USN | | | | | | | | | | |

Internal Assessment Test 3 – July 2024

Sub:	Natural Language Processing					Sub Code:	21AI643	Branch:	AI&DS	
Date:	30/07/2024	Duration:	90 mins	Max Marks:	50	Sem / Sec:	VI		OBE	

Answer any FIVE FULL Questions

		MARKS	CO	RBT
1	Explain the functioning of Latent Semantic Analysis (LSA) feedback system.	[10]	CO3	L2

Latent Semantic Analysis (LSA) Feedback Systems

- Latent Semantic Analysis uses statistical computations to extract and represent the meaning of words. Meanings are represented in terms of the similarity to other words in a large corpus of documents.
- LSA begins by finding the frequency of terms used and the number of co-occurrences in each document throughout the corpus and then uses a powerful mathematical transformation to find deeper meanings and relatio among words.
- When measuring the similarity between text-objects, LSA's accuracy improves with the size of the objects. Hence, LSA provides the most bene finding similarity between two documents.
- The method, unfortunately, does not take into account word order; hence, very short documents may not be able to receive the full benefit of LSA.

LSA corpus matrix

To construct an LSA corpus matrix, a collection of documents are selected. A document may be a sentence, a paragraph, or large text. A term-document frequency (TDF) matrix X is created for those terms that appear in two or more documents. The row entitie correspond to the words or terms (hence the W) and the column entities correspond to the documents (hence the D). The matrix i analyzed using Singular Value Decomposition, that is the TDF matrix X is decomposed into the product of three other matrices:

1. vectors of derived orthogonal factor values of the original row entities W
2. vectors of derived orthogonal factor values of the original column entities D
3. scaling values (which is a diagonal matrix) S

The product of these three matrices is the original TDF matrix.

$$\{X\} = \{W\}\{S\}\{D\}$$

These documents consist of terms, which are represented by term vectors; hence, the document can be represented as a docum vector which is computed as the sum of the term vectors of its terms:

$$D_i = \sum_{t=1}^{n} T_{ti}$$

where Di is the vector for the i th document D, Tti is the term vector for the term t in Di, and n is number of terms in D. The similari between two documents (i.e., the cosine between the two document vectors) is computed as

$$Sim(D1, D2) = \frac{\sum_{i=1}^{d}(D1_i \times D2_i)}{\sum_{i=1}^{d}(D1_i)^2 \times \sum_{i=1}^{d}(D2_i)^2}$$

2	Explain the functioning of word matching feedback system used in iSTART.	[10]	CO3	L2

Literal word matching

Words are compared character by character and if there is a match of the first 75% of the characters in a word in the target sentence (or its association list) we call this a literal match.

This also includes removing suffix -s, -d, -ed, -ing, and -ion at the end of each words. For example, if the trainee's self-explanation contains 'thunderstorm' it counts as a literal match with words in the target sentence since the first nine characters are exactly the same. On the other hand, if it contains 'thunder,' it not get a match with the target sentence, but rather with a word on the associ list.

Soundex matching

This algorithm compensates for misspellings by mapping similar characters to the same soundex Words are transformed to their soundex code by retaining the first character, dropping the vowels, then converting other characters into soundex symbols. If the same symbol occurs more than once consecutively, only one occurrence is retained.

For example,

'thunderstorm' will be transformed to 't8693698';

'communication' to 'c8368.'

If the trainee's self-explanation contains 'thonderstorm' or 'tonderstorm,' both will be matched with 'thunderstorm' and this is called a soundex match. An exact soundex match is required for short w (i.e., those with fewer than six alpha-characters) due to the high number of false alarms when sou used. For longer words, a match on the first four soundex symbols suffices. We are considering re this rough and ready approach with a spell-checker.

Word Matching Feedback Systems

Word matching is a very simple and intuitive way to estimate the nature of a self explanation. In the first version of iSTART, several hand-coded components were built for each practice text.

For example, for each sentence in the text, the "important words" were identified by a human expert and a length criterion for the explanation was manually estimated.

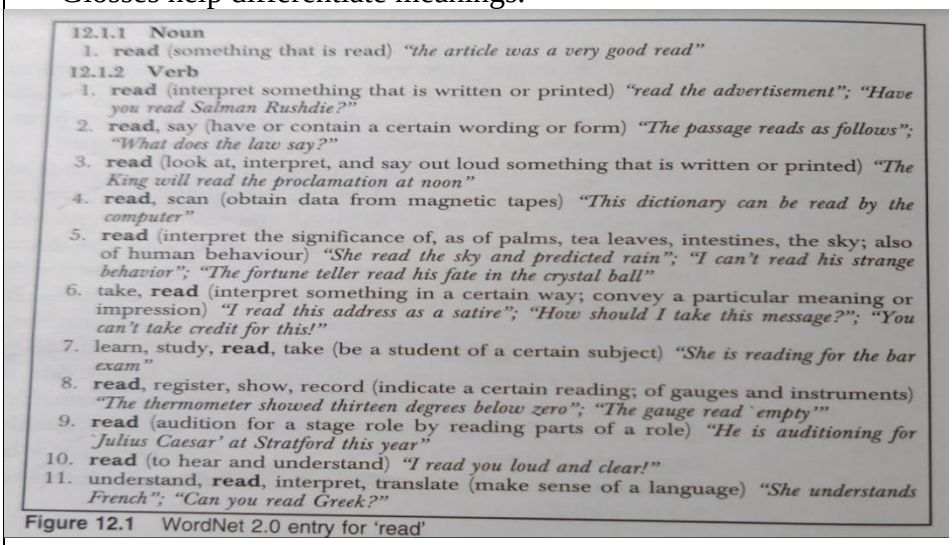
Important words were generally content words that were deemed important to the meaning of the sentence and could include words not found in the sentence.

For each important word, an association list of synonyms and related terms was created by examining dictionaries and existing protocols as well as by human judgments of what words were likely to occur in a self-explanation of the sentence. In the sentence "All thunderstorms have a similar life history," for example, important words are thunderstorm, similar, life, and history. An association list for thunderstorm would include storms, moisture, lightning, thunder, cold, tstorm, t-storm, rain, temperature, rainstorms, and electric-storm. In essence, the attempt was made to imitate LSA.

3a	Write a note on various approaches to analyzing texts. Traditional approaches to categorizing discourse have tended to treat text as if it were a homogeneous whole. These wholes, or bodies of text, are analyzed for various textual features, which are used to classify the texts as belonging to one category or another. To be sure, such approaches have yielded impressive findings, generally managing to significantly discriminate texts into categories such as dialect, domain, genre, or author. Such discrimination is made possible because By forming a picture of the degree to which textual parts inter-relate, we can build a representation of the structure of the texts, a prototypical model that we call the textual signature. Such a signature stands to serve students and researchers alike. For students, their work can be analyzed to see the extent to which their paper reflects a prototypical model. Specifically, a parts analysis may help students to see that sections of their papers are under- or over-represented in terms of the global cohesion. For researchers, a text-type signature should help significantly in mining for appropriate texts. For example, the first ten web sites from a Google search for a text about cohesion (featuring the combined keywords of comprehension, cohesion, coherence, and referential) yielded papers from the field of composition theory, English as a foreign language, and cognitive science, not to mention a disparate array of far less academic sources. While the specified keywords that were entered may have occurred in each of the retrieved items, the organization of the parts of the retrieved papers (and their inter-relatedness) would differ. Knowing the signatures that distinguishes the text types would help researchers to locate more effectively the kind of resources that they require. A further possible benefit of textual signatures involves Question Answering (QA) systems [45, 52]. Given a question and a large collection of texts (often in gigabytes), the task in QA is to draw a list of short answers (the length of a sentence) to the question from the collection. The typical architecture of a modern QA system includes three subsystems: question processing, paragraph retrieval and answer processing. Textual signatures may be able to reduce the search space in the paragraph retrieval stage by identifying more likely candidates.	[5]	CO4	L2
3b	Explain document separation using sequence mapping problem. Large organizations are increasingly confronted with the problem of capturing, processing, and archiving large amounts of data. For several reasons, the problem is especially cumbersome in the case where data is stored on paper. First, the weight, volume, and relative fragility of paper incur problems in handling and require specific, labor-intensive processes to be applied. Second, for automatic processing, the information contained on the pages must be digitized, performing Optical Character Recognition (OCR). This leads to a certain number of errors in the data retrieved from paper. Third, the identities of individual documents become blurred. In a stack of paper, the boundaries between documents are lost, or at least obscured to a large degree.1	[5]	CO4	L2
4	Write short notes on: (i) Word Net (ii) Frame Net. ● WordNet is a large lexical database for the English language. ● Inspired by psycholinguistic theories, it was developed. ● WordNet consists of 3 databases ➢ One for nouns ➢ One for verbs	[5+5]	CO4	L2

- One for both adjectives and adverbs

- ✓ Information is organized into sets of synonymous words called synsets, each representing 1 base concept.
- ✓ The synsets are linked to each other by means of lexical and semantic relations.
- ✓ Lexical relations occur between word-forms (senses).
- ✓ semantic relations occur between word meanings.
- ✓ These relations include synonymy, hypernymy / hyponymy, antonymy, meronymy / holonymy, troponymy, etc.
- ✓ If a word appears in more than 1 synset and in more than 1 part-of-speech.
- ✓ the meaning of a word is called sense.
- ✓ WordNet lists all senses of a word.
- ✓ Each sense belonging to a different synset.
- ✓ WordNet's sense-entries consist of a set synonyms and a gloss.
- ✓ A gloss consists of a dictionary-style definition and examples demonstrating the use of a synset in a sentence, as shown in the figure below.
- ✓ The figure shows the entries for the word 'read'. Read has 1 sense as a noun and 11 senses as a verb.
- ✓ Glosses help differentiate meanings.



12.1.1 **Noun**

1. **read** (something that is read) *"the article was a very good read"*

12.1.2 **Verb**

1. **read** (interpret something that is written or printed) *"read the advertisement"; "Have you read Salman Rushdie?"*
2. **read**, say (have or contain a certain wording or form) *"The passage reads as follows"; "What does the law say?"*
3. **read** (look at, interpret, and say out loud something that is written or printed) *"The King will read the proclamation at noon"*
4. **read**, scan (obtain data from magnetic tapes) *"This dictionary can be read by the computer"*
5. **read** (interpret the significance of, as of palms, tea leaves, intestines, the sky; also of human behaviour) *"She read the sky and predicted rain"; "I can't read his strange behavior"; "The fortune teller read his fate in the crystal ball"*
6. take, **read** (interpret something in a certain way; convey a particular meaning or impression) *"I read this address as a satire"; "How should I take this message?"; "You can't take credit for this!"*
7. learn, study, **read**, take (be a student of a certain subject) *"She is reading for the bar exam"*
8. **read**, register, show, record (indicate a certain reading; of gauges and instruments) *"The thermometer showed thirteen degrees below zero"; "The gauge read `empty'"*
9. **read** (audition for a stage role by reading parts of a role) *"He is auditioning for `Julius Caesar' at Stratford this year"*
10. **read** (to hear and understand) *"I read you loud and clear!"*
11. understand, **read**, interpret, translate (make sense of a language) *"She understands French"; "Can you read Greek?"*

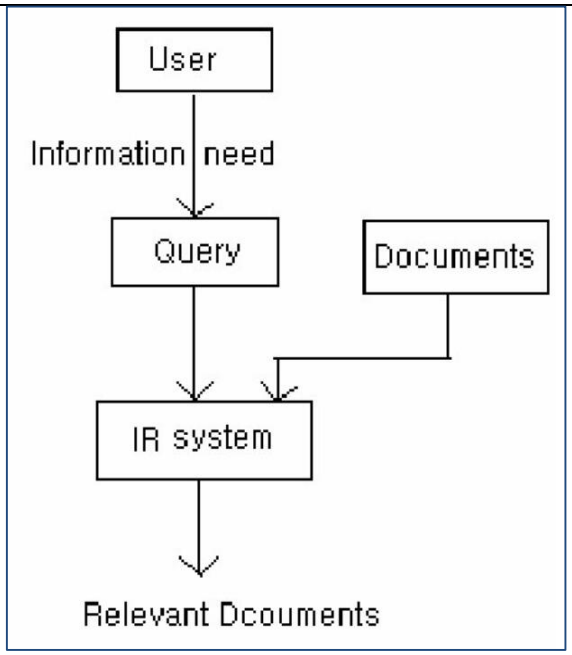
Figure 12.1 WordNet 2.0 entry for 'read'

(i) Frame Net

- ✓ FrameNet is a large database of semantically annotated English sentences.
- ✓ It is based on principles of frame semantics.
- ✓ It defines a tagset of semantic roles called the frame element.
- ✓ Sentences from the British National Corpus are tagged with these frame elements.

	✓ The basic philosophy involved is that each word evokes a particular situation with particular participants. ✓ FrameNet aims at capturing these situations through case-frame representation of words. ✓ The word that invokes a frame is called target word or predicate, and the participant entities are defined using semantic roles, which are called frame elements. ➢ Each frame contains a main lexical item as predicate and associated frame-specific semantic roles, such as AUTHORITIES, TIME, AND SUSPECT in the ARREST frame, called frame elements. ➢ Example: The sentence below is annotated with semantic roles AUTHORITIES AND SUSPECT ➢ [Authorities The police] nabbed [suspect the snatcher] ➢ The COMMUNICATION frame has the semantic roles ADDRESSEE, COMMUNICATOR, TOPIC, and MEDIUM. ➢ A JUDGEMENT frame contains roles such as a JUDGE, EVALUEE, and REASON. ➢ Example: ➢ [judge She] [Evaluee blames the police] [Reason for failing to provide enough protection] ➢ A frame may inherit roles from another frame. Eg., a STATEMENT frame may inherit from a COMMUNICATION frame, it contains roles such as SPEAKER, ADDRESSEE, and MESSAGE. ➢ Example: ➢ [Speaker She] told [Addressee me] [Message 'I'll return by 7:00 pm today']			
5a	Consider a document represented by three terms {tornado, swirl, wind} with the raw tf 4, 1, 1 respectively. In a collection of 100 documents, 15 documents contain the term tornado, 20 contain swirl and 40 contain wind. Find the idf and the term weight of the three terms. idf - tornado -> log(n / ni) = log (100 / 15) = 0.824 Weight - tornado -> tf x idf = 4 * 0.824 = 3.296 idf - swirl -> log(n / ni) = log (100 / 20) = 0.699 Weight - tornado -> tf x idf = 1 * 0.699 = 0.699 idf - wind -> log(n / ni) = log (100 / 40) = 0.398 Weight - tornado -> tf x idf = 1 * 0.398 = 0.398 The following table shows the weights assigned to the three terms using tf x idf weighting scheme	[6]	CO4	L3

	Table 5.2 Computing tf-idf for the [illegible] **Term** | **Frequency (tf)** | **Document frequency (n_i)** | **idf [log(n/n_i)]** | **Weight (tf × idf)** Tornado | 4 | 15 | 0.824 | 0.296 Swirl | 1 | 20 | 0.699 | 0.699 Wind | 1 | 40 | 0.398 | 0.389 1. Some of these				
5b	Explain the benefits of eliminating stop words. Give example in which eliminating stop word may be harmful. **Advantages:** ● Eliminating stop words can result in considerable reduction in number of index terms without losing any significant information. **Disadvantages:** ● The drawback of eliminating stop words is that it can sometimes result in elimination of useful index terms. ● For example, the stop word 'A' in Vitamin A. ● Some phrases like 'to be or not to be' consist entirely of stop words. Eliminating stop words in this case makes it impossible to correctly search a document.	[4]	CO4	L2
6	Explain design feature of IR with a neat diagram and Define precision and recall. ➢ The process of IR begins with the user's information need. ➢ Based on the need, the user formulates a query. ➢ The IR system returns documents that seem relevant to the query. ➢ The retrieval is performed by matching the query representation with document representation. ➢ The actual text of the document is not used in the retrieval process. ➢ Instead documents in a collection are frequently represented through a set of index terms or keywords.	[8+2]	CO4	L2



- Representation of keywords provides a logical view of the document.
- The process of transforming document text, to some representation of it, is known as indexing.
- There are different types of index structures.
- The one commonly used is inverted index.
- An inverted index is a list of keywords, with each keyword carrying pointers to the documents containing that keywords.

Precision:

- ✓ Precision is defined as the proportion of relevant documents in a retrieved set.
- ✓ It is the probability that a relevant document is retrieved.
- ✓ It measures the accuracy of a system.

Recall:

- ✓ Recall is the proportion of relevant documents that are actually been retrieved.

Recall measures the exhaustiveness of the system