NLP – Unit 4 (Information Retrieval using NLP) – END-SEM PYQ Answers

May–June 2023

Q3(a) Information Retrieval System in NLP [4 Marks]

- **Definition:** Information Retrieval (IR) is the activity of obtaining information resources relevant to an information need (query) from a large collection of unstructured text. Unlike database queries (exact match on structured data), IR returns ranked, approximately matching documents.

- **IR System Architecture — 7 Core Components:**
  - 1. Document Collection — the corpus: web pages, PDFs, news articles, emails, legal texts
  - 2. Preprocessing Pipeline: tokenize → lowercase → remove stopwords → stem/lemmatize → index
  - 3. Inverted Index: central data structure — maps each term to a posting list of (doc_id, frequency, positions)
  - 4. Query Processor: parse user's natural language query; apply same preprocessing as documents
  - 5. Retrieval Model: compute relevance score for each candidate document (Boolean, VSM/TF-IDF, BM25, Neural)
  - 6. Ranker: sort documents by relevance score; return top-k results to user
  - 7. User Interface: display ranked results with title, URL, and relevant text snippet (KWIC: keyword in context)

- **Inverted Index Structure:**

```
Term        | Posting List (doc_id: freq)
'machine'   | [(1,3), (4,1), (7,2)]
'learning'  | [(1,5), (2,1), (4,3)]
'NLP'       | [(2,2), (5,1)]

Query: 'machine learning'
→ intersect posting lists → docs containing BOTH: [doc1, doc4]
→ rank by TF-IDF score → return sorted
```

- **Role of NLP in IR:**

NLP Task	How it Helps IR	Example
Tokenization	Splits text into indexable terms	'machine-learning' → ['machine','learning']
Stemming/Lemmatization	Normalizes word forms for unified indexing	'running','ran','runs' → 'run' (same index entry)

Stopword Removal	Removes non-discriminating words, reduces index size	Remove 'the','is','of' — tiny IDF, no retrieval value
NER in Queries	Identifies entity types for semantic search	'Apple stock' → Apple = ORG, not fruit
Query Expansion	Adds synonyms/related terms to improve recall	'car' → also retrieve 'automobile','vehicle'
Semantic Search	Maps query intent to semantically similar documents	'symptoms of cold' retrieves 'rhinovirus signs'
Summarization	Generates query-relevant snippet for result preview	Extract sentences containing query terms

Modern search engines (Google, Bing) use a 2-stage pipeline: (1) fast sparse retrieval with BM25/inverted index to get top-1000 candidates, (2) slow but accurate neural re-ranking with BERT cross-encoder to return top-10. This combination of classical IR with neural NLP is the state-of-art production architecture.

Q3(b) Named Entity Recognition (NER) and Evaluation Metrics [8 Marks]

- **NER Definition:** NER is a sequence labeling task that locates spans (contiguous word sequences) in text that refer to named entities and classifies each span into a predefined category.

- **Standard NER Entity Types:**

Entity Type	Tag	Examples	Detection Signals
Person	PER	Elon Musk, Marie Curie, Narendra Modi	Capitalized, in person-name list
Organization	ORG	Google, SPPU, United Nations, ISRO	Abbreviations, 'Inc.', 'Ltd.' suffixes
Location / GPE	LOC/GPE	Mumbai, India, Silicon Valley	Capitalized, in a gazetteer
Miscellaneous	MISC	iPhone 15, Olympics, English (language)	Doesn't fit PER/ORG/LOC
Date/Time	DATE/TIME	January 5 2025, last Monday, 3 PM	Temporal expressions, month names
Money	MONEY	$500 million, Rs. 10,000	Currency symbols, numeric patterns

- **IOB (Inside-Outside-Beginning) Tagging Scheme:**

```
Sentence: 'Barack Obama visited New York yesterday'

Token:    Barack   Obama    visited  New      York     yesterday
IOB Tag:  B-PER    I-PER    O        B-LOC    I-LOC    O
```

```
B-PER = Beginning of a PERSON entity
I-PER = Inside (continuation) of the same PERSON entity
O     = Outside any entity

⇒ Entities: [Barack Obama] (PERSON),  [New York] (LOCATION)
```

- **BIOES Variant:**

```
B = Beginning of multi-token entity
I = Inside multi-token entity
O = Outside
E = End (last token) of multi-token entity
S = Single-token entity

Example: 'Obama visited Paris'
  S-PER  O  B-LOC  E-LOC
```

- **NER Evaluation Metrics:**

Metric	Formula	Interpretation
Precision (P)	TP / (TP + FP)	Of all predicted entities, what % are correct?
Recall (R)	TP / (TP + FN)	Of all real entities, what % did the system find?
F1-Score	2PR / (P + R)	Harmonic mean — penalizes extreme imbalance

- **Worked NER Evaluation Example:**

```
Gold:      {[Tim Cook: PER], [Apple: ORG], [California: LOC]}
Predicted: {[Tim Cook: PER], [Apple: ORG], [California: ORG]}

Tim Cook   → TP  (correct span + correct type)
Apple      → TP  (correct span + correct type)
California → FP  (correct span, wrong type: predicted ORG, actual LOC)
           → FN  (the correct LOC entity was missed)

Precision = 2/3 = 0.667
Recall    = 2/3 = 0.667
F1        = 0.667
```

> *F1-Score is the standard NER metric. BERT-CRF achieves ~93% F1 on CoNLL-2003 English. Common errors: (1) Boundary errors — predicted 'New York' but actual is 'New York City'. (2) Type errors — correct span, wrong category. (3) Missing entities — common with domain-specific or ambiguous terms.*

Q3(c) Cross-Lingual Information Retrieval (CLIR) [6 Marks]

- **CLIR Definition:** A form of IR where the query is in one language and relevant documents are in a different language. The system must bridge the language gap to retrieve relevant cross-language content.

- **CLIR Approaches:**

Approach	Mechanism	Pros	Cons
Query Translation	Translate query to target language; run standard IR	Fast, cheap — only translate the short query	MT errors propagate; query meaning may shift
Document Translation	Translate all target documents to source language	Can use existing monolingual IR system	Extremely expensive for large corpora
Bilingual Dictionary	Map each source word to target words via dictionary	Simple, no MT system needed	Cannot handle phrases, idioms, OOV
Cross-lingual Embeddings	Map source + target to shared vector space (LaBSE, mBERT)	No explicit translation; captures semantics	Needs multilingual training data
Pivot Language	Source → English → Target (indirect translation)	Leverages large English MT resources	Two translation steps compounds errors

- **Cross-lingual Embedding Approach (Modern Best Practice):**
  - LaBSE: maps sentences from 109 languages into a shared 768-dim space
  - Semantically equivalent sentences cluster together regardless of language
  - CLIR query: encode query with LaBSE → find documents with nearest LaBSE vectors in target language

> *Modern production CLIR (Google multilingual search) uses multilingual neural retrievers (mDPR, LASER, LaBSE) rather than explicit query translation. These handle idiomatic expressions and culturally-specific concepts better than word-for-word translation.*

Q4(a) Vector Space Model (VSM) in Information Retrieval [6 Marks]

- **VSM Concept:** Represents both documents and queries as vectors in a high-dimensional space where each dimension corresponds to a unique vocabulary term. Relevance is the cosine of the angle between document and query vectors.

- **Cosine Similarity Formula:**

$$\text{sim}(d, q) = \cos(\theta) = (d \cdot q) / (|d| \times |q|)$$

$$= [\Sigma^{I}(d^{I} \times q^{I})] / [\text{sqrt}(\Sigma^{I} d^{I2}) \times \text{sqrt}(\Sigma^{I} q^{I2})]$$

Range: 0 (no overlap) to 1 (identical content)
Length-normalised: document length does not affect score

- **BM25 (Best Match 25) — Improved VSM:**

$$BM25(t,d) = IDF(t) \times [TF(t,d) \times (k_1+1)] / [TF(t,d) + k_1 \times (1-b + b \times |d|/avgdl)]$$

$k_1 = 1.5–2.0$ (term saturation: prevents 100 occurrences scoring 100× better than 1)
$b = 0.75$ (length normalisation: penalises unusually long documents)
avgdl = average document length in corpus

Aspect	Strengths	Weaknesses
Ranking	Provides graded relevance (not binary)	Ignores term proximity
Partial match	Docs with some query terms score > 0	Terms with zero overlap score exactly 0
Semantics	None — bag of words only	Synonyms (car/automobile) are separate dimensions
Word order	Not applicable	Ignored completely

BM25 remains the industry standard for sparse retrieval (Elasticsearch, Apache Lucene, TREC evaluations). Dense retrieval (DPR, bi-encoder BERT) has surpassed BM25 on passage retrieval benchmarks but BM25 is still used as the first stage in production systems for its speed and zero training data requirement.

Q4(b) Entity Extraction and Relation Extraction [8 Marks]

- **Entity Extraction:** Identifying and extracting all mentions of named or domain-specific entities from unstructured text. Broader than standard NER — can include domain-specific entities not in standard taxonomies.

- **Entity Extraction Methods:**

Method	Approach	Pros	Cons
Rule-Based / Pattern	Regex, capitalization heuristics, gazetteer lookup	Transparent, fast, no training data	Brittle, misses new/unseen entities
Classical ML (CRF)	Feature engineering + CRF sequence labeler	Good accuracy, generalisation	Feature engineering effort required
Deep Learning (BiLSTM-CRF)	BiLSTM learns contextual features; CRF decodes	High accuracy, learns features automatically	Needs substantial labelled data
Transformer (BERT-NER)	Fine-tuned BERT + token classification head + CRF	State-of-art accuracy; context-aware	Computationally expensive; needs GPU

- **Relation Types and Examples:**

Relation	Entity 1 Type	Entity 2 Type	Example Sentence
WORKS_AT / CEO_OF	PERSON	ORG	Sundar Pichai is the CEO of Google
BORN_IN	PERSON	LOCATION	Marie Curie was born in Warsaw
FOUNDED_BY	ORG	PERSON	Apple was founded by Steve Jobs in 1976
ACQUIRED	ORG	ORG	Google acquired YouTube in 2006 for $1.65B
LOCATED_IN	LOC	LOC	Pune is located in Maharashtra, India
PRODUCT_OF	PRODUCT	ORG	iPhone is manufactured by Apple

- **Relation Extraction Methods:**
  - Hearst Patterns (rule-based): 'X such as Y_1, Y_2' → Y IS-A X. 'fruits such as apple, mango' → apple IS-A fruit.
  - Supervised classification: train binary classifier on (e_1, context, e_2) triples labelled with relation type
  - Distant supervision: use knowledge base (Freebase, Wikidata) to auto-label — if (A,B) linked in KB, label all sentences mentioning both
  - Neural (BERT-RE): insert [E1] and [E2] markers around entity spans; classify relation from [CLS] or entity representations

Entity extraction + Relation extraction = Information Extraction (IE). IE is the foundation for building structured knowledge bases from raw text. Google's Knowledge Graph and Wikidata are populated largely by IE systems. The pipeline: Text → NER → RE → (entity$_1$, relation, entity$_2$) triples → Knowledge Graph nodes and edges.

Q4(c) Coreference Resolution [4 Marks]

- **Definition:** Coreference Resolution is the NLP task of finding all expressions (mentions) in a document that refer to the same real-world entity and grouping them into coreference chains.

Term	Definition	Example
Mention	Any noun phrase referring to an entity	Barack Obama, He, The president, Obama
Antecedent	The earlier (canonical) mention in the chain	Barack Obama — first full-name introduction
Coreference Chain	All mentions of the same entity grouped	Barack Obama → He → The president → Obama
Anaphora	Reference that looks BACK to a prior entity	He (refers back to Barack Obama)
Cataphora	Reference that looks FORWARD to a later entity	Before he left, John packed his bags. (he = John)
Bridging	Implicit reference via semantic association	I saw a car. The engine was loud. (engine OF the car)

- **Worked Examples:**

```
Example 1:
 'Ratan Tata built the Tata Group. He grew it into a global empire.'
 Chain 1: {Ratan Tata, He}
 Chain 2: {the Tata Group, it}

Example 2 (cataphora):
 'Before she left for work, Priya made coffee.'
 Chain 1: {she, Priya}  — 'she' appears BEFORE 'Priya'
```

> *Coreference is critical for downstream tasks: without it, an NLP system reading 'John hired Mary. He said she was great.' cannot connect 'He' to John or 'she' to Mary — losing critical relational information for IE, QA, and summarisation. SpanBERT achieves 83% F1 on OntoNotes.*

November–December 2023

Q3(a) IR Concept and Significance of NLP in IR [4 Marks]

REPEATED — Refer to: May–June 2023 → Q3(a) [IR architecture 7 components + NLP role table]

IR System Type	Description	Example
Web IR	Retrieves web pages from billions of documents	Google Search, Bing
Enterprise IR	Searches internal corporate documents, emails	Elasticsearch, SharePoint search
Legal IR	Specialised for case law and statute retrieval	LexisNexis, Westlaw
Medical IR	Searches clinical notes, medical literature	PubMed, ClinicalKey
Conversational IR	Multi-turn dialogue to refine search	Cortana, Alexa knowledge search

Q3(b) Reference Resolution and Coreference Resolution [8 Marks]

REPEATED — Refer to: May–June 2023 → Q4(c) [Full coreference explanation with terms table and examples]

Type	Description	Example
Pronominal reference	Pronoun refers to entity	She → Mary
Definite NP	Definite description refers to specific entity	The scientist → Albert Einstein
Anaphora	Back-reference to prior entity	John arrived. He was late.
Cataphora	Forward reference to entity stated later	Before he spoke, Obama paused.
Bridging	Implicit association, not identical entity	I read a book. The author was French.

- **Why Reference Resolution Is Hard:**
  - Gender/number ambiguity: 'The nurse called the doctor. They…' — who does 'they' refer to?
  - Pleonastic pronouns: 'It is raining' — 'it' has no antecedent (dummy pronoun)
  - World knowledge required: 'The trophy wouldn't fit in the suitcase because it was too big' — what does 'it' refer to? Requires reasoning about sizes.

Q3(c) Cross-Lingual Information Retrieval [6 Marks]

REPEATED — Refer to: May–June 2023 → Q3(c) [CLIR definition, 5 approaches table, cross-lingual embeddings]

Q4(a) Vector Space Model [6 Marks]

REPEATED — Refer to: May–June 2023 → Q4(a) [VSM full explanation with cosine formula, BM25, comparison table]

Q4(b) Entity Extraction and Relation Extraction [8 Marks]

REPEATED — Refer to: May–June 2023 → Q4(b) [Entity extraction 4 methods + relation types + end-to-end extraction]

Q4(c) Named Entity Recognition (NER) [4 Marks]

REPEATED — Refer to: May–June 2023 → Q3(b) [NER definition, entity types, IOB tagging, evaluation metrics, worked example]

May–June 2024

Q3(a) VSM: Representation, Similarity, Strengths and Weaknesses [9 Marks]

REPEATED — Refer to: May–June 2023 → Q4(a) [VSM full explanation including BM25]

- **IR Evaluation Metrics:**

Metric	Formula	Interpretation
Precision@k	Relevant in top-k / k	Quality of top-k results
Recall	Retrieved relevant / Total relevant	Completeness — did we find all relevant docs?
MAP	Mean of Average Precision across queries	Single number for system-level evaluation
NDCG@k	$\Sigma(rel_i / \log_2(i+1))$ normalised by ideal order	Rewards highly relevant docs ranked higher
MRR	Mean of $1/rank_{first\_relevant}$	For QA — how quickly do we find the answer?

Q3(b) NER Evaluation and Error Analysis [9 Marks]

REPEATED — Refer to: May–June 2023 → Q3(b) [NER types, IOB, metrics, worked example]

Error Type	Symptom	Root Cause	Solution
Low Precision	Many false positives	Model predicts entities too aggressively	Add hard negatives to training
Low Recall	Many entities missed	Model too conservative	Collect more labelled data for rare types
Boundary Errors	Wrong span	Tokenisation issues; compound names	Better tokeniser; more multi-word examples
Type Confusion	Right span, wrong category	Ambiguous entities — 'Apple' as company vs fruit	Add context features; wider context window
Domain Mismatch	High on news, low on medical text	Training distribution mismatch	Domain-adaptive pre-training (BioBERT)

Q4(a) CLIR: Challenges and MT Assistance [9 Marks]

REPEATED — Refer to: May–June 2023 → Q3(c) [CLIR 5 approaches table + cross-lingual embeddings]

Challenge	Description	Solution
Translation Ambiguity	Source word has multiple target translations	Context-aware Neural MT (Transformer seq2seq)
Morphological Variation	Target language inflects words heavily	Stemming/lemmatisation; subword BPE tokenisation
NE Translation	Proper nouns not in MT training data	Transliteration models; named entity lookup tables
Low-resource Pairs	Little parallel training data	Cross-lingual transfer from English; back-translation
Word Reordering	SOV (Japanese) vs SVO (English) order	Transformer MT handles via self-attention
Code-switching	Query mixes two languages	Multilingual BERT; language detection first
Semantic Drift	Translated query changes meaning subtly	Query expansion before/after translation

Q4(b) Entity Extraction: Domain-Specific Applications [9 Marks]

REPEATED — Refer to: May–June 2023 → Q4(b) [Entity extraction methods + domain applications]

Domain	Entity Types Extracted	Tools/Models	Use Case
Healthcare	Drug name, dosage, route, side effect, ICD code, gene	BioBERT, SciSpaCy	Populate EHR, clinical decision support
Finance	Company name, revenue figure, fiscal date, rating	FinBERT, custom CRF	Earnings report analysis
Legal	Case citation, statute number, judge name, ruling date	Legal-BERT, rule-based	Case law mining, due diligence
E-commerce	Product brand, colour, size, material, price	Custom NER + attribute extraction	Product catalogue normalisation
Social Media	Event name, location, hashtag, mention (@user)	TweetNLP, BERTweet	Trend detection, crisis monitoring

May–June 2025

Q3(a) Reference Resolution and Coreference Resolution [8 Marks]

REPEATED — Refer to: Nov–Dec 2023 → Q3(b) [Reference types + coreference algorithms]

- **Extended — Coreference in Knowledge Graph Building:**

```
Text: 'Ratan Tata founded Tata Group. He grew the company into a global conglomerate.
      The industrialist was awarded the Padma Vibhushan.'

Step 1 — ENTITY EXTRACTION:
  [PERSON: Ratan Tata]  [ORG: Tata Group]  [AWARD: Padma Vibhushan]
  [PERSON: He]          [ORG: the company]  [TITLE: The industrialist]

Step 2 — COREFERENCE RESOLUTION:
  Chain 1: {Ratan Tata, He, The industrialist}  → canonical: Ratan Tata
  Chain 2: {Tata Group, the company}            → canonical: Tata Group

Step 3 — RELATION EXTRACTION (after coreference):
  FOUNDED(Ratan Tata, Tata Group)
  AWARDED(Ratan Tata, Padma Vibhushan)
```

Q3(b) CLIR [8 Marks]

REPEATED — Refer to: May–June 2023 → Q3(c) + May–June 2024 → Q4(a) [Full CLIR with challenges table]

Q3(c) What is Information Retrieval? [2 Marks]

- **IR:** Finding relevant documents from a large collection in response to a user's query. Unlike database queries (exact match), IR handles approximate matching on unstructured text and returns results ranked by relevance score.

Q4(a) Entity Extraction in Information Retrieval [8 Marks]

REPEATED — Refer to: May–June 2023 → Q4(b) [Entity extraction methods + domain applications table]

- **Additional — How Entity Extraction Improves IR:**
  - Entity-based indexing: index documents by named entities — 'articles about Elon Musk' matches by entity, not keyword
  - Semantic expansion: query contains 'Tesla' → expand to also search for 'Elon Musk' (CEO), 'SpaceX' (related org), 'Model S' (product)
  - Knowledge-driven retrieval: link query entities to knowledge graph nodes → retrieve documents about related entities

Q4(b) Vector Space Model [8 Marks]

REPEATED — Refer to: May–June 2023 → Q4(a) [Full VSM with formula, BM25, evaluation metrics]

Q4(c) What is NER? [2 Marks]

- **NER:** Named Entity Recognition is a sequence labelling task that identifies and classifies named entity mentions (PERSON, ORGANIZATION, LOCATION, DATE, MONEY, etc.) in text.
  - Example: 'Tim Cook [PER] announced iPhone 15 [PRODUCT] at Apple Park [LOC].'

November–December 2025

Q3(a) Vector Space Model in IR [9 Marks]

REPEATED — Refer to: May–June 2023 → Q4(a) + May–June 2024 → Q3(a) [VSM with BM25 and evaluation metrics]

Model	Term Weighting	Handles Length?	Handles Saturation?	Best For
Boolean	Binary (term present/absent)	No	No	Exact, unranked search
VSM / TF-IDF	TF × IDF	Partially	No	Baseline ranked retrieval
BM25	BM25 score (k_1, b parameters)	Yes	Yes	Standard IR benchmark
Dense Retrieval (DPR)	Dot product of BERT embeddings	Yes	Yes (semantic)	Neural, semantic retrieval

Q3(b) Entity, Relation Extraction and Coreference [8 Marks]

REPEATED — Refer to: May–June 2023 → Q4(b) [entity+relation extraction] + Q4(c) [coreference] + MJ25 Q3(a) [unified example]

Q4(a) NER System Building [9 Marks]

REPEATED — Refer to: May–June 2023 → Q3(b) + Nov–Dec 2023 → Q3(b) [error analysis table]

- **Modern BERT-NER System Building Steps:**
  - Step 1 – Collect domain text + annotate entities using BRAT or Prodigy annotation tools
  - Step 2 – Format as IOB-tagged CoNLL: word tab POS tab chunk tab NER_tag per line
  - Step 3 – Load pre-trained BERT or domain-specific BERT (BioBERT for medical, LegalBERT for law)
  - Step 4 – Add token classification head: linear layer on BERT token representations → logits over IOB tags
  - Step 5 – Optionally add CRF layer for globally consistent tag sequences (prevents invalid I-PER without B-PER)
  - Step 6 – Fine-tune: Adam optimiser, lr = 2×10^{-5}, batch size = 32, 3–5 epochs
  - Step 7 – Evaluate on held-out test set: compute P, R, F1 per entity type
  - Step 8 – Error analysis: inspect FP and FN; retrain with hard examples

Q4(b) CLIR: Challenges and Approaches [8 Marks]

REPEATED — Refer to: May–June 2023 → Q3(c) + May–June 2024 → Q4(a) [Complete CLIR with all challenge tables]

Topic Frequency Analysis — Unit 4

Rank	Topic	Sessions	Priority
1	CLIR — all aspects	MJ23, ND23, MJ24, MJ25, ND25	VERY HIGH — 5× every exam
2	Vector Space Model + Cosine Similarity	MJ23, ND23, MJ24, MJ25, ND25	VERY HIGH — 5× every exam
3	Entity + Relation Extraction	MJ23, ND23, MJ24, MJ25, ND25	VERY HIGH — 5× every exam
4	NER (Definition + IOB + Metrics)	MJ23, ND23, MJ24, MJ25, ND25	VERY HIGH — 5× every exam
5	Coreference / Reference Resolution	MJ23, ND23, MJ25, ND25	HIGH — 4×
6	Information Retrieval Introduction	MJ23, ND23, MJ25	MEDIUM — 3×

EXAM STRATEGY — Unit 4: (1) Cosine similarity formula — memorise and practice computing it. (2) IOB tagging — practise converting sentences to IOB tags manually. (3) CLIR approaches table — know all 5 approaches with pros/cons. (4) Coreference chain examples — always draw the full chain. (5) BM25 formula — understand the k_1 and b parameters.