Using Topic Modeling to Explain Movements in ETF

Author: Qiao Lin | Tian Xia

Abstract

Given the dynamic nature of the financial market and the numerous attempts to capture profits in the stock market, there has been various investment vehicles created to capture profits from the global equity markets; One of the recent developments that spurred changes in investment world is the exchange traded funds; as the exchange traded funds industry became an dominating industry with more than 5 trillion dollars of assets under management, the constructions of index portfolio has as well became encapturing. For this specific research, we will identify the connection of social media to the ups and downs in a movie industry index, PBS. Previous research has been conducted with the use of twitter facebook; however, in order to further the research, we will bring a new perspective into the study with respect to NLP and social media website, Reddit. We propose to study the effect of Reddit comments on ETF performance by, first, identifying a period of a positive return according to historical ETF price data, second, applying various NLP methods to reddit contents to visualize and interpret the changes in the underlying discussions on Reddit posts.

Background

Exchange Traded Fund

Exchange-Traded Funds, (ETFs), are a portfolio of many individual stocks that are selected to match a given stock index. ETFs originally introduced as a way to invest in a diversified selection of stocks through matching a stock index; From the first introduction back in 1993, ETFs have grown to become a dominating industry with more than 5 trillions dollars in asset under management in 2018 [1]. Due to the nature of portfolio strategy, ETFs primarily adopt a passive investment strategy, for it simply purchases assets, stocks, according to an given index. While there are many benefits of ETFs, for our study, we are primarily focused on the diversification in a given index, for it allows the ETF to capture movements in all areas of the industry. Take PBS for example, a index that focuses primarily in the entertainment movie industry; it includes companies that ranges from content giants,like Disney and CBS, to movie distributors, like AMC. Thus, the importance of ETF for our study lies in the fact that it provides sensitivity to the variety of topics that are covered on reddit.

NLP and Reddit

While the financial background is important for comprehensive understanding, our subject of study lies heavily in the field of NLP. NLP is a field of computer science that concerns with machine learning in processing and analyzing large amounts of natural language data. In this particular study, we would like to analyze the application of NLP in relation to movie industry, and in order to do so, we have selected reddit as our primary source of data; reddit is widely social forum in which registered users can submit contents, links, images, and individual opinions. In the proposed study, we value the content from popular subreddits related the movie industry, r/movie, as we believe the user contents from this subreddit will provide us necessary information to gauge the sentiment and volatility in the oil industry.

Topic Modeling

Topic modeling is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents. Latent Dirichlet Allocation(LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions. For our specific study, we would like to apply LDA modeling to discover the underlying topics in Reddit comments. For example, if a popular reddit post was titled “Venom breaks box-office record in opening night”, then using LDA, we would be able to snatch out important vocab such as “Opening”, “Venom”, and etc. This is especially useful for our data, as topic modeling is able to intake large amounts of natural language and process it into meaningful data. The efficiency in identifying meaningful data, however, is subject to finding the right parameters in our processing, i.e. number of desired topics, and training sample size.

Motivation

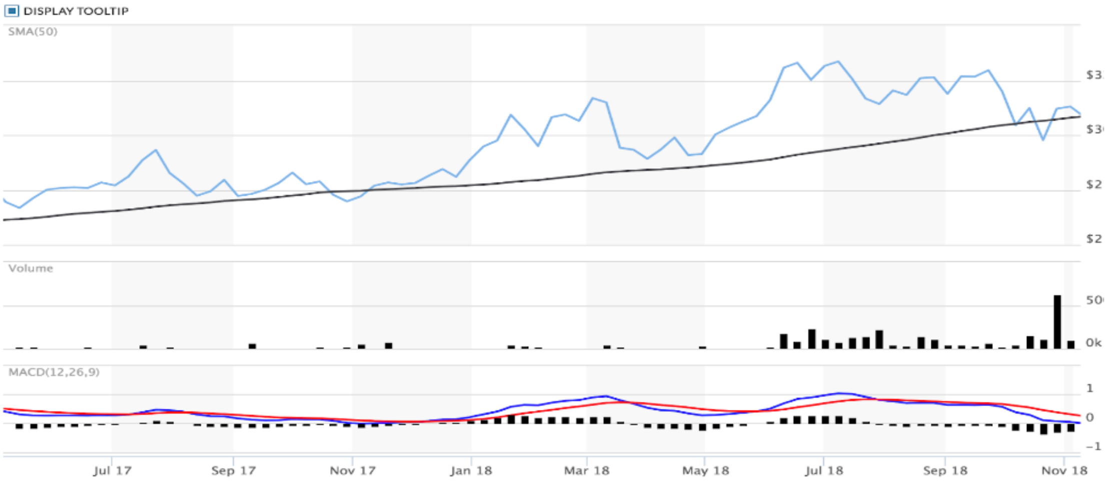
Reddit is an popular online commentary forum that has gathered much attention in recent years. As more and more users participate in online discussions and postings, it not only creates a friendly leisure online environment, it also creates a wonderful source of data for various types of studies due to the massive volumes of user data available on reddit’s platform. One of the recent studies on done on reddit comments with topic modeling was focused on studying anxiety disorders [2]. The authors analyzed data from r/anxiety, r/panicparty, r/healthanxiety, and r/socialanxiety, as well as more than a dozen control subreddits, such as r/askscience, r/books, r/jokes, and etc. The study was conducted on reddit comments after stop words were removed and word tokens are lemmatized. The researchers then tested different methods of feature generation, including vector space embedding: word2vec, and doc2vec, LDA topic modsonal narratives collected from reddit comments. In “Detecting anxiety on Reddit” (Judy, Freling, LIWC, N-gram. Then, analysis was done on the effectiveness of using aforementioned types of features for binary classification, i.e. anxiety or not anxiety. The finals results of the study showed that all models were supported with accuracy >80%, however, with word2vec having 90% accuracy in classification.

To draw a contrast from the previous study, our analysis is focused on a slightly different angle. First of all, in our proposed study, we are primarily focused on analyzing the differences between two subsets of data, from the same subreddit, r/movies. This makes a differences in our analysis because unlike the data from anxiety related subreddits, our data are typically shorter, and absent of long-paragraphed, first person writing, which is typical in anxiety sharing posts. Furthermore, unlike the previous study, we simply want to analyze the differences between two time periods of data; instead of analyzing the effectiveness of multiple feature extraction models, we want to use TF-IDF for the effect of eliminating words that appear too many times without significant meaning, i.e. “movies”, “like”, “we”, and etc., and then apply LDA topic modeling to extract meaningful topics from our data.

Data Collection

Timeline

As far as our timeline is concerned, we want to specifically focus on two periods of timeline, first one targets a period where our index is exhibiting normal returns, second one targets the period, in which our index performs above it’s historical average.



According to online information, we have identified two periods that satisfies our requirements for timeline. The first period entails the timeframe from July 2017 to February 2018, while the second periods entails the timeframe from March 2018 to October 2018, resulting in two timelines of the same length. As depicted in the picture, there is clearly a upward trend in MCAD, which is short for moving average convergence/divergence, it is designed to reveal changes in the strength, direction, momentum, and duration of a trend in a stock's price.

Reddit API

Reddit API allows us to directly access posts by subreddit. Therefore our goal is to collect as data as possible for each of the two periods that we’ve selected. However, due to the limitation of the Reddit API, we are not able to crawl all posts created during each of the time frames. We were able to crawl the all-time top 1000 posts made within the subreddit, r/movies. Therefore, we filtered and selected our desired results out of top 1000 posts in r/movies by dates that matched our time frame.

Comment vs. Topic

While we were not able collect as much data as we wanted, we still obtained a significant amount of data due to the popularity of selected. For each month in our selected periods, we were able to crawl 10 - 15 posts, which is not a significant number. However, if we take into account the number of comments in each post, we were able to gather significant amounts of data. In fact, for the average of 10-15 posts per month, we were able to gather from 7000 - 15000 comments, depending on monthly reddit activity.

Sample data

As listed below, our data comes primarily from r/movies subreddit. Due to the nature of reddit, our data includes title, comments on various topics in the r/movies subreddit. Fortunately, our data does not have a particular focus on recent movies only, our data pertains to all elements related to the film industry: new and old movies, stars, authors, directors, movie discussions, and any news relevant to the movie industry. The diversity in our data gives us advantage when we are connecting the relationship between PBS and r/movies, as the content topics varies in a way that pertains to all the companies included in the PBS index.

- drank himself to death.
- I agree that The Pacific probably catches the ""war is hell"" statement unlike anything ever produced in film
- Sledge's book does really give you a great insight as to how absolutely brutal the fighting was against the japanese. I also highly recommend it.
- From what I heard it was the film makers who didn't want him. Cohen wanted to do a dark gritty biopic of Mercury being a closet homosexual with gay scenes. The band who hold all the rights to the song wanted a pg movie. Ultimately you can't have a behind the music movie without the music so the film makers sided with the band. Honestly a bummer to me

As the sample data suggests, we have a wide variety of data in our dataset, from simple 1 sentence comments, to paragraphed commentaries. The variety of data is beneficial to our study for it offers information that will allow us to discover the underlying topics in each month.

Data preprocessing

Due to the nature of reddit comments, or even more so, human languages, there is inherent flaw in our data for it cannot be readily accepted by computer to conduct analysis. Thus, we’d have to exercise the power of data cleaning on the comments and titles that we will be crawling from subreddit, r/movies, through the Reddit API. Here are the three important steps that will be used to clean our data.

Tokenize and remove Stop words

After packaging and unloading our data, we used the gensim package to first tokenize the sentences and then remove stop words. Due to the nature of stop words (such as “the”, “a”, “and”), they do not contain useful information for our study, and thus had to be removed before the analysis. If we do not remove the stop-words, our data would be filled with irrelevant information. In addition, we tokenize our data into individual words for further data processing.

Lemmatize

Given the data is from Reddit, which is a platform containing posts, and comments made by individual human beings, the data is going to contain different forms of a word, such as organize, organizes, and organizing. Hence, the goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance: am, is, are, are all converted to be.Thus we are able to capture all words of the same meaning, regardless of tense, forms, and other derivations. Lemmatization also in a way helps with identifying words that contain typos.

Dictionary

Then, we used gensim to create a dictionary containing all the words from our data, as well as the number of times a particular word has appeared in our data set. For the purposes of our data, we decided to filter out words that appeared in too many documents, as well as words that appear only in single documents. For our analysis, we filtered out words that appeared below 1 document as well as more than 5 documents, we did this to keep out insignificant words and aimed to keep track of popular words in each month.

BOW

After filtering out the words, we have an dictionary that records how many times each word has occurred. Then we used doc2bow to convert our data in each post to a bag of words vector. This is a particular important step for our analysis, because we want to separate out data by individual posts. Thus, after using doc2bow, we now have a bag of words for each individual posts [3].

TFIDF

Create tf-idf model object using models.TfidfModel on ‘bow_corpus’ and save it to ‘tfidf’, then apply transformation to the entire corpus and call it ‘corpus_tfidf’. Finally we preview TF-IDF scores for our first document. The purpose of TF-IDF is again to remove all irrelevant words. Because popular words, other than stop words, could appear in our posts for multiple times, thus using TF-IDF, we are able to neglect words that appear too many times, and focus only on the words that contain useful information to our study.

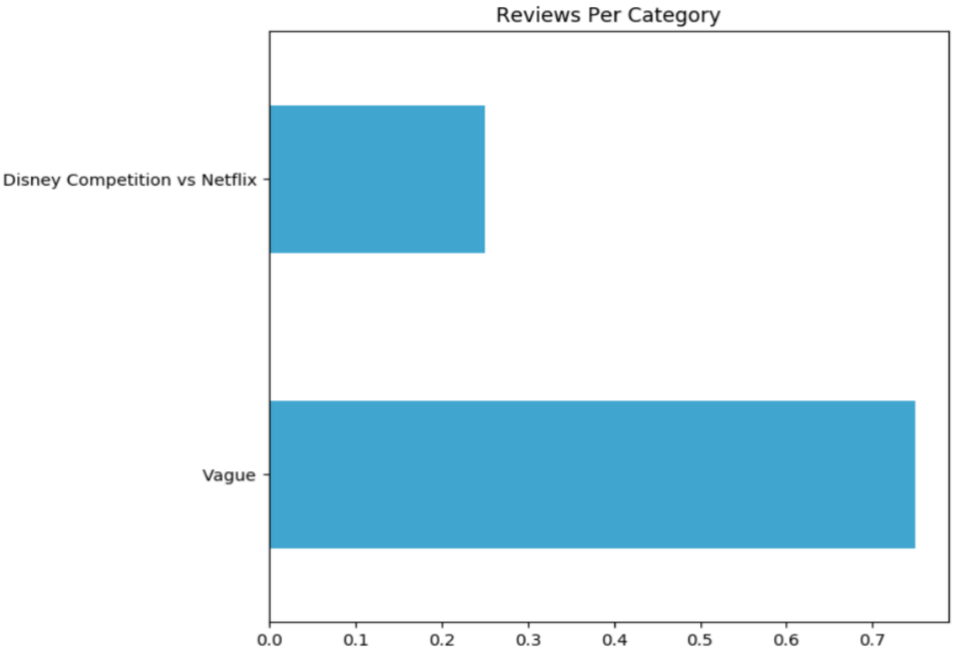
Topic Modeling

For Topic Modeling, we are going to use gensim, a Python toolkit built for Topic modeling and the technique called Latent Dirichlet Allocation. The input will be a bag of words and we aim to figure out the relationship between certain topics and the turmoil of the PBS stock price during two certain periods. The first period is when PBS almost remain the same and the second one when PBS is overall increasing.

We are going to measure the top topics and distribution of topics in each period using LDA. Each document contains comments of movies from a one period. We will then try to find the correlation between the change of topics and the change of PBS within these two periods. Our hypothesis is a frequent change in movie topics usually means a turmoil in movie industry stock price.

Topic Modeling

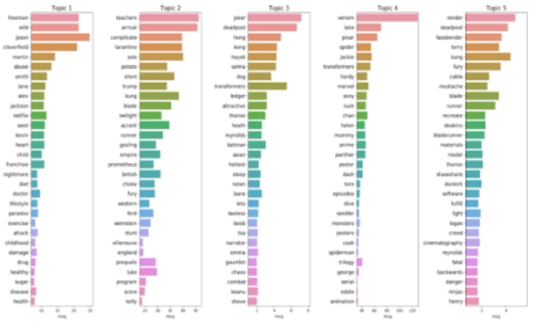
Two challenges were presented to us when we were evaluating results. First, it was difficult to gauge the number of topics that would most closely capture all topics in our data set. For example, if we are only to use two distinct topic models, we have have one topic model encapsulating all different topics and the other encapsulating a small amount of irrelevant topics. Below shows a picture of topic distribution of using only 2 topics.

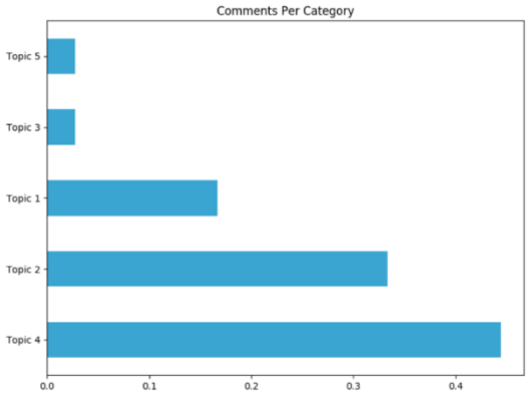


As shown in the picture, the topics in July 2018 could theoretically be partitioned into two categories, however, with one category representing most of the review, and the other only an insignificant amount. Therefore, in order to provide a solution to the first challenge, we decided to include 5 categories of data for each month, for it provided a better distribution of topic encapsulation.

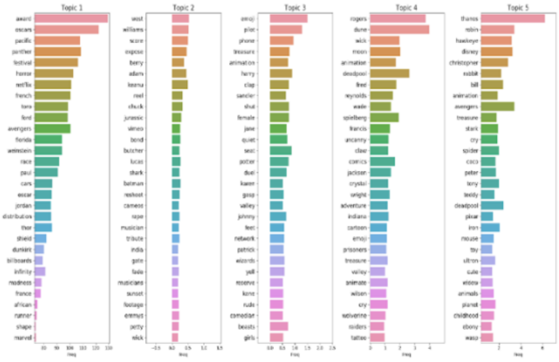
Second, it was a difficult decision to decide how we would break down the data in order to not just provide meaningful topic models, but also meaningful results for our analysis. There were multiple options for time intervals that we could have used, 1 month at a time, 3 months at a time, or whole periods at a time. We tested the validity of each of the options and found that month by month data provided the most legible and meaningful results. Below is a comparison between 1 month interval and 8 month interval, whole period.

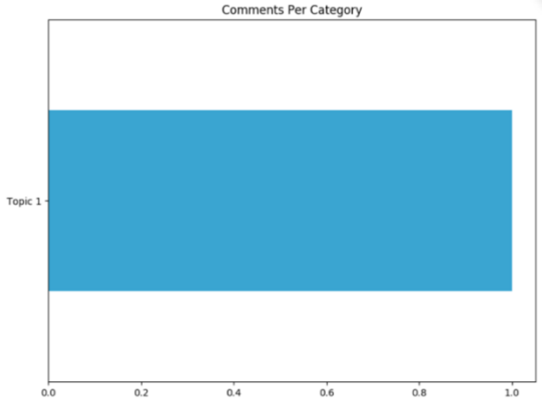
Month by Month





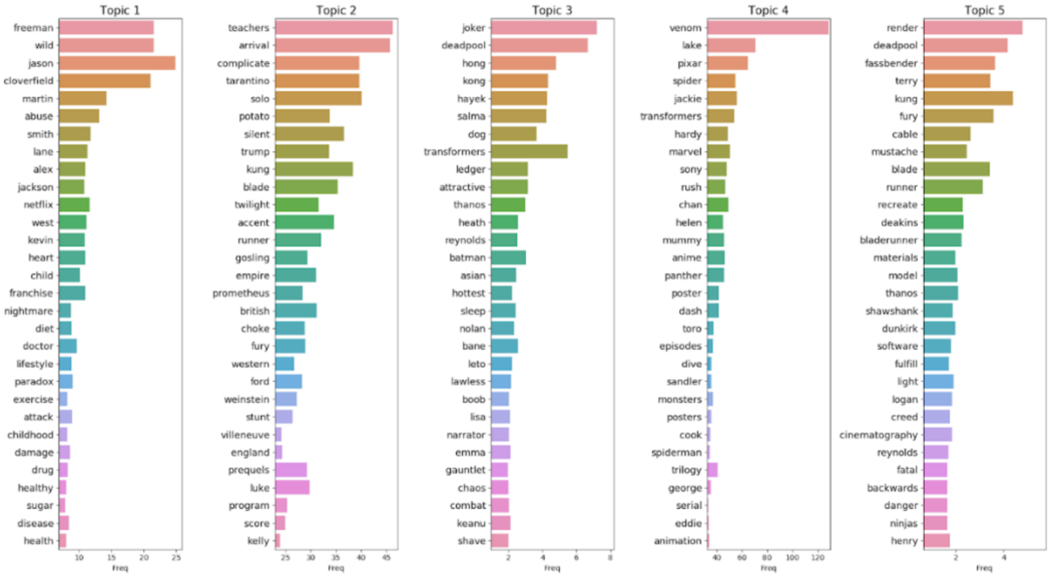
Whole Period



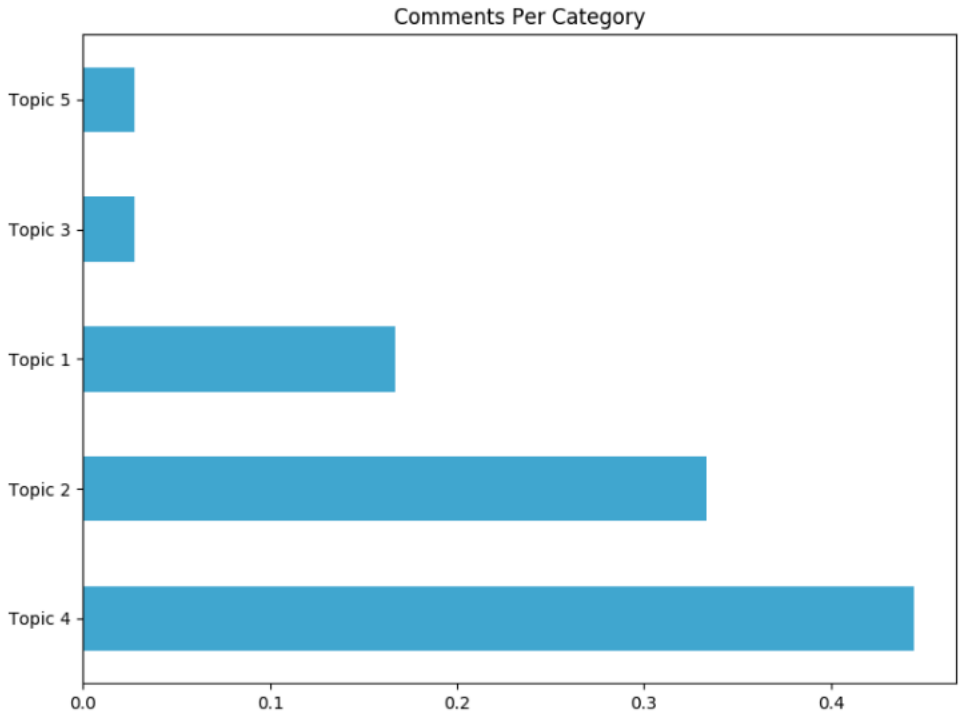


Thus, we made the decision to only use month by month data for our analysis, for first, we were able to better distribute comments among our topic models, and second, we were able to make better capture the changes over entire period as we had more time intervals, i.e. 8 time intervals for each period we want to measure.

Interpreting Results



As show in the illustration above, our analysis over the comments in February 2018 produces 5 different categories of topic for each month, with each individual topic representing an actual underlying discussion around a particular topic, for example, "topic 4" in the picture above is clearly a topic centered around the movie Venom, which makes intuitive sense because Venom was a popular topic back in February 2018. However, there is one particular challenge in our analysis, the inconsistency of the effectiveness of topic model across our time frames. For example, as illustrated below, in February 2018, we are able to extract 5 topics, out of which only 3 represented meaningful topics.



Such inconsistency is not uncommon in our results, in fact, the majority of our time periods contain topic models that were ineffective; Similar to "topic 5" and "topic 3" above, they contain generic unigrams that do not hold concrete meaning. As a solution to this problem, we decided to analyze our results qualitatively.

To fully test our hypothesis, we decided to manually count the total number of meaningful topics in each of the periods that we are examining, and come up with aggregate count number that is comparable across the two periods that are relevant to our overall study, July 2017 - February 2018, and March 2018 - October 2018. The results is shown below.

	1	2	3	4	5	6	7	8	Total Count
7/2017 - 2/2018	Marvel Heroes	Disney Franchise	Indiana Jones	Harvey Weinstein, Jedi	Oscar Award	Russian Documentary, Tarantino film	James Bond	Peter Jedi	10
3/2018 - 10/2018	Adam West	Jedi Solo	Dragon, John Wick	Brad Pitt	Predator, Moana, Venom	Venom	Nazi	Robin Hood, Christian Bale	12

As illustrated in the table and according to our method of measurement, we did observe more meaningful topics in the second period. The result corresponds to our initial hypothesis: during the period in which PBS shows abnormal positive return, there is more activity in the movie industry, which is illustrated by the many topics that we've identified in our analysis.

Broader Impact:

Ultimately our subject has one goal: determining the usefulness of NLP in connecting subreddit sentiment to the changes in our selected stock index, PBS. Our analysis focuses on using NLP methods to most efficiently extract the underlying topics from Reddit comments in r/movies subreddit. However, since our research originates in subreddit content analysis, our research exerts far greater implications than just in the movie related industry. Our method of analysis presents the a logical process that carefully extracts information from subreddits. The broader implications of our study would be seen effective in sociology studies, in which a careful understanding of discussion and topic is necessary.

References:

1. Weinberg, Ari I. (December 6, 2015). "Should You Fear the ETF? ETFs are scaring regulators and investors: Here are the dangers—real and perceived". Wall Street Journal. Archived from the original on December 7, 2015. Retrieved December 7, 2015.
2. Detecting Anxiety on Reddit. Judy Hanwen Shen, Frank Rudzicz. Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—from Linguistic Signal to Clinical Reality (2017), pp. 58-65
3. A gentle introduction to bag of words model. Jason Brownlee