EAS 504 Applications of Data Science : Industrial Overview
Assignment 6

Name : Aboli Rawanhate
UB number : 50374341

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

Enterprise Search is the market sector covered in this lecture by Abhishek Singh Tomar. The term "enterprise search" refers to the process of locating relevant material across different data sources. This approach identifies and allows authorized users across businesses to index, search, and display specified material. Information Retrieval is the most important component of enterprise search. Information Retrieval is finding materials of an unstructured nature like html pages that satisfies an information need from within large collections usually stored on large computers. He also mentioned that Enterprise search is the most widely used sectores in the world and the spread of the internet is responsible for that. As a consequence, any enormous amounts of unstructured data or any other type of information are collected, stored, and analyzed in order to produce appropriate results depending on user inputs. When using Enterprise Search as a tool, there are three important actors to keep in mind are mentioned below:

- User's perspective: is to access information
  - high-quality search results - should receive quality information
  - fast response to queries - should receive response faster

- Search engine's perspective: monetization - how to earn money with the user’s search
  - Attract more users
  - Increase the ad revenue
  - Reduce the operational costs

- Advertiser's perspective: publicity
  - attract more users to their site
  - pay little

Getting desired results on the web might be difficult since it is a dynamic environment with a large amount of material and a diverse collection of individuals with varying interests. So, machine learning helps to carry these tasks. Product, People, and Priority are the three most critical aspects of Enterprise Search (3 P's of Enterprise Search). Product denotes what the client is looking for; persons denote the person conducting the search; and priority denotes the type of search query that is created, whether it is a basic or advanced search.

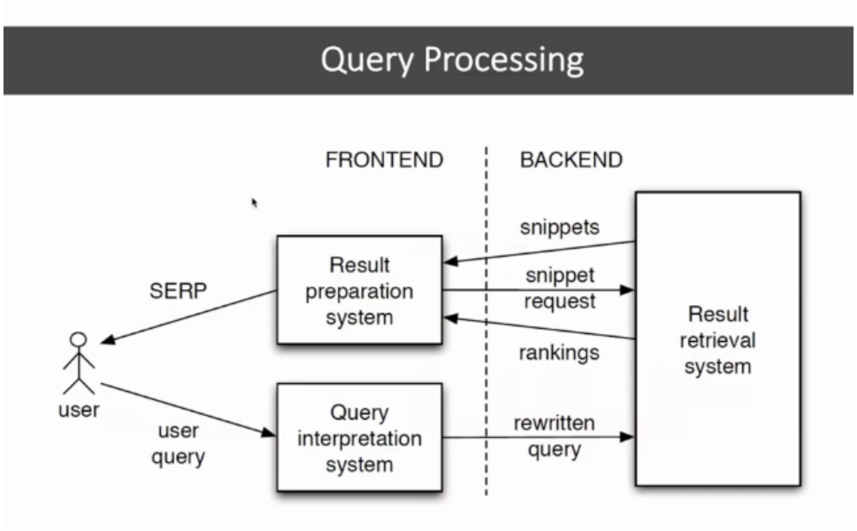
Q2. What data science related skills and technologies are commonly used in this sector?

Machine Learning used in Enterprise sector in following tasks:

1. Transformational HR services
2. Self Driving-Customer Service
    a. Managing IT services
    b. IT tickets classification
3. Conversational bots
    a. Password resets
4. Student services
    a. Registering for courses
    b. Getting transcripts

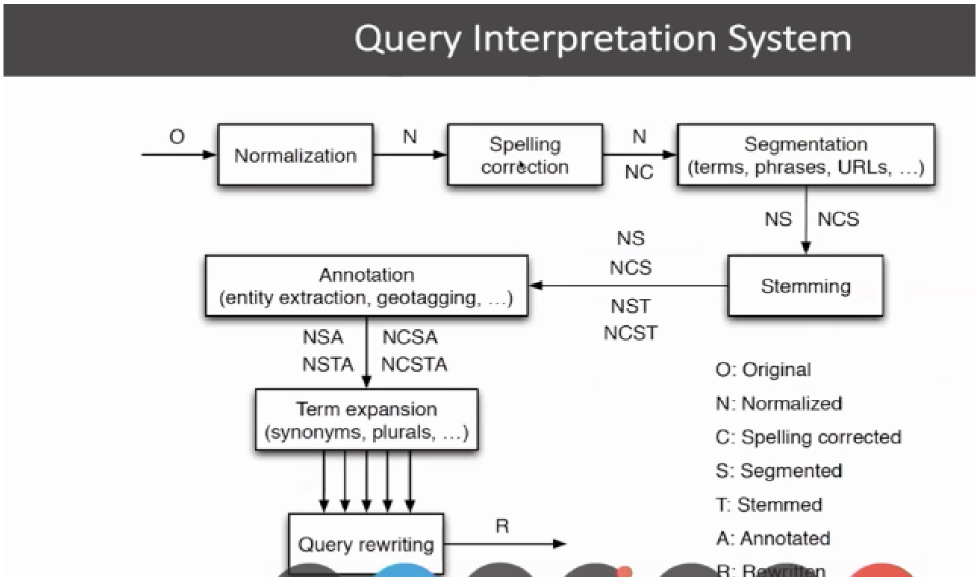
Data science related skills are also involved in query processing.

- Once a query is written on a web browser, the keywords from the query are extracted using tokenization and then relevant documents including keywords based on criteria like geographic location, personal interest are ranked using Page ranking algorithm. The page rank algorithm is used to rank pages in order of importance. The amount of relevant external links connecting to a page is used to measure its relevance using web graphs
- Natural Language Processing is used to extract and parse data.
- Term-Frequency Degerming the web pages that are most relevant to the search query is done using Inverse-Document-Frequency (TF-IDF). It evaluates the frequency of terms and sorts the most essential words in the text to see if the information on the web page is relevant to the search query.
- The indexing system performs information extraction, filtering, and categorization, among other activities.
- The query interpretation system uses machine learning algorithms to comprehend and rephrase the query so that it may be transmitted to the result retrieval system, which utilizes the page rank algorithm to rate the results.
- On downloaded web pages, information is retrieved, filtered, and classified; analysis such as spam detection, duplication identification, mirror site detection, and link quality estimation are to be done.



Query Processing

-
- Machine learning methods are mostly used in this investigation. To further filter results, a number of classifiers and feature extractors may be constructed on top of this, such as a spam classifier and a text quality estimator.
- Snippets that are generated to offer an overview of material on a page are chosen using a machine learning algorithm to assist users comprehend the context of the information on the page and assess whether it is relevant to the search they conducted.
- To execute tasks like normalization, spell correction, segmentation, stemming, annotation, and term expansion, query rewriting employs machine learning and deep learning techniques.



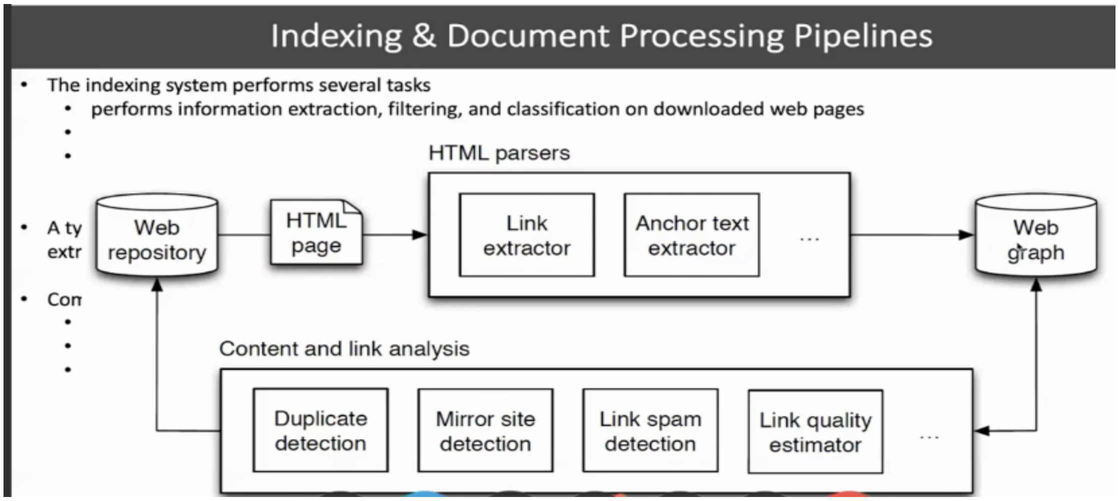
Query Interpretation System

-

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

The standard Enterprise Search process varies depending on the activities that are being completed. Indexing system performs several tasks like information extraction, filtering and classification on downloaded web pages.
In Indexing, the weblink is collected from the web repository where the data was deposited by the web crawler. The anchor text and links will then be extracted from the link using HTML parsers. A web graph is a data structure that describes the directed links between pages of the World Wide Web. Then content and link analysis is done. Under that, duplicate detection using comparable hashing or machine learning algorithms, mirror-site detection - if someone has created the exact same site just by changing url, link spam detection, and link quality estimates are all possible using the web graph.

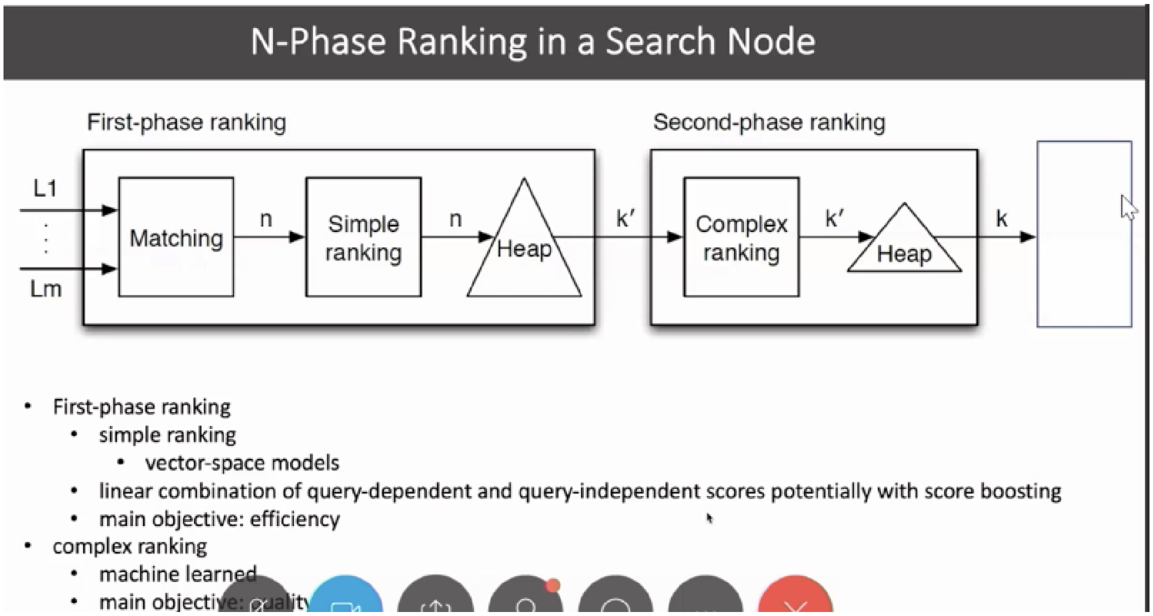
-



Indexing & Document Processing Pipelines

- The indexing system performs several tasks
    - performs information extraction, filtering, and classification on downloaded web pages
    -
    -
- A ty
  extr
- Con
    -
    -
    -

The query is processed and rewritten and the query is interpreted using the ML algorithm. The results are then ranked and also snippets with meaningful less information are created.
Ranking is another example that demonstrates the data and processing procedures.
In a Search node, consider the N-phase ranking algorithm. For a particular search query, this method will assist in getting the most relevant web link. The initial phase of ranking entails keyword matching and a simple ranking utilizing vector space models. The final score will be a linear mix of question dependent and query independent scores potentially with score boosting. Heap storage is used to store all of the results. The main objective of this phase is efficiency. The second phase of ranking is then carried out, which entails complicated ranking based on machine learning algorithms. The end output is of higher quality and may be stored in a smaller heap. The procedure is repeated until the best outcomes are achieved.



N-Phase Ranking in a Search Node

- First-phase ranking
  - simple ranking
    - vector-space models
  - linear combination of query-dependent and query-independent scores potentially with score boosting
  - main objective: efficiency
- complex ranking
  - machine learned
  - main objective: quality

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

The management of the web and users is the most difficult aspect of doing searches. With the dynamic and diverse type of data expanding exponentially, it's difficult to extract relevant information from them. With the world's largest data repository estimated to be billions and billions of pages, retrieving any information from such a large database may be a difficult undertaking. The material in these repositories is different in terms of both content and data types, making searching more complex. Concurrently handling multiple users. As search engines are the widely used application, there are millions of users who query at the same time. So serving these requests and giving them results with minimum latency is important.

Also, extract results based on the geographic location, their personal interest and the search results for each of these phrases may differ. The syntax of a query can vary greatly, and some questions might be quite brief, making it difficult to grasp the user's meaning.
Enterprise search is concerned with conducting a search for a highly particular query. A significant number of people contribute to the creation of content in any given business. It is explained in the lecture with a comparison to LinkedIn, where material is created by 4000+ engineers spread over 8 different offices.

Other departments, such as sales, finance, HR, BizOps, Analytics, and product management, provide content. Finding the correct material for consumers is difficult due to the large range of information available. Finding the perfect information to show a user might feel like looking for a needle in a haystack. Tribal wisdom is applied to solve this challenge. When conducting an enterprise search, there are a lot of constraints, such as documentation being obsolete rapidly, documentation being difficult to locate, and so on. Documents that are redundant, Tribal wisdom radiance, unsustainable support load.

Keeping in mind the proficiency of users, are they technical, non-technical or very technical. Results should be based on that. Also initial results should be very relevant because barely anyone visits the second page of results. Understanding the ecosystem of users that are search experts, application developers. However, these difficulties may be overcome by recognizing the most important priority while getting information. It is necessary to assess if efficiency or quality is more essential. This may be accomplished by including features such as location-based search results, a skill-based expert search system, and integration with chatbots like Slack.

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

RIght from web crawling, indexing, query processing to ranking, all the components of Enterprise search involve data science and machine learning opportunities.
An effective and well-tested search engine is one that not only finds the most important and relevant information among all the irrelevant ones, but also ranks that information according to its relevance. Building or coming closer to such a search engine can improve the user experience. Users expect the most relevant results in the shortest amount of time. To improve this, Data Science and Machine Learning approaches may be used to increase browsing quality or personalize the results with extra aspects like sorting, spell correction, suggested searches, and more.
The query interpretation system, I find the most interesting. Following steps are involved in query interpretation:

1. Query typed by the user is the original query and it is normalized and then passed to the spell checker to correct incorrect spellings of the query.
2. Query with correct spellings and normalization is then forwarded to the segmentation. In that, terms, phrases and urls are extracted from the query.
3. This query is then passed to the stemming engine where words are stemmed e.g. running will stem to run.
4. This stemmed query then forwarded to the Annotation module. Entities are extracted from the query, also other information like geotagging is included in the query to filter out the required information.
5. Later it is forwarded for term expansion i.e. synonyms and plurals are replaced.

6. This reformed query is passed to query rewriting and a new rewritten query is acquired. To make the query comparable to those in the online index, relevant synonym, plural, and term expansion are used.

The rewritten query may then be sent into deep learning classifiers to produce an optimum search query.

Q6. What's the difference between a forward index and an inverted index?

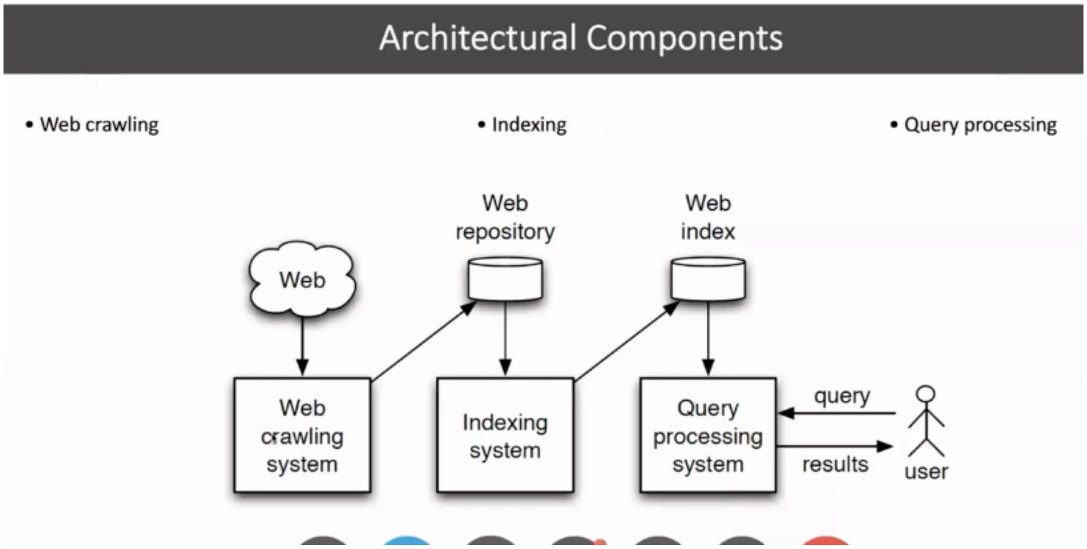
Forward index with page content is passed to Text processors. In text processors, tokenization, stop word removal takes place like - is , that, these words removed. Then the query is converted to lowercase. This query is then stemmed. Words from the query are shortened e.g. living converted to liv. These are then passed to Inverted Index Builder and output it produces as inverted index. The search is slow in the forward index.

1. **Forward Indexing :** The document name is used as an index, and the words are used as mapped references. In the forward index, indexing is fast as keywords are appended when found. In the forward index, duplicate keywords can be present in an index.
   Steps involved in Forward Index Building Process:
   - Scan the document for unique terms and make a list.
   - As an index, map all of the terms to the document.
   - Follow the instructions above for all of the papers.

2. **Inverted Indexing :** The words are stored as indexes, while the document names are stored as mapped references in Inverted Indexes. In the Inverted index, indexing is slow as each word has to be checked before preparing the index. In the Inverted index, no duplicate keyword is stored in an index. The search is quite fast in inverted index
   Steps involved in Inverted Index Building Process:
   - Scan the document and make a list of all the terms that aren't found anywhere else.
   - Make a list of all unique terms' indexes and map them to document search.
   - Follow the instructions above for all of the papers.

Q7. Describe the high level architectural components of web search.

There are three main components of Enterprise architecture:



Architectural Components

- Web crawling
- Indexing
- Query processing

Web Crawling: An online crawler visits web sites, searches for content, and saves it to a repository. It parses the data to generate the next set of URLs, which are then added to the buffer of pages to be crawled, and the procedure is repeated for all URLs in the queue. Limits are placed as to which sites can be crawled as well as which websites should be disregarded when this becomes a repeating operation. Performance, politeness, implementation issues, quality metrics, and external issues are utilized to determine whether or not a URL should be crawled.

Indexing: Web crawling is used to convert and clean data so that information can be retrieved quickly and results may be displayed to the user in milliseconds. On downloaded web pages, the indexing system conducts duties such as information extraction, filtering, and categorization, as well as analyses such as spam detection, duplication identification, mirror site detection, and link quality estimate. It feeds meta-data, analytics, and other types of information to crawling and query processing systems, as well as converting pages in the web repository into index structures that make finding the textual content of pages easier. Various document processing pipelines, each performing various normalization or extraction tasks on web pages, make up a typical indexing system.

Query Processing: In a basic query processor, the user enters in a query, which is sent to an interpreter system, which rewrites it and sends it to a result retrieval system, which rates the document and returns the result.

Q8. MCQs:

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