EAS 504 - Applications of Data Science - Industry Overview

Assignment 5

Name - Aboli Rawanhate

UB number - 50374341

Q1. Describe the market sector or sub-space covered in this lecture

E-commerce is the market sector or sub-space covered in this lecture by Mr. ManojKumar Kannadasan. Ecommerce is the purchasing and selling of goods and services through the internet. Since its inception, the usage of e-commerce has grown dramatically. This area covers a wide range of activities, including online banking, payments, shopping, auctions, and more. The e-commerce website is developed in such a way that it generates the most money for the organization, and data science is utilized extensively to build each of these features. Any e-commerce company's purpose is to search and discover things for consumers to buy, with the goal of not just assisting users but also maximizing revenue or profit per session. Data Science methodologies are used in a variety of departments such as Search, SEO, Trust/Fraud/Abuse, Selling, Shipping, Pricing, Merchandising, Ads / Marketing, Structured Data, Inventory Management, Machine Translation, Coupons & Rewards, Customer Service, and Infrastructure to provide a seamless way to purchase, increase revenue, and engage more users on the platform.

Along with text search, the e-commerce website aids with faceted search, image search, voice search, and recommendations. Many approaches, such as query classification, query autocompletion, and spelling corrections, are used to enhance the search on websites. Statistics, Linear algebra, Probability, Machine Learning, Natural Language Processing to extract intent-entities from query, Deep Learning, Recommendation Engine and Data Mining are all used in these strategies.

Q2. What data science related skills and technologies are commonly used in this sector?

Linear Algebra, Probability Theory, Machine Learning, Deep Learning, and Data Mining are some of the data science abilities that are often employed in E-commerce. FastCat is an efficient text classification library that takes less time to train and produces high inference outputs. It is one of the data science-related technologies that is often used in E-commerce. Word2Vec, fastText, and GloVe are semi-supervised machine learning technologies that aid in comprehending the user context by learning features based on prior queries.

- Operations like Word hashing, Query transitions, and Semantic feature extraction are conducted using Linear algebra notions like vectors, rank, similarity, and linear transformations.

- Language and Error models employ probability ideas like Bayes Theorem, Posterior Probability, and Markov assumption to improve the search engine with sophisticated spell correction procedures.
- To forecast appropriate product categories given a query, deep learning techniques such as the Deep semantic similarity model and the Convolutional Latent semantic model are used.
- Clustering and classification are two data mining techniques that are used for a variety of objectives.
- Because the volume of data in this field is so large, techniques like Spark and Hadoop are utilized to cope with it.
- Recommendation system is used to recommend products to users.
- The enormous database is handled via cloud computing, and database query language is used to access the database.
- In subspaces like shipping, selling, search, customer support, pricing, and advertising, exploratory data analysis, predictive analysis, and regression are applied.
- In Text Mining, NLP techniques are used to extract data to match terms in a user's search query or to analyze customer reviews.
- For spelling corrections and query autocompletion, Convolutional Neural Networks are employed.

Q3. How are data and computing related methods used in typical workflows in this sector? Illustrate with an example.

Data science process followed in sector is -

- Capture - Data collection from various sources
    - Data Acquisition
    - Data Entry
    - Signal Reception
    - Data Extraction

- Maintain - maintaining data to perform further operations
    - Data Warehousing
    - Data Cleaning
    - Data Staging
    - Data Processing
    - Data Architecture

- Process - Most important step in data science cycle to identify correct sample of data and build model on that

    - Data Mining
    - Clustering / Classification
    - Data Modeling
    - Data Summarization

- Analyze - Analyze the data to answer some business questions
    - Exploratory / Confirmatory Data Analysis
    - Predictive Analysis
    - Regression
    - Text Mining
    - Qualitative Analysis

- Communicate - How to communicate with client/leadership
    - Data Reporting
    - Data Visualization
    - Business Intelligence
    - Decision Making

The technique of maximizing the posterior probability of a category given a query is known as query categorization. To provide the best results, query classification requires a scalable approach. To show relevant products, a recommendation system is used to recommend top k items to the user based on their previous searches and product similarity. Ranking method is used to rank listing of items based on the query and user feedback. The cosine similarity and posterior probability estimated by softmax are used to assess the relevance to other categories of product. Another example of systematic data and computational interactions is query autocompletion. Query auto completion is a mechanism that assists users in avoiding spelling mistakes and obtaining search results more quickly. The query data is utilized as training data for machine learning models that are either supervised or semi-supervised. Query understanding is also done and attribute extraction is used to extract entities from the queries and provide relevant results.

Also, incorrectly answered queries are again retrained to improve accuracy of model. User typed queries are processed using NLP and answered. Language translation also takes place if the user types a query in French and the search result is in Japanese, he will get a response in French. Users can also search using audio records, this will then process to text and give the results. Chatbot is also in function on e-commerce websites to communicate with the user what they need and provide responses based on that. Conversational AI is used in this.

Faceted search means filter criteria, narrow results, identify right categories, customer conversion is predicted using a metric and Machine Learning algorithm.

Q4. What are the data science related challenges one might encounter in this domain?

The data science challenges we encounter in this domain are:

- The massive volume of data, as well as the unstructured nature of the data, has significant consequences for processing and obtaining appropriate outcomes.
- In this industry, data science is also utilized to identify clients who should be rewarded with discounts and presents. However, anticipating faithful purchases is challenging, making it tough to identify clients in whom to invest.
- Identifying fraud seller or buyers might be challenging if we don't have any previous data about them
- Memory-intensive methods and non-compressed suffix trees will be the slowest to execute. Using them will cause latency issues.
- Platform should be able to serve users simultaneously and process every transaction and request.
- Terms that have never been seen before can cause issues during query classification since there will be less prediction across those words.
- Long searches will have weak probability estimates during query autocompletion, resulting in incorrect predictions.
- Maintaining trust between users by providing money back policy if needed but identifying genuine customers is difficult.

Q5. What do you find interesting about the nature of data science opportunities in this domain?

There are many interesting use of data science in this sector, some of them are below:

- Search engine - searching something from search engine and then landing on e-commerce website. For that data science is used to organize information such that for particular search products from that site are displayed.
- SEO - Using data science for driving traffic from various internet components to the website
- Trust/Fraud - To provide money back guarantee to users and to provide security to customer data, to identify fraud sellers/buyers various ML models are used.
- Selling - Various data science techniques are used to estimate demand for the seller also to suggest techniques to maximize revenue
- Shipping - Estimating the right delivery date, for that data science is used
- Pricing - How to show users affordable and quality products compared to other platforms, this is done using machine learning.
- Merchandising - To promote items to the users and to increase sales it is used.
- Advertisement - Another way of sponsored recommendations, email campaigns are carried using data science.

- Structured Data - iphone should be mapped to apple, memory - information from items and products, attributes extracted from product and mapped so that if user searches on the basis of attribute, ML model suggests items based on that. Inventory management - Product graph, relationship between product, needed for graph search
- Machine Translation - For International customers, local language translation is available.
- Coupons and Rewards - Customer acquisition improves. To whom we should suggest coupons.
- Customer service - Making this process seamless, chatbots are used to answer customer queries.
- Infrastructure - optimizing site performance, managing virtual machines, automatically ramping up and ramping down traffic, monetary systems, identifying bots - bots try to crawl your website and slow the system down, building internal platforms. For all these, data science is used.

Q6. Please discuss how sellers and buyers may need different data features in an ecommerce platform such as eBay.

EBay is one of the most well-known ecommerce giants. This marketplace has a significant amount of duty for both customers and sellers. It must improve each interaction in order to improve the next one. Ebay's goal is to provide the greatest possible experience for both consumers and sellers.

Sellers :

- To estimate the requirement of the product so that they will be able to keep up with the demand
- Inventories can be managed, product is mostly sold in which region that can be identified and respective warehouses can be provided with maximum items of the products.
- Based on the preferences of clients, the vendor can adjust prices, offer discounts, and sell in volume as needed.
- Suggesting different ways to maximize revenue, improve quality or pricing based on reviews or comparing other sellers for similar products.
- Identifying fraud purchase and being fair to sellers also in case of product return by checking condition of the product and policies by the seller.

Buyers :

- To provide a product at the lowest cost and provide a good quality product.
- Providing filter categories to refine and narrow their searches
- Recommend products based on their previous purchase
- Suggesting reduced pricing on the items in their wishlist
- Providing auto completion of the queries they are typing

- Also providing spell correction and results even on incorrect spellings
- Showing results quickly and showing relevant products based on their search to keep them engaged and providing good service
- Providing Trust by keeping product return policies
- Being transparent by providing the product's features and functioning.
- Providing reviews and ratings, helping them to make correct decisions
- On online platforms, coupons and rewards are frequently offered, enticing buyers to purchase more things.

Q7. Describe briefly the algorithmic steps involved in query correction as described in the lecture.

The query correction algorithm assists consumers in entering difficult-to-spell product names. Extracted keywords from the user's text is used as an input query in this method. The Bayes theorem is used to choose choices that are comparable to the input query. The candidate's efficiency is increased by selecting relevant terms from a list of established options. We employ models like DAWG and Suffix trees to increase efficiency. The memory footprint of DAWG and non-compressed suffix trees will be large, while the execution time of nave and compressed suffix trees will be the slowest. The Language Model is used to find comparable terms, using Naive Bayes theorem and Markov's assumption. Error models are used, which employ measurements such as logs, keyboard distances, and phonetic distances to identify the proper word to replace the incorrectly spelled one. The words obtained from the error model are then subjected to a precision ranking algorithm. The top-ranking term will be substituted for the incorrectly spelled word.

Q8. MCQs:

Q1**.** C
Q2. B
Q3. C
Q4. B
Q5. D