WSCAR: Web Scraping and Contextual analysis of Reviews

1. GOAL

The objective of this project is to create an application that can process the reviews given by several customers for a product on Amazon and provide recommendations based on the product reviews. Everytime a customer looks for a product on amazon, they spend a lot of time going through the reviews and deciding if the product is as good as advertised. Our application would cut short that time giving automatic recommendations about the product to the customer so that they can make an informed decision. Our application will make use of web scraping and Natural Language Processing (NLP) algorithms to analyze the reviews. It is an interesting project because millions of customers access amazon to shop for their products and our application provides a real world case use. Judging the reviews of the customers is a challenge for our project where we intend to use advanced ML algorithms to process the reviews and make suggestions.

2. LITERATURE REVIEW

Introduced in the 1950s, Natural Language Processing (NLP) is a branch of artificial intelligence and linguistics. Its focus is on investigating computer-assisted natural language understanding and manipulation. For this project, we will be referencing the journal paper authored by Arwa S. M. AlQahtani named " Product Sentiment Analysis for Amazon reviews'' which is one of the applications of Natural Language Processing. The primary focus of this paper is opinion mining based on different aspects of the customer reviews.

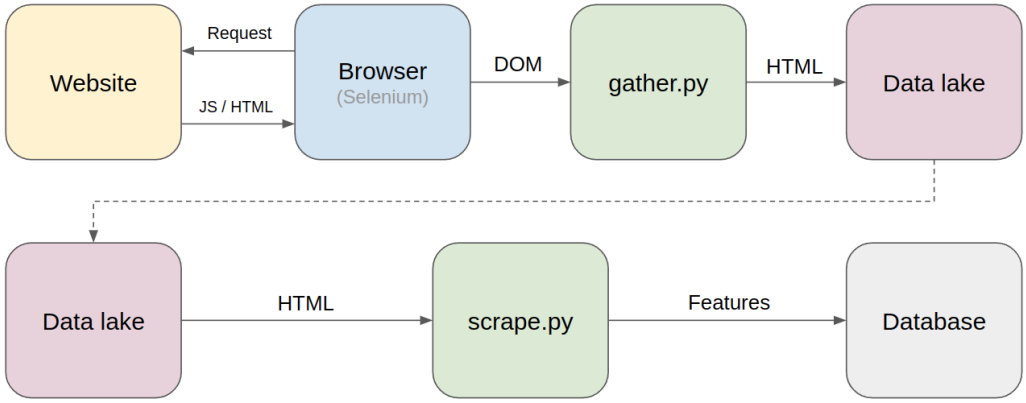
Opinion mining is identifying interests and preferences in regards to a specific product. Pang, Lee, and Vaithyanathan proposed a sentiment categorization using machine learning algorithms in the dataset of movie reviews in the publication cited in 2002. On the basis of the Naive Bayes, Max Entropy, and Help Vector Machine models, they investigated sentiment analysis on monographs and data bigrams. In their experiment, SVM and unigram function extraction produced the greatest results. They recorded 82.9% accuracy. For our application we will be implementing this algorithm to get suggestions for a specific product based on the customer reviews.

3. METHODOLOGY

Our implementation is a combination of two different algorithms, the first being a web scraper (i.e to scour the internet/particular website for a given set of parameters/product) and the second a NLP algorithm to determine contextual clues regarding a particular attribute for a given product. (i.e: to determine wether the searched has positive or negative reviews).The pipeline for web scraping and the Natural Language Processing is explained in further detail below

a. Web Scraping:

Web scraping requires two parts, namely the crawler and the scraper. The crawler is an artificial intelligence algorithm that browses the web to search for the particular data required by following the links across the internet. The scraper, on the other hand, is a specific tool created to extract data from the website. The design of the scraper can vary greatly according to the complexity and scope of the project so that it can quickly and accurately extract the data.

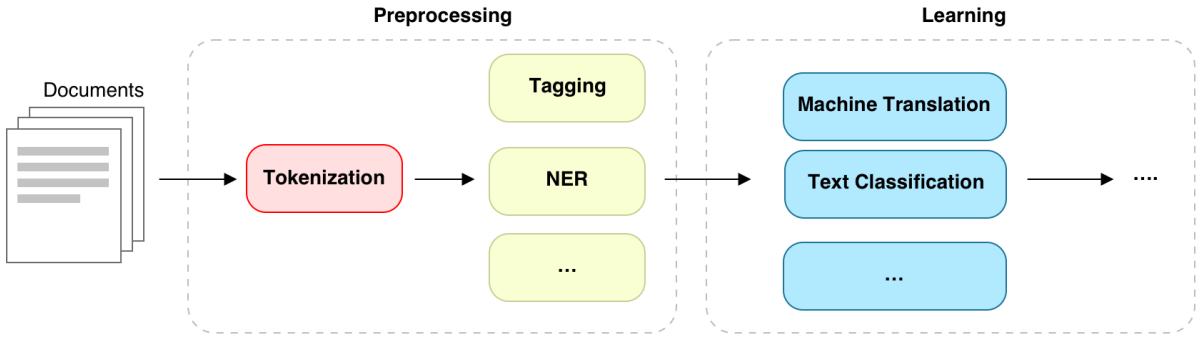


b. Natural Language Processing (NLP):

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

In particular we are applying this technology to product reviews by searching for two primary events. The first being the occurence of the product's attribute that we are interested in (eg: an electronic device's battery life/durability), the second being the contextual clues present in the user reviews. Wether these are positive or negative, if positive then by what margin and if negative then by what margin. Collecting such contextual information from a vast number of reviews gives us the ability to generate a prediction.



4. RESULTS

Our application will get the desired product name from the user and search through amazon website using web scraping. It will automatically fetch reviews from multiple products and process them to analyze if it's a positive or negative review. Based on the positiveness of the reviews of several products that would match the user's keyword, appropriate recommendations can be made to the user. Since the amazon review dataset is openly available we can easily validate the results by comparing the purchase quantity and the rating's review.

5. FUTURE IMPROVEMENTS

We can make our application further accessible to the public by implementing a website using Django or Flask framework were users can input their product keywords and get immediate recommendations.

6. TEAM MEMBERS

a. Sri Sai Charan Velisetti (117509755)
b. Pranav Limbekar (118393711)
c. Vaishant Ramraj (118154237)