Team Amazon: Seller-Forums Information Filtering Tool (SIFT)

Project Plan

Spring 2015

Amazon Contacts
John Marx
Darren Krape
Poornachandra Pesala

MSU Team Members
Max Goovaerts
Carl Johnson
Luke Pritchett
Benjamin Taylor
Johnny Zheng

Table of Contents

1. Executive Summary

Amazon.com is an electronic commerce company based in Seattle, Washington. It started as an online bookstore, but quickly diversified and is currently the largest internet-based retailer in the United States.

Along with selling their own products, it features the ability for third party sellers to use Amazon as a web front to list and sell their merchandise. One of the features provided to third party sellers is Seller Forums, an online forum wherein users may post questions and replies related to selling on Amazon. Roughly sixty-five thousand questions and two million replies are posted annually. The topics of these posts and replies vary significantly.

There is currently no method to allow Amazon employees to monitor Seller Forums; short of manually scanning through each post. This is not feasible at the scale at which Amazon operates. If there is a widespread and trending issue, Amazon would like to know so that they may address it. Along with identifying topic spikes, a classifying system would be beneficial to help keep Seller Forums organized.

The purpose of our project, “Seller-Forums Information Filtering Tool,” (SIFT) is to address these issues and unlock the value behind Seller Forums. This is done by identifying and classifying the topic and sentiment of conversations. SIFT’s dashboard allows Amazon employees to drill down into specific date ranges to analyze topic trends and view intuitive graphs. The system also has a notification system to update employees of the current status of Seller Forums.

This dashboard, accessible by Amazon employees, streamlines analysis and searching of the Seller Forums. This allows Amazon to quickly address potential issues and provide a better quality of service to the many third party sellers on Amazon.

2. Functional Specifications

The functional specifications provide an overview of our project. They outline the problems at hand and our approach to solving them.

2.1 Overview

Since Amazon is a worldwide leader in e-commerce, it consequently provides a great service for third party sellers of all different experience levels. The sellers have a large community on Amazon Seller Forums, processing millions of posts annually. On these forums, the sellers discuss selling their goods through Amazon, whether it be novice questions, expert analysis, or somewhere in the middle.

Currently, Amazon has no way of analyzing these posts. Analyzing all of this data manually would be extremely tedious or perhaps impossible. However, the data is important enough to develop a solution to this problem. Through the forums, sellers are providing Amazon with great ideas for growth by identifying their pain points with Amazon's services, praising the great features, and even pinpointing areas of opportunity. All of this data is currently encrypted by its sheer size, and the decryption technique is to develop software to select only the useful information.

The focal point of the project is to unlock the value of this data. The application gives a visualization of the forums' topics and what sentiment the sellers are using. By identifying the users' sentiments, Amazon will be able to properly align their business plan to cater to the users.

2.2 Data Organization and Analysis

The forums data is stored in an Amazon Web Services relational database, and the application runs several algorithms against it. These algorithms will analyze and organize all of the data from the forums allowing Amazon to make sense of it.

Text clustering groups forum posts based on similarity. These posts are brought together on the basis of a subset of all words present in Seller Forums, called features. Words that are eligible to be used as features may be altered through the use of a stop list. The stop list is a collection of words that are distinct from the set of words that may appear in the features list. This allows words like "Amazon", "Selling", "Help", etc. that are present in most posts, and provide no uniqueness as compared to other posts, from causing noise within the clustering algorithm.

Through this process, new categories of discussion that the Seller Forums staff may not have been previously aware of may be brought to light, and addressed as seen fit. The title of each cluster is temporarily formed by the top three features found in that cluster by count, these words typically are a fairly descriptive label for the cluster e.g. "Shipping, Tax, Order." Clustering is more powerful than text classification because it allows new categories to be formed automatically based on popularity, as opposed to user defined categories present in classification. Twitter's trending topics is a great example of this in action.

Because the content of Seller Forums is fairly volatile in terms of currently trending topics, users may adjust several aspects of the clustering algorithm such as number of clusters, date range, etc. and will receive results regarding the efficacy of the clustering. Due to the somewhat random nature of clustering and the capricious nature of new posts within the Seller Forums this allows the Seller Forums staff to dictate how clustering is run, ensuring the results are accurate, useful, and desirable. Once it is apparent that the accuracy of clustering has diminished, the user may experiment with new values until certain scores (detailed in technical specifications) are at a sufficient value. Upon this correction, clustering may be re-run over all posts by these user defined values.

2.2.2 Text Classification

Text classification organizes posts into a static set of predefined topics through keywords. The analysis will be sorting posts into the topics found by clustering.

Text Classification expands upon clustering by delegating each post to a previously created cluster. While clustering is useful in terms of determining previously unknown categories of posts, it will never generate the exact same output. With the combination of clustering and classification, the Seller Forums team can carefully select the number of clusters they wish to have, run clustering, ensure the results are to their liking, and from that point forward run classification on any new posts uploaded in the database.

If clustering were run each time new posts were added, the categories would constantly be changing, and require renaming and understanding of the underlying content in each category. With this approach the power of clustering's ability to find new topics may be paired with classification's ability to keep constant categories.

2.2.3 Sentiment Analysis

Sentiment Analysis refers to the use of natural language processing and computational linguistics to identify and extract disposition within text. SIFT provides sentiment as an additional metric for forum activity.

2.3 Report

To visualize the extracted data from text classification, text clustering and sentiment analysis, the application displays various charts and graphs. The home page displays information about the forums in a broad perspective: such as the volume of posts in topics over time, and the sentiment distribution for each individual topic. There are also individual cluster detail pages allowing the user to specify what time range they wish to view and displaying a sampling of posts within that topic over that period of time. Sentiment data for the selected topic will also be displayed.

2.4 Settings

Administrative settings are kept separate from SIFTs clean dashboard. For the users that wish to tinker with the inner workings of SIFTs clustering there are several pages that allow for this.

One of these pages allow changing of cluster names and also displays a list of the top 100 words per cluster with the ability to add each word to the list of “stop words.” Another page allows users to run diagnostic clustering on a set date range with other specified settings. Once the diagnostic clustering is completed, the results will be populated in a table and an option to set these settings to the entirety of SIFTs data is available. On separate page, users can download a specified set of data into csv format for more freedom to analyze this set of data.

Our project also allows for notifications. Daily email notifications keep users up to date on the status of the topics, including sample posts, top key words, and the “health” of the clusters through silo score and normalized inertia. A notification page allows for easy adding and removal of users.

3. Design Specifications

The design specifications articulate the project's user interface features and elaborates on the design by providing and explaining screenshots.

3.1 Overview

SIFT is designed for Amazon employees who will use the data from the seller forums to make business decisions. It is accessible through any major browser. The dashboard will aid in visualizing the data analysis portion. Users can search for specific keywords and see what the sellers are saying about them via graphs and other reports.

3.2 User Interface

The index of the SIFT website is a general homepage, where the user can get an overview of how all forum topics are performing. The user can select an individual topic and it will link to that topic's page, where graphs specific to that topic and a table with sample posts (see Section 2.3) are shown. Additionally, there are several settings pages, where users can rename cluster titles, adjust settings for clustering, run diagnostic clustering, download samples of data and adjust the email list for SIFT's notifications each topic. The screenshots below (*Figure 1-7*) display all aspects of SIFT.

3.2.1 General Analytics

Figure 1 displays SIFTs homepage. The left column is populated with available topics. The rest of the page provides information about the volume of posts in different topics and changes in topic volume over time, as well as per-topic sentiment information.

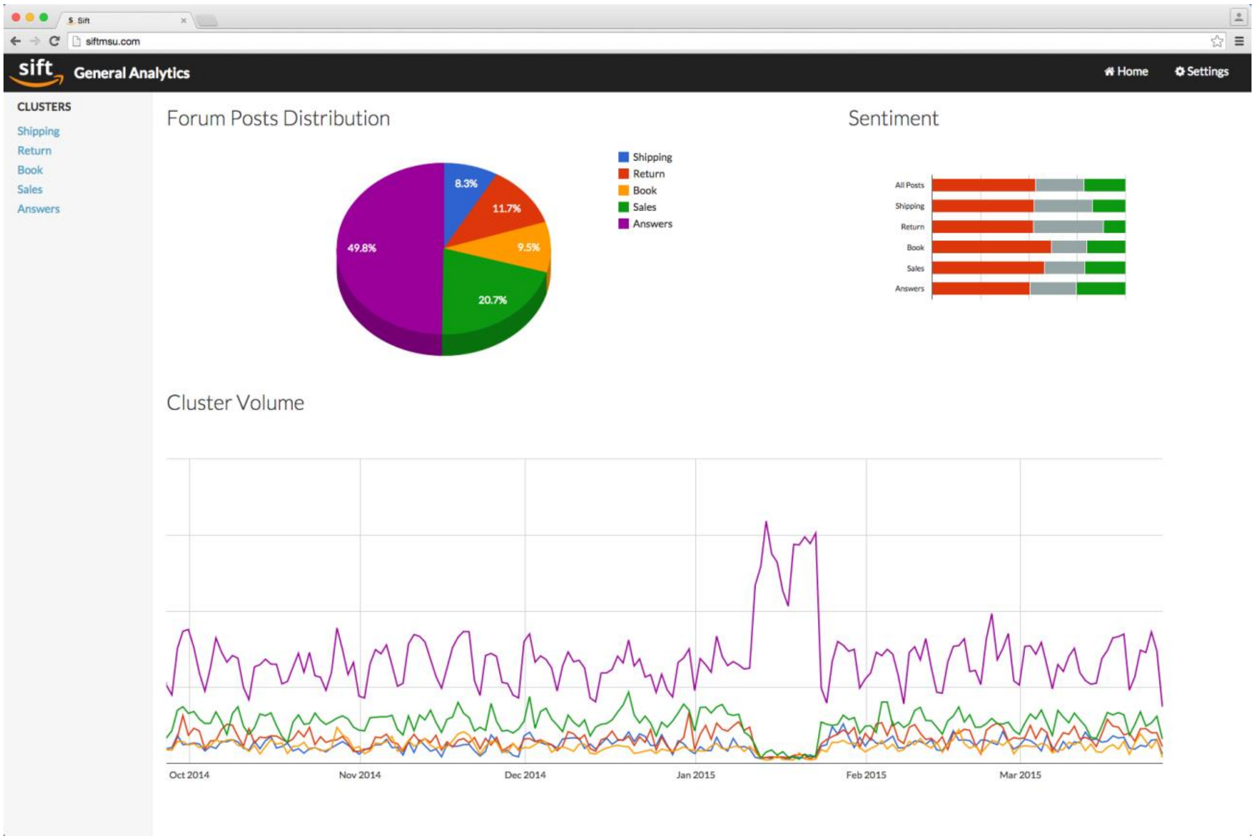
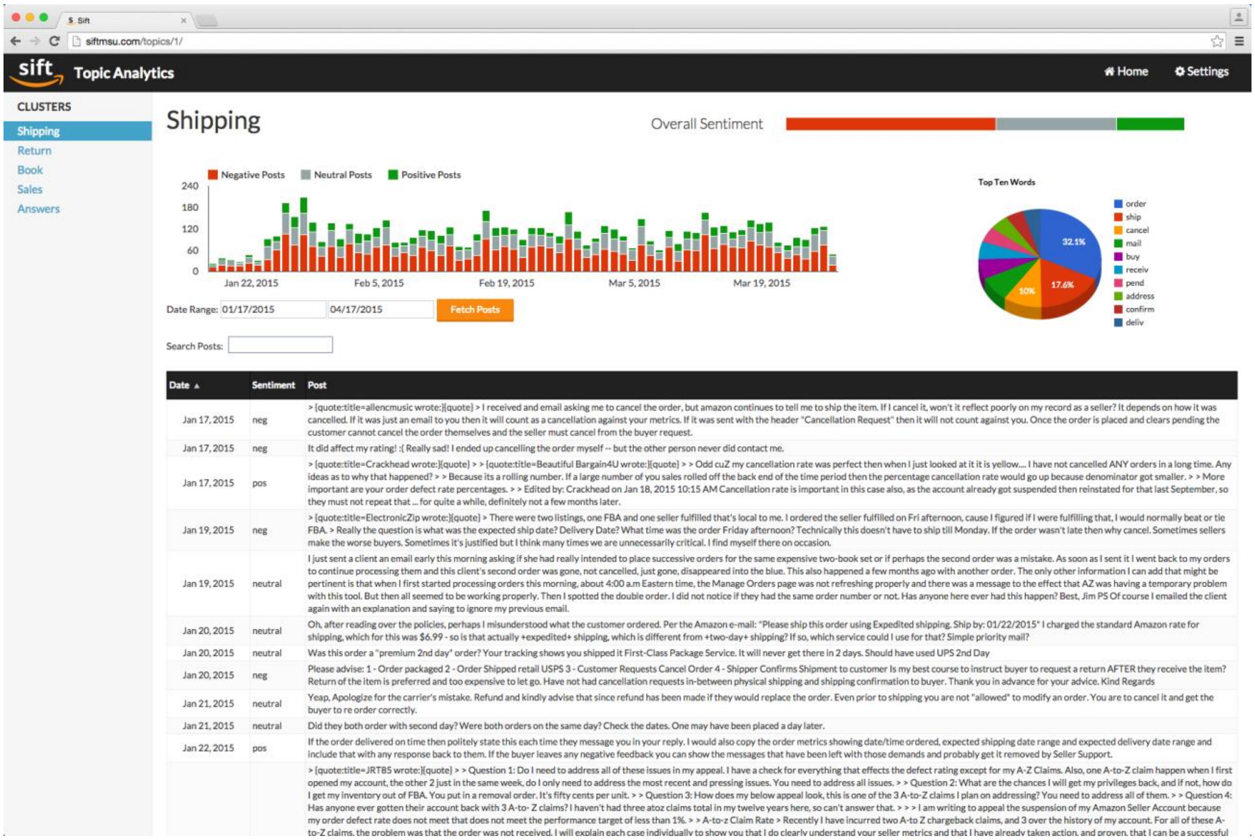


Figure 1 - General Analytics

3.2.2 Topic Analytics

The Topic Analytics page (*Figure 2 and 3*) focuses on one of the topics selected from the left column. Users can query SIFT for a larger date range using the form and "Fetch Posts" button on the page. There are several graphs depicting how the topic is performing, such as the "Sentiment Analysis" bar at the top right and the "Top Ten Words" pie chart just below it.

The column graph, to the left, displays the sentiment breakdown of posts on a day-to-day basis. The table below, which contains a random sampling of five hundred posts from the selected date range, can be sorted by any column but defaults to show older posts at the top. Finally, the search bar allows users to search for posts containing a specific keyword.

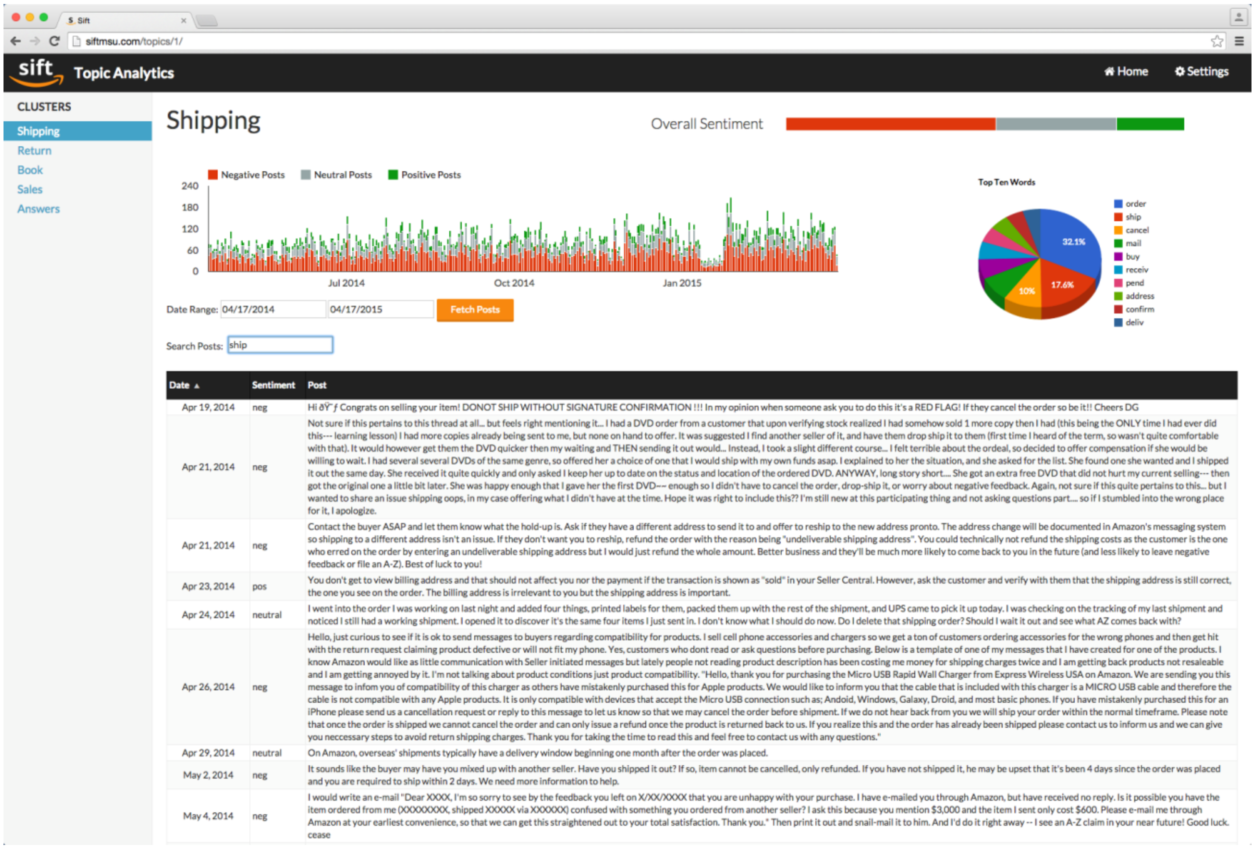


Date Range: 01/17/2015 04/17/2015

Search Posts:

Date ▲	Sentiment	Post
Jan 17, 2015	neg	> {quote:title=allencmusic wrote:}{quote} > I received and email asking me to cancel the order, but amazon continues to tell me to ship the item. If I cancel it, won't it reflect poorly on my record as a seller? It depends on how it was cancelled. If it was just an email to you then it will count as a cancellation against your metrics. If it was sent with the header "Cancellation Request" then it will not count against you. Once the order is placed and clears pending the customer cannot cancel the order themselves and the seller must cancel from the buyer request.
Jan 17, 2015	neg	It did affect my rating! :(Really sad! I ended up cancelling the order myself -- but the other person never did contact me.
Jan 17, 2015	pos	> {quote:title=Crackhead wrote:}{quote} > > {quote:title=Beautiful Bargain4U wrote:}{quote} > > Odd cuZ my cancellation rate was perfect then when I just looked at it it is yellow.... I have not cancelled ANY orders in a long time. Any ideas as to why that happened? > > Because its a rolling number. If a large number of you sales rolled off the back end of the time period then the percentage cancellation rate would go up because denominator got smaller. > > More important are your order defect rate percentages. > > Edited by: Crackhead on Jan 18, 2015 10:15 AM Cancellation rate is important in this case also, as the account already got suspended then reinstated for that last September, so they must not repeat that ... for quite a while, definitely not a few months later.
Jan 19, 2015	neg	> {quote:title=ElectronicZip wrote:}{quote} > There were two listings, one FBA and one seller fulfilled that's local to me. I ordered the seller fulfilled on Fri afternoon, cause I figured if I were fulfilling that, I would normally beat or tie FBA. > Really the question is what was the expected ship date? Delivery Date? What time was the order Friday afternoon? Technically this doesn't have to ship till Monday. If the order wasn't late then why cancel. Sometimes sellers make the worse buyers. Sometimes it's justified but I think many times we are unnecessarily critical. I find myself there on occasion.
Jan 19, 2015	neutral	I just sent a client an email early this morning asking if she had really intended to place successive orders for the same expensive two-book set or if perhaps the second order was a mistake. As soon as I sent it I went back to my orders to continue processing them and this client's second order was gone, not cancelled, just gone, disappeared into the blue. This also happened a few months ago with another order. The only other information I can add that might be pertinent is that when I first started processing orders this morning, about 4:00 a.m Eastern time, the Manage Orders page was not refreshing properly and there was a message to the effect that AZ was having a temporary problem with this tool. But then all seemed to be working properly. Then I spotted the double order. I did not notice if they had the same order number or not. Has anyone here ever had this happen? Best, Jim PS Of course I emailed the client again with an explanation and saying to ignore my previous email.
Jan 20, 2015	neutral	Oh, after reading over the policies, perhaps I misunderstood what the customer ordered. Per the Amazon e-mail: "Please ship this order using Expedited shipping. Ship by: 01/22/2015" I charged the standard Amazon rate for shipping, which for this was $6.99 - so is that actually +expedited+ shipping, which is different from +two-day+ shipping? If so, which service could I use for that? Simple priority mail?
Jan 20, 2015	neutral	Was this order a "premium 2nd day" order? Your tracking shows you shipped it First-Class Package Service. It will never get there in 2 days. Should have used UPS 2nd Day
Jan 20, 2015	neg	Please advise: 1 - Order packaged 2 - Order Shipped retail USPS 3 - Customer Requests Cancel Order 4 - Shipper Confirms Shipment to customer Is my best course to instruct buyer to request a return AFTER they receive the item? Return of the item is preferred and too expensive to let go. Have not had cancellation requests in-between physical shipping and shipping confirmation to buyer. Thank you in advance for your advice. Kind Regards
Jan 21, 2015	neutral	Yeap, Apologize for the carrier's mistake. Refund and kindly advise that since refund has been made if they would replace the order. Even prior to shipping you are not "allowed" to modify an order. You are to cancel it and get the buyer to re order correctly.
Jan 21, 2015	neutral	Did they both order with second day? Were both orders on the same day? Check the dates. One may have been placed a day later.
Jan 22, 2015	pos	If the order delivered on time then politely state this each time they message you in your reply. I would also copy the order metrics showing date/time ordered, expected shipping date range and expected delivery date range and include that with any response back to them. If the buyer leaves any negative feedback you can show the messages that have been left with those demands and probably get it removed by Seller Support.
		> {quote:title=JRT85 wrote:}{quote} > > Question 1: Do I need to address all of these issues in my appeal. I have a check for everything that effects the defect rating except for my A-Z Claims. Also, one A-to-Z claim happen when I first opened my account, the other 2 just in the same week, do I only need to address the most recent and pressing issues. You need to address all issues. > > Question 2: What are the chances I will get my privileges back, and if not, how do I get my inventory out of FBA. You put in a removal order. It's fifty cents per unit. > > Question 3: How does my below appeal look, this is one of the 3 A-to-Z claims I plan on addressing? You need to address all of them. > > Question 4: Has anyone ever gotten their account back with 3 A-to- Z claims? I haven't had three atoz claims total in my twelve years here, so can't answer that. > > > I am writing to appeal the suspension of my Amazon Seller Account because my order defect rate does not meet that does not meet the performance target of less than 1%. > > A-to-z Claim Rate > Recently I have incurred two A-to Z chargeback claims, and 3 over the history of my account. For all of these A-to-Z claims, the problem was that the order was not received. I will explain each case individually to show you that I do clearly understand your seller metrics and that I have already taken action, and proven, that I can be a successful

Figure 2 - Topic Analytics



Date Range: 04/17/2014 04/17/2015

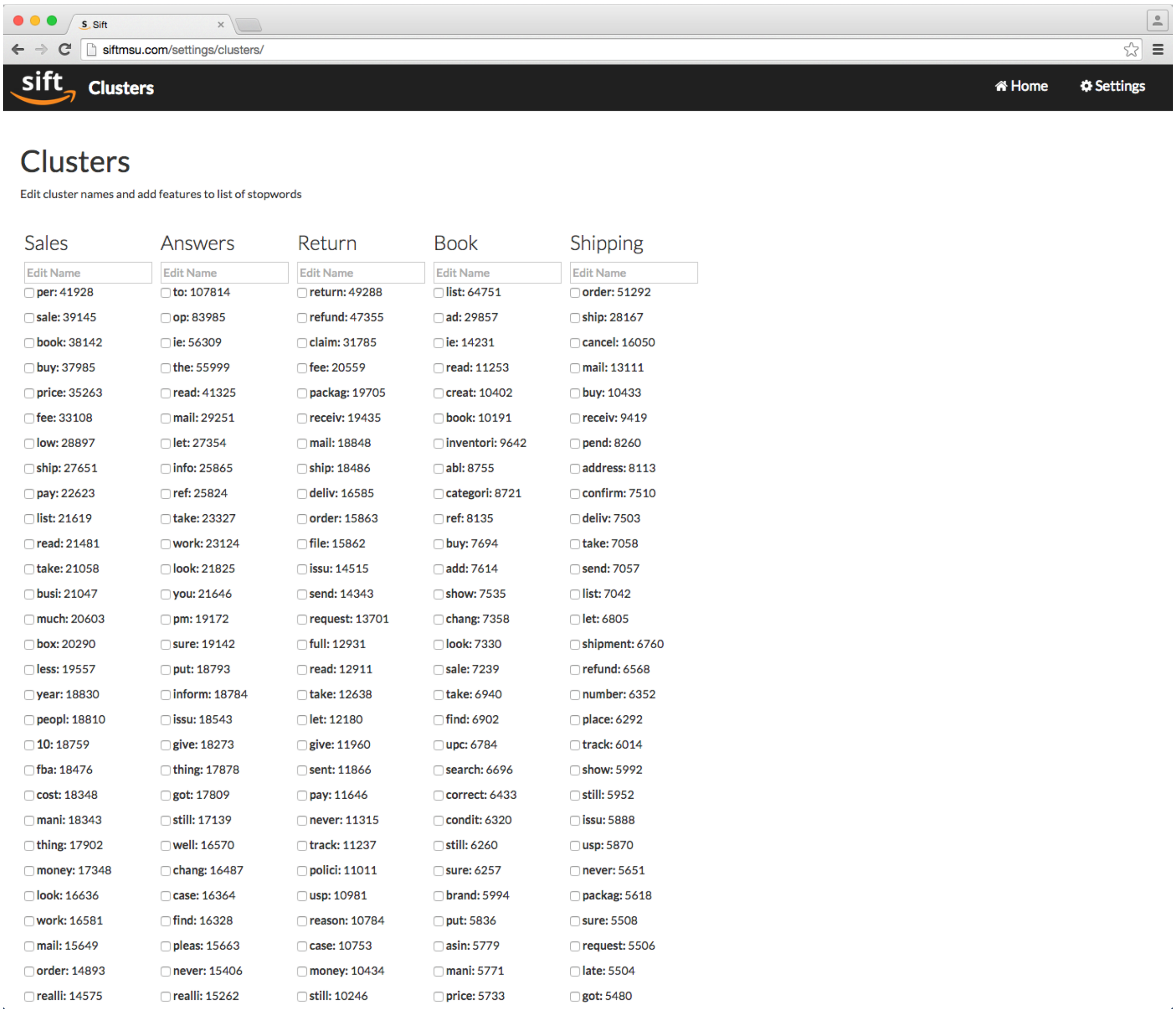
Search Posts: ship

Date ▲	Sentiment	Post
Apr 19, 2014	neg	Hi ðŸ˜ƒ Congrats on selling your item! DONOT SHIP WITHOUT SIGNATURE CONFIRMATION !!! In my opinion when someone ask you to do this it's a RED FLAG! If they cancel the order so be it!! Cheers DG
Apr 21, 2014	neg	Not sure if this pertains to this thread at all... but feels right mentioning it... I had a DVD order from a customer that upon verifying stock realized I had somehow sold 1 more copy then I had (this being the ONLY time I had ever did this--- learning lesson) I had more copies already being sent to me, but none on hand to offer. It was suggested I find another seller of it, and have them drop ship it to them (first time I heard of the term, so wasn't quite comfortable with that). It would however get them the DVD quicker then my waiting and THEN sending it out would... Instead, I took a slight different course... I felt terrible about the ordeal, so decided to offer compensation if she would be willing to wait. I had several several DVDs of the same genre, so offered her a choice of one that I would ship with my own funds asap. I explained to her the situation, and she asked for the list. She found one she wanted and I shipped it out the same day. She received it quite quickly and only asked I keep her up to date on the status and location of the ordered DVD. ANYWAY, long story short.... She got an extra free DVD that did not hurt my current selling--- then got the original one a little bit later. She was happy enough that I gave her the first DVD~~ enough so I didn't have to cancel the order, drop-ship it, or worry about negative feedback. Again, not sure if this quite pertains to this... but I wanted to share an issue shipping oops, in my case offering what I didn't have at the time. Hope it was right to include this?? I'm still new at this participating thing and not asking questions part.... so if I stumbled into the wrong place for it, I apologize.
Apr 21, 2014	neg	Contact the buyer ASAP and let them know what the hold-up is. Ask if they have a different address to send it to and offer to reship to the new address pronto. The address change will be documented in Amazon's messaging system so shipping to a different address isn't an issue. If they don't want you to reship, refund the order with the reason being "undeliverable shipping address". You could technically not refund the shipping costs as the customer is the one who erred on the order by entering an undeliverable shipping address but I would just refund the whole amount. Better business and they'll be much more likely to come back to you in the future (and less likely to leave negative feedback or file an A-Z). Best of luck to you!
Apr 23, 2014	pos	You don't get to view billing address and that should not affect you nor the payment if the transaction is shown as "sold" in your Seller Central. However, ask the customer and verify with them that the shipping address is still correct, the one you see on the order. The billing address is irrelevant to you but the shipping address is important.
Apr 24, 2014	neutral	I went into the order I was working on last night and added four things, printed labels for them, packed them up with the rest of the shipment, and UPS came to pick it up today. I was checking on the tracking of my last shipment and noticed I still had a working shipment. I opened it to discover it's the same four items I just sent in. I don't know what I should do now. Do I delete that shipping order? Should I wait it out and see what AZ comes back with?
Apr 26, 2014	neg	Hello, just curious to see if it is ok to send messages to buyers regarding compatibility for products. I sell cell phone accessories and chargers so we get a ton of customers ordering accessories for the wrong phones and then get hit with the return request claiming product defective or will not fit my phone. Yes, customers who dont read or ask questions before purchasing. Below is a template of one of my messages that I have created for one of the products. I know Amazon would like as little communication with Seller initiated messages but lately people not reading product description has been costing me money for shipping charges twice and I am getting back products not resaleable and I am getting annoyed by it. I'm not talking about product conditions just product compatibility. "Hello, thank you for purchasing the Micro USB Rapid Wall Charger from Express Wireless USA on Amazon. We are sending you this message to infom you of compatibility of this charger as others have mistakenly purchased this for Apple products. We would like to inform you that the cable that is included with this charger is a MICRO USB cable and therefore the cable is not compatible with any Apple products. It is only compatible with devices that accept the Micro USB connection such as; Andoid, Windows, Galaxy, Droid, and most basic phones. If you have mistakenly purchased this for an iPhone please send us a cancellation request or reply to this message to let us know so that we may cancel the order before shipment. If we do not hear back from you we will ship your order within the normal timeframe. Please note that once the order is shipped we cannot cancel the order and can only issue a refund once the product is returned back to us. If you realize this and the order has already been shipped please contact us to inform us and we can give you neccessary steps to avoid return shipping charges. Thank you for taking the time to read this and feel free to contact us with any questions."
Apr 29, 2014	neutral	On Amazon, overseas' shipments typically have a delivery window beginning one month after the order was placed.
May 2, 2014	neg	It sounds like the buyer may have you mixed up with another seller. Have you shipped it out? If so, item cannot be cancelled, only refunded. If you have not shipped it, he may be upset that it's been 4 days since the order was placed and you are required to ship within 2 days. We need more information to help.
May 4, 2014	neg	I would write an e-mail "Dear XXXX, I'm so sorry to see by the feedback you left on X/XX/XXXX that you are unhappy with your purchase. I have e-mailed you through Amazon, but have received no reply. Is it possible you have the item ordered from me (XXXXXXXX, shipped XXXXX via XXXXXX) confused with something you ordered from another seller? I ask this because you mention $3,000 and the item I sent only cost $600. Please e-mail me through Amazon at your earliest convenience, so that we can get this straightened out to your total satisfaction. Thank you." Then print it out and snail-mail it to him. And I'd do it right away -- I see an A-Z claim in your near future! Good luck. cease

Figure 3 - Topic Analytics: Search + Date

3.2.3 Settings: Clusters

The Clusters page (*Figure 4*) allows users to rename cluster topics. Since they are originally set as the top three features/words of the cluster this can be used in order to better represent them. This page also lists the top 100 features/words of each cluster, and allows users to easily select individual words to add to the stop words list.



Clusters

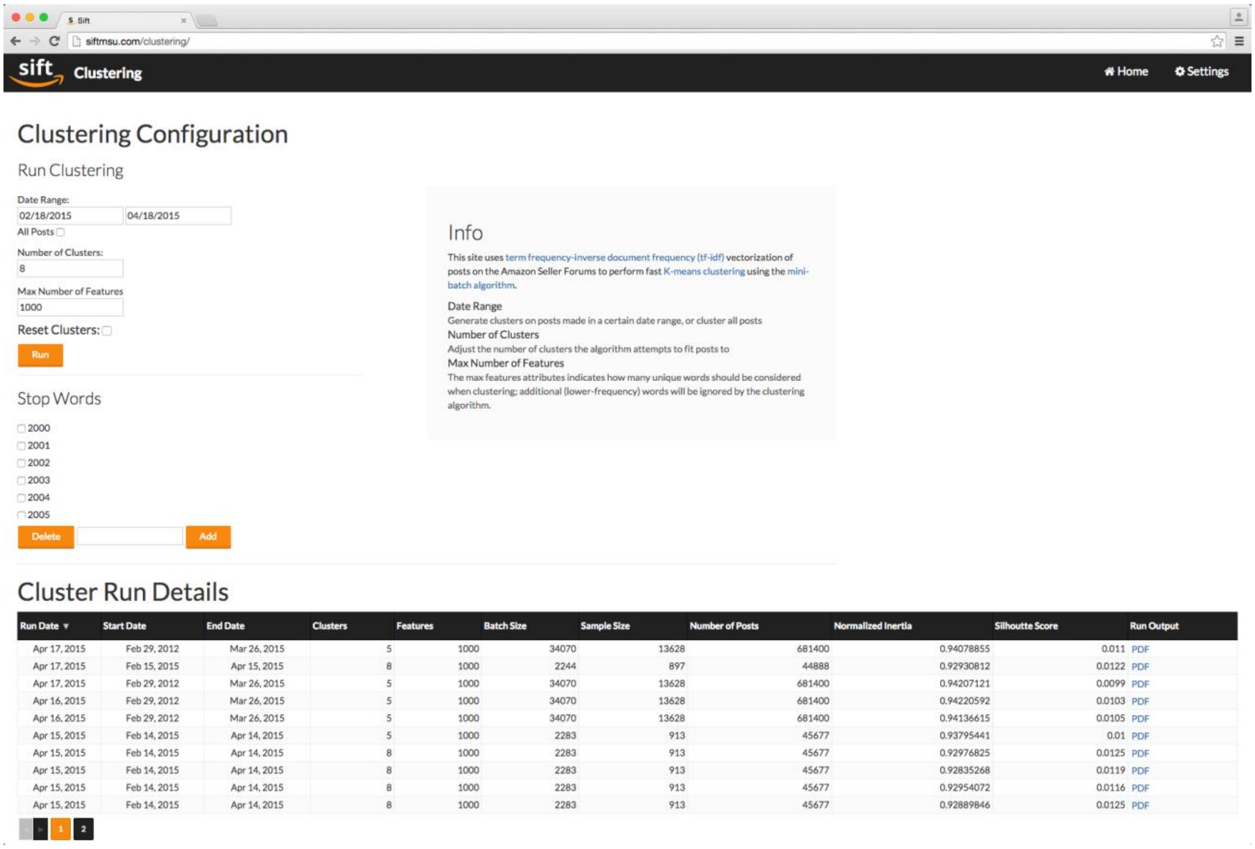
Edit cluster names and add features to list of stopwords

Sales	Answers	Return	Book	Shipping
Edit Name	Edit Name	Edit Name	Edit Name	Edit Name
□ per: 41928	□ to: 107814	□ return: 49288	□ list: 64751	□ order: 51292
□ sale: 39145	□ op: 83985	□ refund: 47355	□ ad: 29857	□ ship: 28167
□ book: 38142	□ ie: 56309	□ claim: 31785	□ ie: 14231	□ cancel: 16050
□ buy: 37985	□ the: 55999	□ fee: 20559	□ read: 11253	□ mail: 13111
□ price: 35263	□ read: 41325	□ packag: 19705	□ creat: 10402	□ buy: 10433
□ fee: 33108	□ mail: 29251	□ receiv: 19435	□ book: 10191	□ receiv: 9419
□ low: 28897	□ let: 27354	□ mail: 18848	□ inventori: 9642	□ pend: 8260
□ ship: 27651	□ info: 25865	□ ship: 18486	□ abl: 8755	□ address: 8113
□ pay: 22623	□ ref: 25824	□ deliv: 16585	□ categori: 8721	□ confirm: 7510
□ list: 21619	□ take: 23327	□ order: 15863	□ ref: 8135	□ deliv: 7503
□ read: 21481	□ work: 23124	□ file: 15862	□ buy: 7694	□ take: 7058
□ take: 21058	□ look: 21825	□ issu: 14515	□ add: 7614	□ send: 7057
□ busi: 21047	□ you: 21646	□ send: 14343	□ show: 7535	□ list: 7042
□ much: 20603	□ pm: 19172	□ request: 13701	□ chang: 7358	□ let: 6805
□ box: 20290	□ sure: 19142	□ full: 12931	□ look: 7330	□ shipment: 6760
□ less: 19557	□ put: 18793	□ read: 12911	□ sale: 7239	□ refund: 6568
□ year: 18830	□ inform: 18784	□ take: 12638	□ take: 6940	□ number: 6352
□ peopl: 18810	□ issu: 18543	□ let: 12180	□ find: 6902	□ place: 6292
□ 10: 18759	□ give: 18273	□ give: 11960	□ upc: 6784	□ track: 6014
□ fba: 18476	□ thing: 17878	□ sent: 11866	□ search: 6696	□ show: 5992
□ cost: 18348	□ got: 17809	□ pay: 11646	□ correct: 6433	□ still: 5952
□ mani: 18343	□ still: 17139	□ never: 11315	□ condit: 6320	□ issu: 5888
□ thing: 17902	□ well: 16570	□ track: 11237	□ still: 6260	□ usp: 5870
□ money: 17348	□ chang: 16487	□ polici: 11011	□ sure: 6257	□ never: 5651
□ look: 16636	□ case: 16364	□ usp: 10981	□ brand: 5994	□ packag: 5618
□ work: 16581	□ find: 16328	□ reason: 10784	□ put: 5836	□ sure: 5508
□ mail: 15649	□ pleas: 15663	□ case: 10753	□ asin: 5779	□ request: 5506
□ order: 14893	□ never: 15406	□ money: 10434	□ mani: 5771	□ late: 5504
□ realli: 14575	□ realli: 15262	□ still: 10246	□ price: 5733	□ got: 5480

Figure 4 - Settings: Clusters

3.2.4 Settings: Clustering

The Clustering page (*Figure 5*) allows users to run diagnostic clustering over a specified date range, number of clusters, clustering type, and number of features. All currently saved stop words are displayed and users can choose to add or remove words. Finally, a table showing the details of previous diagnostic cluster runs is available for easy comparison.



Clustering Configuration

Run Clustering

Date Range: 02/18/2015 04/18/2015

All Posts □

Number of Clusters: 8

Max Number of Features 1000

Reset Clusters: □

Info

This site uses term frequency-inverse document frequency (tf-idf) vectorization of posts on the Amazon Seller Forums to perform fast K-means clustering using the mini-batch algorithm.

Date Range
Generate clusters on posts made in a certain date range, or cluster all posts

Number of Clusters
Adjust the number of clusters the algorithm attempts to fit posts to

Max Number of Features
The max features attributes indicates how many unique words should be considered when clustering; additional (lower-frequency) words will be ignored by the clustering algorithm.

Stop Words

□ 2000
□ 2001
□ 2002
□ 2003
□ 2004
□ 2005

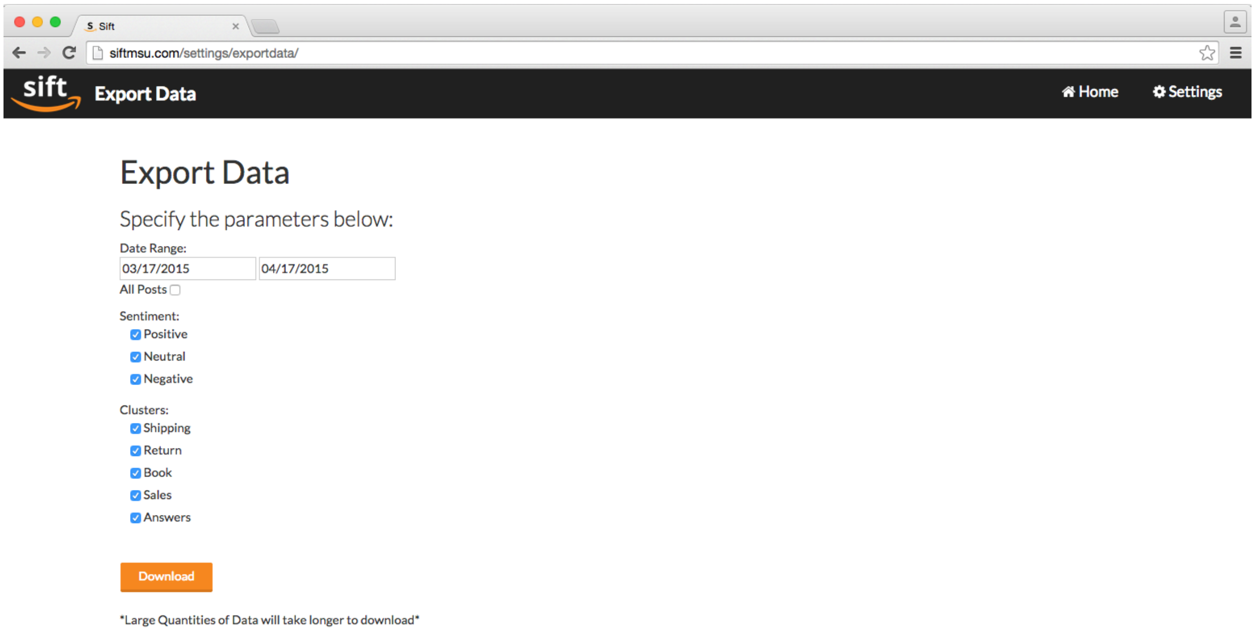
Cluster Run Details

Run Date ▼	Start Date	End Date	Clusters	Features	Batch Size	Sample Size	Number of Posts	Normalized Inertia	Silhoutte Score	Run Output
Apr 17, 2015	Feb 29, 2012	Mar 26, 2015	5	1000	34070	13628	681400	0.94078855	0.011	PDF
Apr 17, 2015	Feb 15, 2015	Apr 15, 2015	8	1000	2244	897	44888	0.92930812	0.0122	PDF
Apr 17, 2015	Feb 29, 2012	Mar 26, 2015	5	1000	34070	13628	681400	0.94207121	0.0099	PDF
Apr 16, 2015	Feb 29, 2012	Mar 26, 2015	5	1000	34070	13628	681400	0.94220592	0.0103	PDF
Apr 16, 2015	Feb 29, 2012	Mar 26, 2015	5	1000	34070	13628	681400	0.94136615	0.0105	PDF
Apr 15, 2015	Feb 14, 2015	Apr 14, 2015	5	1000	2283	913	45677	0.93795441	0.01	PDF
Apr 15, 2015	Feb 14, 2015	Apr 14, 2015	8	1000	2283	913	45677	0.92976825	0.0125	PDF
Apr 15, 2015	Feb 14, 2015	Apr 14, 2015	8	1000	2283	913	45677	0.92835268	0.0119	PDF
Apr 15, 2015	Feb 14, 2015	Apr 14, 2015	8	1000	2283	913	45677	0.92954072	0.0116	PDF
Apr 15, 2015	Feb 14, 2015	Apr 14, 2015	8	1000	2283	913	45677	0.92889846	0.0125	PDF

Figure 5 - Settings: Clustering

3.2.5 Settings: Export Data

The Export Data page (*Figure 6*) allows users to specify a set of data they would like to download. The page will compile these posts from the database and download them in csv format. This allows users to have more freedom in analyzing posts if they so choose.



Export Data

Specify the parameters below:

Date Range:
03/17/2015 04/17/2015
All Posts □

Sentiment:
☑ Positive
☑ Neutral
☑ Negative

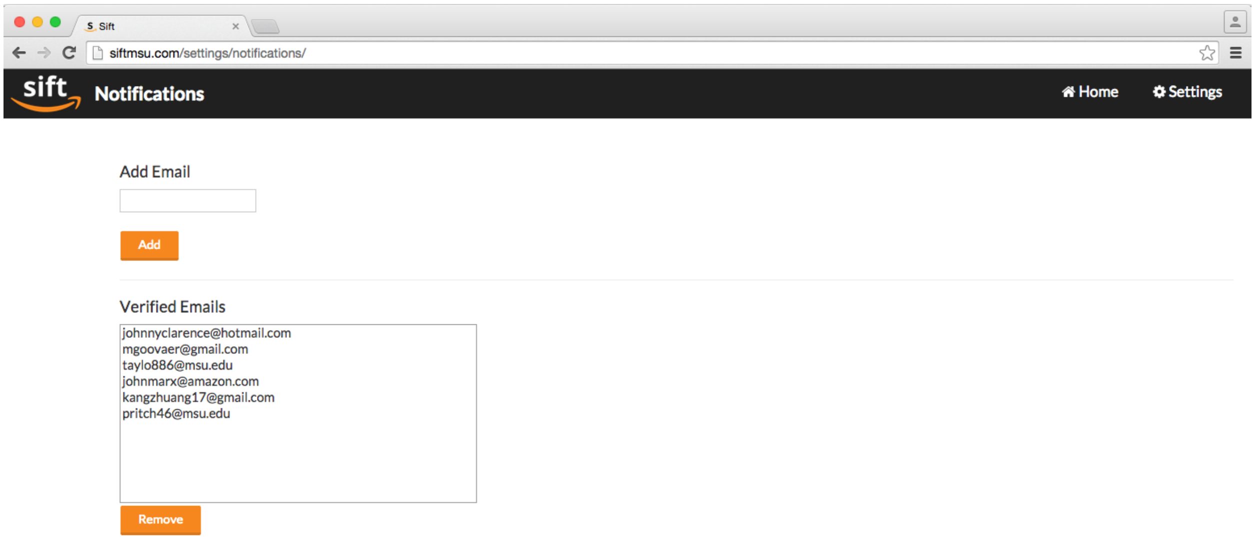
Clusters:
☑ Shipping
☑ Return
☑ Book
☑ Sales
☑ Answers

Large Quantities of Data will take longer to download

Figure 6 - Settings: Export Data

3.2.6 Settings: Notifications

The Notifications page (*Figure 7*) allows users to add or remove emails from the mailing list of SIFTs notifications.



johnnyclarence@hotmail.com
mgoovaer@gmail.com
taylo886@msu.edu
johnmarx@amazon.com
kangzhuang17@gmail.com
pritch46@msu.edu

Figure 7 - Settings: Notifications

4. Technical Specifications

The technical specifications articulate the project's system architecture and what technologies are used.

4.1 System Architecture

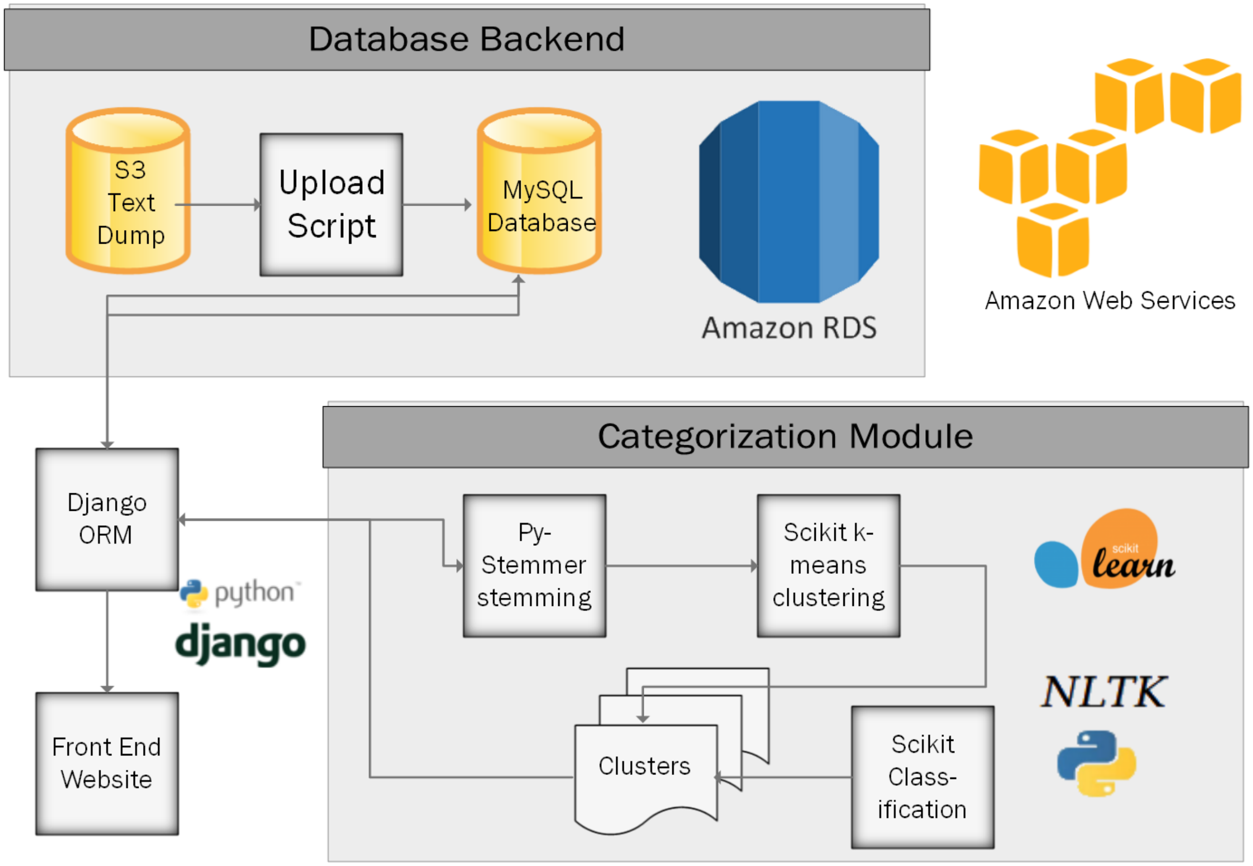


Figure 8 – Overview of System Architecture

Figure 8 details the overall architecture of our system. There are three main components to SIFT: the database backend, the categorization module, and the Django front-end. These all interact through Django's Object Relational Models, and enable content to be easily transferred in-between.

4.1.1 Clustering and Classification Schedule

The clustering schedule details the process of new information being added to the database. Every night the cron job will kick off a series of processes. First, the parser script is used to take all data located in the S3 data dump, and convert it to rows in the MySQL Database. Following this, each post is stemmed, analyzed for sentiment and this new information is uploaded to the database. Following the addition of data to the database, these new posts are placed into categories using the classification script.

Upon completion of the classification script, diagnostic clustering is run. This is run over every post, and while it has no effect on the current clusters residing in the database, its output can be used to determine whether or not adjustments are necessary. This decision is based on the normalized inertia and silhouette score. The normalized inertia is the sum of squared distances from each post to its cluster center. While the silhouette score is another distance based metric that takes into account the distance from each post to its cluster and second nearest cluster. These are metrics regarding the accuracy of the clustering, and can determine if a larger or smaller number of clusters may be necessary.

Following these processes, a notification email is sent to all subscribed users detailing the results of diagnostic clustering.

4.1.2 MySQL Database

Each post featured in the MySQL database (*Figure 9*) contains a variety of information regarding the posts, and is either associated with a cluster, or will be via the nightly classification. The ClusterRun table displays metrics regarding the results of each clustering. This helps show the health of clustering each night, as well as give users the opportunity to run clustering for themselves with personalized input parameters. Finally the ClusterWord table contains the top features associated with each cluster – to assist in recognizing the topics being discussed in the cluster.

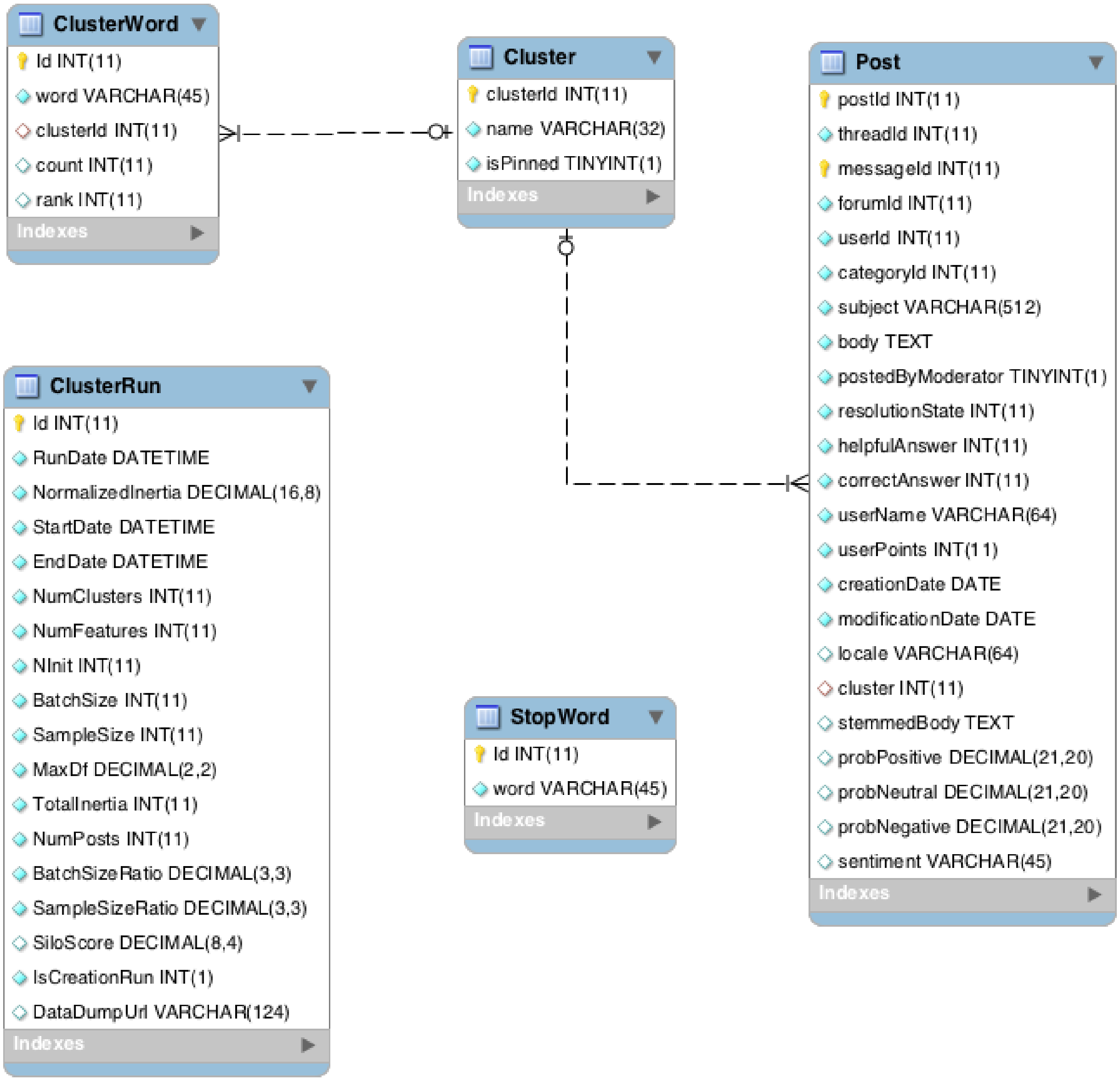


Figure 9 - Database Schema

4.2 Back-End Technologies

4.2.1 AWS S3 Text File Data

All data regarding posts in the third party forums is available via a key-value based text file. These files will be parsed on a scheduled basis (to be determined) and imported to the MySQL database. Since data regarding previous posts does not change, this only needs to occur for the most recent forum posts.

4.2.2 AWS Relational Database Service

RDS provides a scalable cloud solution on which our MySQL database resides. RDS can automatically scale the number or capability of servers the database is operating on if the number of accesses to it quickly rises or decreases.

4.2.3 Django ORM

Django Object Relational Mapping acts as an interface between the three non-database portions of the application and the database. Django ORM maps database tables to objects and provides a convenient format in which to access the database.

4.2.4 AWS EC2 Virtual Server

SIFT is hosted on an Amazon Web Services Elastic Compute Cloud (AWS EC2) virtualized server running Ubuntu 14.04. Static assets are served using the Nginx webserver, while the Django application is run by the Green Unicorn (Gunicorn) WSGI server. Background tasks are enabled via Celery and the Django-Celery module, with task queueing handled by the Redis key-value cache store. These technologies were chosen for their open source licenses and easy integration with Python and Django applications.

4.2.5 Scikit-learn Clustering

Clustering is performed using calls to Scikit-Learns's mini-batch k-means clustering. K-means clustering takes an input of n clusters, where n is the number of clusters to be created via the algorithm. Mini-batch k-means is an optimization on this algorithm wherein clustering is performed on separate groups, with the same cluster centroids. The result is a faster algorithm with almost identical results. As described previously, posts are fit into clusters via a list of features where features are words contained within posts in high frequency throughout the seller forums.

Once the feature list is created the module takes each post and converts it into an n-dimensional point in space where n is the number of features within the feature list. By doing this the clustering algorithm is able to perform its primary function, fitting points into clusters by optimizing for inertia. A cluster is defined as a group of points with an n-dimensional point defining its center. Inertia is the sum of squared distances from each point to its cluster. By optimizing for inertia it can be certain that at a large scale posts will be assigned to clusters that are "closest" to them in terms of n-dimensional feature space. Which, in this case means the post has words found often in the cluster.

4.2.7 Running Clustering

Clustering can be run on a scheduled or ad-hoc basis. Each night clustering is run over all posts via an on server application cron job, this does not reset all clusters, and is run simply as a diagnostic tool. However, upon user request clustering may be run to create entirely new clusters based on given input parameters. Upon being run, the module will consume all new forum posts present in the MySQL database, clustering them into relevant and popular topics. The now clustered posts are then stored in a table within the MySQL database to be accessed by the front end website.

4.2.8 Scikit-learn Classification

Classification uses Nearest-Neighbor classification provided by Scikit-Learn. By supplying the algorithm with a training set of posts already associated with a cluster, the Nearest-Neighbor algorithm determines the best category, of non-classified posts, based on likeness in words in the same way as the clustering algorithm.

4.2.9 Sentiment Analysis

The sentiment analysis is performed on a post by post basis. The post body is sent to mashape, a third party API service which uses NLTK’s sentiment implementation. An overall sentiment tag is returned along with a positive, negative and neutral score.

4.2.10 AWS SES Email Sending Service

AWS Simple Email Service (SES) is an outbound-only email-sending service. SES is used for notifications within SIFT. When diagnostic clustering and routine classification complete it sends an email with details of the run, including the inertia score, and silhouette score.

4.3 Front-End Technologies

4.3.1 Django Web Application

The application utilizes the Model-View web framework Django. The web application queries the MySQL database for information regarding the top clustered posts, which are displayed on the web portal. The website is able to query the database in order to perform searches via user input. It uses this information to generate graphs and visualizations.

4.3.2 Front-End Libraries

Numerous libraries are used to make the user interface. Bootstrap, a front-end framework that contains UI templates, is used throughout the application. SIFT's graphs are displayed with Google Charts. Font-Awesome is used for icons on the navigation bar to give users a visualization of the link (i.e. a house icon for the Homepage).

4.4 Testing and Integration

4.4.1 Nose Unit Testing

Our software makes use of unit testing and the Nose testing framework for Python to ensure correct and reproducible behavior of each module and component of the software.

Nose allows us to write simple, loosely-coupled tests for our code, and automates the running of these tests. It presents the output in an easily readable format, and is capable of generating more detailed information about code coverage at the user's request.

5. Risks

Risks are events that could lead to delayed development or even significant changes in project plan. They can be either known or unknown and vary in severity. This section details a few risks that may affect our progress.

5.1 Unfamiliarity with Technologies

None of us have used SciKit or NLTK. We were not sure if they would work as we intend them to: for Clustering, Categorization and Sentiment Analysis. We mitigated this risk by setting up a test environment early on in order to test these unfamiliar technologies.

5.2 Machine Learning

Machine Learning is a tricky topic. It can get difficult and we did not exactly know how it tied into our project. We mitigated this risk by using existing libraries, NLTK, rather than developing our own implementations of Machine Learning.

5.3 Feature Creep

Feature creep is the ongoing expansion or addition of new features in our project. We mitigated this risk by developing wireframes early and using them to solidify future project features and capabilities. Along with sticking to our detailed timeline (see Section 6).

5.4 Scalability

SIFT has to be scalable in order to properly load and interact with the large quantity of data provided by Amazon. We mitigated this risk by making sure that the technologies we used are scalable, and that SIFT does not attempt to load more data than is necessary. For instance, on the Topic Analytics page, SIFT loads a sampling of data (500 posts) over the specified date range rather than all posts in order to keep page load times reasonable.

6. Timeline

Below is our plan for development with tasks organized in a week by week manner.

Week 1: (1/12 - 1/16)

- Receive project description
- Individual research on project technologies
- Meet team and assign roles
- First client meeting: January 14th

Week 2: (1/19 - 1/23)

- Set up code repository
- Explore given data
- Install and configure computers with specific technologies
- Complete status report presentation and rough draft of Design Specification

Week 3: (1/26 - 1/30)

- Status Report Presentation: January 26th
- Meet with client, solidify technical specifications/technologies to use. Show first wireframes and get initial thoughts.
- Complete Project Plan Rough Draft, email to Amazon on January 28th
- In person visit with Amazon on January 30th go over Project Plan, fix any issues

Week 4: (2/2 - 2/6)

- Project Plan due: February 2nd
- Develop skeleton of front end
- Setup backend: database and simple python script for data analysis

Week 5: (2/9 - 2/13)

- Make website live
- Hookup front end and backend
- Rework front end based on input from Amazon
- Work on *classification* – both front end and backend

Week 6: (2/16 - 2/20)

- Refine Project
- Work on *clustering* – both front end and backend
- Test all features prior to presentation
- Prepare for Alpha Presentation

Week 7: (2/23 - 2/27)

- Alpha Presentation due: February 23rd

- Add *sentiment* functionality – both front end and backend

Week 8: (3/2 - 3/6)

- Add *notification* functionality
- Include “admin” functionality in dashboard to adjust notification features
    - Would include logging in – develop simple authentication system

Week: (3/9 - 3/13)

- Spring Break
- Continue work from previous week

Week 9: (3/16-3/20)

- Identify current standing in project
    - If on schedule begin work on “nice to have”
    - Otherwise complete essential functionality

Week 10: (3/30-4/3)

- Begin rigorous second round of testing, bring in unfamiliar users
- Prepare for Beta Presentation

Week 11: (4/6-4/10)

- Beta Presentation due: April 6th
- Time cushion – plan for this time to be taken for emergencies and unforeseen problems

Week 12: (4/13-4/17)

- Wrap up any final glitches and features
- Do final round of testing and acceptance

Week 13: (4/20-4/24)

- Only critical defects to be worked on this week, no new functionality
- Develop script for project video and record
- Prepare for Design Day presentation

Week 14: (4/27-5/1)

- All Deliverables due: April 27th and 29th
- Design Day Setup: April 30th
- Design Day: May 1st

7. Additional Information

7.1 Point of Contact

For further information regarding this document and project please contact Dr. Wayne Dkysen at Michigan State University. All materials in this document are free of proprietary data. The students and instructor gratefully acknowledge the participation of our industrial collaborators.

7.2 Acknowledgements

We would like to acknowledge the Amazon team for their contribution and effort spent working with us on our Senior Design Capstone Project at Michigan State University. We would like to specifically thank John Marx, Darren Krape and Poornachandra Pesala.

Additionally, we would like to thank the Department of Computer Science at Michigan State University, Malcom Doering and Dr. Wayne Dyksen for preparing us for applying our skills in a business environment.