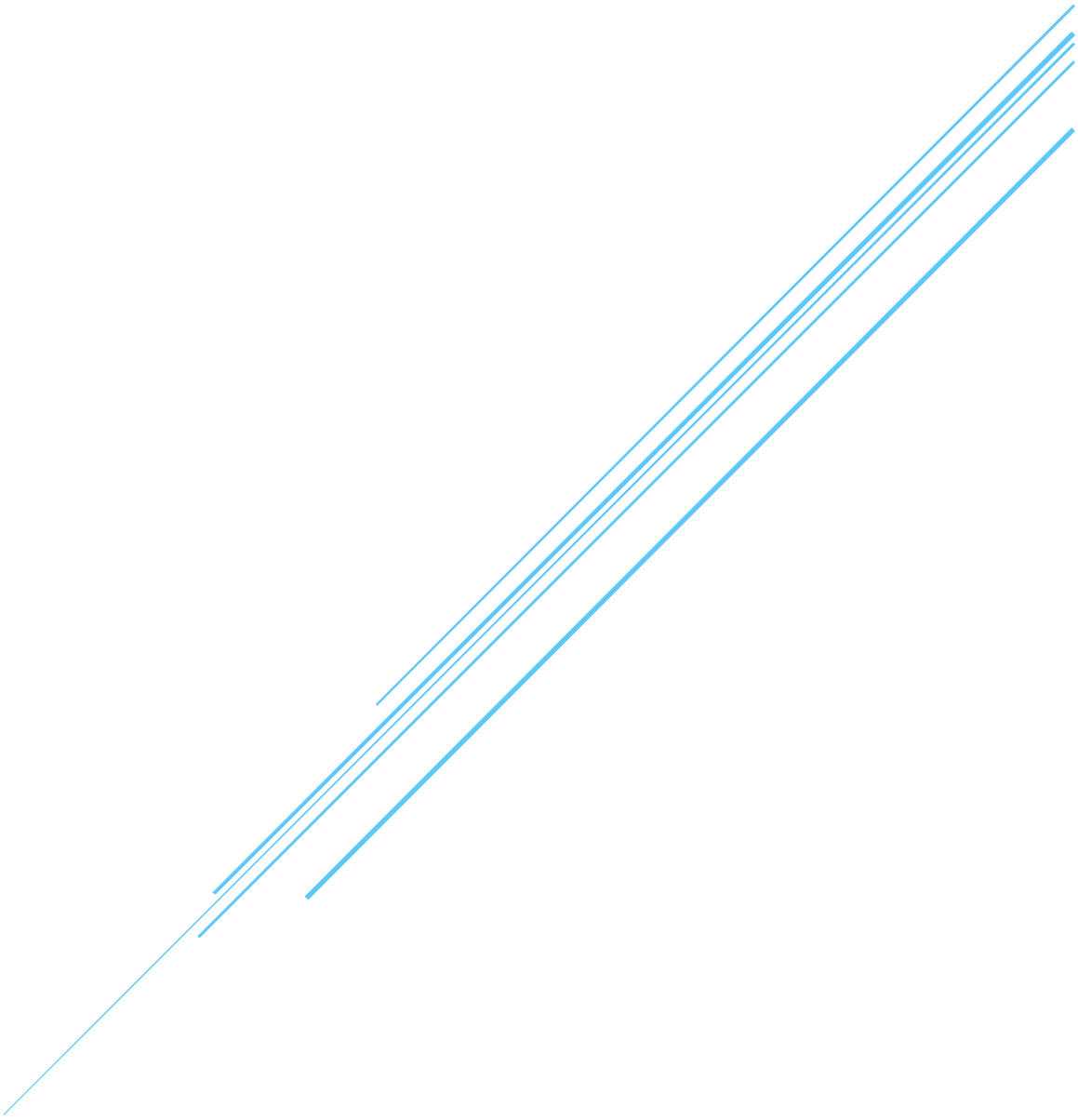
SENTIMENT ANALYSIS (AMAZON REVIEWS)



Sanat Chugh

Graduate Student - CCIS - NEU

CONTENTS

INTRODUCTION

AN ABSTRACT

Extremely often we come across a situation where we have to make a choice, and our choice at this point is influenced by a number of factors, mostly dominated by our senses.

Let's consider for example, you decide to buy a book on Amazon. The rating for the product includes a lot of data, what works, what doesn't, pros, cons, ratings, unnecessary comments, etc. How do we know what are people really feeling about the product, is their review based on their sentiment, or is it just a 5 star they preferred to give cause they were too lazy to type.

Often there are products that we see have a lot of 5 star ratings, but on what basis? If out of 10 reviews, 9 are just a 5-star rating, which would mean that people love the product. What do we do next, probably read a reviews to see what people have to say. If there are reviews, great! Otherwise, there is a higher probability we would consider looking at other products as we might get to read what people have to say to make that smarter decision , what is the best for me.

To aid in such a day, when you want to be a smart buyer, where the review of a product is not only based on a star rating but also on the sentiment towards the product, such a sentiment analysis comes of great use.

While my idea works specifically on Amazon Review Data, a sentiment analysis can be done on any text of appropriate size. I believe Sentiment Analysis is a great tool that can be used to help make a smarter judgement, and adds great value to a rating system. It is a great tool to work more closely with people really want and what they feel about your product.

Sentiment Analysis is a great tool than can be used in a variety of industries, only to provide a better nd smarter insight into your product, review, essay, etc.

I plan to implement a simple MapReduce task to query the right data from the data set. I use Pig to do my analysis, sort, and filter my data. I run the analysis against a dictionary of pre-defined words.

I will show that when I run the job using the overall rating vs sentimental analysis based rating, they will not match.

SENTIMENT ANALYSIS

WHAT IS IT?

First let's look at the definition of the Sentiment Analysis [Wikipedia]:

"**Sentiment analysis** or **opinion mining** refers to a broad (definitionally challenged) area of natural language processing, computational linguistics and text mining. Generally speaking, it aims to determine the attitude of a speaker or a writer with respect to some topic. The attitude may be their judgment or evaluation (see appraisal theory), their affective state (that is to say, the emotional state of the author when writing) or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader)." And why you should care?

Social Media and Social Networking have fueled the online space. Ratings, reviews, comments, etc – are everywhere. From NYT articlehttp://www.nytimes.com/2009/08/24/technology/internet/24emotion.html : "This is more than just an interesting programming exercise. For many businesses, online opinion has turned into a kind of virtual currency that can make or break a product in the marketplace."

We tell our clients – you have to listen to what people are saying about your brand, products, services… and more importantly you should react, respond. Nicely said than done. Simple Twitter search on iPhone will give you tons of results. Is it possible for a brand to manually look at the every single mention and respond?? Of course not! Automation is the strategy… But – smart automation. As a consumer I do not want to get some irrelevant auto-response from a brand.

Solution – analysis of the unstructured texts. Not just on a set of keywords, but also on emotions. Not an easy task to do, but there are visionary companies who are working on tools/products that can help brands to deal with all these amounts of unstructured content and help them to make sense of the emotions hidden behind customer's feedback.

SOURCE:
http://customerthink.com/what_is_sentiment_analysis_and_why_you_should_care/

MAPREDUCE

WHAT IS IT?

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Conceptually similar approaches have been very well known since 1995 with the Message Passing Interface standard having reduce and scatter operations.

A MapReduce program is composed of a Map() procedure (method) that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name) and a Reduce() method that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies). The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance.

The model is inspired by the map and reduce functions commonly used in functional programming, although their purpose in the MapReduce framework is not the same as in their original forms. The key contributions of the MapReduce framework are not the actual map and reduce functions, but the scalability and fault-tolerance achieved for a variety of applications by optimizing the execution engine once. As such, a single-threaded implementation of MapReduce will usually not be faster than a traditional (non-MapReduce) implementation, any gains are usually only seen with multi-threaded implementations. The use of this model is beneficial only when the optimized distributed shuffle operation (which reduces network communication cost) and fault tolerance features of the MapReduce framework come into play. Optimizing the communication cost is essential to a good MapReduce algorithm.

MapReduce libraries have been written in many programming languages, with different levels of optimization. A popular open-source implementation that has support for distributed shuffles is part of Apache Hadoop. The name MapReduce originally referred to the proprietary Google technology, but has since been genericized. By 2014, Google was no longer using MapReduce as their primary Big Data processing model, and development on Apache Mahout had moved on to more capable and less disk-oriented mechanisms that incorporated full map and reduce capabilities.

SOURCE:
https://en.wikipedia.org/wiki/MapReduce

PIG

WHAT IS IT?

Pig is a high-level platform for creating MapReduce programs used with Hadoop. The language for this platform is called Pig Latin. Pig Latin abstracts the programming from the Java MapReduce idiom into a notation which makes MapReduce programming high level, similar to that of SQL for RDBMSs. Pig Latin can be extended using UDF (User Defined Functions) which the user can write in Java, Python, JavaScript, Ruby or Groovy and then call directly from the language.

Pig was initially developed at Yahoo! to allow people using Apache Hadoop® to focus more on analyzing large data sets and spend less time having to write mapper and reducer programs. Like actual pigs, who eat almost anything, the Pig programming language is designed to handle any kind of data—hence the name!

Pig is made up of two components: the first is the language itself, which is called PigLatin (yes, people naming various Hadoop projects do tend to have a sense of humor associated with their naming conventions), and the second is a runtime environment where PigLatin programs are executed. Think of the relationship between a Java Virtual Machine (JVM) and a Java application. In this section, we'll just refer to the whole entity as Pig.

The programming language

Let's first look at the programming language itself so you can see how it's significantly easier than having to write mapper and reducer programs.

1. The first step in a Pig program is to LOAD the data you want to manipulate from HDFS.
2. Then you run the data through a set of transformations (which, under the covers, are translated into a set of mapper and reducer tasks).
3. Finally, you DUMP the data to the screen or you STORE the results in a file somewhere.

Few important commands:

LOAD

As is the case with all the Hadoop features, the objects that are being worked on by Hadoop are stored in HDFS. In order for a Pig program to access this data, the program must first tell Pig what file (or files) it will use, and that's done through the LOAD 'data_file' command (where 'data_file' specifies either an HDFS file or directory). If a directory is specified, all the files in that directory will be loaded into the program. If the data is stored in a file format that is not natively

accessible to Pig, you can optionally add the USING function to the LOAD statement to specify a user-defined function that can read in and interpret the data.

TRANSFORM

The transformation logic is where all the data manipulation happens. Here you can FILTER out rows that are not of interest, JOIN two sets of data files, GROUP data to build aggregations, ORDER results, and much more.

DUMP and STORE

If you don't specify the DUMP or STORE command, the results of a Pig program are not generated. You would typically use the DUMP command, which sends the output to the screen, when you are debugging your Pig programs. When you go into production, you simply change the DUMP call to a STORE call so that any results from running your programs are stored in a file for further processing or analysis. Note that you can use the DUMP command anywhere in your program to dump intermediate result sets to the screen, which is very useful for debugging purposes.

SOURCE:

https://en.wikipedia.org/wiki/Pig_(programming_tool)

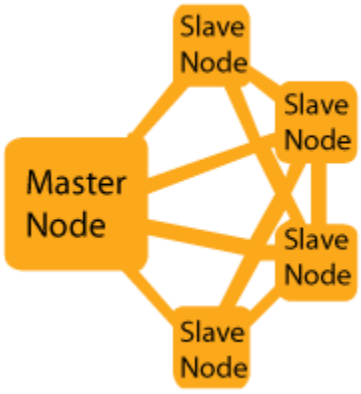
https://www.ibm.com/software/data/infosphere/hadoop/**pig**/

AMAZON EMR

WHAT IS IT?

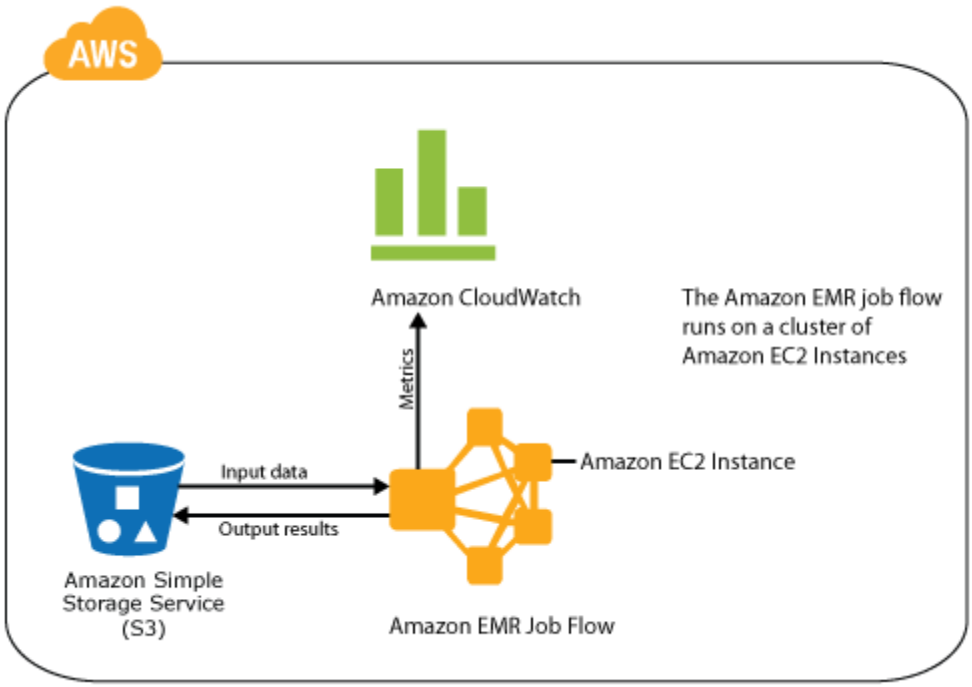
With Amazon Elastic MapReduce (Amazon EMR) you can analyze and process vast amounts of data. It does this by distributing the computational work across a cluster of virtual servers running in the Amazon cloud. The cluster is managed using an open-source framework called Hadoop.

Hadoop uses a distributed processing architecture called MapReduce in which a task is mapped to a set of servers for processing. The results of the computation performed by those servers is then reduced down to a single output set. One node, designated as the master node, controls the distribution of tasks. The following diagram shows a Hadoop cluster with the master node directing a group of slave nodes which process the data.



Amazon EMR has made enhancements to Hadoop and other open-source applications to work seamlessly with AWS. For example, Hadoop clusters running on Amazon EMR use EC2 instances as virtual Linux servers for the master and slave nodes, Amazon S3 for bulk storage of input and output data, and CloudWatch to monitor cluster performance and raise alarms. You can also move data into and out of DynamoDB using Amazon EMR and Hive. All of this is orchestrated by Amazon EMR control software that launches and manages the Hadoop cluster. This process is called an Amazon EMR cluster.

The following diagram illustrates how Amazon EMR interacts with other AWS services.



Open-source projects that run on top of the Hadoop architecture can also be run on Amazon EMR. The most popular applications, such as Hive, Pig, HBase, DistCp, and Ganglia, are already integrated with Amazon EMR.

By running Hadoop on Amazon EMR you get the benefits of the cloud:

- The ability to provision clusters of virtual servers within minutes.
- You can scale the number of virtual servers in your cluster to manage your computation needs, and only pay for what you use.
- Integration with other AWS services.

SOURCE:
http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-what-is-emr.html

THE DATA SET

AMAZON REVIEWS

For my project, I decided to use the Amazon Review Data Set. The data set has reviews relating to various product reviews on the amazon website. I use this data set as it has all the appropriate fields that I require in-order to reach the goal of my project.

I obtained the data set form SNAP Standard via Julian McAuley. After getting in touch with him and emailing him he provided me the latest data set. This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014. I worked on the data set that handles all the review data, containing no duplicates.

The link to the data set can be obtained from below:

http://jmcauley.ucsd.edu/data/amazon/
https://snap.stanford.edu/data/web-Amazon.html

The data set that suited my analysis the best was per-category based. It covers the following categories:

Electronics, Movies and TV, CDs and Vinyl, Clothing, Shoes and Jewelry, Home and Kitchen, Kindle Store, Sports and Outdoors, Cell Phones and Accessories, Health and Personal Care, Toys and Games, Video Games, Tools and Home Improvement, Beauty, Apps for Android, Office Products, Pet Supplies, Automotive, Grocery and Gourmet Food, Patio, Lawn and Garden, Baby, Digital Music, Musical Instruments, Amazon Instant Video.

Each of these categories have the review fields in the following format:

(reviewerID, asin, reviewerName, helpful, reviewText, overall, summary, unixReviewTime, reviewTime) where,

- reviewerID - ID of the reviewer
- asin - ID of the product
- reviewerName - name of the reviewer
- helpful - helpfulness rating of the review
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)

For my analysis the required fields I used are:

- reviewerID
- reviewText
- overall

The data was provided in a [NAME].json.gz format. I converted all the data into a CSV format for easy of analysis using a self-developed python script. The python script will be added to the solution zip.

Total Data I used is divided as shown below:

CATEGORY	SIZE (MB)
Amazon Instant Video	228
Apps for Android	728
Automotive	560
Baby	476
Beauty	850
CDs and Vinyl	2,975
Cell Phones and Accessories	1,418
Clothing, Shoes and Jewelry	2,104
Digital Music	427
Electronics	1,750
Grocery and Gourmet Food	550
Health and Personal Care	1,349
Home and Kitchen	1985
Kindle Store	1754
Movies and TV	3252
Musical Instruments	287
Office Products	604
Patio, Lawn and Garden	447
Pet Supplies	589
Sports and Outdoor	1.476
Tools and Home Improvement	914
Toys and Games	993
Video Games	946

The total amount of data reviewed: **25,217 MB (~25.2 GB)**

All the data was uploaded to Amazon S3 Storage. This provides easy access to the Amazon EMR setup that is recommended for multi cluster implementation.

ANALYSIS TASK AND OUTPUT

SO WHAT DID I ANALYZE?

The idea of the task is to show the Top 20 and Worst 20 reviews in each category based on the overall rating and sentiment analysis individually.

This will give us a picture that while aggregating on the overall rating vs sentiment analysis, we will see that the reviewer IDs would not match. This shows us that it is better to use sentiment analysis to analyze the positivity and negativity towards the product. This can be very valuable to product designers as they can then judge that sentiment towards the product. Not just based on the 5-star rating that the reviewer might provide.

For the analysis task, I ran the MapReduce using Pig against each category of the data set. In each category I used the **reviewerID** field to map against the rating and the review text.

The **overall** field help me do my analysis without considering the sentiment of the reviewer and the **reviewText** field help me do my analysis considering the sentiment of the reviewer.

The Sentiment Analysis is performed against a pre-defined dictionary of words that have a sentiment rating attached. Each word in the **reviewText** was tokenized and mapped against each word in the dictionary. The sentiment of each word is then added and averaged to form the sentiment for the entire review. It is important to note that the dictionary restricts the level to which the sentiment can be analyzed and so it is important to use a good, updated dictionary.

The MapReduce task is run for the entire data set, only to limit the results to the top in-order to see the extremities in sentiment.

I implemented the entire task on Amazon EMR, using a x3 cluster setup (1 master, 2 nodes). Amazon has a good setup to use the online storage s3, and I used the same to store my csv file once I had converted the same from json.gz using a python script.

The results and output as pertaining to each category of products in the Amazon Data Set is as shown in the next few pages:

OUTPUT

CATEGORY	**MATCHES** (in reviwerID between overall and sentiment analysis)
Amazon Instant Video	0
Apps for Android	0
Automotive	0
Baby	0
Beauty	0
Books	0
CDs and Vinyl	0
Cell Phones and Accessories	0
Clothing, Shoes and Jewelry	0
Digital Music	0
Electronics	0
Grocery and Gourmet Food	0
Health and Personal Care	0
Home and Kitchen	0
Kindle Store	0
Movies and TV	0
Musical Instruments	0
Office Products	0
Patio, Lawn and Garden	0
Pet Supplies	0
Sports and Outdoor	0
Tools and Home Improvement	0
Toys and Games	0
Video Games	0

AMAZON INSTANT VIDEO

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A1HK9952KM25SL	5.0	AGVRO5K0KXRIZ	5.0
AD6E1LY97KGC2	5.0	A4NVEFRAJNDIO	5.0
AONFFMO9QY0G9	5.0	AZV1W3DXKVYMR	5.0
A1IWX2XBFTCMSR	5.0	A2TWUCVFHHA7J6	5.0
A37IYQC8EKLDYJ	5.0	A1II4DTERTT3RS	5.0
AJE4MXRQ5W9JU	5.0	A2SC33FQHZE31S	5.0
A9MBSKL8LTFN9	5.0	A3AKC4WN7L5FH1	5.0
A2YXWWVABHWIXN	5.0	A2Q7SD9R0B3HKA	5.0
A1UMMMN8QG2LCD	5.0	AUIOLUC88BNTR	5.0
A2CGJQAQ5A7MDW	5.0	A1MDVAKYSPFOJC	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

APPS FOR ANDROID

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A1AL90WSPY73Z	5.0	A1K6ZW3PGLTXSF	5.0
A1EO8KJWPRQ5G	5.0	A1ZRNDHIXFJJL4	5.0
A28NZPURMBP7D	5.0	A1ZB3WPNLUAKGX	5.0
A1M8Y0AP1A6VA	5.0	A18O3MMWHRYGQN	5.0
A2YI4W42DMST4	5.0	A274E9USNTICGF	5.0
A1R5W0LJQ9XK1	5.0	A19C9AK1NBU1CD	5.0
AUI0OLXAB3KKT	5.0	A252G8M3B60WBD	5.0
A2AS1YOW3H2SK	5.0	A1UTT6QH0NNBQ2	5.0
AY98O2C2YYSOO	5.0	A1C6F1RXUQHNEZ	5.0
A1AOL4US72HIS	5.0	A24GD9PB0LA6SH	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

AUTOMOTIVE

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A2S6405KCSIJPA	5.0	ARN7DZOC9V6JZ	5.0
A1K3BF6U9GO6F8	5.0	A120KLWTFR8HMJ	5.0
ABI7KXVCTSS8F	5.0	AFCGTOGNL0UM5	5.0
A3BG7Q6NNZCEGP	5.0	APTYZKSL8L8UV	5.0
A8BXQ7Z2M01AG	5.0	ALLB5J0GW5NBR	5.0
A1TCZF6H0ZULVU	5.0	A1AJ5IT6GVSZD0	5.0
A1QFCZCVEA4Y2Q	5.0	A13XZMQMQQQ5SA	5.0
A1MUILE70L0N2J	5.0	AA85VRYO9SGZ7	5.0
A3EQBCBFR5VUQB	5.0	A18KU24V8P5MQT	5.0
A2DF7SAGI92GS8	5.0	ADKKYDLIW5N1E	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

BABY

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A2B3WBNVQSLM6V	5.0	A9PY38P4YI9GG	5.0
A1FSRXEF9D588F	5.0	A5RZYLCOB7L9	5.0
A2YDMQCHYJIYPG	5.0	AHP8HZ4RAUHAR	5.0
A3O8O1EMXIB3T9	5.0	AE4IX0NBQCRTG	5.0
A3U7438NXFL16A	5.0	A5FLEL8XUQVGL	5.0
A1K0FYAEF3H0UP	5.0	AWOU0ULDOTUG7	5.0
A25DMFZ9UR90S6	5.0	A1BHH0P2VS3256	5.0
AVH3CQ9L55WW0	5.0	APU0JWBDX15MY	5.0
A2R67GZM2HCVL2	5.0	AEVXRGD91MH5U	5.0
A1ASGRF2DF23SDG	5.0	ARVZ355LGS7W8	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

BEAUTY

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A1RYQPQ01T5D5R	5.0	A1LLTN59TUDU0H	5.0
A3MQDRRGC9070R	5.0	A11WNKBT5TWG0Y	5.0
A2LB743IV4NS2W	5.0	A1JGCO5P7FB655	5.0
A1YSYUGR6XRMIW	5.0	AWEZ5Q1BFSES4	5.0
A3DEHKPFANB8VA	5.0	A19X60U0J70L96	5.0
AG9TJLJUN5OM3	5.0	A1OREC9M1GPQP1	5.0
AYBIB14QOI9PC	5.0	A1HZ3E0K3OSXNH	5.0
A1PB6OT0OOPKNQ	5.0	A15Y9JFJ30F9UI	5.0
A2HDFHID3BDI23	5.0	AYQ08PKR9I85F	5.0
AY3AHSV39DFG7	5.0	A1J45YBMGE7IMR	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

CD AND VINYLS

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A1K1NMCQ1LHLXT	5.0	ASFGCABOBYZU7	5.0
A2WPG31TDR63HY	5.0	ADGEA74NVEQS4	5.0
A12R54MKO17TW0	5.0	A93AHWEVXFC12	5.0
APXN90S1CEZO8	5.0	ALVWH4WN6E2T1	5.0
A1S8962OZCEV73	5.0	AEFFD94K6IF3E	5.0
A2AMGO0BQU8IPA	5.0	AHRFC5D4ZNHFR	5.0
A3T9IX8EDFDX4G	5.0	A61AI6ZVVEO0M	5.0
A207T421LSA2N6	5.0	A54BV49KNDT4F	5.0
ASF235FSGSDF23	5.0	AKAZT5193KFR1	5.0
A1DGJNKSF84NEE	5.0	A65W284FBARC5	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

CELL PHONES AND ACCESSORIES

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A10K98QSJKNXM7	5.0	A3F2NFB3RGH5SK	5.0
A3315VWQGK0DO4	5.0	A39ML27D3DEV1D	5.0
A12DT4IRKYMD1Z	5.0	A3DPIQCY6UF5YQ	5.0
A1SHKKG5D93TP8	5.0	A3DTV6JNHCR6FO	5.0
A140HTQ9B38BL4	5.0	A3LS6HNRWC3Q36	5.0
A13XYYFUXCC1TR	5.0	A3IPT762Q9F43L	5.0
AUKHZBW58JEGE2	5.0	A3RBLMVX5M4A2D	5.0
AJN5FX35AU4TK3	5.0	A3ARBLDJF8DW55	5.0
A1NR37AER89LHF	5.0	A3UO3NHMJ1ZNGC	5.0
A1SDF3Y14BFWSL	5.0	A3IHFNRYNA5H4D	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

CLOTHING, SHOES AND JEWELRY

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A38NAMLUM1Z8RM	5.0	A2EMZW4BMHVBHN	5.0
A2VWA7WVP7YXW	5.0	A22TXV06QLJ0OS	5.0
A3H4HZ4G683M8H	5.0	A2FSKA4HF26ANY	5.0
A3MPJDZ7W6MI7U	5.0	A1UJZMM8ZB641V	5.0
A2H06DH6AEBUBD	5.0	A1WED58QNU7VKI	5.0
AZSKAC8ONNCVG	5.0	A23PZ96N8531B5	5.0
A1R5Z784N97LWC	5.0	A2B4MISEOEGM4K	5.0
ADUMM8JB9H3ZH	5.0	A29S1RRT9OFJNF	5.0
A321Z8TLS9WTZE	5.0	A1Y3F5F74C2OFS	5.0
A1YOUEB23HWUF	5.0	A24FZ7RX6GX0Y8	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

DIGITAL MUSIC

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A8X1YI90JPSGP	5.0	A37PPQAV77OUS2	5.0
A2GXML0X74Z3C0	5.0	A3BJ1ILXDLW07L	5.0
AQL13WF76ZQQJ	5.0	A3RVJL6QW1KHOM	5.0
AA89Y0XAI0LV6	5.0	A3D802QWHMDK92	5.0
AUW8U9U79Z361	5.0	A3VDCW8FNPUNT9	5.0
A1M1DL7XTT61KA	5.0	A3PEZQQ3I0B036	5.0
A3C124NOK7Y041	5.0	A3V4DHCK9UABYC	5.0
AOAIPA2KZHAY1	5.0	A3S3ICWC7VX53D	5.0
AW7JSI0YNMD7Q	5.0	A3AC1G3P3IWS4K	5.0
AF3ASLLM3IU5F	5.0	A3QPKOPXGE5A8D	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

ELECTRONICS

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A2ZQLZXGRI12WT	5.0	A2GAO01NSINHIO	5.0
A245DGQK96JZ2Y	5.0	A276QYIC89IOK6	5.0
A1INUW36QMRLGG	5.0	A1QZU0HYRJZLYF	5.0
A2MNRB08H3GO96	5.0	A26BIHOSX364CN	5.0
A3M61LT8L48ZR	5.0	A22OSC4OPGUUWL	5.0
A1XOQHWF7QSRIX	5.0	A1RVV3ZYJUXCUQ	5.0
AFFKMCVRORBKT	5.0	A2CGJQAQ5A7MDW	5.0
A1O3N248CHQON9	5.0	A1PDO53N2OKR9O	5.0
ABZ4W51650B0S	5.0	A2FIUZ04XWO74F	5.0
AY2AOLDNBTO24	5.0	A1Z2FPIG4L4D4L	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

GROCERY AND GOURMET FOOD

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A12XIXE9S64TF7	5.0	A259YUQUD6ZJZ1	5.0
A2DG81T8YR0PTM	5.0	A200AXVX568PMK	5.0
ACF0NFD95VRUS	5.0	A1S7LK9M1U3Q4T	5.0
A6B61KJ12Z9SV	5.0	A24OPT0FNR172C	5.0
A1DGBQQIBHTLPF	5.0	A1BZMMQGZOJ5GL	5.0
A3DYJ90MXMH04P	5.0	A2E4KG7TFYE4ER	5.0
AWQYAJV2QWLWA	5.0	A22CA0O8G2DG2H	5.0
A2AT0AI6QKSY2C	5.0	A1DXTSHQEVQJTR	5.0
A2NSZZ7Y0RAE45	5.0	A27R5MYDPG5YP7	5.0
A3SJ39BDKJD3S	5.0	A1R475Y222EJGJ	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

HEALTH AND PERSONAL CARE

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A1BDUJ3BJ3B2N	5.0	A1MWV6RYE4O75V	5.0
A2NTYNU6ZUQN21	5.0	A21HHHLBMMJY8M	5.0
A2GONTH7GYUFVQ	5.0	A29YQFD5D1YHJN	5.0
A1V8926JUKCKJS	5.0	A1F20ZI2K2NJ39	5.0
A2FHWROYY84TEC	5.0	A264CP9B5WBYCX	5.0
ADOYO1MGBTIQF	5.0	A1ORO2RPF78G0O	5.0
A3UUOF03BQRON0	5.0	A1C5CD216F3CI4	5.0
A3ESFYD819NWUE	5.0	A1T29QAVCDBRFR	5.0
AMZRC92A56UF2	5.0	A21BT40VZCCYT4	5.0
AK3JSBDFKJ295	5.0	A1CVDCHKM0ABTI	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

HOME AND KITCHEN

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A3R7LVIW1WKH3D	5.0		5.0
ACJHQDHN6SVMD	5.0		5.0
A2ISTEFZQ198Z0	5.0		5.0
A3Q7Y2C7VJNJIU	5.0		5.0
A3SF7HJWZOAG4M	5.0		5.0
AKI76B2IJAW6X	5.0		5.0
A1PZPRUZ3CS5F2	5.0		5.0
A1EIEF008WIZNT	5.0		5.0
ABSLTYYMEPZGO	5.0		5.0
A1DJBO9PEMXIT	5.0		5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

KINDLE STORE

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
AQSL89CSKJYB1	5.0	AXAPE4PEG8F1Y	5.0
A20KO0BPMNREJL	5.0	AUSLF67W3SNLB	5.0
A1BQO66R6OLCCW	5.0	A1BZCS1VHIIFPN	5.0
A1CI87FHKK6HVC	5.0	A1FYL6MXXL0187	5.0
A8W4BR3HGGS3C	5.0	A1EJE7ELS5L1X9	5.0
A2NRGE3CSFY2TQ	5.0	AX2CTQZ5CZSD7	5.0
A1JGVEK72KIDB	5.0	A1C5SIPSVKU5XH	5.0
AKB3NSDKL8PKE	5.0	A18HFBKH7KAERB	5.0
A28HYWBFDLQ80	5.0	A17CQ7ED9EUNE2	5.0
AJSKJB3FL2SN3	5.0	A14BRFJX7AGZNJ	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

MOVIES AND TV

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A2JBSKGJW2KLNS	5.0	A1DQN05ESO75TZ	5.0
A2A7NHE5HTK79N	5.0	A1AK9361RO53XG	5.0
A2516WRCXALQNG	5.0	A1PT4KQ6QU3AYB	5.0
AU5AMGERWZYXC	5.0	A14ZN3E694KLSK	5.0
A1GHGAK6PSGPLR	5.0	A1IT3AS41XI2LR	5.0
A3223CXH9BSXQU	5.0	AWQ5LEP2OO4W4	5.0
AHJAE899CPR41	5.0	A1LDYCL4Z04QAK	5.0
A32SVRW6ZF5YMA	5.0	A1KNM8MSNO7XD6	5.0
A147KLEYVIHSL8	5.0	A1BNJPZ5J659Y4	5.0
A2JLKSDBKJB23	5.0	A1AT09N1FA34OC	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

MUSICAL INSTRUMENTS

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A1OEENCY5VM88K	5.0	A1VLEHY1GHDOOS	5.0
A1IKLLSRP5U0M1	5.0	A29J3DIF5ERKJ	5.0
ANGH2JWPH9TDR	5.0	A1L0VZSDV2M2A8	5.0
AVEUBZDN94U1G	5.0	A28Z02AWRNBMG	5.0
AYOUOUJBQUKTB	5.0	A1GCU0WR2YUJ3	5.0
A23WGPOTP6PXGW	5.0	AHPGQBPVJZMJD	5.0
A1IUAWLKGXP8LU	5.0	A1L0BNFTB45ODR	5.0
A2GTAJT2ZQM28Y	5.0	A151EUL2C4GJNE	5.0
A1CKXKLE92TD6U	5.0	ARGR4XZMTE2EF	5.0
A1JSDFJK3BJKB3	5.0	A1GGAMK5HX4MZ4	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

OFFICE PRODUCTS

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
JA7020P8HQ1T	5.0	A2BEUP7E33IYPO	5.0
A2YRKGKYIDH3D3	5.0	A1OJPRISTOQ35K	5.0
A2YZNXRIU8WIYU	5.0	A2EMHXW68CW6KK	5.0
AX0NGZ1UYSARW	5.0	A1U5NWJOYH2QQH	5.0
A14GAIZ7ZGADPX	5.0	A1Q5SFCHNI7YTB	5.0
AJLW1DZSHOVGW	5.0	A2HDY3NW5914XL	5.0
A1SHHQSPOWR00F	5.0	A24549X3MTCNYF	5.0
A2CO4LGCFM8TH1	5.0	A1V26K6I0MG3JB	5.0
AC1K4OQOZ90RS	5.0	A2AZTCSMNSJWR8	5.0
A2J4NS23NOLOP	5.0	A1CI2ZUQJ49QXK	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

PATIO, LAWN AND GARDEN

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A2V247ZQT9OJIT	5.0	A1A3S3DBCPL9LT	5.0
AF30DHAVSYNZ0	5.0	A1DWGWB2QI5NO8	5.0
A34KTJ4VWNBOHI	5.0	A1H6PRK6HB80RR	5.0
A4XPE5UGK51IA	5.0	A1ACSTJZI83WSP	5.0
A2G7RZI15ZPCTH	5.0	ANUK5WSNBJPET	5.0
A2V247ZQT9OJIT	5.0	AXVGCDVS7TCMI	5.0
AF30DHAVSYNZ0	5.0	A1IY6238K3LHVC	5.0
A34KTJ4VWNBOHI	5.0	A1HPMVA3BG3XB	5.0
A4XPE5UGK51IA	5.0	AC8ZYQ7Y5U0FR	5.0
A2BDO3VFOQAO4	5.0	A1M9ZOFVUXLDEP	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

PET SUPPLIES

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A1FLCT5K0O5IP0	5.0	A1B4QXKSI1BOL0	5.0
A10G2K6SCAWKAG	5.0	A2EM2Z1271IBSX	5.0
A352KPAPZ8LDYY	5.0	A1B4I332SBOV3N	5.0
A1D6IVEJPR2SGE	5.0	A1FAATTE5SRN96	5.0
A33AQPJYH7UUXR	5.0	A1E3FY47K1Y5TM	5.0
A1KXZ1GMY2M6TH	5.0	A2JP7HQGEZMV3U	5.0
A1E6T4KRPL4HJW	5.0	A2GRLLJC2JK9FN	5.0
A1K2ZB3WDKPCLI	5.0	A2E2EMMV5E91U9	5.0
A269SKUWC7YEYU	5.0	A2BD7EBXPBP08L	5.0
A1JKFD3IB3O12N	5.0	A2F4UM3A2Q4WMK	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

SPORTS AND OUTDOORS

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A2JB9139E6RZDY	5.0	A1DVH7TX3ZWJ2N	5.0
A1NYBDCM68IHAB	5.0	A1J78ZQ2S9CKT4	5.0
A1QP8Y9RS96PEE	5.0	A1G3Y5UED5JED2	5.0
AI0Y5QFSF881V	5.0	A1MSDMTIJLUV1D	5.0
A1UD0BQO53CYAM	5.0	A1NAPKOIW9FMUV	5.0
A2SMS3BCOVK4FE	5.0	A1QKC9PRWAHWHU	5.0
A3FM2HAV42XUB5	5.0	A1RAPS91IMJYII	5.0
A2PQBJP8ZMUI5I	5.0	A14F7SO433DX8G	5.0
AAPOA500V04QW	5.0	A1VU47WSSAE1Z6	5.0
AH10DSBFIJ23BS	5.0	A10IYFZI1991J4	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

TOOLS AND HOME IMPROVEMENT

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A3NYGF4BWP25RF	5.0	AEG3FUW798TD2	5.0
A1UG07FAXWLBNU	5.0	A9NKSOPCLTWFF	5.0
A3Q7TY0T4PQULD	5.0	A78537UQRSMF4	5.0
A2I6A8OSTUB1LD	5.0	ABCAXWBS0XDQP	5.0
A1FP180T76ZC8X	5.0	AQW359RJN19XX	5.0
AJZY5DA5UBCV0	5.0	AQEO5RKPD0BQY	5.0
ASB8HV1YQNSON	5.0	AS55KMYP7X6R4	5.0
A1KEBMYM2ZFNE1	5.0	A6OQEU20L3F95	5.0
A2Y0L72IWWLI2I	5.0	AOYX3WKN4LTCH	5.0
A3BKWJEB23LJN2	5.0	AFUOEHCF2JK3S	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

TOYS AND GAMES

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
AFLVOJKZWXBA7	5.0	A2SYF4XLE33K4T	5.0
AU6HK20YZQATN	5.0	A2ELLJ8J4VM9XZ	5.0
A29FZGOSFJU1RP	5.0	A2KZF08DNTN0LR	5.0
AVJ4N5LBKAOG5	5.0	A2G189TSX6WKUE	5.0
A2HJSR00979T5Z	5.0	A2I30ZWPX4NQT3	5.0
AME8WUWDTACFF	5.0	A1I86PUW08S7S2	5.0
AJ062DSRHW9RX	5.0	A1GIYUHNDA2F2R	5.0
A3JB0C3QWIN61Q	5.0	A22FQQSKC1GG61	5.0
AVJ84HKPIGBMH	5.0	A2KHLFE89AIMMU	5.0
ADJLQBBRLK2IR	5.0	A1I81M4QDRQ1V9	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

VIDEO GAMES

So who are the top 10?

reviewerID	RATING (based on Overall rating)	reviewerID	RATING (based on Sentiment Analysis)
A2WMZ9L3CXLVR9	5.0	A6RNSF0MYJ5Y7	5.0
A3FN6G6K4GN9OH	5.0	A566YWIZIBET7	5.0
ASNVXQQJC86F1	5.0	ADWX7YCYDL4EZ	5.0
A3TKD71L42GBPE	5.0	A993VFUVU6SD8	5.0
A3RAAQ4HNATH2X	5.0	A5TZ3LYV40IJP	5.0
A37GQ0HWA2AK2B	5.0	A14H0VLX6QQXT	5.0
A3LHGRYGZ7QH3I	5.0	ANWAQFCSN8M4U	5.0
A3GKMQFL05Z79K	5.0	AKDFLFI7RZ9A9	5.0
A3IRPXHQVN0DZU	5.0	AC2VI1QYHPSO2	5.0
A1NJKWBDFJB1SD	5.0	ANGT25Z1J4RH	5.0

Matches: 0

Inference:

No two reviewerIDs match, Overall Top 10 reviewers in both columns are different when we take sentiment into account. So how important is the star rating compared to the Sentiment Rating now that we can do such an analysis. I believe Sentiment Rating > Overall Rating.

CONCLUSION

SO WHAT DID I LEARN

Through this project I have come to realize a lot of applications where Big Data can lead to obtaining more accurate results.

This analysis provides a platform to understand how sentiment analysis can be effective in determining the "vibe" around a product. Just aggregating only over numbers can be one way of concluding the general feel. But, sentiment analysis is a great tool to achieve more accurate results as it directly relates to what the reviewer is expressing.

Interesting results lead us to think more about how we can work to make things better. I feel this project is a great base to start at with a good vertical to into a deeper analysis. Some of my suggested future scopes are on the next page.

If you want your customer to be attached you have to understand what the customers likes, dislikes, and feels. Sentiment Analysis is that tool to bring you closer to understanding the customer. It helps you build a product more aligned to the needs of the consumer, and also gives you extremely valuable feedback on how you can improve, and what the market really thinks about your product.

Sentiment Analysis brings you closer to the user perspective through analyzing the user's thoughts and sentiment.

FUTURE WORK

SO WHERE CAN THIS GO

Sentiment Analysis can be used in a variety of industries and applications. It can be used in a variety of applications:

- Business Applications
- Politics
- Recommender System
- Expert Finding
- Summarization
- Government Intelligence
- Technology Development
- Enterprise Review
- Defense Operations
- Manufacturing
- Employee Feedback Systems, etc.

These are just few of the industries that can apply sentiment analysis. In a world full of people, with cognitive computing being at its peak, sentiment analysis can have applications to understand people across domains. In turn, closely connecting the consumer to the client, and building smarter solutions.

In this project I have provided a simple insight into the advantage of Sentiment Analysis. The project can be taken to greater levels as it forms a generic base to perform much deeper analysis, and going into specific solutions.

ACKNOWLEDGMENTS

I would like to firstly thank my professor, Rachel Lomasky, for being a great guide to introducing me into the world of Big Data. It has opened doors that I never imagined. Today, I realize and understand the importance of dealing with Data rightly to make smarter decisions. Thank you ma'am and I look forward to being a part of the Big Data industry for a while.

I would like to also thank my University, and the College of Computer Science for providing me with a platform to learn this technology.

I would like to thank my family and friends for their valuable insights and support as I worked through this project.

REFERENCES

SO WHERE DID I TAKE HELP

1. http://customerthink.com/what_is_sentiment_analysis_and_why_you_should_care/
2. https://en.wikipedia.org/wiki/Pig_(programming_tool)
3. https://www.ibm.com/software/data/infosphere/hadoop/**pig**/
4. https://en.wikipedia.org/wiki/MapReduce
5. https://pig.apache.org/docs/r0.7.0/tutorial.html
6. http://www.rohitmenon.com/index.php/apache-pig-tutorial-part-1/
7. http://www.rohitmenon.com/index.php/apache-pig-tutorial-part-2/
8. http://www.tutorialspoint.com/apache_pig/index.htm
9. http://hortonworks.com/hadoop-tutorial/how-to-use-basic-pig-commands/
10. http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-pig-launch.html
11. http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/kinesis-pig-generate-data.html
12. http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-pig.html
13. https://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-README.txt