Project 9

Customer Review Analytics

Major Project

Information Retrieval and Extraction

Team 12

Mentor : Anurag Tyagi

Aniket Jain
Kaveri Anuranjana
Sayantan Hensh

Problem Statement

Understand the sentiment of user reviews and provide useful information for the end-user as well as the product manufacturer regarding public opinion of the product.

Abstract

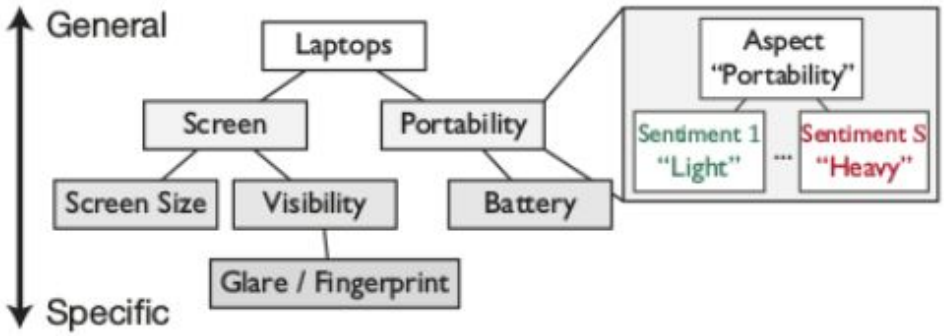
Sentiment Analysis is a widely addressed Natural Language Processing task wherein the semantic orientation of a text unit is adjudged. However, a major challenge in Sentiment Analysis is the identification of entities towards which the opinion is expressed. Sentitool (Aspect Based Sentiment Analysis system) receives as input a set of texts (product reviews) discussing a particular entity (e.g., a new model of a mobile phone). The systems attempt to detect the main (e.g., the most frequently discussed) aspects (features) of the entity (e.g., 'battery', 'screen') and to estimate the average sentiment of the texts per aspect (e.g., how positive or negative the opinions are on average for each aspect). It involves the extraction of the aspect term from a sentence and secondly the polarity of the opinion corresponding to that aspect is adjudged. We adopted an approach based on Probabilistic Graphical Models(PGMs). A linear-chain CRF is trained with features based on word vectors and text processing techniques(POS, dependency parse) to sequentially label the aspect term in a sentence. SVM classifier then identities the polarity corresponding to the aspect, with features based on cosine similarity with words from sentiwordnet.

Introduction

The term aspect refers to the features or aspects of a product, service or topic being discussed in a text. Sentiment analysis refers to identification and extraction of subjective impressions from text sources. It aims to determine the attitude of a person with respect to something in particular or the overall contextual polarity of a document.

In general, a binary composition of opinions is assumed: for/against, like/dislike, good/bad etc. However, an opinion may also be categorized into a neutral sentiment.

When a review or a social media post talks about a product or service, the user might want to discuss multiple aspects or sub-topics related to the product or service being discussed. For example, in a restaurant review, while the customer might have good things to say about the food quality offered at a restaurant, she might be disappointed with the service offered to her, and she might think the decor needs to be revamped. So a general sentiment analyzer that determines the overall sentiment towards the product or service might not be able to capture the full essence of the review. Hence the need for Aspect-based Sentiment Analysis, for better and more fine-grained analysis of user feedback, which would enable service providers and product manufacturers to identify those business aspects that needs improvement.



Recent years has seen rapid growth of research on sentiment analysis. Sentiment analysis has both business importance and academic interest. So far, most sentiment analysis research has focused on classifying the overall sentiment of a document into positive or negative. We would, however, often like to understand what are the specific sentiments towards different aspects of an entity, e.g. a restaurant review ”Food is decent but service is so bad.” contains positive sentiment towards aspect food but

strong negative sentiment towards aspect service . Classifying the overall sentiment as negative would neglect the fact that food was actually good.

The ultimate End goal is to be able to generate summaries listing all the aspects and their overall polarity. The outcome will be average sentiment for each aspect of an entity.

Related Work

We came across various works that were on the same lines as our project.
Our basic inspiration for this problem statement is from [4]. We further found a blend of other ideas in [1] and [2] in their papers and use the same framework as theirs albeit with some changes. Further we have taken some considerations based on [3] as they present a new approach to phrase-level sentiment analysis which determines if an expression is neutral or polar and then disambiguates the polarity of the polar expressions.

Challenges

- Storing and retrieving huge corpus efficiently.
- Also, reviews are sometimes noisy and filled with grammatical mistakes. Theses problem are tough to handle. Dataset was cleaned and duplicates were removed in order to handle this.
- Problems with identifying the aspect, which is being talked about in a given review and its corresponding sentiment. Example – *“The new iPhone has a bad camera but a long lasting battery.”* , Here the sentiment for camera is negative while that of the battery is positive.

- If the review is comparative in nature i.e talks about multiple products in a single review. Example – "*I recently bought an iPhone. But my Samsung Grand has a better processor."* , here iPhone was the original product but the review does not talk about it directly.
- Problems with anaphora resolution. Example – *"The new iPhone has a lithium battery. It is really bad"*, here its difficult for the system to answer the question what does the "it" refer to ?

Applications

Parameter-based sentiment analysis of user reviews would allow us to give a detailed feedback to the manufacturer. Such a feedback would help them understand if the general public is unhappy with a certain aspect of their product and hence can help them modify it accordingly. For example, users may be unhappy with the screen resolution in the new iPhone 6s mobile). It can also be used to develop new products with emphasis on those particular parameters.

Such an analysis also helps us provide a targeted recommendation system for the users. For example, we can provide suggestions for products with good sentiment on screen resolution to users who might have complained about the same in their previous reviews.

Dataset

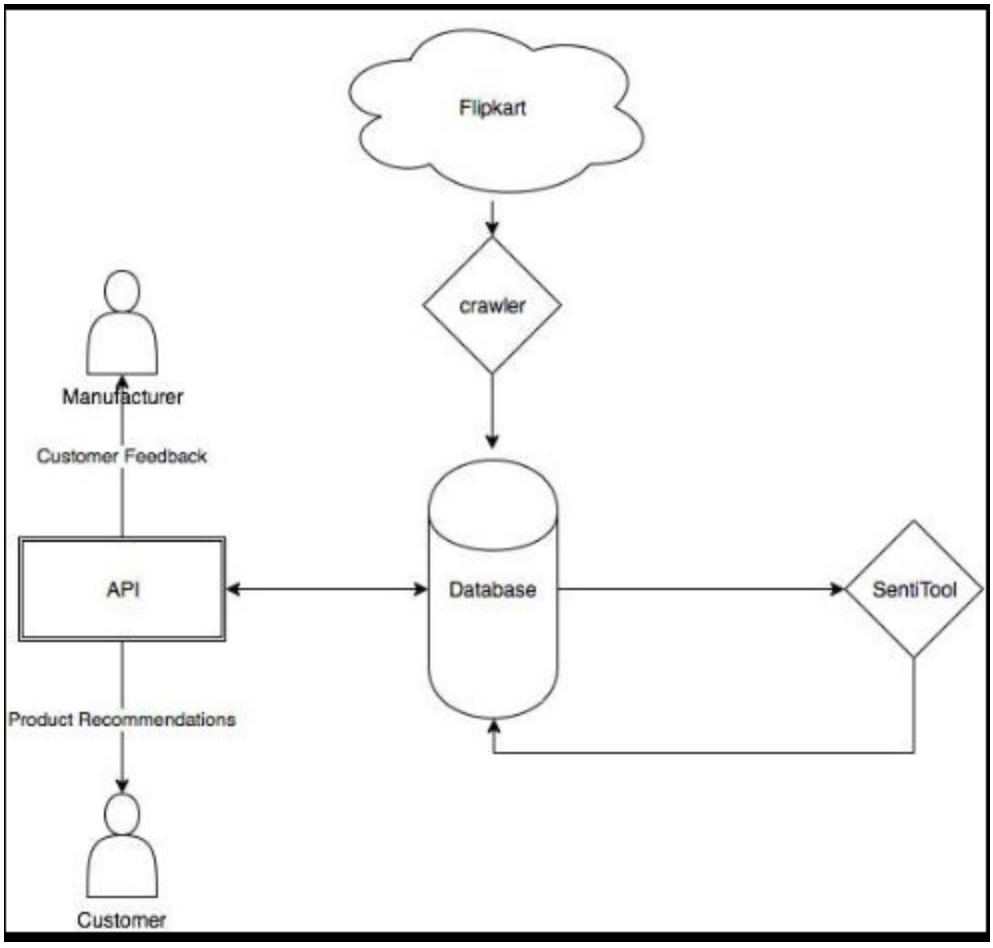
The site was crawled to obtain the reviews. The data extracted had to be cleaned again to remove reduplicated data (over 2gb). It is divided into various categories (books, toys, movies, etc.)

Approach

The project can be divided into three major tasks namely data extraction and processing, aspect and its category detection and assigning sentiment polarity.

Data Extraction involves collecting data (user reviews and other meta-data) from popular ecommerce websites.

Processing step converts unstructured data (raw html) into a structured format (relational tables) which can be used by our tool to determine the various aspects and their corresponding sentiments for each product.



Aspect Category (Entity and Attribute). Identify every entity E and attribute A pair E#A towards which an opinion is expressed in the given text. E and A should be chosen from predefined inventories of Entity types (e.g. laptop, keyboard, operating system, restaurant, food, drinks) and Attribute labels (e.g. performance, design, price, quality) per domain. Each E#A pair defines an aspect category of the given text.

Sentiment Polarity. Each identified E#A pair of the given text has to be assigned a polarity, from a set P = {positive, negative, neutral}.

Data Extraction and Processing

For the purposes of this project, user reviews were collected from e-commerce website: flipkart.

Tools used :

Scrapy : contains mechanisms for crawling and scraping

www.scrapy.org

Selenium : contains mechanisms to render javascript and ajax enabled web pages

http://www.seleniumhq.org/docs/

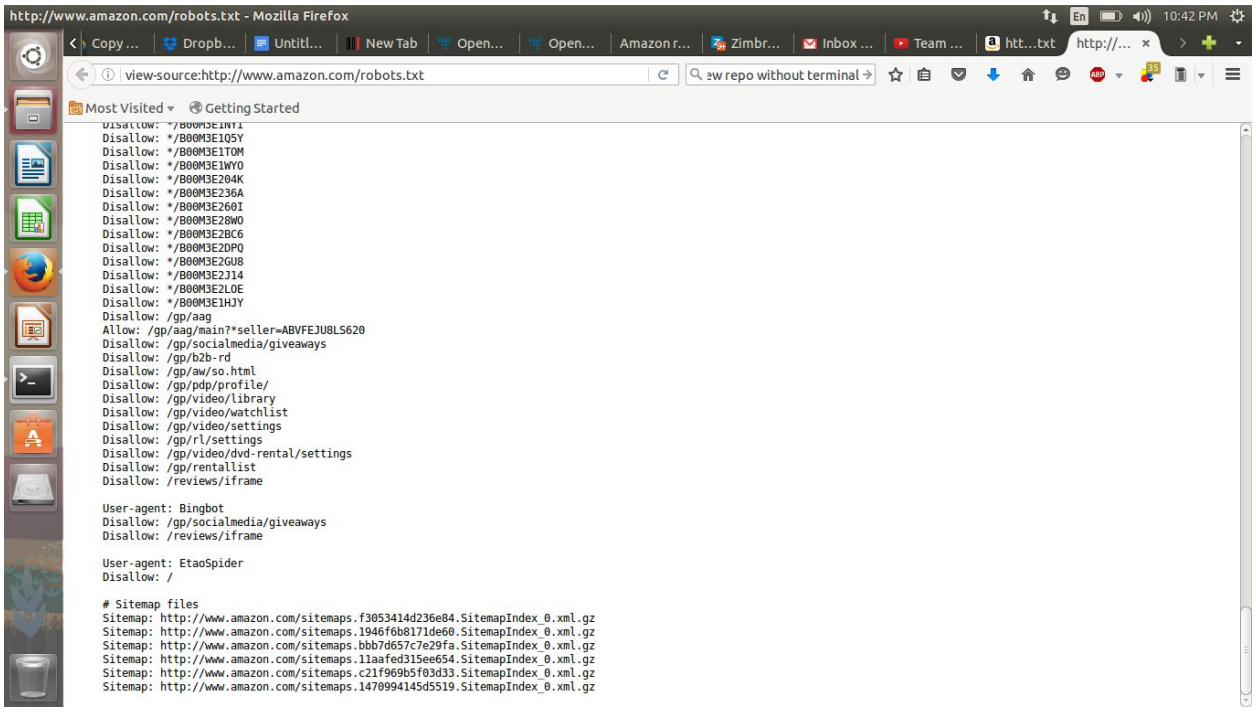
The crawling process was divided into 2 steps:

1. Collect a complete list of products under various categories (electronics, clothing, home appliances etc)
2. Collect all the user reviews for each product.

The data was stored in a relational database to allow for easy access in the future.

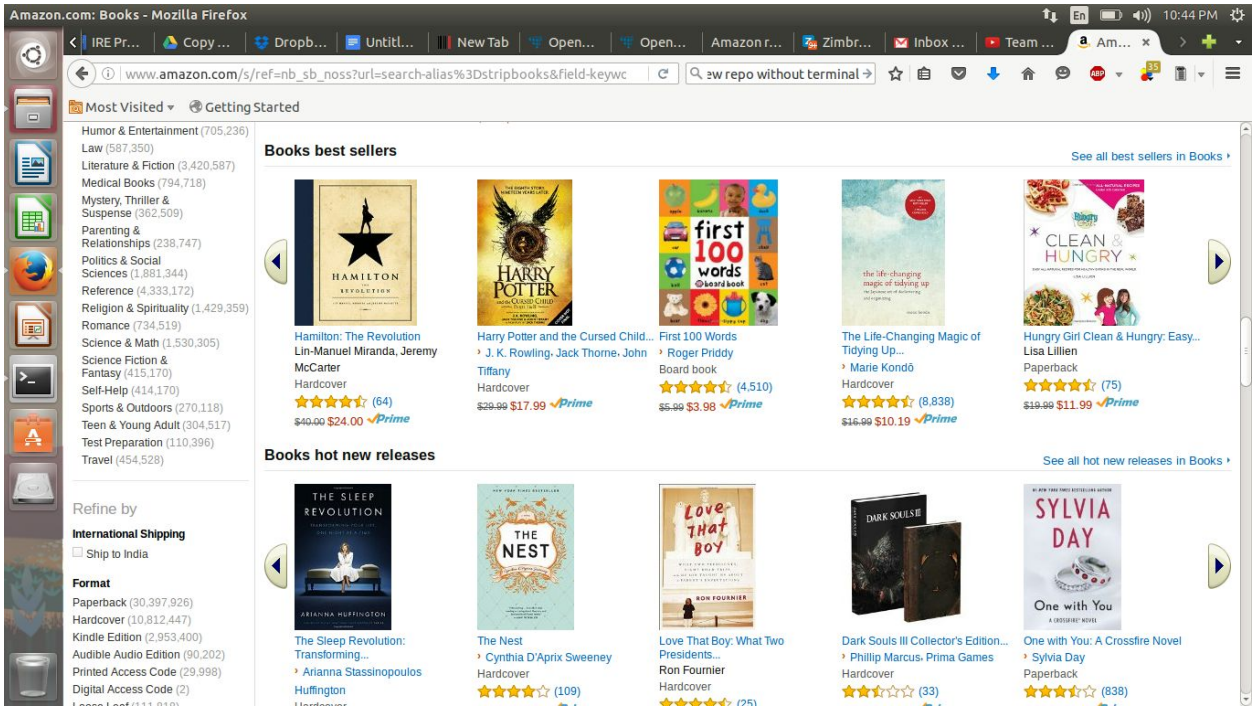
The following section explains the process of crawling in detail using the example of flipkart:

1. http://www.amazon.com/robots.txt was used to obtain the sitemap. The various urls provided in the sitemap were used to create the seed set to start the crawl.



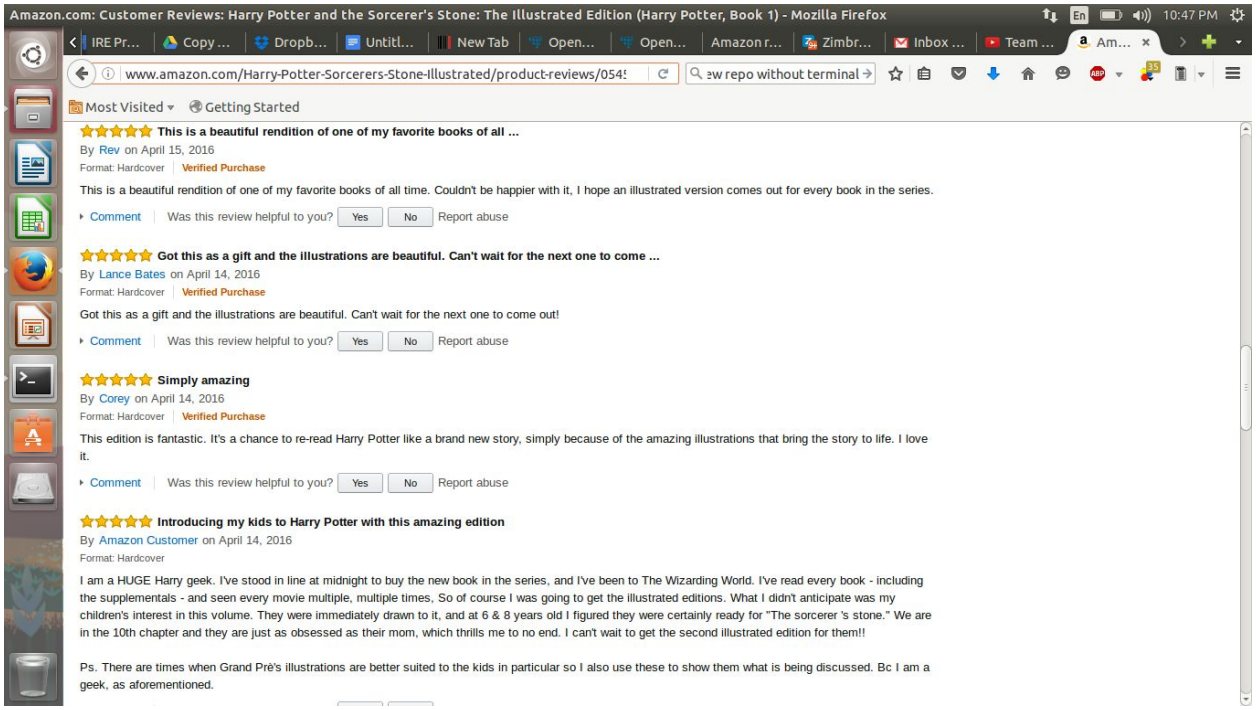
sitemap for amazon.com

2. Each of the seed urls provide a list of all products for a particular category.



All products in the books category

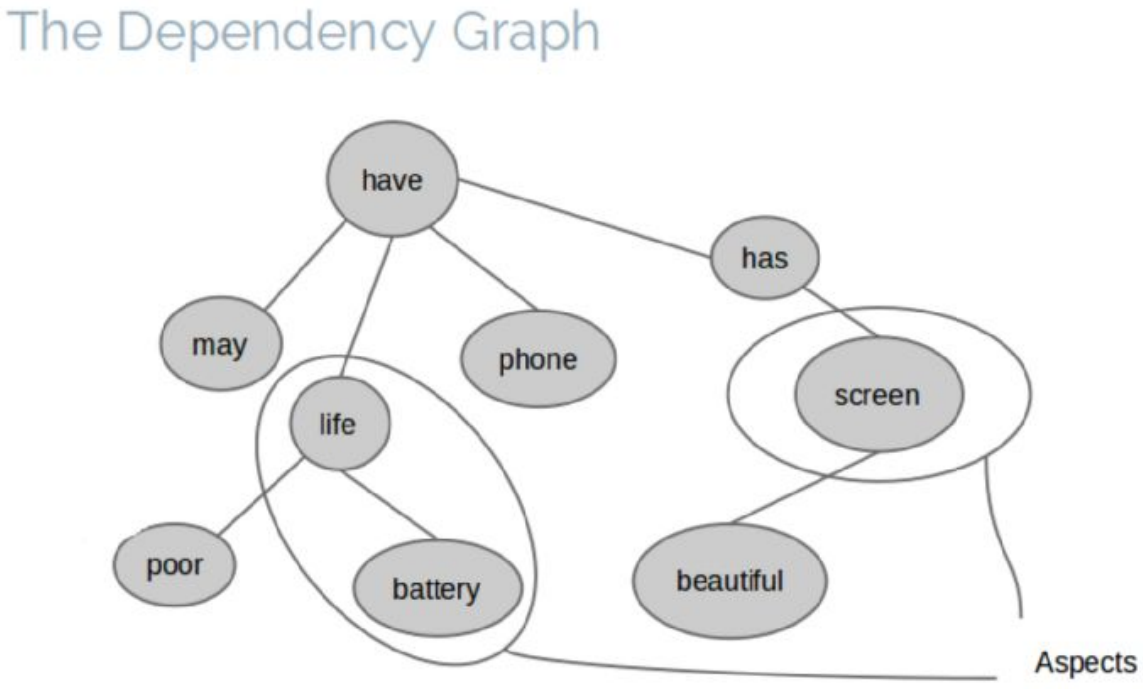
3. The urls for each product were then followed to obtain the complete list of user reviews.



All the user reviews for *Harry Potter Sorcerer's Stone Illustrated* phone

4. The crawled data was stored in a relational database. We decided to use MongoDB to store our data.
    - It is an industry standard capable of storing large volumes of data while still maintaining the speed of retrieval.
    - The SentiTool will use the reviews (along with meta data such as user rating of the product, reliability of the user, rating of the review itself etc) to provide sentiment to the various aspects of the product being discussed.

Aspect Detection



- Identify every Aspect or (Entity E) towards which an opinion is expressed in the given text. To solve this problem, a combination of both Machine Learning and rule based approaches was used.
- An Named Entity Recognition (NER) model was implemented using Conditional Random Fields (CRFs) making use of different types of contextual information along with a variety of features such as word prefixes and shapes that are helpful in predicting the different named entity (NE) classes.
- The model uses inputs of the form - The O service ASPBEGIN at O Saul ASPBEGIN Martin ASPCONT is O the O best O. These interdependencies are learnt and tagged according to the CRF model. In case this fails to identify aspects (generally for vague reviews), a rule-based model which identifies all noun-phrases and then filters out the common ones was implemented.
- When multiple aspects are present in a sentence, general purpose sentiment analysers are not quite useful as various aspects may portray conflicting polarities. Example: “The service was splendid but the food was inedible”. Here,

the extracted entities give out opposite polarities and hence a decentralized approach has to be taken for this task.

- When there exists only 1 aspect in the review, a general purpose and normally more accurate sentiment analysis model can be used

Aspect	Representative Words	Aspect	Representative Words
Performance	power, performance, mode, fan, quiet	Mouse	mouse, right, touchpad, pad, buttons, left
Hardware	drive, wireless, bluetooth, usb, speakers, webcam	General	great, little, machine, price, netbook, happy
Memory	ram, 2GB, upgrade, extra, 1GB, speed	Purchase	amazon, purchased, bought, weeks, ordered
Software	using, office, software, installed, works, programs	Looks	looks, feel, white, finish, blue, solid, glossy
Usability	internet, video, web, movies, music, email, play	OS	windows, xp, system, boot, linux, vista, os
Portability	around, light, work, portable, weight, travel	Battery	battery, life, hours, time, cell, last
Comparison	netbooks, best, reviews, read, decided, research	Size	screen, keyboard, size, small, enough, big

The above table presents the expected aspects and their representative words.

Category Detection

Category should be chosen from predefined inventories of Attribute labels (e.g. design, performance, price, quality) per domain.

- Using a pre-tagged corpus, an SVM model was used to learn the corpus to predict the category to which the extracted corpus belonged.
- This is not exactly equal to categorizing sentences directly as the aspect plays a vital role in deciding the output. In SVM, the TF values for the aspects manually increases to a value where it affects the output as per requirements.
- The feature space considering all the n-grams and eliminating stop words comes to around 8000 dimensions.

Some examples highlighting these annotations are given below :

(1) It fires up in the morning in less than 30 seconds and I have never had any issues with it freezing. → {LAPTOP#OPERATION_PERFORMANCE}

(2) Sometimes you will be moving your finger and the pointer will not even move. → {MOUSE#OPERATION_PERFORMANCE}

(3) The backlit keys are wonderful when you are working in the dark. → {KEYBOARD#DESIGN_FEATURES}

(4) I dislike the quality and the placement of the speakers. {MULTIMEDIA DEVICES#QUALITY}, {MULTIMEDIA DEVICES#DESIGN_FEATURES}

(5) The applications are also very easy to find and maneuver. → {SOFTWARE#USABILITY}

(6) I took it to the shop and they said it would cost too much to repair it. → {SUPPORT#PRICE}

(7) It is extremely portable and easily connects to WIFI at the library and elsewhere. → {LAPTOP#PORTABILITY}, {LAPTOP#CONNECTIVITY}

Sentiment Polarity

To obtain metric for features of the products we need to identify the opinion of that feature from various reviews available. We view the goal of reading multiple reviews as finding widely-held opinions and weighing the positive against the negative, and we wish to automate this sort of task using NLP and machine-learning techniques.
Each identified Aspect and Category pair of the given text has to be assigned a polarity, from a set P = {positive, negative, neutral}. The neutral label applies to mildly positive or mildly negative sentiment as in examples 3 and 4 below. When the polarities of the aspects are found, the corresponding polarities of the categories is tuned accordingly and the final polarity of a category is maintained and updated with each review.

(1) The applications are also very easy to find and maneuver.

{SOFTWARE#USABILITY, positive}

(2) We were planning to get dessert, but the waitress basically through the bill at us before we had a chance to order.

{SERVICE#GENERAL, waitress, negative}

(3) It does run a little warm but that is a negligible concern. → {LAPTOP#QUALITY neutral}

(4) The cameras are nothing out of the ordinary” → {PHONE#CAMERA, “cameras”, neutral}

(5) I bought this laptop yesterday. → {}

(6) The yoga pads are their first such products → {}

Challenges

- The dataset has over 18gb worth data. Here, the size becomes and issue and an opitimised code is required (the data not only has to be read, but learned as well).
- The amount of data to be crawled from amazon was in gbs. Using no proxies leads to the ip being blacklisted. Measures needed to be taken for that.

Conclusion

We ran our model on Amazon Review dataset crawled by us. The tool works perfectly on any type of reviews single or compound reviews. It generates summaries listing all the aspects and their overall polarity. The outcome is be average sentiment for each aspect of an entity.

The only drawback is that it uses SentiWordNet which gives polarity of the aspects and the polarity accuracy is not so good. This work can be extended to include other data sets. A more accurate polarity generator can be used instead of SentiWordNet.

References

[1] Minqing Hu, Bing Liu ., ”Mining Opinion Features in Customer Reviews.” Proceedings of Nineteeth National Conference on Artificial Intellgience (AAAI-2004).

[2] Minqing Hu, Bing Liu ., ”Mining and summarizing customer reviews.” Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper)

[3] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann ,. ”Recognizing contextual polarity in phrase-level sentiment analysis” In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT 05.

[4] Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, Saif Mohammad., ”Detecting aspects and sentiment in customer reviews” In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014).

[5] Wagner J. et al. (2014) DCU: Aspect-based Polarity Classification for SemEval Task 4, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) pp. 223 - 229, Dublin, Ireland, August 23-24, 2014.

[6] Kiritchenko S. et al. (2014) NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) pp. 437 - 442, Dublin, Ireland, August 23-24, 2014.

[7] Brychcin T. et al. (2014) UWB: Machine Learning Approach to Aspect-Based Sentiment Analysis, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) pp. 817 - 822, Dublin, Ireland, August 23-24, 2014.

[8] Brun C. et al. (2014) XRCE: Hybrid Classification for Aspect-based Sentiment Analysis, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) pp. 838 - 842, Dublin, Ireland, August 23-24, 2014.