RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based On AICTE Flexible Curricula

CSE-Data Science/Data Science, VI semester

Open Elective CD604 (A) Information Extraction and Retrieval

UNIT-I

Introduction - History of IR- Components of IR - Issues -Open source Search engine Frameworks - The Impact of the web on IR - The role of artificial intelligence (AI) in IR – IR Versus Web Search - Components of a search engine, Characterizing the web.

UNIT -II

Boolean and Vector space retrieval models- Term weighting - TF-IDF weighting- cosine similarity - Preprocessing - Inverted indices - efficient processing with sparse vectors Language Model based IR - Probabilistic IR -Latent Semantic indexing - Relevance feedback and query expansion.

UNIT- III

Web search overview, web structure the user paid placement search engine optimization, Web Search Architectures - crawling - meta-crawlers, Focused Crawling - web indexes - Nearduplicate detection - Index Compression - XML retrieval.

UNIT -IV

Link Analysis -hubs and authorities - Page Rank and HITS algorithms -Searching and Ranking - Relevance Scoring and ranking for Web - Similarity - Hadoop & Map Reduce - Evaluation - Personalized search - Collaborative filtering and content-based recommendation of documents And products - handling invisible Web - Snippet generation, Summarization. Question Answering, Cross- Lingual Retrieval.

UNIT -V

Information filtering: organization and relevance feedback - Text Mining- Text classification and clustering - Categorization algorithms, naive Bayes, decision trees and nearest neighbor - Clustering algorithms: agglomerative clustering, k-means, expectation maximization (EM).

References:

1. C. Manning, P. Raghvan and H Schutze: Introduction to Information Retrieval, Cambridge University Press, 2008.
2. Ricardo Baeza -Yates and Berthier Ribeiro –Neto, Modern Information Retrieval. The Concepts and Technology behind Search 2nd Edition, ACM Press Books 2011.
3. Bruce Croft, Donald Metzler and Trevor Strohman Search Engines Information Retrieval in Practice 1st Edition Addison Wesley, 2009
4. 4.Mark Levene, An Introduction to Search Engines and Web Navigation, 2nd Edition Wiley 2010.

RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based On AICTE Flexible Curricula

CSE-Data Science/Data Science, VI semester

Open Elective CD604 (B) Agile Software Development

Pre-Requisite: Software Engineering
Course Outcomes:
After completing the course student should be able to:

1. Describe the fundamental principles and practices associated with each of the agile development methods.
2. Compare agile software development model with traditional development models and identify the benefits and pitfalls.
3. Use techniques and skills to establish and mentor Agile Teams for effective software development.
4. Apply core values and principles of Agile Methods in software development.

Course Contents:

Unit-I: Fundamentals of Agile Process: Introduction and background, Agile Manifesto and Principles, Stakeholders and Challenges, Overview of Agile Development Models: Scrum, Extreme Programming, Feature Driven Development, Crystal, Kanban, and Lean Software Development.

Unit-II: Agile Projects: Planning for Agile Teams: Scrum Teams, XP Teams, General Agile Teams, Team Distribution; Agile Project Lifecycles: Typical Agile Project Lifecycles, Phase Activities, Product Vision, Release Planning: Creating the Product Backlog, User Stories, Prioritizing and Estimating, Creating the Release Plan; Monitoring and Adapting: Managing Risks and Issues, Retrospectives.

Unit-III: Introduction to Scrum: Agile Scrum Framework, Scrum Artifacts, Meetings, Activities and Roles, Scrum Team Simulation, Scrum Planning Principles, Product and Release Planning, Sprinting: Planning, Execution, Review and Retrospective; User story definition and Characteristics, Acceptance tests and Verifying stories, Burn down chart, Daily scrum, Scrum Case Study.

Unit-IV: Introduction to Extreme Programming (XP): XP Lifecycle, The XP Team, XP Concepts: Refactoring, Technical Debt, Timeboxing, Stories, Velocity; Adopting XP: Pre-requisites, Challenges; Applying XP: Thinking- Pair Programming, Collaborating, Release, Planning, Development; XP Case Study.

Unit-V: Agile Software Design and Development: Agile design practices, Role of design Principles, Need and significance of Refactoring, Refactoring Techniques, Continuous Integration, Automated build tools, Version control; Agility and Quality Assurance: Agile Interaction Design, Agile approach to Quality Assurance, Test Driven Development, Pair programming: Issues and Challenges.

Recommended Books:

1. Robert C. Martin, Agile Software Development- Principles, Patterns and Practices, Prentice Hall, 2013.
2. Kenneth S. Rubin, Essential Scrum: A Practical Guide to the Most Popular Agile Process, Addison Wesley, 2012.
3. James Shore and Shane Warden, The Art of Agile Development, O'Reilly Media, 2007.
4. Craig Larman, ―Agile and Iterative Development: A manager's Guide, Addison-Wesley, 2004.
5. Ken Schawber, Mike Beedle, Agile Software Development with Scrum, Pearson, 2001.
6. Cohn, Mike, Agile Estimating and Planning, Pearson Education, 2006.
7. Cohn, Mike, User Stories Applied: For Agile Software Development Addison Wisley, 2004.

Online Resources:

1. IEEE Transactions on Software Engineering
2. IEEE Transactions on Dependable and Secure Computing
3. IET Software
4. ACM Transactions on Software Engineering and Methodology (TOSEM)
5. ACM SIGSOFT Software Engineering Notes

RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based On AICTE Flexible Curricula

CSE-Data Science/Data Science, VI semester

Open Elective CD604 (C) Natural Language Processing

COURSE OBJECTIVES: Students should develop a basic understanding in natural language
processing methods and strategies and to evaluate the strengths and weaknesses of various Natural
Language Processing (NLP) methods & technologies and gain an insight into the application areas
of Natural language processing.

Detailed Contents:

UNIT I: Introduction: Origins and challenges of NLP, Human languages, Application and Goals of NLP, Main approach of NLP, Knowledge in speech and language processing, Ambiguity, Models and algorithms, Formal language and Natural Language, Regular Expression, and automata.

UNIT II: Text Pre-processing, Tokenization, Feature Extraction from text Morphology: Inflectional morphology, Derivational morphology, Finite state morphological parsing, Morphology, and Indian languages. Part of Speech Tagging: Rule based, Stochastic POS, HMM, Transformation based tagging (TBL), N-Grams: Simple N-grams, Smoothing, Backoff, Entropy. Handling of unknown words, Named entities, Multi word expressions.

UNIT III: Parsing: Syntactic and statistical parsing, parsing algorithms, hybrid of rule based and probabilistic parsing, scope ambiguity and attachment ambiguity resolution, Tree banks. Discourse and dialogue: discourse and dialogue analysis, anaphora resolution, named entity resolution, event anaphora, Information extraction and retrieval. Hidden Markov and Maximum Entropy models, Viterbi algorithms and EM training.

UNIT IV: Semantic analysis, Semantic attachments – Word Senses, Relations between Senses, Word Sense Disambiguation, WSD using Supervised, Dictionary & Thesaurus, Bootstrapping methods –Word Similarity using Thesaurus and Distributional methods. Compositional semantics.

Speech Processing: Speech and phonetics, Vocal organ, Phonological rules and Transducer, Probabilistic models: Spelling error, Bayesian method to spelling, Minimum edit distance, Bayesian method of pronunciation variation.

UNIT V: Application of NLP: intelligent work processors: Machine translation, user interfaces,Man-Machine interfaces, natural language querying, tutoring and authoring systems, speechrecognition, and commercial use of NLP.

Text Books:

1. Daniel Jurafsky, James H. Martin―Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech, Pearson Publication.
2. Steven Bird, Ewan Klein and Edward Loper, ―Natural Language Processing with Python, OReilly Media.
3. Manning and Schutze "Foundations of Statistical Natural Language Processing", MIT Press.

Reference Books:

1. Breck Baldwin, Language Processing with Java and LingPipe Cookbook, Atlantic Publisher.
2. Richard M Reese, Natural Language Processing with Java, OReilly Media.
3. Nitin Indurkhya and Fred J. Damerau, Handbook of Natural Language Processing, Chapman and Hall/CRC Press.
4. Tanveer Siddiqui, U.S. Tiwary, Natural Language Processing and Information Retrieval, Oxford University Press.

UNIT I: Introduction to Natural Language Processing (NLP)

1. **Origins and Challenges of NLP**:

- Historical background: Evolution from early language processing attempts to modern NLP.

- Challenges: Ambiguity, context sensitivity, syntactic and semantic complexities.

2. **Human Languages and Application Goals**:

- Characteristics of human languages: Syntax, semantics, pragmatics.

- NLP applications: Machine translation, sentiment analysis, information retrieval.

3. **Main Approaches of NLP**:

- Rule-based systems: Knowledge engineering, expert systems.

- Statistical models: Probabilistic language models, hidden Markov models.

- Deep learning: Neural networks, recurrent neural networks, transformer models.

4. **Knowledge in Speech and Language Processing**:

- Linguistics fundamentals: Phonetics, phonology, syntax, semantics.

- Computational linguistics: Corpus linguistics, annotation, linguistic resources.

5. **Ambiguity, Models, and Algorithms**:

- Types of ambiguity: Lexical, syntactic, semantic.

- NLP models: Finite state machines, context-free grammars, neural networks.

- Algorithms: Part-of-speech tagging, parsing, machine translation.

6. **Formal Language and Natural Language**:

- Formal language theory: Regular expressions, context-free grammars.

- Challenges in natural language: Context dependence, ambiguity, non-linearity.

7. **Regular Expression and Automata**:

- Regular expression basics: Metacharacters, quantifiers, character classes.
- Automata theory: Finite automata, pushdown automata, Turing machines.

UNIT II: Text Pre-processing and Morphology

1. **Tokenization and Feature Extraction**:

- Tokenization techniques: Word-based, character-based, subword tokenization.
- Feature extraction methods: Bag-of-words, TF-IDF, word embeddings.

2. **Morphology**:

- Inflectional and derivational morphology: Affixation, compounding, inflectional paradigms.
- Finite state morphological parsing: Transducers, morpheme segmentation.

3. **Part of Speech Tagging (POS)**:

- POS tagging algorithms: Rule-based, stochastic, deep learning.
- N-Grams and smoothing: Language modeling, perplexity, backoff techniques.

4. **Handling Unknown Words and Named Entities**:

- Out-of-vocabulary handling: Morphological analysis, context-based inference.
- Named entity recognition: Rule-based, statistical, deep learning approaches.

UNIT III: Parsing, Discourse, and Dialogue

1. **Parsing**:

- Syntactic parsing algorithms: Top-down, bottom-up, chart parsing.

- Statistical parsing models: PCFG, dependency parsing, transition-based parsing.

2. **Discourse and Dialogue Analysis**:

- Discourse coherence: Cohesion, coherence relations, rhetorical structure theory.

- Dialogue act classification: Intent recognition, slot filling, dialogue state tracking.

3. **Information Extraction and Retrieval**:

- Information extraction tasks: Named entity extraction, relation extraction, event extraction.

- Information retrieval models: Vector space model, BM25, language models.

UNIT IV: Semantic Analysis and Speech Processing

1. **Semantic Analysis**:

- Word sense disambiguation: Supervised, unsupervised, knowledge-based methods.

- Semantic role labeling: Predicate-argument structures, FrameNet, PropBank.

2. **Speech Processing**:

- Speech signal processing: Pre-emphasis, framing, windowing, feature extraction.

- Speech recognition models: Hidden Markov models, deep neural networks, sequence-to-sequence models.

UNIT V: Applications of NLP

1. **Intelligent Work Processors**:

- Machine translation systems: Statistical MT, neural MT, domain adaptation.

- Natural language understanding: Intent classification, entity recognition, sentiment analysis.

2. **Commercial Use of NLP**:

- Industry-specific applications: Healthcare (clinical NLP), finance (algorithmic trading), customer service (chatbots).

- Ethical considerations: Bias mitigation, privacy preservation, transparency in AI systems.

By expanding on each of these topics with explanations, examples, algorithms, and real-world applications, you can create comprehensive notes exceeding 10,000 words.