]	['polite installation' , 'same day' , '2nd flo	[]	['very good installation delivery person' , 'good installation delivery person' , 'quick installation']
31	['i buy part']	[]	[]	[]	[]
32	['corners hurt legs']	[]	['under 10k' , 'good fit']	[]	['once assembled' , 'under price tag' , 'cleverly designed' , 'well designed' , 'take apart']

FEATURE SELECTION

TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

TF-IDF is the multiplication of the TF and IDF

10 Features : money, service, installation, delivery, wood, storage, design, quality, bed, product

NEXT →

FEATURE SELECTION

TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

TF-IDF is the multiplication of the TF and IDF

10 Features : money, service, installation, delivery, wood, storage, design, quality, bed, product

PHRASES FEATURES USING FP GROWTH ALGORITHM

price range, bed size, product installation, product design etc.

NEXT →

FEATURE SELECTION

TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

TF-IDF is the multiplication of the TF and IDF

10 Features : money, service, installation, delivery, wood, storage, design, quality, bed, product

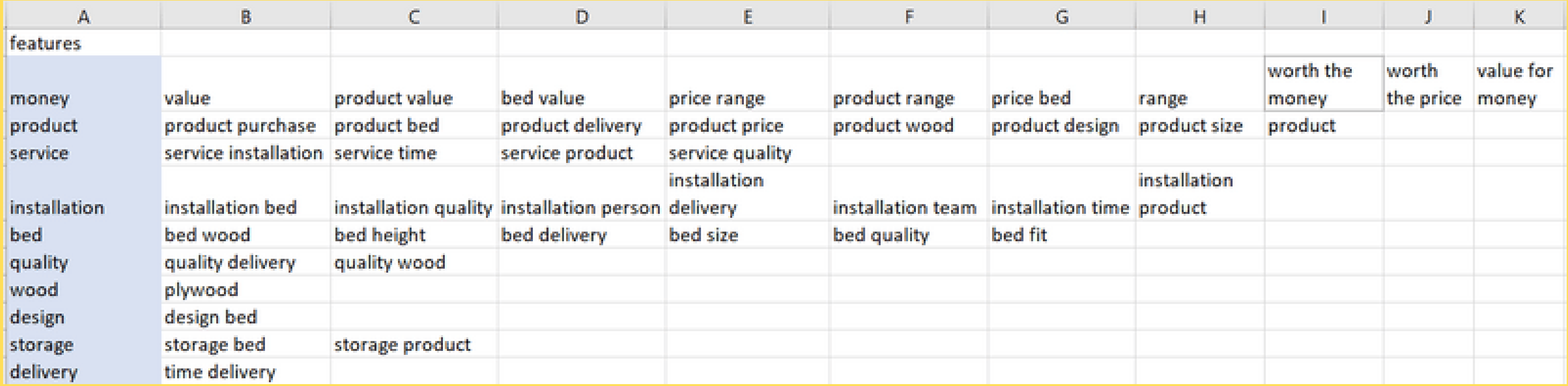
PHRASES FEATURES USING FP GROWTH ALGORITHM

price range, bed size, product installation, product design etc.

CLUBBED PHRASES VS PHRASE FEATURES (EXPERIMENT)

NEXT →

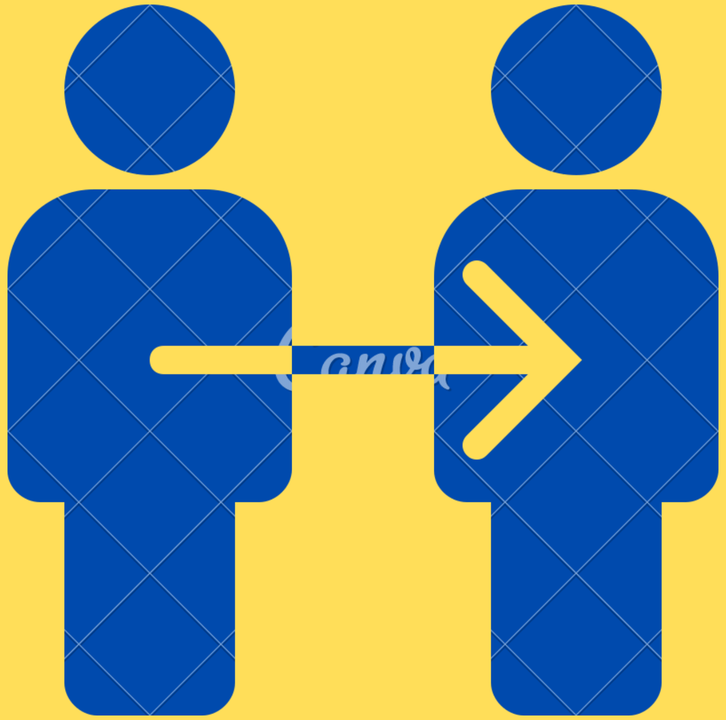
EXPERIMENT RESULT



A	B	C	D	E	F	G	H	I	J	K
features										
money	value	product value	bed value	price range	product range	price bed	range	worth the money	worth the price	value for money
product	product purchase	product bed	product delivery	product price	product wood	product design	product size	product		
service	service installation	service time	service product	service quality						
installation	installation bed	installation quality	installation person	installation delivery	installation team	installation time	installation product			
bed	bed wood	bed height	bed delivery	bed size	bed quality	bed fit				
quality	quality delivery	quality wood								
wood	plywood									
design	design bed									
storage	storage bed	storage product								
delivery	time delivery									

42 PHRASED FEATURES CLUBBED INTO 10 MAIN FEATURES

SIMILAR WORDS



BED --> COT
BED --> FURNITURE

WORD 2 VEC

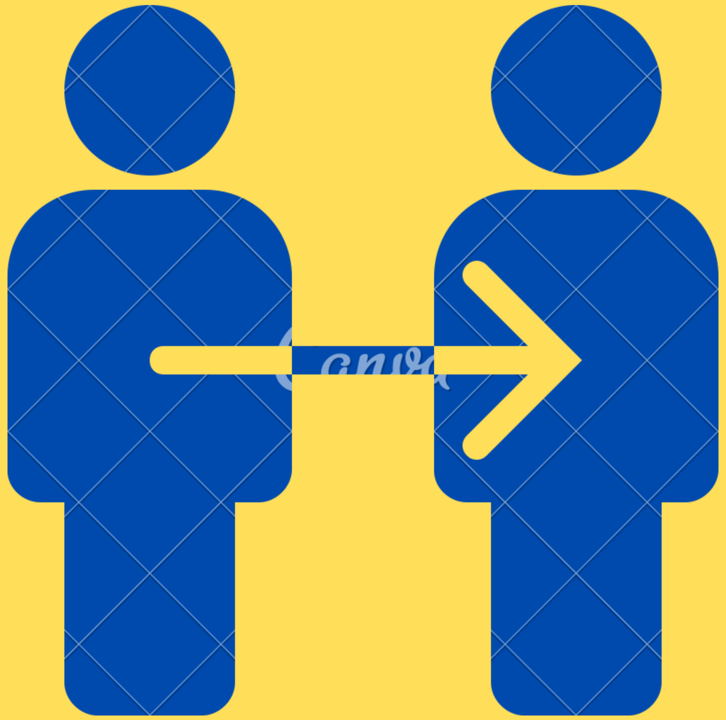
LEVENSHTEIN DISTANCE

NEXT →

LEVENSHTEIN DISTANCE

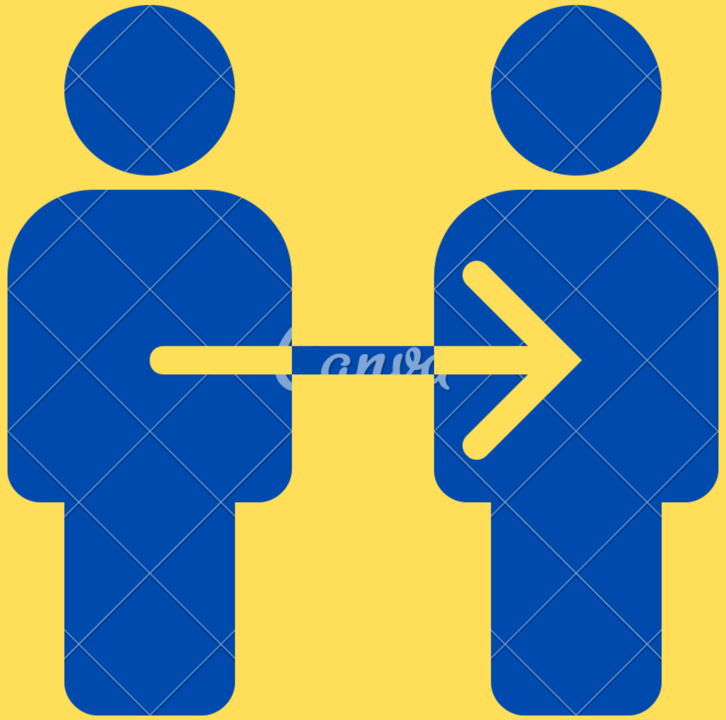
The Levenshtein distance is a number that tells you **how different two strings are.** The higher the number, the more different the two strings are.

For example, the Levenshtein distance between “kitten” and “sitting” is 3 since, at a minimum, 3 edits are required to change one into the other.



NEXT →

SIMILAR WORDS



BED --> COT
BED --> FURNITURE

WORD 2 VEC

LEVENSHTEIN DISTANCE

COSINE SIMILARITY

WEB SCRAPPING

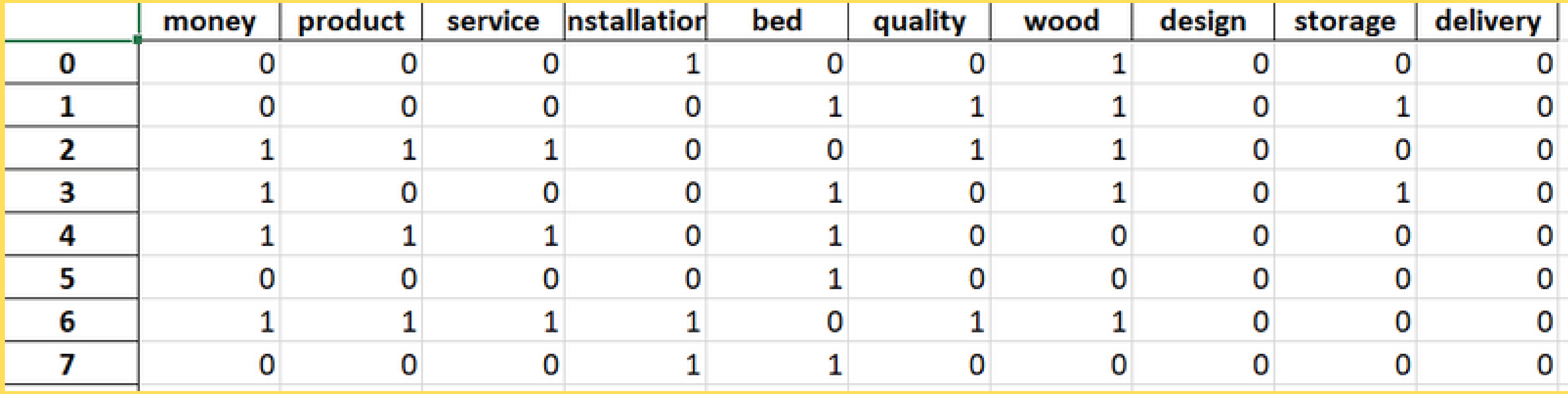
from synonyms.com

NEXT →

POLARITY AND BINARY DATA PREPARATION

FUZZYWUZZY LIBRARY
AND
COSINE SIMILARITY

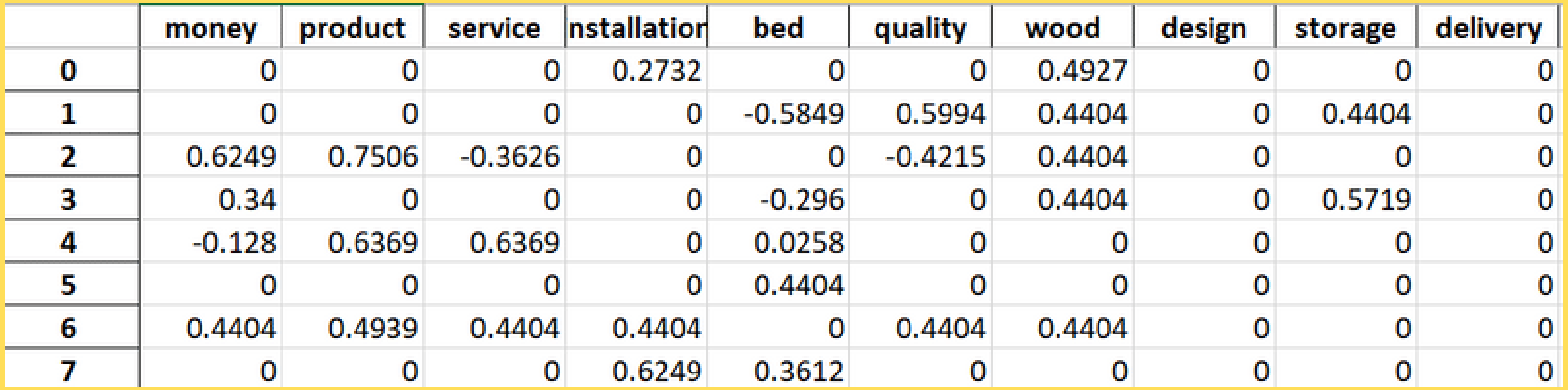
REVIEW ID



	money	product	service	nstallation	bed	quality	wood	design	storage	delivery
0	0	0	0	1	0	0	1	0	0	0
1	0	0	0	0	1	1	1	0	1	0
2	1	1	1	0	0	1	1	0	0	0
3	1	0	0	0	1	0	1	0	1	0
4	1	1	1	0	1	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0
6	1	1	1	1	0	1	1	0	0	0
7	0	0	0	1	1	0	0	0	0	0

SENTIMENT EVALUATION TECHNIQUE

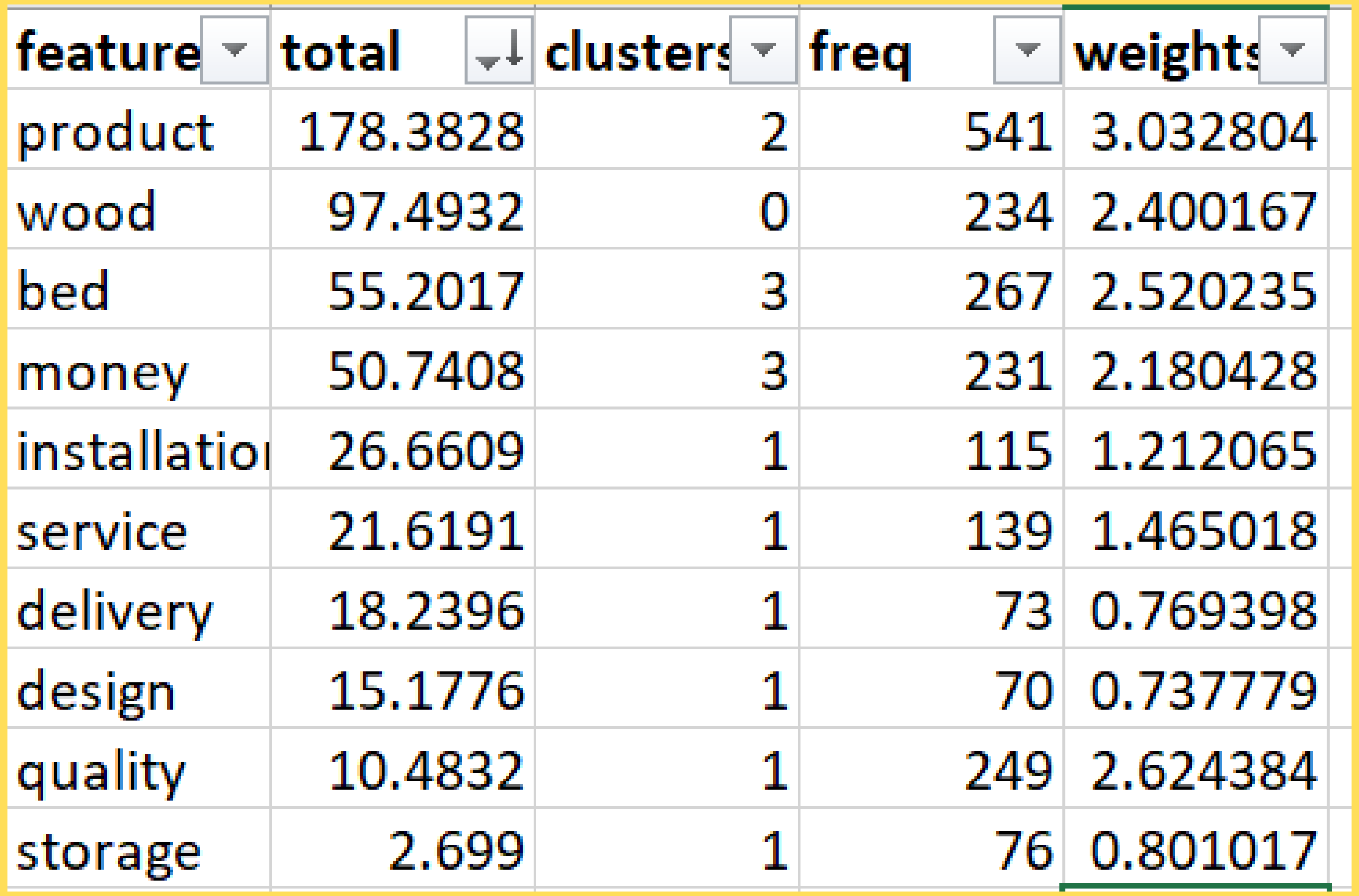
VADER SENTIMENT



	money	product	service	nstallation	bed	quality	wood	design	storage	delivery
0	0	0	0	0.2732	0	0	0.4927	0	0	0
1	0	0	0	0	-0.5849	0.5994	0.4404	0	0.4404	0
2	0.6249	0.7506	-0.3626	0	0	-0.4215	0.4404	0	0	0
3	0.34	0	0	0	-0.296	0	0.4404	0	0.5719	0
4	-0.128	0.6369	0.6369	0	0.0258	0	0	0	0	0
5	0	0	0	0	0.4404	0	0	0	0	0
6	0.4404	0.4939	0.4404	0.4404	0	0.4404	0.4404	0	0	0
7	0	0	0	0.6249	0.3612	0	0	0	0	0

NEXT →

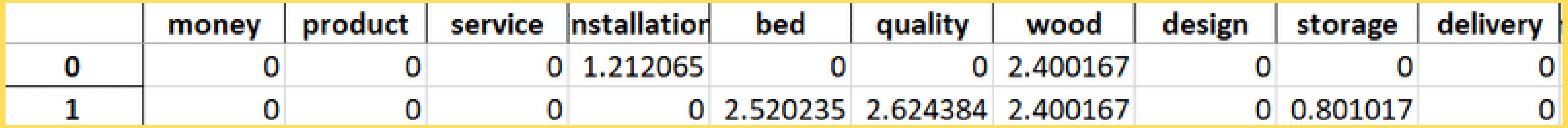
CALCULATION OF WEIGHTS AND CLUSTERING



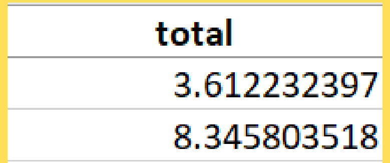
feature	total	clusters	freq	weights
product	178.3828	2	541	3.032804
wood	97.4932	0	234	2.400167
bed	55.2017	3	267	2.520235
money	50.7408	3	231	2.180428
installatio	26.6609	1	115	1.212065
service	21.6191	1	139	1.465018
delivery	18.2396	1	73	0.769398
design	15.1776	1	70	0.737779
quality	10.4832	1	249	2.624384
storage	2.699	1	76	0.801017

BINARY DATA FILE

Weight Wix for the feature fi (Wix) = (frequency of fi in Cx) / (sum of the frequencies of features fi in Cx)



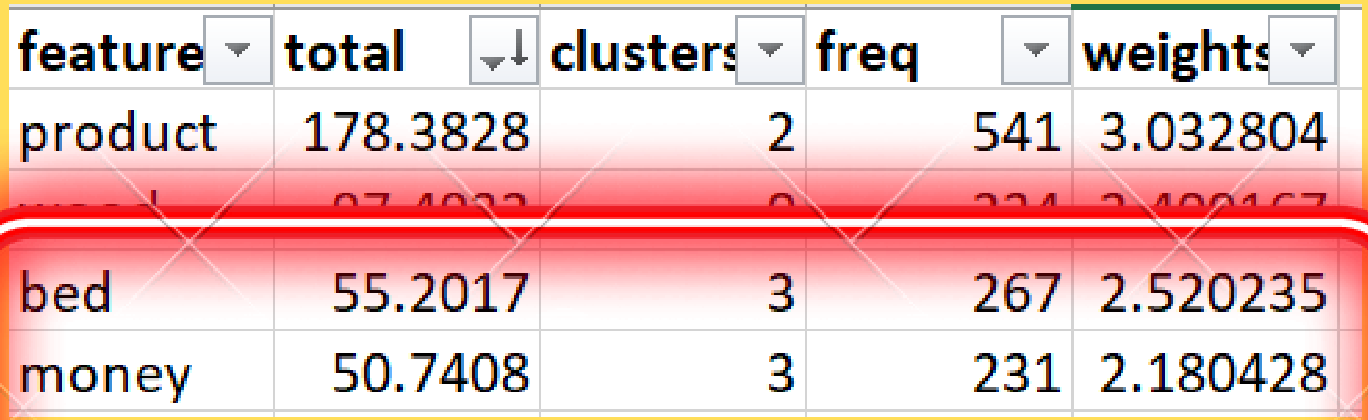
	money	product	service	nstallatior	bed	quality	wood	design	storage	delivery
0	0	0	0	1.212065	0	0	2.400167	0	0	0
1	0	0	0	0	2.520235	2.624384	2.400167	0	0.801017	0



total
3.612232397
8.345803518

NEXT →

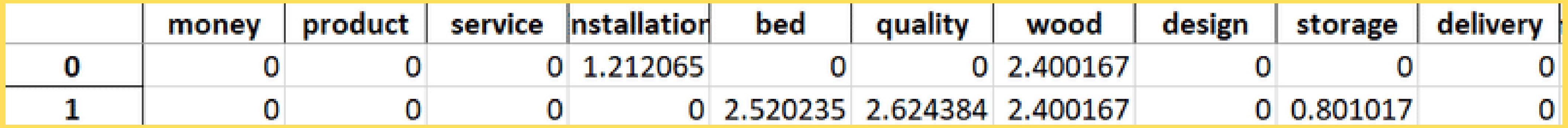
CALCULATION OF WEIGHTS AND CLUSTERING



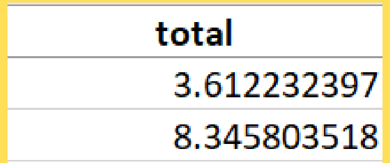
feature	total	clusters	freq	weights
product	178.3828	2	541	3.032804
wood	97.4922	2	224	2.400167
bed	55.2017	3	267	2.520235
money	50.7408	3	231	2.180428

BINARY DATA FILE

Weight Wix for the feature fi (Wix) = (frequency of fi in Cx) / (sum of the frequencies of features fi in Cx)



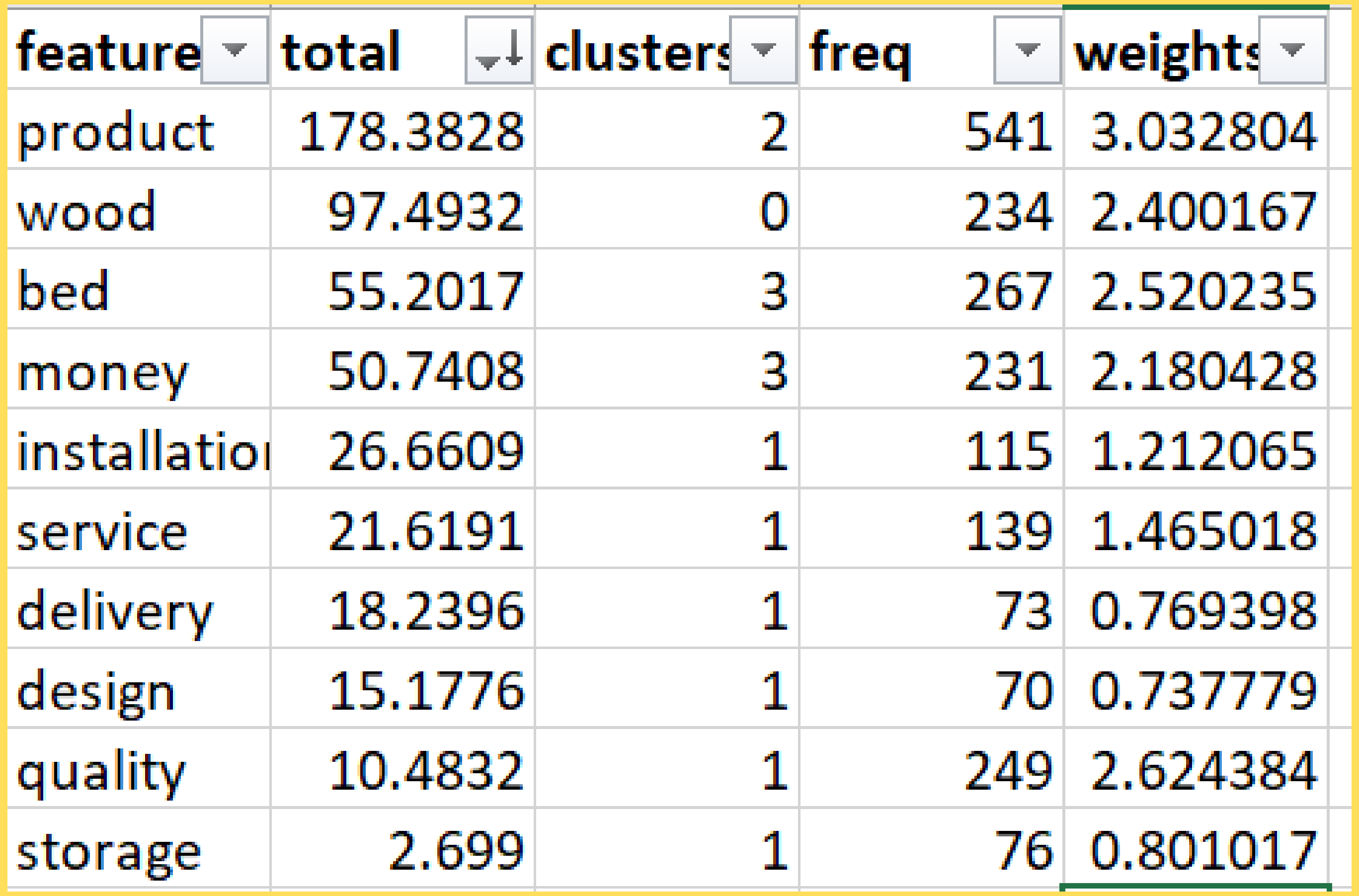
	money	product	service	nstallatior	bed	quality	wood	design	storage	delivery
0	0	0	0	1.212065	0	0	2.400167	0	0	0
1	0	0	0	0	2.520235	2.624384	2.400167	0	0.801017	0



total
3.612232397
8.345803518

NEXT →

CALCULATION OF WEIGHTS AND CLUSTERING

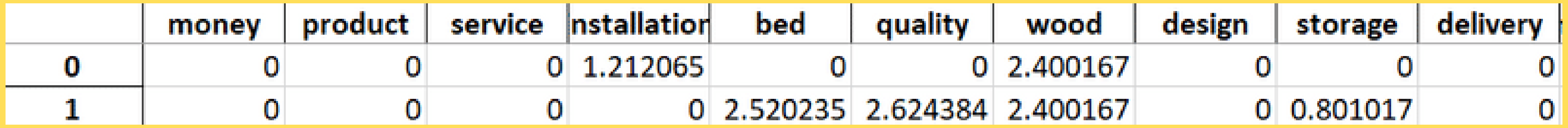


feature	total	clusters	freq	weights
product	178.3828	2	541	3.032804
wood	97.4932	0	234	2.400167
bed	55.2017	3	267	2.520235
money	50.7408	3	231	2.180428
installatio	26.6609	1	115	1.212065
service	21.6191	1	139	1.465018
delivery	18.2396	1	73	0.769398
design	15.1776	1	70	0.737779
quality	10.4832	1	249	2.624384
storage	2.699	1	76	0.801017

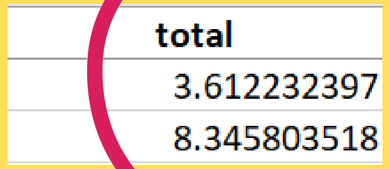
BINARY DATA FILE

Weight Wix for the feature fi (Wix) = (frequency of fi in Cx) / (sum of the frequencies of features fi in Cx)

TOTAL EVALUATION OF CONSUMER SENTIMENT



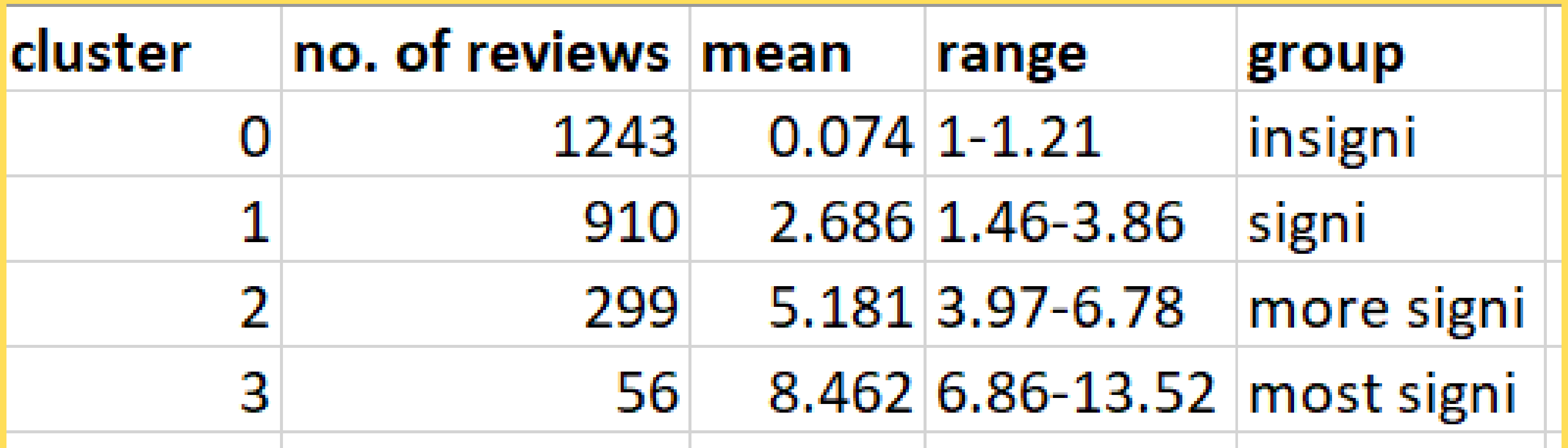
	money	product	service	nstallatior	bed	quality	wood	design	storage	delivery
0	0	0	0	1.212065	0	0	2.400167	0	0	0
1	0	0	0	0	2.520235	2.624384	2.400167	0	0.801017	0



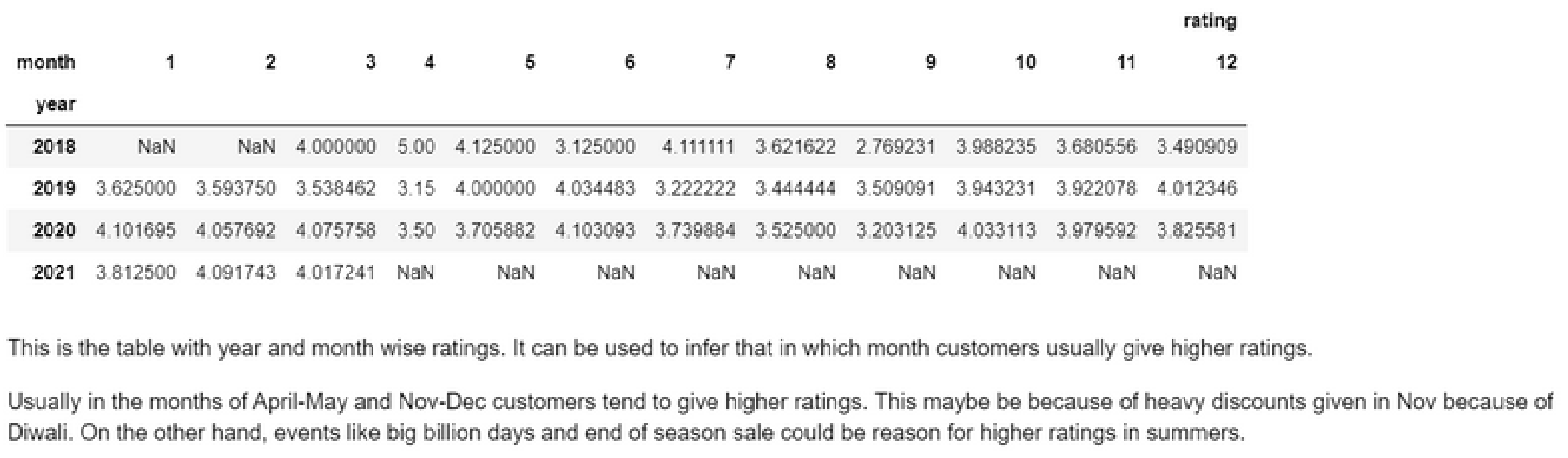
	total
	3.612232397
	8.345803518

NEXT →

Results



cluster	no. of reviews	mean	range	group
0	1243	0.074	1-1.21	insigni
1	910	2.686	1.46-3.86	signi
2	299	5.181	3.97-6.78	more signi
3	56	8.462	6.86-13.52	most signi



												rating
month	1	2	3	4	5	6	7	8	9	10	11	12
year												
2018	NaN	NaN	4.000000	5.00	4.125000	3.125000	4.111111	3.621622	2.769231	3.988235	3.680556	3.490909
2019	3.625000	3.593750	3.538462	3.15	4.000000	4.034483	3.222222	3.444444	3.509091	3.943231	3.922078	4.012346
2020	4.101695	4.057692	4.075758	3.50	3.705882	4.103093	3.739884	3.525000	3.203125	4.033113	3.979592	3.825581
2021	3.812500	4.091743	4.017241	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

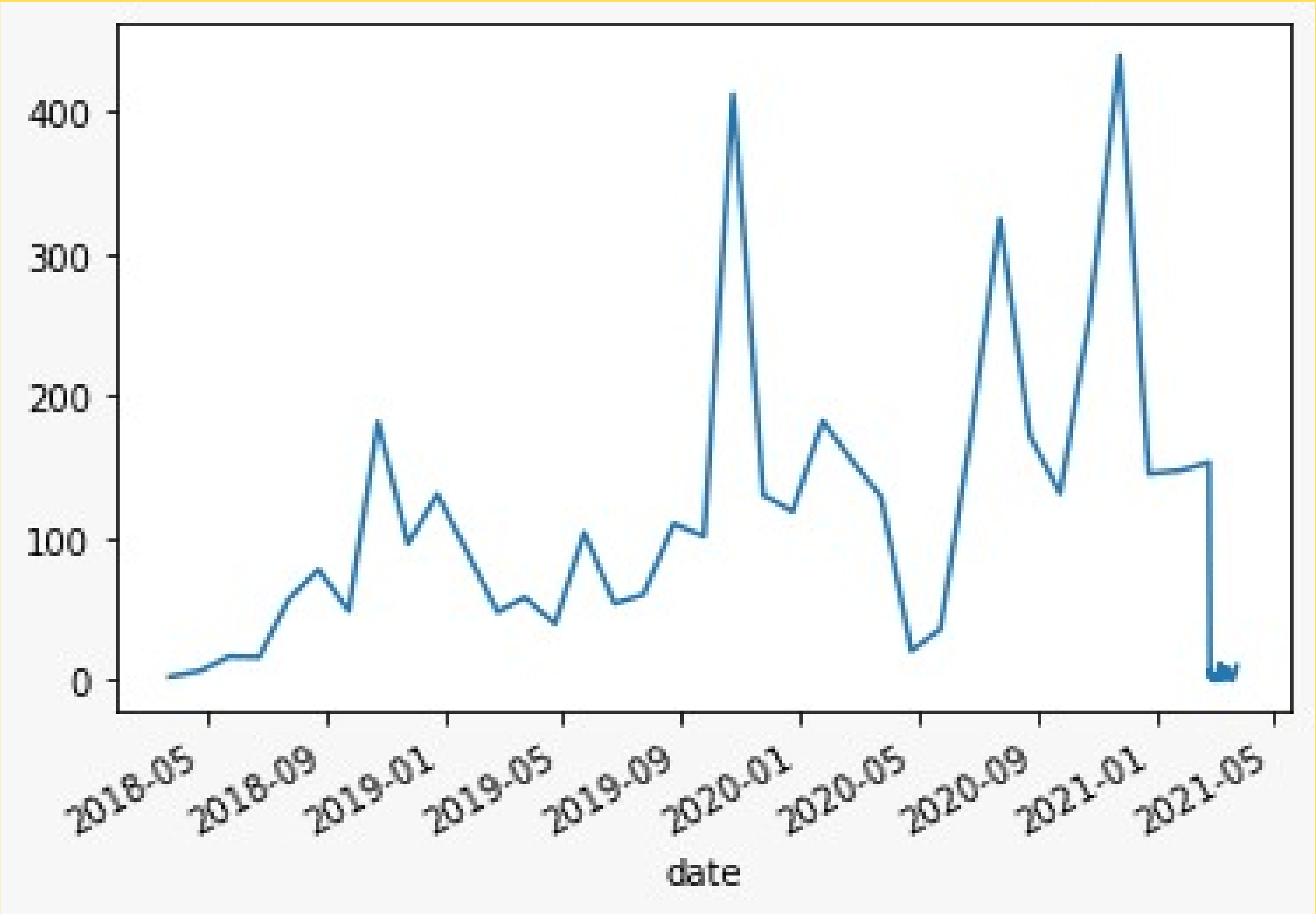
This is the table with year and month wise ratings. It can be used to infer that in which month customers usually give higher ratings.

Usually in the months of April-May and Nov-Dec customers tend to give higher ratings. This maybe be because of heavy discounts given in Nov because of Diwali. On the other hand, events like big billion days and end of season sale could be reason for higher ratings in summers.

OVERALL PRODUCT RATING

★★★★☆

MONEY

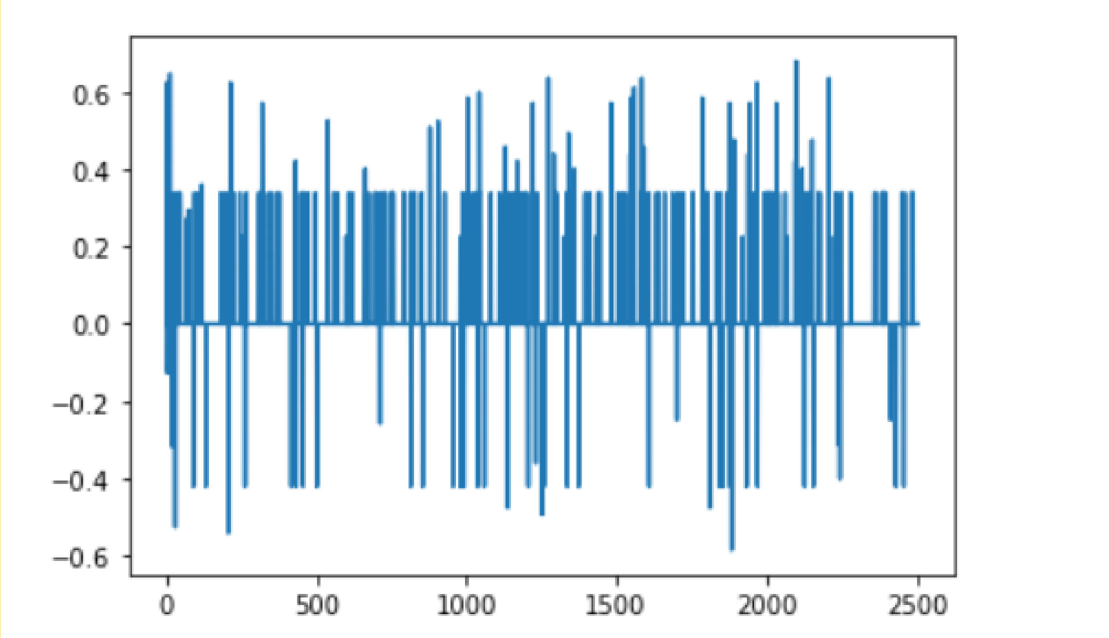


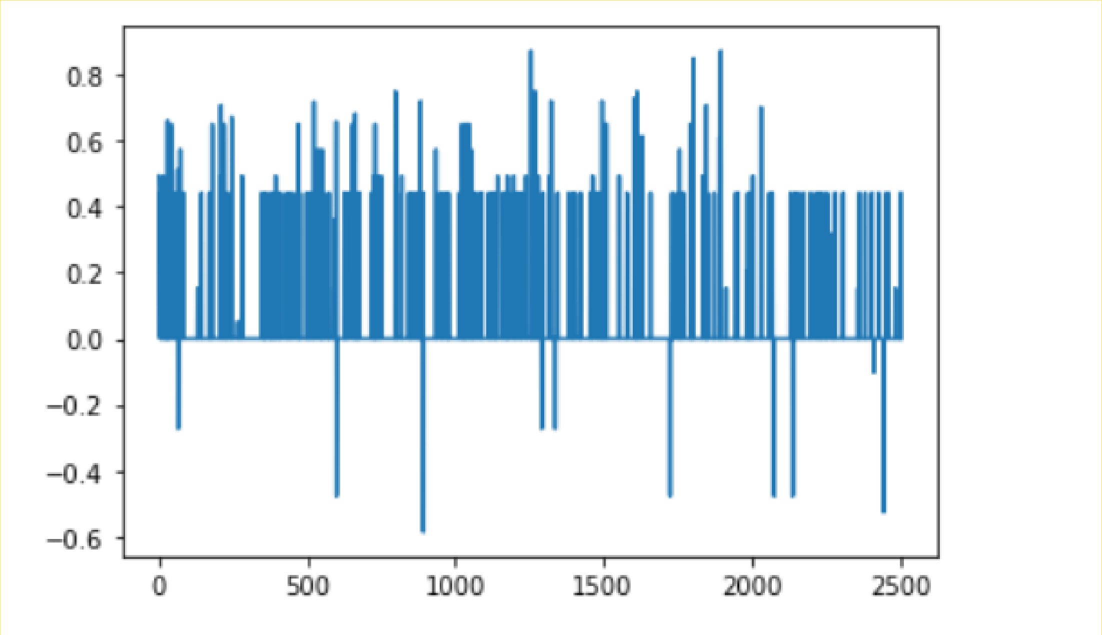
NEXT →

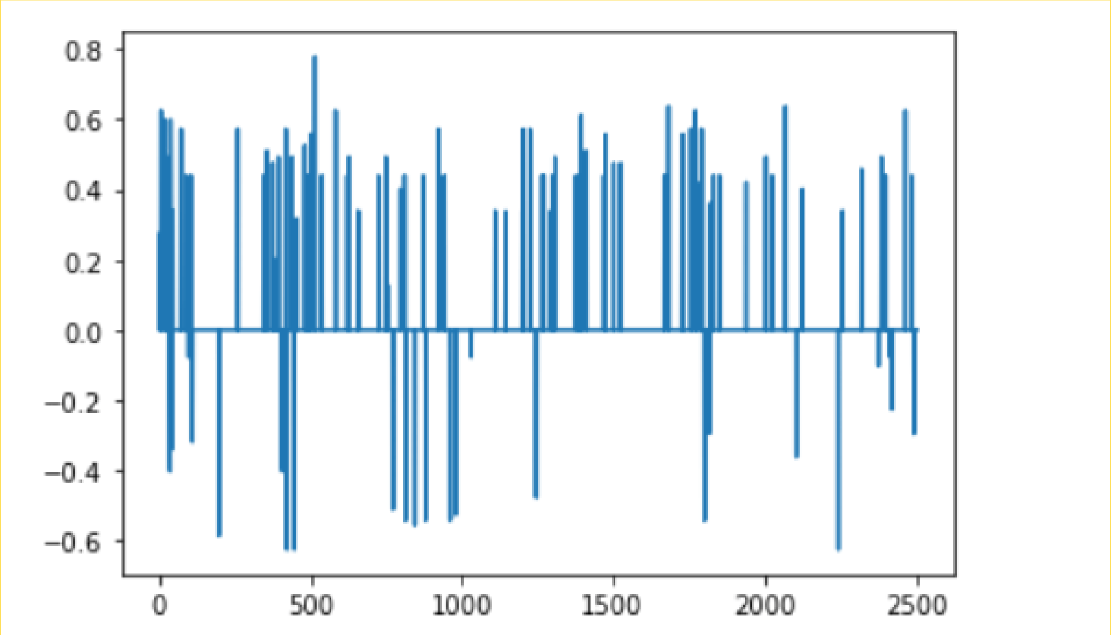
Features that have
positive polarity

Features that have
negative polarity

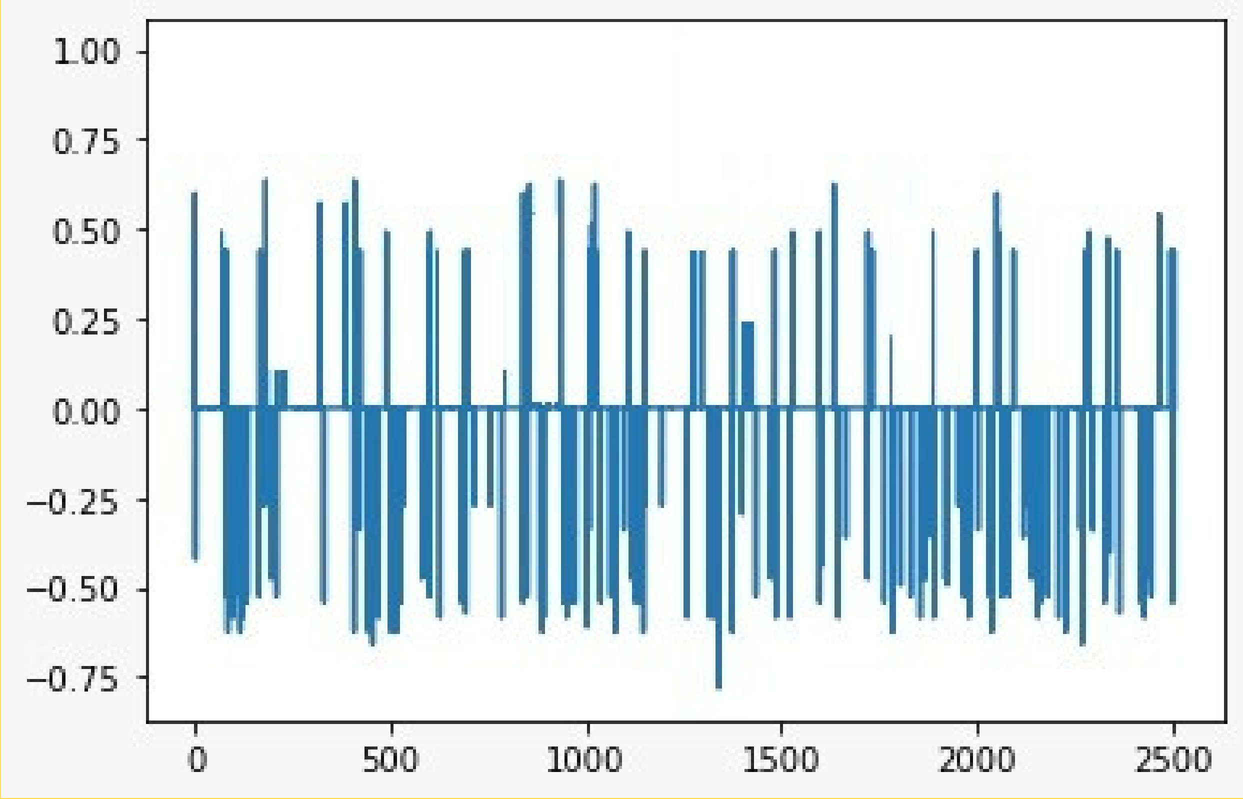
MONEY
INSTALLATION
WOOD







QUALITY



NEXT →

QUICK DEMO - DEPLOYMENT USING HTML+FLASK

Scope of Work

01

Pros

- **easy** to understand and implement
- computationally **cheap**
- **dynamic** in nature
- **reduced manual intervention**

Cons

- not feasible on newly launched products with **less reviews**
- not very effective on **poorly written reviews**
- ratings cannot be used since consumers **do not have a protocol to follow** while rating a product

NEXT →

Scope of Work

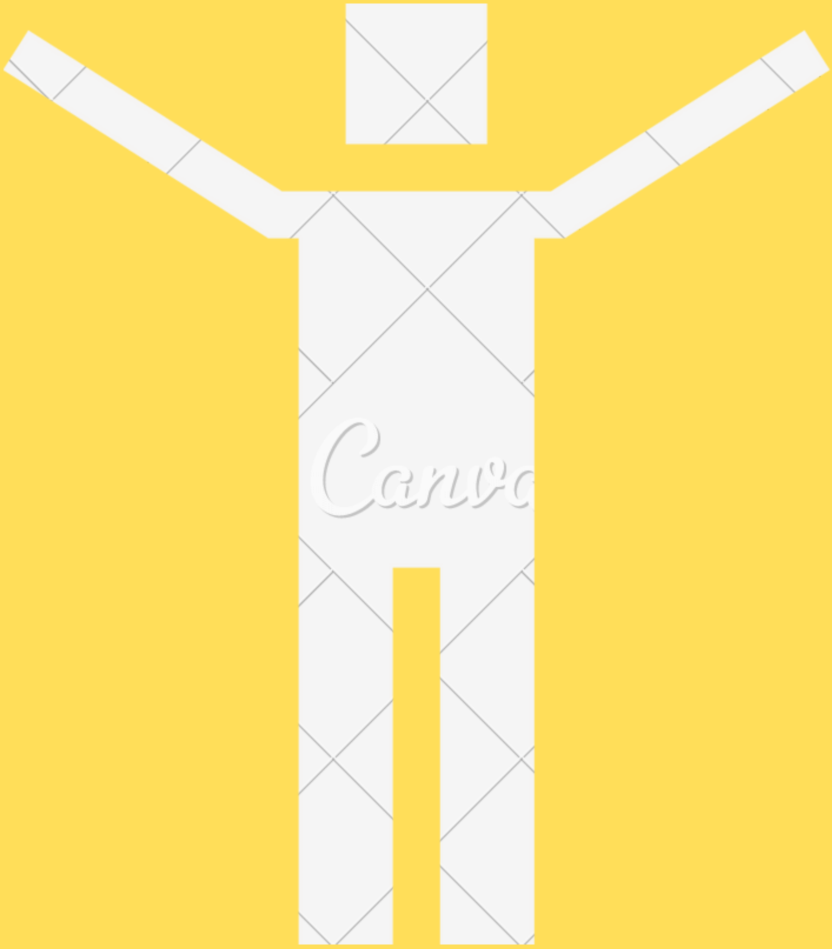
02

Win-Win for everyone

Clustering reviews makes it **easier for buyers** to them. On the other hand, **extracting information** from these reviews via EDA, Time Series, etc. will help sellers to make the product better than before!

NEXT →

Scope of Work



03

Wider Adaptability

The project design is **flexible** to be implemented on any product. It is **adaptive** to capture all consumer experiences.

NEXT →

Some Things That We Learnt

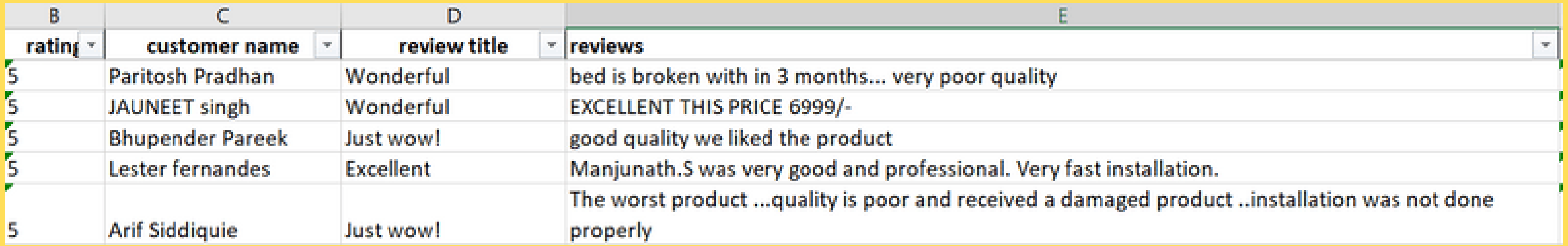
WHEN IN DOUBT, EXPERIMENT AND DECIDE

NEXT →

Some Things That We Learnt

WHEN IN DOUBT, EXPERIMENT AND DECIDE

EVEN THE SIMPLEST DATA CAN HELP A LOT



B	C	D	E
rating	customer name	review title	reviews
5	Paritosh Pradhan	Wonderful	bed is broken with in 3 months... very poor quality
5	JAUNEET singh	Wonderful	EXCELLENT THIS PRICE 6999/-
5	Bhupender Pareek	Just wow!	good quality we liked the product
5	Lester fernandes	Excellent	Manjunath.S was very good and professional. Very fast installation.
5	Arif Siddiquie	Just wow!	The worst product ...quality is poor and received a damaged product ..installation was not done properly

NEXT →

Some Things That We Learnt

WHEN IN DOUBT, EXPERIMENT AND DECIDE

EVEN THE SIMPLEST DATA CAN HELP A LOT

THINGS MIGHT NOT GO ACCORDING TO THE PLAN SO ONE SHOULD LEARN TO IMPROVISE

Thank You

GROUP 7

ABHIRUP SARKAR, NIPUN MOHINDRA, RUSHIKESH BADGUJAR, TANYA MANGATH AND VAMSITEJ GADIVEMUELA