USN											

Internal Assessment Test III – January 2024

Sub:	Natural Language Processing					Sub Code:	18CS743	Branch:	ISE
Date:	03/01/24	Duration:	90 mins	Max Marks:	50	Version/ Sem / Sec:	C/VII/A, B, C		OBE

Answer any FIVE FULL Questions

		MARKS	CO	RBT
1.	Define Indexing. List and explain methods used to reduce the set of representative keywords during indexing.	[10]	CO3	L2
2.	Explain briefly about Research Corpora	[10]	CO3	L2
3.	(a) Explain in detail the high-level representation approaches in text mining (5) (b) Explain SVM learning method in Sequence Model Estimation. (5)	[10]	CO4	L2
4	Explain classical and non-classical information retrieval models with suitable examples	[10]	CO4	L2
5	Write short notes on: (i) Word Net (ii) Frame Net	[10]	CO4	L2
6	(a) Explain design feature of IR with a neat diagram. (5) (b) Define precision and recall. Explain the trade-off between them in evaluation of IR systems. (5)	[10]	CO4	L3

0.25	1.0
0.4	0.67
0.55	0.8
0.8	0.6
1.0	0.5

USN											

Sub:	Natural Language Processing					Sub Code:	18CS743	Branch: ISE
Date:	03/01/24	Duration:	90 mins	Max Marks:	50	Version/ Sem / Sec:	C/VII/A, B, C	

Answer any FIVE FULL Questions MARKS

1.	Define Indexing. List and explain methods used to reduce the set of representative keywords during indexing. ➢ The process of transforming a collection of raw documents into an easily accessible representation is known as indexing. ➢ Indexing technique involves identifying good document descriptors such as keywords or terms. ➢ It helps in: 1. Describing the content of the document and 2. Discriminating it from other documents in the collection. ➢ Methods used to reduce the set of representative keywords during indexing are as follows: 1. Eliminating Stop Words 2. Stemming 3. Zipf's Law	[10]
2.	Explain briefly about Research Corpora ➢ Research carpora is developed for a number of NLP related tasks. ➢ Following are the available standard document collections : 1. IR Test Collection 2. Summarization Data 3. Word Sense Disambiguation 4. Asian Language Corpora	[10]
3.	(a) Explain in detail the high-level representation approaches in text mining (5) ➢ Another main stream in KDT involves using more structured or higher-level representations to perform deeper analysis so to discover more sophisticated novel / interesting knowledge. Although in general, the different approaches have been concerned with either performing exploratory analysis for hypothesis formation, or finding new connections/relations between previously analysed natural language knowledge, it has also involved using term-level knowledge	[10]

for other purposes than just statistical analysis.

- Some early research by Swanson used an augmented low-level representation (the words in the titles) and exploratory data analysis to discover hidden connections leading to very promising and interesting results in terms of answering questions for which the answer was not currently known. He showed how chains of causal implication within the medical literature can lead to hypotheses for causes of rare diseases, some of which have received scientific supporting evidence.
- Other approaches using Information Extraction (IE) which inherited some of Swanson's ideas to derive new patterns from a combination of text fragments, have also been successful. Essentially, IE is a Natural-Language (NL) technology which analyses an input NL document in a shallow way by using defined patterns along with mechanisms to resolve implicit discourse-level information to match important information from the texts. As a result, an IE task produces an intermediate representation called "templates" in which information relevant has been recognised, for example: names, events, entities, etc., or high-level linguistic entities: noun phrases, etc.
- This technique is meant to be useful as an automated or semi-automated aid for lexicographers and builders of domain-dependent knowledge bases. Also, it does not require an additional knowledge base or specific interpretation procedures in order to propose new instances of WordNet relations [9]. Once the basic relations are obtained, they are used to find common links with other "similar" concepts in WordNet [9] and so to discover new semantic links [18].
- However, there are tasks which need to be performed by hand such as deciding on a lexical relation that is of interest (i.e., hyponym) and a list of word pairs from WordNet this relation is known to hold between. One of the main advantages of this method is its low cost for augmenting the structure of WordNet and its simplicity of relations.
- However, it also has some drawbacks including its dependence on the structure of a general-purpose ontology which prevents it from reasoning about specific terminology/concepts, the restricted set of defined semantic relations , its dependence on WordNet's terms (i.e., only terms present in WordNet can be related and any novel domain-specific term will be missed), the kind of inference enabled etc

(b) Explain SVM learning method in Sequence Model Estimation. (5)

- The learning method adopted here for estimating the sequence model is a Support Vector Machine[10] (SVM). It is commonly known that Support Vector Machines are well suited for text applications given a small number of training examples.

- This is an important aspect for the commercial use of the system, since the process of gathering, preparing, and cleaning up training examples is time consuming and expensive. Support Vector Machines solve a binary classification problem.
- The SVM score associated with an instance of the considered events is its signed distance to the separating hyperplane in units of the SVM margin. In order to solve multiclass problems, a series of Support Vector Machines have to be trained, e.g., in the case of a one-vs-all training schema, the number of SVMs trained is given by the number of classes.
- The scores between these different machines are not directly comparable and the scores must be calibrated such that at least for a given classification

	instance the scores are on an equal scale. In this application, the scores not only must be comparable between classes for a given classification instance (page), but also between different classification instances (pages), i.e., the SVM scores must be mapped to probabilities. ➢ Platt uses SVM scores that are calibrated to class membership probabilities by adopting the interpretation of the score being proportional to the logarithmic ratio of class membership probability. He determines the class membership probability as a function of the SVM score by fitting a sigmoid function to the empirically observed class membership probabilities as a function of the SVM score. ➢ The fit parameters are the slope of the sigmoid function and/or a translational offset. The latter parameter, given the interpretation of the SVM scores discussed above, is the logarithmic ratio of the class prior probabilities. The method used here fixes the translational offset and only fits the slope parameter. ➢ In addition, the Support Vector Machines are trained using cost factors for the positive as well as for the negative class and optimize the two costs independently. ➢ Empirical studies performed by the authors showed that cost factor optimization in conjunction with fitting the slope parameter of the mapping function from SVM scores to probabilities yields superior probability estimates than fitting the slope and the translational offset without cost factor optimization, fitting the slope and the translational offset with cost factor optimization, and fitting the slope only	
4	Explain classical and non-classical information retrieval models with suitable examples classical information retrieval models ➢ These are based on mathematical knowledge that is easily recognized and well understood. ➢ They are simple, efficient and easy to implement. ➢ The 3 classical IR models are: 1. Boolean model 2. Probabilistic model 3. Vector model ➢ Non-classical IR models are based on principles other than similarity, probability, Boolean operations, etc., on which classical IR models are based. ➢ Non-classical IR models are: 1. Information logic model 2. Situation theory model	[10]

	3. Interaction model.	

5	Write short notes on: (i) Word 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 ➢ 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.	[10]	CO4	L2

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'

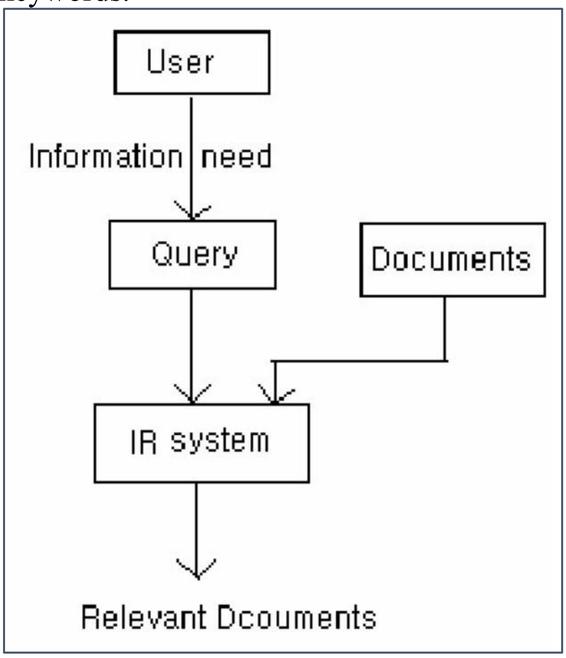
(ii) Frame Net

- ✓ 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’]

Q	Question	Marks	CO	Level
6	(a) Explain design feature of IR with a neat diagram. (5)	[10]	CO4	L3

- 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.



- 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.

(b) Define precision and recall. Explain the trade-off between them in evaluation of IR systems. (5)

0.25	1.0
0.4	0.67
0.55	0.8
0.8	0.6
1.0	0.5

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.
- ✓
- ✓ Interpolated Average Precision = 0.745