***Argin B. Borinaga BSIT 3-A***

***PRE-TEST***

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*MODULE3:***

*1. What is structured data and what is unstructured data? Give an example of each from your experience.*

*Structured data is highly-organized and formatted in a way so it's easily searchable in relational databases. Unstructured data has no pre-defined format or organization, making it much more difficult to collect, process, and analyze. Structured data is most often categorized as quantitative data, and it's the type of data most of us are used to working with. Think of data that fits neatly within fixed fields and columns in relational databases and spreadsheets.*

*Examples of structured data include names, dates, addresses, credit card numbers, stock information, geolocation, and more.*

*Unstructured data is most often categorized as qualitative data, and it cannot be processed and analyzed using conventional tools and methods.*

*Examples of unstructured data include text, video, audio, mobile activity, social media activity, satellite imagery, surveillance imagery – the list goes on and on.*

*2. Give a general definition of information retrieval (IR). What does information retrieval involve when we consider information on the Web?*

*Information retrieval (IR) deals with the construction of automatic systems that allow users to inquire about textual data of any kind through natural language queries. Information retrieval (IR) is the activity of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds.*

*3. Discuss the types of data and the types of users in today’s information retrieval systems.*

*The IR system assists the users in finding the information they require but it does not explicitly return the answers to the question. It notifies regarding the existence and location of documents that might consist of the required information. Information retrieval also extends support to users in browsing or filtering document collection or processing a set of retrieved documents.*

*4. What is meant by navigational, informational, and transformational search?*

*Navigation-oriented searches are targeting a certain kind of website. The user is not seeking a particular product or company and has no clear intent to purchase. He simply wishes to visit a particular type of website and uses the search engine to navigate there. Navigational searches are also conducted when the user does not know the exact web address or the correct spelling of the brand name or company he is looking for.*

*Information-oriented searches are probably the earliest form query because search engines were first developed to filter specific information from a dataset that had become unmanageably large (the World Wide Web). Due to this, users employ informational searches whenever they need guidance, background information, or specific information about a topic or product without having any concrete intention to purchase or any wish to seek out a certain landing page.*

*Transactional search queries*

*Search requests that indicate a clear intention to buy or concern a planned purchase are regarded as transactional queries. Here, searchers are looking to acquire a (digital) product, either free or in return for payment. And they will use the search engine to select this product and visit the relevant website to make their purchase.*

*5. What are the two main modes of interaction with an IR system? Describe and provide and examples.*

*Retrieval - Extract relevant information from document repository.*

*Example:*

*Recalling something important matter.*

*Browsing - Exploratory activity based on user’s assessment of relevance*

*Example:*

*Browsing something in the internet.*

*6. Explain the main differences between the database and IR systems.*

*Databases deal with structured information retrieval through well-defined formal languages for representation and manipulation based on the theoretically founded data models and have fixed schemas defined in some data model such as the relational model while IR deals with unstructured search with possibly vague query or search semantics and without a well-defined logical schematic representation. No fixed data model.*

*7. Describe the main components of the IR system.*

*1) Query /Collections: store only a representation of the document or query which means that the text of a document is lost once it has been processed for the purpose of generating its representation.*

*2) IR System: Involve in performing actual retrieval function, executing the search strategy in response to a query.*

*3) Ranked Results: a set of documents which improves the subsequent run after information retrieval.*

*8. What are digital libraries? What types of data are typically found in them?*

*Digital libraries- Collections of electronic resources and services for the delivery of materials in variety of formats.*

*Libraries were the first to adopt IR systems for information retrieval. In first-generation, it consisted, automation of previous technologies, and the search was based on author name and title. In the second generation, it included searching by subject heading, keywords, etc. In the third generation, it consisted of graphical interfaces, electronic forms, hypertext features, etc.*

*9. Name some digital libraries that you have accessed. What do they contain and how far back does the data go?*

*Internet Archive - The Internet Archive allows the public to upload and download digital material to its data cluster, but the bulk of its data is collected automatically by its web crawlers, which work to preserve as much of the public web as possible. Its web archive, the Wayback Machine, contains hundreds of billions of web captures. The Archive also oversees one of the world's largest book digitization projects.*

*Open Library - Open Library claims to have over 20 million records in its database. Copies of the contents of tens of thousands of modern books have been made available from 150 libraries and publishers for ebook digital lending. Other books including in-print and in-copyright books have been scanned from copies in library collections, library discards, and donations, and are also available for lending in digital form.*

*Google Books -More than 100,000 books for consultation, download or on-line purchase and it is a service from Google Inc. that searches the full text of books and magazines that Google has scanned, converted to text using optical character recognition (OCR), and stored in its digital database*

*10. Give a brief history of IR and mention the landmark developments in this field.*

*The idea of using computers to search for relevant pieces of information was popularized in the article As We May Think by Vannevar Bush in 1945. It would appear that Bush was inspired by patents for a 'statistical machine' - filed by Emanuel Goldberg in the 1920s and '30s - that searched for documents stored on film.[7] The first description of a computer searching for information was described by Holmstrom in 1948,[8] detailing an early mention of the Univac computer. Automated information retrieval systems were introduced in the 1950s: one even featured in the 1957 romantic comedy, Desk Set. In the 1960s, the first large information retrieval research group was formed by Gerard Salton at Cornell. By the 1970s several different retrieval techniques had been shown to perform well on small text corpora such as the Cranfield collection (several thousand documents). Large-scale retrieval systems, such as the Lockheed Dialog system, came into use early in the 1970s.*

*In 1992, the US Department of Defense along with the National Institute of Standards and Technology (NIST), cosponsored the Text Retrieval Conference (TREC) as part of the TIPSTER text program. The aim of this was to look into the information retrieval community by supplying the infrastructure that was needed for evaluation of text retrieval methodologies on a very large text collection. This catalyzed research on methods that scale to huge corpora. The introduction of web search engines has boosted the need for very large scale retrieval systems even further.*

*11. What is the Boolean model of IR? What are its limitations?*

*The (standard) Boolean model of information retrieval (BIR) is a classical information retrieval (IR) model and, at the same time, the first and most-adopted one. It is used by many IR systems to this day.[citation needed] The BIR is based on Boolean logic and classical set theory in that both the documents to be searched and the user's query are conceived as sets of terms (a bag-of-words model). Retrieval is based on whether or not the documents contain the query terms.*

*LIMITATIONS:*

*• The first relates to the formulation of search statements.*

*• It has been noted that users are not able to formulate an exact search statement by the combination of AND, OR and NOT operators, especially when several query terms are involved.*

*• In such cases either the search statement becomes too narrow or too broad.*

*• The second limitation relates to the number of retrieval items.*

*• It has been noted that users cannot predict a priori exactly how many items are to be retrieved to satisfy a given query.*

*• If the search statement is broad, the number of retrieved items may sometimes be several hundreds and thus it may be quite difficult to find out the exact information required.*

*• The third limitation is that it identifies an item as relevant by finding out whether a given query term is present or not in a given record in the database.*

*12. What is the vector space model of IR? How does a vector get constructed to represent a document?*

*The Vector-Space Model (VSM) for Information Retrieval represents documents and queries as vectors of weights. Each weight is a measure of the importance of an index term in a document or a query, respectively. The index term weights are computed on the basis of the frequency of the index terms in the document, the query or the collection. At retrieval time, the documents are ranked by the cosine of the angle between the document vectors and the query vector. For each document and query, the cosine of the angle is calculated as the ratio between the inner product between the document vector and the query vector, and the product of the norm of the document vector by the norm of the query vector. The documents are then returned by the system by decreasing cosine.*

*13. Define the TF-IDF scheme of determining the weight of a keyword in a document. Why is it necessary to include IDF in the weight of a term?*

*Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.*

*14. What are probabilistic and semantic models of IR?*

*Probabilistic Approach to IR Data Basic Probability Theory Probability Ranking Principle Extension Probabilistic Approach to Retrieval Given a user information need (represented as a query) and a collection of documents (transformed into document representations), a system must determine how well the documents satisfy the query An IR system has an uncertain understanding of the user query, and makes an uncertain guess of whether a document satisﬁes the query Probability theory provides a principled foundation for such reasoning under uncertainty Probabilistic models exploit this foundation to estimate how likely it is that a document is relevant to a query PhD Comprehensive presentation Part 1: Probabilistic Information Retrieval.*

*Semantic Information Retrieval deals with the usage of knowledge sets to annotate documents and solve queries. The idea behind this research track lies in the following hypothesis: using only computed word distribution statistics, may not be sufficient to find the semantic links that join query to relevant documents in some IR situations.*

*15. Define recall and precision in IR systems.*

*Recall is defined as ratio of the number of retrieved and relevant documents (the number of items retrieved that are relevant to the user and match his needs) to the number of possible relevant documents (number of relevant documents in the database).*

*Precision measures one aspect of information retrieval overhead for a user associated with a particular search. If a search has 85 percent precision then, then 15(100-85) percent of user effort is overhead reviewing non-relevant items.*

*16. How is an F-score defined as a metric of information retrieval? In what way does it account for both precision and recall?*

*The F-score, also called the F1-score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into ‘positive’ or ‘negative’. The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model’s precision and recall.*

*Precision is the ratio of the correctly +ve labeled by our program to all +ve labeled.*

*Precision = TP/(TP+FP)*

*numerator: +ve labeled diabetic people.*

*denominator: all +ve labeled by our program (whether they’re diabetic or not in reality).*

*Recall is the ratio of the correctly +ve labeled by our program to all who are diabetic in reality.*

*Recall = TP/(TP+FN)*

*Numerator : +ve labeled diabetic people.*

*Denominator : all people who are diabetic (whether detected by our program or not)*

*17. What are the different types of queries in an IR system? Describe each with an example.*

*Keyword-based queries are the simplest and most commonly used forms of IR queries: the user just enters keyword combinations to retrieve documents. The query keyword terms are implicitly connected by a logical AND operator. A query such as ‘database concepts’ retrieves documents that contain both the words ‘data-base’ and ‘concepts’ at the top of the retrieved results.*

*Boolean Queries allow using the AND, OR, NOT, ( ), + , and – Boolean operators in combinations of keyword formulations. AND requires that both terms be found. OR lets either term be found. NOT means any record containing the second term will be excluded. ‘( )’ means the Boolean operators can be nested using parentheses. ‘+’ is equivalent to AND, requiring the term; the ‘+’ should be placed directly in front of the search term. ‘–’ is equivalent to AND NOT and means to exclude the term; the ‘–’ should be placed directly in front of the search term not wanted. Complex Boolean queries can be built out of these operators and their combinations, and they are evaluated according to the classical rules of Boolean algebra. No ranking is possible, because a document either satisfies such a query (is “relevant”) or does not satisfy it (is “non relevant”).*

*Phrase Queries represented using an inverted keyword index for searching, the relative order of the terms in the document is lost. In order to perform exact phrase retrieval, these phrases should be encoded in the inverted index or implemented differently (with relative positions of word occurrences in documents). A phrase query consists of a sequence of words that makes up a phrase. The phrase is generally enclosed within double quotes. Each retrieved document must contain at least one instance of the exact phrase.*

*Proximity Queries refers to a search that accounts for how close within a record multiple terms should be to each other. The most commonly used proximity search option is a phrase search that requires terms to be in the exact order. Other proximity operators can specify how close terms should be to each other. Some will also specify the order of the search terms. Each search engine can define proximity operators differently, and the search engines use various operator names such as NEAR, ADJ (adjacent), or AFTER. In some cases, a sequence of single words is given, together with a maximum allowed distance between them. Vector space models that also maintain information about positions and offsets of tokens (words) have robust implementations for this query type.*

*Wildcard Queries is generally meant to support regular expressions and pattern matching-based searching in text. In IR systems, certain kinds of wildcard search support may be implemented—usually words with any trailing characters (for example, ‘data\*’ would retrieve data, database, data point, dataset, and so on). Providing support for wildcard searches in IR systems involves preprocessing over-head and is not considered worth the cost by many Web search engines today. Retrieval models do not directly provide support for this query type.*

*Natural Language Queries aim to understand the structure and meaning of queries written in natural language text, generally as a question or narrative. This is an active area of research that employs techniques like shallow semantic parsing of text, or query reformulations based on natural language under-standing. The system tries to formulate answers for such queries from retrieved results. Some search systems are starting to provide natural language interfaces to provide answers to specific types of questions, such as definition and factoid questions, which ask for definitions of technical terms or common facts that can be retrieved from specialized databases. Such questions are usually easier to answer because there are strong linguistic patterns giving clues to specific types of sentences—for example, ‘defined as’ or ‘refers to’. Semantic models can provide support for this query type.*

*18. What are the approaches to processing phrase and proximity queries?*

*Phrase queries play an important role in web search and other applications. Traditionally, phrase queries have been processed using a positional inverted index, potentially augmented by selected multi-word sequences. The concept of phrase queries has proven easily understood by users.*

*Proximity queries is a clearly, positional indexes can be used for such queries; biword indexes cannot.*

*Positional index size*

*• We can compress position values/offsets, later in index compression*

*• Nevertheless, a positional index expands postings storage substantially*

*• Nevertheless, a positional index is now standardly used because of the power and usefulness of phrase and proximity queries whether used explicitly or implicitly in ranking retrieval system.*

*19. Describe the detailed IR process*

*The working of Information Retrieval process is explained below*

*• The Process of Information Retrieval starts when a user creates any query into the system through some graphical interface provided.*

*• These user-defined queries are the statements of needed information for example, queries fork by users in search engines.*

*• In IR single query does not match to the right data object instead it matches with the several collections of data objects from which the most relevant document is taken into consideration for further evaluation.*

*• The ranking of relevant documents is done to find out the most related document to the given query.*

*• This is the key difference between the Database searching and Information Retrieval.*

*• After the query is sent to the core of the system. This part has the access to the content management module which is directly linked with the back-end i.e. the large collections of data objects.*

*• Once results R are generated by the core system then it is returned to the user by some graphical user interfaces.*

*• The process repeats and results are modified until the user satisfied for what he is actually looking for.*

*20. What is stop word removal and stemming? Why are these processes necessary for better information retrieval?*

*Stop words are very common words that appear in the text that carry little meaning; they serve only a syntactic function but do not indicate subject matter. These stop words have two different impacts on the information retrieval process. They can affect the retrieval effectiveness because they have a very high frequency and tend to diminish the impact of frequency differences among less common words, affecting the weighting process. The removal of the stop words also changes the document length and*

*Subsequently affects the weighting process.*

*Stop-word removal is an important preprocessing techniques used in Natural Language processing applications so as to improve the performance of the Information Retrieval System, Text Analytics & Processing System, Text Summarization, Question-Answering system, stemming etc.*

*Stemming is a part of linguistic studies in morphology and artificial intelligence (AI) information retrieval and extraction. Stemming and AI knowledge extract meaningful information from vast sources like big data or the Internet since additional forms of a word related to a subject may need to be searched to get the best results. Stemming is also a part of queries and Internet search engines.*

*21. What is a thesaurus? How is it beneficial to IR?*

*The prime function of a thesaurus is to support information retrieval by guiding the choice of terms for indexing and searching. The traditional aim of a thesaurus is to guide the indexer and the searcher to choose the same term for the same concept a thesaurus should first list all the concepts that might be useful for retrieval purposes in a given domain. The concepts are represented by terms, and for each concept, one of the possible representations is selected as the preferred term Secondly, a thesaurus should present the preferred terms in such a way that people will easily identify the one(s) they need. This is achieved by establishing relationships between terms and/or between concepts and using the relationships to present the terms in a structured display.*

*22. What is information extraction? What are the different types of information extraction from structured text?*

*Information extraction is the process of extracting information from unstructured textual sources to enable finding entities as well as classifying and storing them in a database. Semantically enhanced information extraction (also known as semantic annotation) couples those entities with their semantic descriptions and connections from a knowledge graph. By adding metadata to the extracted concepts, this technology solves many challenges in enterprise content management and knowledge discovery.*

*Free text: This type of document contains unstructured text such as news, stories, etc. The extraction process in this kind of text document is difficult because it contains variant information with a week relation.*

*Semi-structured text: The text that is presented and formatted in a high quality manner in a specific domain; for instance, information about the economy, education, medicine and so forth. A lot of work has been carried out on semi-structured text. [2] has proposed an automatic IE*

*Method: Their proposed system was very effective and valuable in many semi-structured types of sources. Furthermore, in 1997, [3] proposed an automatic system that automatically generated wrappers from a variety of internet sources.*

*23. What are vocabularies in IR systems? What role do they play in the indexing of documents?*

*Vocabularies are information organization systems whose purpose is to increase the effectiveness of the processes of indexing, search, or navigation of collections of digital resources with informational content (hereinafter, simply resources). Every term in a vocabulary set will have all the associated or combined collection of information of the document like the document id and counts of the occurrence.*

*24. Describe the process of constructing the result of a search request using an inverted index.*

*The inverted index data structure is a central component of a typical search engine indexing algorithm. A goal of a search engine implementation is to optimize the speed of the query: find the documents where word X occurs. Once a forward index is developed, which stores lists of words per document, it is next inverted to develop an inverted index. Querying the forward index would require sequential iteration through each document and to each word to verify a matching document. The time, memory, and processing resources to perform such a query are not always technically realistic. Instead of listing the words per document in the forward index, the inverted index data structure is developed which lists the documents per word.*

*25. Define relevance feedback.*

*Relevance feedback is a feature of some information retrieval systems. The idea behind relevance feedback is to take the results that are initially returned from a given query, to gather user feedback, and to use information about whether or not those results are relevant to perform a new query.*

*26. Describe the three types of Web analyses.*

*Three types of Web analyses are web analytics, web spamming and web security.*

*27. What are the three categories of agent-based Web content analyses mentioned in this chapter?*

*Intelligent Web agents*

*Personalized Web agents*

*Information filtering/categorization*

*28. What is the database-based approach to analyzing Web content? What are Web query systems?*

*The database approach for Web mining tries to develop techniques for organizing semi structured data stored in the Web into more structured collections of information resources. Standard database querying mechanisms and data mining techniques can be used to analyze those collections then.*

*Web-base query systems and languages developed recently that attempt to utilize standard database query languages such as SQL, structural information about web documents, and even natural language processing for accommodating the types of queries that are used in World Wide Web searches. We mention a few examples of these Web-base query systems here. W3QL combines structure queries, based on the organization of hypertext documents, and content queries, based on information retrieval techniques. WebLog is a logic-based query language for restructuring extracted information from Web information sources. Lorel and UnQL support querying of heterogeneous and semi-structured information on the Web using a labeled graph data model. TSIMMIS helps to extract data from heterogeneous and semi-structured information sources and correlates them to generate an integrated database representation of the extracted information.*

*29. What algorithms are popular in ranking or determining the importance of Web pages? Which algorithm was proposed by the founders of Google?*

*Pagerank(PR) is an algorithm used by Google Search to rank web page in their search engine results. PageRank is a way of measuring the importance of website pages. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.*

*30. What can you learn from Web usage analysis? What data does it generate?*

*Web usage analysis is the measurement, collection, analysis and reporting of web data to understand and optimize web usage. Web analytics is not just a process of measuring web traffic but can be used as a tool for business and market research and asses and improve website effectiveness.*

*31. What mining operations are commonly performed on Web usage data? Give an example of each.*

*Data preprocessing and Pattern discovery. The huge data present in the web is a collection of raw data, so to get the user needed information the web data preprocessing should be done. The different phases in web usage mining include data cleaning, data preparation, user identification, session identification, data integration, data transformation, pattern, and discovery and pattern analysis.*

*32. What are the applications of Web usage mining?*

*Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis.*

*33. What is search relevance? How is it determined?*

*Search relevance is a measurement of how closely related a document is to a query. Search relevance can be determined in a wide variety of ways, ranging from simple binary relevance to a weighted relevance algorithm such as TF-IDF, which assigns a relevance score to documents.*

*34. Define and explain conversational search.*

*Conversational search is distinct from voice search which allows allows users to submit spoken queries, but returns answers in text, voice, or other formats that don’t resemble a conversation. Conversational search initiates a dialogue through asking click-based questions to help consumers find what they’re interested in. By narrowing down choice in a personalized and interactive way, conversational search is evolving how consumers engage online.*

*35. Define topic modeling.*

*In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discove*