EAS 504: Applications of Data Science – Industrial Overview – Spring 2023

-Lecture by Abhishek Singh Tomar

Name: Tananki Ranga Sai Saran Rohit

UB ID: 50441793

UB Email: rangasai@buffalo.edu

Ques 1 : Describe the market sector or sub-space covered in this lecture :

The market sector or sub-space covered in this lecture is how data and ML used in Enterprise with an example of search engine. Information retrieval in computing and information science is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Information retrieval is the process of extracting relevant information from collections of unstructured or structured data sources. This includes finding, retrieving, and displaying information in ways that are meaningful and useful to users. Finding data patterns makes information retrieval crucial in the fields of data science and machine learning. Finding data patterns through supervised or unsupervised learning is the foundation of machine learning. There are several ways to do this, one of which is through employing information retrieval techniques to locate pertinent material. it is a collection of algorithms that improves the relevancy of presented materials to searched queries. A search engine's efficacy is influenced by the quality of its data retrieval methods, as well as the fineness and volume of the material it has indexed. The success of search engines and other information-based applications depends on information retrieval.

Ques 2 : What data science related skills and technologies are commonly used in this sector?

In this sector the main architectural components covered are web crawling, query processing system, indexing system . Queries submitted by users are translated into pertinent search results in search engines through a process known as query processing. Understanding the user's intent and providing the most pertinent search results are the objectives of a search engine's query processing system. There are various ways that query processing is related to data science abilities. For instance, machine learning techniques are used by search engines to enhance their query processing capabilities. These algorithms may be created by data scientists, which will

increase the precision of the search results. Indexing is the process of creating a database of web pages that can be swiftly searched and retrieved in response to user queries in search engines. Few External factors of web crawl factors mentioned are site behavior, social objectives ,Spatial locality ,Web growth , web change. The indexing system is in charge of gathering information about web sites, categorizing it, and storing it in such a way that it can be searched quickly and efficiently. Indexing and data science are related in a number of ways. Natural language processing algorithms, for instance, may be used by data scientists to decipher the meaning of content on web pages. In order to find trends in the data and enhance the indexing system's accuracy, they can also employ machine learning methods. Inquiry-based systems are an important part of search engines, and data science abilities like machine learning, natural language processing, and data analysis are required for their creation and optimization. These abilities allow search engines to deliver more tailored and relevant search results, enhancing the user experience and increasing engagement. Data scientists may also contribute to the improvement of the recommendation engine by monitoring user behavior and determining the most efficient strategies for displaying search results. This entails examining click-through rates, bounce rates, and other data to identify which search results are effective and which may be improved.

Ques 3 : How are data and computing related methods used in typical workflows in this sector? Illustrate with an example.

In general, data and computational methods are critical to the success of search engines, from data collection and analysis to search results generation, search results optimization and performance analysis. The ability to work with large data sets, use machine learning and predictive modeling, and understand n optimization tools is key to success in information retrieval and search engine optimization. Overall, data and computing-related technologies are crucial in the creation and optimization of search engine ad targeting algorithms. They let search engines to provide relevant advertisements to users, increasing engagement and income. Search engines utilize ad targeting algorithms to offer appropriate advertising to users based on their search queries and activity. These algorithms enhance the accuracy and efficacy of ad targeting by utilizing a number of data and computing-related methodologies. Search engines have become an essential component of our daily lives in the digital era. We utilize them to get information on almost everything, from new recipes to breaking news. Yet, advertising is the search engine's primary source of revenue. Search engines take user data, evaluate it using machine learning algorithms, locate relevant adverts, and present them in an easy-to-read style to display ads that are relevant to the user. Google , Bing ,Yahoo and other search engines capture user data such as search history, location, and activity. To understand the user's preferences and interests, this data is examined using machine learning techniques. If a person repeatedly looks for recipes and visits cooking websites, the search engine will recognize the user as someone who enjoys cooking. Advertisers selling cooking-related items or services bid on cooking-related keywords, and the search engine's algorithm chooses the most relevant adverts to show to the user. These advertisements are provided in an easy-to-read style and are clearly labeled as advertisements. Those who do not want to see targeted adverts can opt out or delete their search history.

Ques 4 : What are the data science related challenges one might encounter in this domain?

In Web search ,huge amounts of data needs to be continuously searched and optimized from various other sources like different social media platforms, browsers, plug-ins, extensions, mobile phones, applications etc…The 3 main things which makes the web search difficult are Size, Dynamicity and Diversity. Some of the data science challenges that can be encountered are data variability - different types of data , quantity - data size, speed - processing speed where data can often be complex and difficult to process and data can contain missing values, which are some of the obstacles that can be encountered in Web search. The internet is a worldwide platform with material in several languages. Indexing web pages in several languages provides a huge issue for data scientists, necessitating the development of algorithms capable of processing and understanding content in multiple languages. Machine learning algorithms, which are used to analyze user data and show advertisements, might be biased, resulting in unjust treatment of specific users. This might happen as a result of biased training data or algorithmic judgments. The internet's information is constantly changing, and data scientists must create algorithms that can keep up with the rate of change. This necessitates regular index updates as well as algorithms capable of detecting and prioritizing the most recent information. Data scientists must create algorithms that can give users with relevant search results. Understanding user intent, studying user behavior, and continually refining ranking algorithms to give the greatest search experience are all required.

Ques 5 : What do you find interesting about the nature of data science opportunities in this domain?

Increasing demand in the use of data is exponentially growing from past few years in online shopping, use of mobile phone apps , social media ,tech devices etc…. For instance, the amount of data in 2010, which was being created every two days, and in 2021 it was being created just for every 40 minutes. Data scientists have many intriguing opportunities to work on projects requiring cutting-edge machine learning and natural language processing in the field of web search and search engines. With the potential for data-driven decision making, scalability and efficiency, cognitive search, and an emphasis on security and privacy, this is an interesting and challenging field for data scientists to work on. Machine learning algorithms, which are used to analyze user data and show advertisements, might be biased, resulting in unjust treatment of specific users. The internet's information is constantly changing, and data scientists must create algorithms that can keep up with the rate of change. Natural language processing (NLP) may be used to create conversational search interfaces, which allow users to engage with search engines using natural language. This necessitates the creation of complicated algorithms capable of understanding and processing natural language in real time. NLP may be used to assess text sentiment, allowing search engines to better grasp the tone and emotion underlying search requests and site content. This can help to inform customized suggestions and improve user experience overall.

(i) What's the difference between a forward index and an inverted index? (10 pts of the 80 C+R points in the rubric))

The fundamental difference between a forward index and an inverted index is in their structure and capability for query processing. A forward index is built around documents and their words, whereas an inverted index is built around terms and the papers in which they occur. In search engines, inverted indices are often more efficient, although forward indices are more widely utilised in text processing applications. Query processing in a forward index entails scanning each item to locate terms that match the query, which may be time-consuming for huge collections of documents. In contrast, because the index is arranged by words, query processing in an inverted index entails recognizing the pages that contain the query terms, which is typically quicker.In applications that need positional information, such as text processing and information retrieval, a forward index is frequently utilized. In search engines, an inverted index is often employed to promote rapid and efficient document retrieval. A search engine displays a list of online pages where you may enter keywords and receive results.A forward index is a listing of papers and the terms that occur in them. A online search engine scans the web, compiling a list of documents and determining which terms exist on each page. Comparing the page ID to the Word identifier, the reverse index provides a map of the relationship between the two. The inverted index contains a list of terms as well as the publications in which they appear. In the web search example, your search query and the web search yield the documents. An inverted index, the term is the primary unit of retrieval, and each term contains a list of documents in which it appears.

(ii) Describe the high level architectural components of web search. (10 pts of the 80 C+R points in the rubric))

The high-level architectural components of web search can be broadly classified into 3 categories: crawling, indexing, query processing. Crawling is the technique through which search engines dispatch a team of robots (referred to as crawlers or spiders) to find new and updated material. The crawler begins by retrieving a set of seed URLs before following the links on those pages to locate additional URLs to retrieve. This is an iterative process, and the crawler may view millions of web sites in a short period of time. The crawler stores the retrieved pages and metadata in a database, including the URL, title, and latest changed date. Search engines process and store information they find in an index, a huge database of all the content they’ve discovered and deem good enough to serve up to searchers. Creating an index of words and documents requires parsing and evaluating the gathered pages. The indexer analyzes the pages it has retrieved, extracts the text, and tokenizes it into distinct words. A web index is a repository for Internet data. When your page has been crawled, it will then be indexed. These databases are used by search engines to hold billions of pages of information. Based on a user's search query, query processing entails obtaining and ranking documents. The search engine initially examines the user's query to determine their purpose when they input a query.