(a) Key Themes

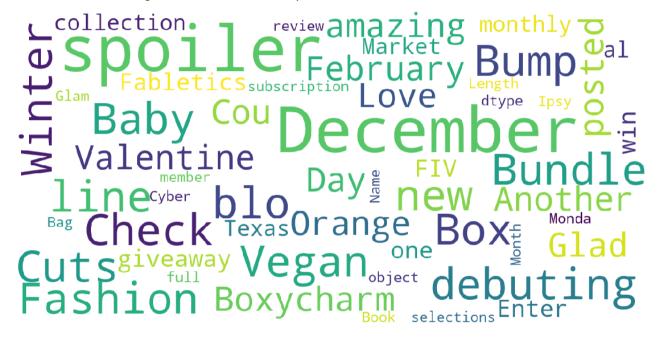
- luxe hair care
- unite
- happy mother's day
- island adventure
- toons box ship
- travel essential
- welcome tauerville
- myths legends
- leather suede
- home holidays
- modern man
- cracking case
- deadpool home
- moodboard year
- rush relax
- this mean war
- hvdrate
- magical mystery
- operation mystery box
- the irish cottage
- galactic
- women power
- mad scientist
- hamilton
- pool party
- good morning sunshine
- darkness
- robots
- natural beauty
- grateful thankful
- japan's iconic snacks
- wonderland
- sorting hat
- banned books
- watching
- happy happy halloween
- christmas
- roques rebels
- hello sanrio
- serenity
- legends
- yule ball
- happy new year
- the world traveler wonder
- book musical

(b) Analysis Plan

- Preprocess data set :-
 - From the given DOM webpages, we create the BeautifulSoup object by converting it into lxml format
 - We extract all the reviews, urls, keywords, tags and other metadata from the retrieved lxml model
 - And store them in the csv file processed.csv
- Text processing :
 - change case to lower
 - tokenize text and remove puncutation
 - remove words that contain numbers
 - remove stop words

- remove empty tokens
- lemmatize text
- remove words with only one letter
- Feature extraction
 - Apply sentiment analysis on the text processed data to get score for positive, neutral, negative and compound features
 - Adding number of characters and words as a feature for all samples
 - Adding document feature and term frequency feature for all samples
- Classification
 - Since, the problem at hand is unsupervised, we cannot apply random forest or svm for classification, as we don't know the labels
 - ∘ We can proceed by :-
 - Clustering apply k-means clustering to cluster positive and negative reviews based on the new features
 - Vader model Use vader model to analyze the overall sentiment of the review to predict if it is positive or negative
 - Vader proves to be better as it applys LSTM in the backend to get the meaning of the review, and not just words frequencies

WORDCLOUD showing the different topics modelled :-



- (c) Additional Data of Importance :-
 - Keywords, Tags, Categories
 - This will help in better classification :-
 - We can divide the dataset into different subsets based on the tags, so that we have reviews of similar kind of products altogether.
 - Then, the features like doc2vec and if-tdf will have less variance, and classification of similar data samples gives better accuracy.
 - Theme
 - Getting the theme of the products will directly help in knowing which kind of product is performing well and which is not and thus can be easily improved upon.