Name: Sahithi Dodda
Person Number: 50441731
UB Email: sahithid@buffalo.edu

Assignment 6

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

Lecture by Abhishek Singh Tomar – Enterprise Search

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

Ans: This lecture examined the market sector or sub-space of how data and machine learning are employed in enterprises using search engine as example. Information retrieval (IR) is the field of computer science that deals with the processing of documents containing free text, so that they can be rapidly retrieved based on keywords specified in a user's query. Finding important information from a variety of structured or unstructured data sources is known as information retrieval (IR). It involves looking for, finding, and giving users-friendly information that has meaning and value. Information retrieval is essential to tasks like text categorization, sentiment analysis, targeting ads and recommendation systems in data science and machine learning. These activities entail pulling pertinent data from huge databases and making judgments based on that data. For customers to receive results that are relevant to their searches, search engines like Google and Yahoo significantly rely on information retrieval. When a user types in a search query, the search engine pulls pertinent data from its index and presents it in a way that is helpful and understandable to the user.

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

Ans: Search engine query processing is a crucial component, and data science abilities like sentiment analysis, machine learning, and natural language processing are crucial for its optimization and enhancement. Moreover, query processing may be used by data scientists to do sentiment analysis and natural language processing. This enables them to comprehend the underlying meaning of the search query and acquire insights into the user's intent. In many respects, data science abilities are linked to indexing. Data scientists, for example, can utilize natural language processing techniques to extract meaning from content on web pages. Machine learning techniques may also be used to find patterns in data and increase indexing system accuracy. Few web crawl factors for performance include scalability, URL mapping, Repartitioning, Load balancing. The search results may then be enhanced, and the user can access more pertinent material thanks to data scientists. Search engine marketing strategies may be optimized by data scientists using query processing. They can determine the most efficient keywords and phrases to target in their marketing efforts by examining user inquiries and search results. Building a database of web pages for speedy search and retrieval in response to user queries is the process of indexing in search engines. The indexing system has the role of gathering data about web sites, categorizing it, and storing it so that quick and effective searching is possible. Data scientists, for example, can utilize natural language processing techniques to extract meaning from content on web pages. They can also employ machine learning techniques to find trends in data and increase the indexing system's accuracy. In search engines, inquiry-based systems relate to the capability of providing users with individualized search results based on their search history, activity, and other characteristics. These systems are intended to recognize the user's intent and deliver search results that are most relevant to their individual requirements.

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

Ans: When a user submits a query into a search engine, the search engine first parses and understands the query using natural language processing (NLP) techniques. NLP assists the search engine in comprehending the structure and meaning of the question, as well as identifying the most relevant keywords and phrases. After parsing and understanding the question, the search engine employs indexing to fetch a list of web sites that are relevant to the user's query. Indexing is the act of gathering and arranging information about web pages, such as their text, links, and metadata, so that it may be searched quickly and efficiently. Then, the search engine employs machine learning techniques to rank the search results based on relevance and significance after getting a list of relevant web sites. To assess the relevance and value of each search result, machine learning algorithms examine a range of criteria, including the content of the web page, its popularity and authority, and the user's search history and activity. Lastly, the search engine presents the user with the search results in an easy-to-read manner, such as a list of links with brief descriptions or snippets. The objective is to present the most relevant and significant information connected to the user's query so that they can quickly locate what they're looking for. Ultimately, the process of processing and comprehending queries, obtaining relevant web pages, rating search results, and presenting them to users is fundamental to search engine functionality. In this process, NLP, indexing, machine learning, and user interface design are all significant data and computer related approaches.

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

Ans: As Discussed in the lecture, Size, Dynamicity and Diversity are the main things which makes the web search difficult. Indexing websites is a difficult process that incorporates various data science difficulties, such as information overload, unstructured data, data quality, multilingualism, freshness, and search relevancy. To address these issues, sophisticated algorithms, systems, and tools that can handle and analyze huge volumes of data while maintaining accuracy, efficiency, and relevance must be developed. Indexing websites is a critical operation for search engines that entails various data science issues. The quantity of material available on the internet is massive and indexing it all is a major task. Search engines acquire massive amounts of data on a regular basis, necessitating the use of scalable and efficient data processing technologies. Designing and implementing distributed systems for data storage, retrieval, and processing falls under this category. Data scientists must create efficient algorithms to crawl and index online pages in a timely and accurate manner. Because the material on the internet is unstructured and varied, extracting useful information is tough. Ad fraud is a big issue in the internet advertising industry. Malicious actors can generate bogus clicks or impressions, inflating ad performance numbers. Data scientists must create algorithms to detect and prevent ad fraud. Data scientists must create algorithms that can extract and interpret data from a variety of sources, such as text, photos, audio, and video. The quality of the data available on the internet varies greatly. Some online sites may include faulty or obsolete information, while others may be purposely deceptive. Data scientists must create algorithms that can filter out unnecessary or low-quality data while prioritizing high-quality information.

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

Ans: The area of web search and search engines offers various exciting prospects for data scientists to work on projects requiring advanced machine learning techniques and natural language processing. The potential for data-driven decision making is one of the most intriguing features of data science prospects in the realm of enterprise search and search engines. Businesses may obtain significant insights into consumer behavior, tastes, and requirements by analyzing search data and employing machine learning algorithms. This can help to guide product development, marketing initiatives, and other company choices, ultimately leading to greater performance and revenue growth. In terms of security and privacy, the realm of business search and search engines provides a particular difficulty. Data scientists must design algorithms and systems that emphasize privacy and security when sensitive corporate information and consumer data are stored and handled. In web search , natural language processing (NLP) techniques may be utilized to create models that can process and interpret material in different languages. This necessitates a thorough grasp of linguistics as well as the capacity to create models that can handle a variety of writing systems and grammatical patterns. Search engines can also utilize NLP techniques to customize search results depending on a user's preferences, search history, and activity. This necessitates the creation of algorithms capable of analyzing massive volumes of data in real time and providing customized suggestions. This is an intriguing and hard topic for data scientists to work in, with the potential for data-driven decision making, scalability and efficiency, cognitive search, and an emphasis on security and privacy.

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

Ans: To efficiently fetch pages based on queries, search engines employ two different types of data structures: a forward index and an inverted index. This data structure takes us from document to word by storing mapping from documents to words. The following are the steps to establish a forward index: Gather all the keywords and retrieve the document. The inverted index is a data structure that allows efficient, full-text searches in the database. The document serves as the main retrieval unit in a forward index, and each document includes a list of words and their places. A term is the fundamental unit of retrieval in an inverted index, and each term includes a list of the documents where it occurs. Forward indexing takes longer than inverted indexing to do a search because it filters all the mapping data from all of the web pages that have the phrase in question. In contrast, inverted indexing does not filter this data. Forward indexing is relatively quick since it only adds keywords as it moves ahead, but searching is challenging because it must search through every page of the index to find all pages that contain a certain term. As a forward index maintains details about each document's terms, it is often bigger than an inverted index. As just the terms and their occurrences are stored, an inverted index is often smaller.

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

Ans: Crawling, indexing, and query processing can be classified as high-level architectural components of online search. The method through which a crawler finds new and updated sites to be added to a search engine index is called crawling. Web crawling and spidering are other terms for crawling. A list of web addresses from earlier crawls and sitemaps given by website owners serve as the basis for the crawling process. Links on those sites are used by crawlers to find other pages. Crawlers provide the data they collected to servers, where it is cataloged in a search index. Indexing is the process of parsing and analyzing gathered pages to produce an index of words and texts. The indexer goes over the retrieved pages, extracting the content and tokenizing it into individual words. The indexer then generates an inverted index, mapping each word to the documents that contain it. The index is compressed and optimized for quick and efficient query processing. A search engine, query processor is responsible for figuring out how to respond to user information requests in the best way possible. The search engine then gets a group of potential documents by scanning the index for entries that match the query terms. After evaluating each document's relevancy using a variety of ranking algorithms, the search engine shows the user the top-ranked papers.

(iii) Also, answer the following multiple-choice questions: You can list the question number and the letter corresponding to the correct choice as Answer in your report, (2x5 = 10 pts of the 80 C+R points in the rubric)

Ans:

Q1) D
Q2) B
Q3) B
Q4) B
Q5) D