Tauro, Bruce H00228269 – Salah, Tamer H00343334 – Itani, Tarek H00292565 - Serry, Mohamed H00313456

**29th February**

20



08

**Fall**

F21AA Coursework 1

Table of Contents

[Table of Contents 2](#_Toc33902355)

[Introduction 2](#_Toc33902356)

[Q1 Data Exploration and Visualization 3](#_Toc33902357)

[Q2 Text Processing and Normalization 6](#_Toc33902358)

[Q3 Vector space Model and feature representation 7](#_Toc33902359)

[Q4 Model training, selection and hyperparameter tuning and evaluation 8](#_Toc33902360)

[Q5 Topic Modelling of high and low ratings 9](#_Toc33902361)

[Conclusions 11](#_Toc33902362)

[Appendix 12](#_Toc33902363)

# Introduction

# Q1 Data Exploration and Visualization

For this task we began by exploring our training data set. It has 10 attributes and 426,340 examples.

The **first** **task** was understanding the data types:

Data columns (total 10 columns):

Id 426340 non-null int64

ProductId 426340 non-null object

UserId 426340 non-null object

ProfileName 426326 non-null object

HelpfulnessNumerator 426340 non-null int64

HelpfulnessDenominator 426340 non-null int64

Score 426340 non-null int64

Time 426340 non-null int64

Summary 426320 non-null object

Text 426340 non-null object

As seen from above, we have 6 numerical attributes and 5 categorical or non-numeric attributes in total with Score as our class attribute.

The **second task** was the **statistical analysis of our data set** [Appendix A], [Appendix B] showed us that **most of the data is biased towards the score 5.0** that means our examples are not **stratified** or evenly distributed, we began to analyse the data considering the score distributions. As seen from [Appendix C], **272492** examples have a class value of 5.0. this over **50% of our data.**

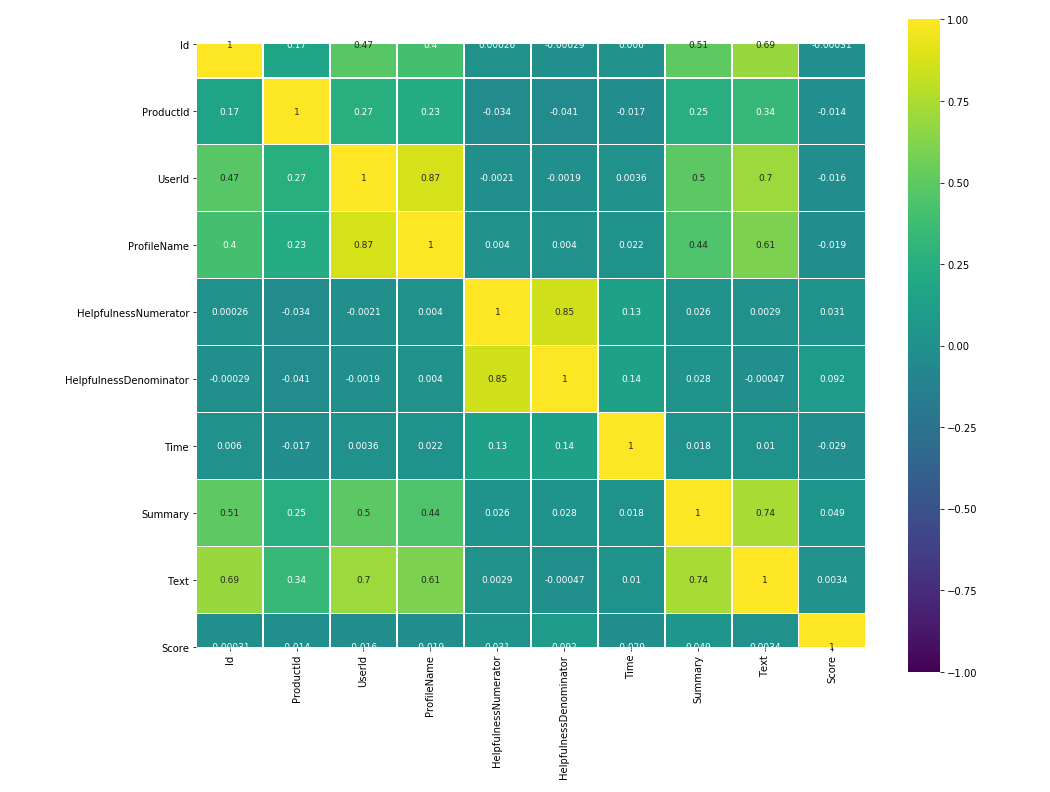
A screenshot of a cell phone

Description automatically generatedOur next step was split into 2 stages: numerical correlation and categorical correlation. We first began by pair plotting our numerical attributes with our score to see if we can find any interesting relationships appear.

A close up of text on a white background

Description automatically generatedThe above plots gave no indicative relationship, calculating the correlation values of each of the numerical attributes with our class attribute [Appendix D] shows the correlation values confirm that the plots have no distinct correlation with the class, factorizing the categorical attributes allowed us to calculate a correlation score.

The **inter-attribute** correlation matrix provided some interesting results, the **text is highly positively correlated with product name and summary while the summary is highly positively correlated with text as well.** From this we concluded that among the categorical attributes **text** and **summary** seem to be strong attributes. We then plotted a full correlation matrix to verify our assumptions.



From the correlation matrix, we discovered that none of the attributes had a strong correlation to the class value. Instead we found more evidence that the best candidates are the **Text** and **Summary** attributes which we merged. In the next stage of our analysis we began by exploring the **values** in our data set, first we searched for null values locating some in the **summary and profile name** attributes [Appendix E].

Duplicate reviews were found in the data set; those were removed to avoid noise in our data. [Appendix F]

After further analysis we found the following undesirable properties in our text attributes:

1. Html tags
2. Accents
3. Punctuation

# Q2 Text Processing and Normalization

We divided this into 2 stages **Text processing** and **Data** **Normalization**, we found *HTML tags* and *punctuations* in the Corpus that we removed using a function we created called **tokenizer**, it is likely that HTML tags and punctuation will not add any value to the model accuracy , on the contrary they will have a negative impact on the accuracy and execution speed of our model as they do not contribute to the text.

**Data** **Normalization** involved **stemming** and **lemmatization**, these are two common techniques used for text normalization, stemming cuts the word into a shorter term or representation without looking into its lexical meaning of the word, lemmatization does the same thing by simplifying the word to a simpler lexical word. We used the **Port Stemmer** and **Worldnet Lemmatizer** from the *NLTK* library.

In [Appendix G] you can an example of a corpus document after applying our processing and normalization functions. We noticed several pros and cons for each technique, stemming in many cases gave a shorter representation but the word lost its meaning, lemmatization gave a better more appropriate word however it was not consistent.

To prove which technique is better for our corpus, we did a comparison using a Pipeline of CountVectorizer, TFIDF Transformer and then **LogisticRegression** , this was done without the use of Grid Search to fix all variables, the input corpus was tokenized removing punctuation and HTML tags. The kernel was restared after each experiment to prevent model retraining.

In our Experiment, the use of Lemmatization or Stemming decreased our overall model score, slightly: [Appendix H]

|  |  |
| --- | --- |
| Technique | Score |
| None | **0.761** |
| Stemming | **0.756** |
| Lemmatization | **0.756** |

# Q3 Vector space Model and feature representation

For this segment of the course work we aimed to experiment with **count-based** representations, we utilized **term frequency** and **inverse-document term frequency** to evaluate the relative importance of each word or feature.

We represented each review as a matrix of **token counts** through the process called Count Vectorization.

<https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html>

Term frequency is simply accounting for the number of times a term (a word) occurs in each corpus (collection of documents) it does not differentiate between the context of the term or the document it is located in.

tf(*t*,*d*) = *ft*,*d*

The relative weight of each term is its relative frequency which is calculated by dividing the term frequency by the cumulative frequency.

In contrast inverse document frequency inverts the frequency of the term in the documents Under the assumption that the most common words would be the least indicative of the documents overall meaning.

Idff=log10(N/dff)

When used in combination turn frequency inverse document frequency indicates the relative weight or importance of a term regarding the entire corpus of documents. This is crucial in feature extraction as it identifies the most important terms in the corpus that can provide insight into the documents overall meaning.

[Appendix I] shows a sample of the output of the term frequency matrix representation of the data. Each row represents a document in the corpus in our case a review with each column representing a term and its frequency.

[Appendix J] shows a sample of the output of the TF-IDF matrix representation of the data. In review we conclude that in the context of food reviews TF-IDF is a more informative vector space representation then turn frequency as it indicates the relative value of each word regarding our corpus of reviews.

Furthermore, we experimented with n-grams. By default, unigrams were used to represent features then we attempted to try bigrams and trigrams and quad grams.

[Appendix K]

As seen from the above examples, n-grams affected …..

# Q4 Model training, selection and hyperparameter tuning and evaluation

In this section we were trying to figure out the best model with the best parameters possible, to do that we have combined **Pipeline** with **the Grid Search** where the Pipeline would be using CountVectorizer and TFIDF transformer and one of 3 estimators

LogisticRegression

MultinomialNB

SGDClassifier

**Pipeline** is a process that enabled us to create all the steps needed for our text analytics model to work starting from text vectorization, TF-IDF and then applying our statistical model. We have also used **GridSearch** which enabled us to test several parameters using several setting combinations and compare results in one run. We tested multiple combination of models, estimators and parameters, using dictionaries to load estimators and the parameters, this enabled us to add or remove estimators and parameters as needed.

The drawback of that process is it was a computationally intensive process and complex to scale up we went through several iterations of processor and memory optimization to make the code work on larger servers ( the use of **n\_jobs** on several cores lead to known symptom of *memory explosion* , we had to optimize using **pre\_dispatch** to control the number of jobs that get dispatched during parallel execution).

Our experiments revealed that many of the parameters did not contribute positively or impact the results, in fact in many cases using the default parameters were the best option.

[Appendix L] shows a sample of the model output using *LogisticRegression* with a GridSearch n-gram of (1,1), (1,2), (1,3)

# Q5 Topic Modelling of high and low ratings

In this section we use topic modelling to find common threads between reviews of score 5 and reviews of score 1, with the aim of discerning any patterns in the terms that contribute to the topic.

We explored several techniques for topic modelling namely NMF, SVD and LDA.

Non-negative matrix factorization

The NMF model decomposes its input matrix into two smaller approximate product matrices that only contain nonnegative values these are iteratively adjusted until they more closely result into the input matrix due to this process the features are clustered as the error value is reduce during each iteration.

NMF can take input matrices that have been processed by both term frequency and TF-IDF

Singular value decomposition

The SVD model decomposes the input matrix into its constituent parts in the form of 3 matrices. SVD acts as a feature reducer removing terms that are not important to the overall corpus.

Latent Dirichlet Allocation

LDA assumes that all topics follow a Dirichlet distribution across the documents in the corpus this leads to the probabilities of context between words being preserved. LDA groups together terms that occur together often into a topic which at times may not lead to topical grouping.

Observations

Appendix M

NMF TF

When using NMF with a TF input there is a clear trend in the topics it is very easy to find an overarching commonality between the terms and possible “topic header” for example topic 9 of the score 5 group and topic 11 of the score 1 group are both about hair products. There is also a trend in the use of positive words in the Score 5 group and negative words in the Score 1 group, while the words No and Not appear in the Score 5 Group more telling words like bad, disappointment, weak and burn are seen in the Score 1 Group in comparison Score 5 group has no such words instead words like loves, like, quality and favourite are seen but some positive terms are seen in Score 1’s topics.

NMF IDF

The results of NMF IDF are similar to those seen in NMF TF with terms being grouped in similar topic groups. Score 5 group Topic 2 of NMF models are almost the same.

SVD

SVD’s topics tends to have several improperly assigned terms in the Score 5 group topic 16 which seems to cover to topic of beverages has a out of place term “dog” this improperly assigned terms occurs in all topics with some topics having no clear topic.

LDA

LDA has mixed results there is no clear positive or negative arrangement of terms in the Score groupings. Some topics like Score 5 group topic 0 is clearly about Oil while theory topics like topic 0 of Score 1 Group has many unrelated terms in them making topic assignment unclear.

Overall Observation

The NMF model has the most coherent results all topics have a clear theme and the Score groups do not have any contamination of positive and negative terms. LDA picks up word groups that occur commonly with one another leading to mis grouped terms while SVC approximates the groups very loosely.

# Conclusions

In this course work the team collectively got together trying to solve the IMDB problem using statistical Natural Language Processing techniques , we have to admit that the questions and their chronological sequence helped shape our thinking to come out with a good working framework to solve this challenge , to the best of our ability.

As we started to understand the data that we were given , visualizing it gave us valuable insights on the action plan that we took during the coursework, we became very clear on the data bias toward a specific score, the relevance of some features , and the diversity and magnitude of normalization that we need to do (HTML Tags, null values, duplicates. Etc), which we did in question 2 and we have documented several functions in our code to deal with those issues.

After normalizing the data, we did a research on the best text vectorization technique we can use for this data, we experimented several techniques and several libraries to reach an optimal vectorization technique, we did document our findings and also tested them with our model in later questions.

On then, we tried to came up with a method to apply some of the techniques we have learned during the course to come out with a appropriate model , from estimators selections, Pipelines generation…etc, We tested that on the validation data set as a true measure of our model effectiveness

We finished this by applying Topic Modeling to the data and generating further insights, you will find our all our research summary detailed in section 6.

Note from the team: There were a great deal of learning and collaboration during the coursework, from applying principals we have learned in the class and researching things helped us conclude what is presented in this document. It was a great learning and experience for all team members

# Appendix

Appendix A – Statistical Table

A screenshot of a cell phone

Description automatically generated

Appendix B – Histogram, per column

A close up of a device

Description automatically generated

Appendix C – Count group by Score

A screen shot of a social media post

Description automatically generated

Appendix D – Correlation Scores

HelpfulnessDenominator: -0.07724218886413477

Time: -0.06318610909302598

HelpfulnessNumerator: -0.022805615821503325

Id: 0.009985657224340675

Appendix E – null Values

Id 0

ProductId 0

UserId 0

ProfileName 14

HelpfulnessNumerator 0

HelpfulnessDenominator 0

Time 0

Summary 20

Text 0

Score 0

Appendix F – Outliers

count 426340.000000

mean 435.396902

std 443.943421

min 12.000000

25% 179.000000

50% 301.000000

75% 526.000000

max 21409.000000

Name: length, dtype: float64

A close up of a logo

Description automatically generated

Appendix F – Duplicate Values. HTML Tags, Accents, Numbers, Conjugates

A screenshot of a social media post

Description automatically generated

A picture containing window

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Appendix G – Stemming / Lemmatization

**#1 Original Corpus Document**

'BEST CHIPS and GLUTEN FREE! These chips are so good they are addictive! Extremely fresh and crispy. Even potato chips can contain gluten, so when I noticed Gluten Free marked on the bag, I had to give them a try. Now these are the only potato chips I will purchase--Thanks for making a GF product that rocks!!'

**#2 After applying tokenizer (HtmlTags and punctuation removal)**

'BEST CHIPS and GLUTEN FREE These chips are so good they are addictive Extremely fresh and crispy Even potato chips can contain gluten so when I noticed Gluten Free marked on the bag I had to give them a try Now these are the only potato chips I will purchase Thanks for making a GF product that rocks'

**#3 Applying Stemming**

'best chip and gluten free these chip are so good they are addict extrem fresh and crispi even potato chip can contain gluten so when I notic gluten free mark on the bag I had to give them a tri now these are the onli potato chip I will purchasethank for make a GF product that rock'

**#4 Applying Lemmatization**

'best chip and gluten free these chip be so good they be addictive extremely fresh and crispy even potato chip can contain gluten so when i notice gluten free mark on the bag i have to give them a try now these be the only potato chip i will purchase thanks for make a gf product that rock'

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Original | Tokenizer | Stemming | Lemmatization | Comments |
| addictive! | addictive | addict | addictive | Lemmatization didn’t do any changes while stemming was was shorter |
| noticed | noticed | notic | notice | Stemming gave a shorter token while lemmatization gave the correct lexical output |

Appendix H – Stemming / Lemmatization Scores

Appendix I – Count Vector Representation

Appendix J – TFIDF Vector Representation

Appendix K – TF/TFIDF n-gram results

Appendix L – Logistic Regression Results

Appendix M – Topic Modeling Results

NMF-TF

Score 5

topic 0 topic 1 topic 2 topic 3 topic 4

-------- -------- -------- -------- --------

not coffee tea food salt

one cup organic cat river

get starbucks teas cats murray

flavor vanilla white old store

would flavored numi canned small

eat cups spice no fingers

little taste flavor one time

want french orange wellness bowl

try blend green loves cooking

honey best black dry seems

found roast bags chicken table

since ground used eat food

work tully chai get one

think stars taste healthy use

still keurig wonderful needs amount

every not herbal weight sea

also go caffeine baby stuff

many drink get two salts

first coffees hibiscus time also

small cafe without feed keeps

topic 5 topic 6 topic 7 topic 8 topic 9

-------- -------- -------- -------- --------

great cookie product day shampoo

price cookies quality mph clear

product mother organic calories men

taste not cans speed scalp

tastes vanilla liver miles dandruff

amazon chocolate even peak not

snack amazon use exercise hair

easy good reviews drink ounce

make chip give work conditioner

years delicious best caffeine fluid

get taste skin avg feel

butter try stars celsius smell

loved creme amazon good therapy

tasting favorite company time one

really like would like like

chips company consistency not need

pack tea arrived works strong

year glad dog something head

right back control maybe mint

pasta used purchased mins shoulders

topic 10 topic 11 topic 12 topic 13 topic 14

-------- -------- -------- -------- --------

dog good like free love

dogs quality really gluten dogs

food well taste delicious buy

treats taste bread best cheese

variety price it chips powder

no flavor add pasta kids

price would real cookies absolutely

treat really best eat find

nature healthy cheese grams favorite

grain high tastes sugar little

small make soup tried drink

amazon made need soy milk

months bread much find my

free think beans ever cats

also products cook one healthy

buy pretty try would easy

picky cereal ham dairy awesome

dry wine make know keep

past fresh better amazon fruit

loves crackers pot bag tried

topic 15 topic 16 topic 17 topic 18 topic 19

-------- -------- -------- -------- --------

water ginger hot olive chocolate

coconut vernor sauce oil taste

tried soda flavor one bar

drink ale use oils bars

also use heat tried sugar

zico sugar frank full healthy

calories syrup also bodied dark

added michigan noodles give fat

best years not sure fruit

coco many perfect impressed better

vita amazon add way sweet

energy drink put different snack

buy since blend smooth whole

taste taste say low awesome

well corn tofu salads make

no gold beans somewhat tasting

brand cane tobasco turkey milk

little known nice turkish beans

sports green enjoy taste kind

way gnome really thought no

Score 1

topic 0 topic 1 topic 2 topic 3 topic 4

-------- -------- -------- -------- --------

not food product like tea

buy cat flavor tastes chai

could cats new flavor green

even ingredients beef try spice

money eat order tried not

get dry still tasted leaves

know meat use even flavor

company vet see lemon use

it bad spicy different blend

waste pet nong really much

anything also shim trying spices

store products packaging stuff teas

right fish said awful black

real dogs purchase get loose

sure natural page ever put

worth menadione made nothing box

either canned received drink tried

way time old something total

honey years get smelled return

disappointed including noodles bad weak

topic 5 topic 6 topic 7 topic 8 topic 9

-------- -------- -------- -------- --------

dog water coffee amazon taste

dogs coconut cup order better

problems natural even ordered bitter

poison concentrate cups item awful

treat plastic flavored get hot

chlorhexidine not mountain received smell

using added vanilla shipping drink

pee tried green package tasted

smell drink half disappointed first

vet brand dark customer tasting

not bottle price halva could

first zico drink return got

post made beans company honey

like tap caff purchase fruit

lick little grounds price brand

go get flavor opened mouth

used flavors instant box much

body one pods great something

long ever aroma sent artificial

chemical another bitter pack nasty

topic 10 topic 11 topic 12 topic 13 topic 14

-------- -------- -------- -------- --------

diet hair sugar made bag

science oil ingredient china popcorn

cats use artificial treats smell

tiki scalp drink chicken time

food acid calories dogs bags

evo bottle ingredients products microwave

bag greasy grams buy old

months sodium much jerky chips

always used stevia dog flavor

cat ci high usa since

get sensitive better not every

kitten using natural pet come

time product low sick burnt

weight tried healthy pets whole

pound natural per say strong

ago laureth aspartame another put

research clear free number open

not almond powder sweet pop

got sunflower contains us case

need might cereal read may

topic 15 topic 16 topic 17 topic 18 topic 19

-------- -------- -------- -------- --------

would good no one love

eat bad chocolate box tuna

cats really even bought eat

never candy cup time not

thought free donut buy us

money stuff way two vietnam

half gluten close cups stomach

great mix ever star wanted

even something money bad packed

since baking glazed yeast caught

caff tried give thought mercury

something pot name never really

use make think first thought

put tastes much get finish

pieces vitamins well better ounce

much looking flavor away california

maybe nothing might try wild

could foods little opened higher

green got bought even sardines

recommend time flavoring cup sandwich

NMF-TF-IDF

Score 5

topic 0 topic 1 topic 2 topic 3 topic 4

-------- -------- -------- -------- --------

not coffee tea food product

good cup green cat great

like bold teas cats excellent

one great bitter old arrived

flavor strong white canned use

taste cups drink chicken taste

really decaf iced dry time

would dark black wellness years

get starbucks wonderful loves condition

use taste flavor feeding really

also blend great no service

eat good chai wet quickly

tried mountain bags feed fast

much flavor peach weight quality

delicious drink smooth problems smooth

well roast caffeine diet substitue

sweet maker loved month easy

little like loose vet cappuccino

make flavored peppermint needs buy

every favorite numi eats using

topic 5 topic 6 topic 7 topic 8 topic 9

-------- -------- -------- -------- --------

treats free chips sauce love

dogs gluten potato hot kids

training cookies bag heat drink

smell know flavor goes dogs

liver eat case say absolutely

maltese gf chip great my

give delicious mmmm anything favorite

freeze pasta ordering wings buy

small dairy stock frank cats

love granola fat try yummy

dried would whole tried bold

no rocks sons enjoyed energy

easy snacks family garlic time

size products addictive buy taste

three oatmeal crunch hooked milk

biscuit soy eating looked makes

perfect tasting alternative back little

pet muffins store given sweet

dog great madhouse awesome every

salmon disease pop sauces corgis

topic 10 topic 11 topic 12 topic 13 topic 14

-------- -------- -------- -------- --------

bars amazon chocolate best coconut

kind stores dark ever water

bar find milk the zico

fruit shipping hot mix electrolytes

candy local bar tasted refreshing

tasting glad delicious tried tastes

yum order sugar around sweet

healthy delicious company order sports

more found rich fast drink

order com again tasting potassium

blueberry grocery chocolates hcg added

nut high taste family brandsf

natural buy coffeemate makes thailand

flavors ordered creamer it nirvana

almond free cocoa brand brand

good each bit vanilla vita

larabars available error decaf mango

thinkthin difficult carrying pasta pure

cherry supermarket turbinado delivery calories

going thank sweetness healthiest coco

topic 15 topic 16 topic 17 topic 18 topic 19

-------- -------- -------- -------- --------

snack price dog peanut oil

healthy great loves butter olive

kids good treat skippy tuna

great high food great quality

quick buy my zuke oils

breakfast buying teeth trusted tins

tasty outstanding treats powdered cost

protein choking stomach delivered wonderful

crunchy though bathroom low taste

popchips even them cookies fresh

eating bulk pill around canola

fat trader little rawhide flash

easy wow gets name avocado

make happy sensitive reeses tried

sweet joe goes go several

satisfying case small too rice

bag picky jerky salmon acids

eaters sky perfect peanutty fatty

happy pack last tried monounsaturated

daughter hello likes purposes high

Score 1

topic 0 topic 1 topic 2 topic 3 topic 4

-------- -------- -------- -------- --------

not coffee cup amazon tea

like flavored cups product chai

taste weak explode order green

product beans keurig ordered leaves

would vanilla mountain shipping stash

buy aroma instant received spice

one bitter filter package teas

flavor grounds green beware spices

good strong box item black

even price again return flavor

tastes pods many refund iced

could starbucks kcups box weak

no instant reason date like

tried flavor hot expired flavors

bad ground brand disappointed tastes

money wolfgang grounds description seasonings

get roast one arrived celestial

bought dark coffees never strong

tasted raspberry twinings get lemon

try club alot customer honey

topic 5 topic 6 topic 7 topic 8 topic 9

-------- -------- -------- -------- --------

china coconut chocolate sugar cookies

made water old grams pamela

treats plastic raspberry healthy cookie

usa concentrate hot calories ginger

chicken natural melted added butt

products zico bar mango pamelas

dogs added white high bakery

pet bottle fresh tsp snap

dingo taste cacao artificial day

not pure dark low cases

read brand where much exact

pets drinking bloomed stevia version

trust milk favorite ingredient pain

no straight disappointed quaker found

sick come the coke oatmeal

dog real taste oatmeal according

buying close gritty taste restrictions

many made curry no reduction

fda coco neither you purchased

buy juice sticks carbs apparently

topic 10 topic 11 topic 12 topic 13 topic 14

-------- -------- -------- -------- --------

dog food stale cat bars

treat cats old touch box

dogs eat time treats chalky

treats would vendor not bar

toy dry sitting sniffs mango

ball pet line eat arrived

not ingredients rancid hates macadamia

refuses one purchase loves eating

touch vet guess maybe almond

eat day purchased near clif

chew dogs long my one

would cat disappointmentsticks old

meat feeding spices kitty think

bone sick fresh shelter larabar

vet bad badly oh kind

my old past extreme eat

it picky gave dry vosges

thing eating big times pie

wellness science ground buy pecan

lick diet now particular dipped

topic 15 topic 16 topic 17 topic 18 topic 19

-------- -------- -------- -------- --------

chips bag drink awful gluten

kettle popcorn drinks based free

brand bags sip strawberries semolina

chip pop ginger th barley

protein butter bad taste not

completely time lemon teas celiac

rancid covered artificial tasting discription

oil popped tasting brands baking

sour microwave iced love member

try lb actually reviews wheat

salt overpriced milk item ingredient

cream store glass purchased disease

chance plastic bitter sprayed mix

onion every taste smelling know

apple pet mixes cologne list

online pound drinking mouth search

style trash beer medicinal malt

salty full shampoo away chex

buying sealed immediately believe family

parmesan michaels coco thought that

SVD

Score 5

topic 0 topic 1 topic 2 topic 3 topic 4

-------- -------- -------- -------- --------

not coffee tea salt salt

like tea organic food great

great cup teas coffee product

good taste white tea use

coffee flavor numi cat store

taste starbucks spice cup murray

one vanilla flavor dog river

tea flavored orange river taste

product teas bags murray fingers

food cups black dogs cooking

love great used no small

flavor drink green organic table

no green ginger cats time

get blend chai fingers good

would numi herbal small amount

also french hibiscus time bowl

best white get bowl stuff

tried like mu seems seems

amazon best lime store sea

use black wonderful also chips

topic 5 topic 6 topic 7 topic 8 topic 9

-------- -------- -------- -------- --------

product good good love great

great cookie product like day

dog taste organic day good

coffee cookies quality mph food

price chocolate day drink mph

love love mph dogs shampoo

food delicious cookie calories cat

quality food liver water hair

tea mother cans dog like

organic vanilla calories speed scalp

dogs free speed work men

amazon bars company coconut calories

free healthy like treats clear

liver bar reviews exercise speed

cans really stars peak drink

cup sugar miles miles dandruff

stars amazon exercise use exercise

even eat numi caffeine peak

best chip cookies avg miles

give gluten peak celsius conditioner

topic 10 topic 11 topic 12 topic 13 topic 14

-------- -------- -------- -------- --------

dogs taste like like love

dog organic hot free flavor

cookie like sauce gluten product

shampoo shampoo love ginger shampoo

cookies sugar good use chocolate

like bars use sauce free

love one flavor vernor one

clear bar cookie cookies day

men chocolate product cookie clear

scalp best cats hot best

great men food soda men

treats scalp cheese ale scalp

dandruff clear add amazon gluten

hair coconut cat best dandruff

feel no taste vanilla hair

amazon dandruff heat food milk

mother white beans used ounce

conditioner water cook sugar chips

fluid ounce quality delicious would

try healthy powder pasta sugar

topic 15 topic 16 topic 17 topic 18 topic 19

-------- -------- -------- -------- --------

water use love one chocolate

coconut ginger organic olive hot

chocolate vernor cats oil sauce

cookie sugar great flavor bar

drink soda spice sauce dark

cookies ale white hot beans

vanilla taste not no make

way years day tried one

one many cat good dog

amazon skin one use bars

mother treats numi organic day

energy no cookies dog heat

also syrup cookie full great

tried michigan eat chips perfect

zico still used best quality

little healthy orange oils granola

beans since ginger always noodles

no get mph ginger fat

get size many give kind

back dog olive treats oil

Score 1

topic 0 topic 1 topic 2 topic 3 topic 4

-------- -------- -------- -------- --------

not food product like tea

product dog amazon tea chai

like product order product not

taste cat new taste food

would cats made flavor spice

one diet received good green

food amazon ordered would leaves

no science still one amazon

good time beef sugar use

even dogs use no return

tea one purchase food blend

buy made item tastes ingredients

water eat spicy much box

dog order nong green spices

get ingredients shim water loose

could vet packaging bad total

flavor meat page try teas

much pet disappointed tried quantity

coffee chicken buy cats cat

bad products china time label

topic 5 topic 6 topic 7 topic 8 topic 9

-------- -------- -------- -------- --------

dog water water product would

like coconut dog would sugar

dogs taste coffee cats taste

made natural coconut like one

china sugar natural cat eat

treats product made coffee better

product added would diet artificial

problems diet tea use good

treat concentrate dogs science honey

poison food sugar eat calories

chlorhexidine science use not love

smell ingredients cup dry free

using plastic china food tuna

sick cat ingredients flavor could

chicken drink ingredient lemon dog

pee bottle drink caff low

used tried treats hair much

lick also green beef grams

chemical zico many animals ingredient

look noodles problems ingredients thought

topic 10 topic 11 topic 12 topic 13 topic 14

-------- -------- -------- -------- --------

diet one sugar made bag

taste would no china would

science good ingredients one smell

dog cats ingredient buy no

sugar diet natural treats popcorn

no science chocolate love taste

bag get like would flavor

good box made bag dog

tiki hair even chicken food

evo time list eat money

corn using products tuna water

always even china flavor opened

better use chicken say put

got never way sauce brand

months could drink products chips

kitten oil high hot never

long noodles hair box ingredients

calories put use dogs another

pound used contains noodles great

need size much never price

topic 15 topic 16 topic 17 topic 18 topic 19

-------- -------- -------- -------- --------

no good amazon many chocolate

one no tuna first bad

food love love used no

product tuna even ingredients even

flavor eat no natural cats

tuna flavor get also got

cup hair cup love box

love oil taste use honey

try really diet much taste

dog us science people cup

meat use green box much

yeast scalp us taste cat

ever stuff like using bought

tried vietnam brand ingredient green

make chicken thought order never

even packed wanted popcorn well

eat bad vietnam time time

really tried well tuna two

sugar wanted hair coffee would

bought little could bad money

LDA

Score 5

topic 0 topic 1 topic 2 topic 3 topic 4

-------- -------- -------- -------- --------

oil cheerios pouch tropical popcorn

olive tree pouches cytomax molasses

tins tassimo pills fruit pops

oils marshmallows donuts moles theater

virgin bunny dunkin hole packaging

zoe baja kittens tunnel honey

jug bonsai popper traps potassium

massive disc warm mole daughter

having pb baby set poper

ummmm uncooked pillpockets light ride

yourself kosher andy lindt tbsp

taster shape senseo black consume

decadence cooker granules trap chews

lark walmart spreading dirt sports

bertolli make expiration dug style

slimey love decided passion nutritional

velvety take my proof makes

roses dyed licked able popper

exceptional peeps open juice styrofoam

accepted tempting ordeal mounted tasteless

topic 5 topic 6 topic 7 topic 8 topic 9

-------- -------- -------- -------- --------

focus dog bars shortening ginger

jittery treat best spectrum vernor

engineering greenies bar basil ale

majoring vinegar like luscious soda

shots toy ever holy beer

mechanical momo sauce gobbles michigan

crutches capsule not unknowing upset

studying end noodles shoprite spicier

exams pet chocolate classmates gnome

burst gelatin one crisco isbn

metabolism bone chips allergan cane

energy minutes calories talking requested

beneficial seniors kind previous promptly

offer company mustard wegmans zest

staple canine tasting members drops

feeling sage garlic sorry exactly

breakfast terriers protein flavored delivered

benefits pectin mph butter accounting

home knee honey family page

morning pain snack fillers microbrews

topic 10 topic 11 topic 12 topic 13 topic 14

-------- -------- -------- -------- --------

treats creamer not bread hair

remember pump great baby tea

tangerine bottle product bubbles used

actually lid good toast day

beverage lift love biscuits echinacea

crisp pushing like gras anti

hips turning food boxes see

race bailey one foie organic

hansen voila taste loaf slim

hansens illy would breakage medicine

wax energized amazon babycook pantry

carbonated insomniac no create take

cable jumpy flavor box white

deserve nirvana also lightly weight

film cranberry get cooked months

creepy ito free device years

movie polluting delicious top almost

cheer brandsf buy broken remedy

possessed en price previous helped

character thailand dog de symptoms

topic 15 topic 16 topic 17 topic 18 topic 19

-------- -------- -------- -------- --------

coffee pasta mg cat shampoo

tea hooves per hot clear

not quinoa gm food scalp

like shopped red sauce men

cup structure kettle cats dandruff

great al black old conditioner

taste dente onion wellness fluid

flavor celiac sour heat therapy

love based spheres chicken shoulders

drink disease noticeable canned clean

water prague gelatinous subscription controlling

chocolate peat raspberries flavor separate

best wargs mono dry tree

good boggy bakes time itchy

one trapsing iron texas eucalyptus

tried scotch sliced not cooling

cups lounging represent feed pyrithione

coconut peaty center rub sudsy

favorite cologne german goes therapeutic

get hordes barely eats lasting

Score 1

topic 0 topic 1 topic 2 topic 3 topic 4

-------- -------- -------- -------- --------

product not not not not

not product product tea money

would taste even like buy

dog like date taste taste

no water cup water product

half no good flavor waste

post coconut expired would like

taste good like sugar no

review order no good would

one one expiration product box

caff get coffee green love

pee even natural no tea

dogs bought disappointed free one

buy much yeast mix hot

even could one tried bought

coffee made many tastes stuff

green price food even bag

mountain candy amazon drink get

read flavor get chai good

better amazon ordered buy horrible

topic 5 topic 6 topic 7 topic 8 topic 9

-------- -------- -------- -------- --------

one not product coffee not

bad product not not taste

not like stale product like

cups bag amazon amazon much

pizza no like cup would

raw food chocolate buy get

artificial taste old taste flavor

love one one even drink

review even order two water

bunnies get box like food

diet much love no tried

product would would one product

thought diet tastes disappointed good

amazon old arrived flavor popcorn

food eat good cups love

purchased amazon time bad first

terrible buy flavor never price

reviews order taste would licorice

flavors know ever pods try

really science tried received hard

topic 10 topic 11 topic 12 topic 13 topic 14

-------- -------- -------- -------- --------

not not food chocolate not

like made not not dog

product china dog sugar product

amazon would cat msg amazon

taste like cats bars would

one dog eat much shipping

no treats would get package

tried one one salt could

try buy dogs litter even

cookies no bad regular store

get know time raspberry made

food dogs like line one

good could chicken buying like

even product good would treats

really products even product no

bad say product really bottle

chips give meat better get

know better first healthy food

bought tried ingredients grams old

day something no like sweet

topic 15 topic 16 topic 17 topic 18 topic 19

-------- -------- -------- -------- --------

not not not not like

like like no would not

one taste would one tastes

box sugar like say cup

product would item buy taste

taste artificial brand taste cups

time good beans like good

really honey way hot sweet

flavor product product bad flavor

good tastes buy never one

peanut much free many even

star bad try made coffee

opened drink get could really

bad could box two boxes

tried one much tabasco product

amazon smell received first could

ordered awful coffee even tasting

butter stuff small thought box

going also open get horrible

tasted never shipping type pretty

**Doc flow:**

1. **introduction**
   1. **scope**
   2. **objective**
2. **data exploration**
   1. **outliers**
   2. **data distribution skewness**
   3. **duplicates**
   4. **null values**
   5. **correlation**
3. **data preprocessing**
   1. **stop words**
   2. **accents**
   3. **lemmatization**
   4. **stemming**
   5. **word conjugates**
   6. **punctuation**
   7. **numbers removal**
4. **vector space model representations**
   1. **count based**
   2. **tf-idf weight based**
   3. **n-grams**
   4. **influence of ngrams**
5. **model EVALUATION, TRAINING, hyperparameters**
   1. **models used**
   2. **hyperparameter selection**
   3. **evaluation**
6. **topic modeling**
   1. **lda**
   2. **nmf**
   3. **svd**
   4. **observations**
7. **conclusion**
8. **appendix**