UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
Coordenação do Curso de Pós-Graduação em Informática Aplicada (PPGIA) - DEINFO/UFRPE
Rua Dom Manoel de Medeiros, s/n, Dois Irmãos – Recife/PE
CEP: 52171-900 – www.deinfo.ufrpe.br

PROGRAMA DE DISCIPLINA

IDENTIFICAÇÃO
DISCIPLINA: Mineração de Textos **CÓDIGO: PPGIA 7345**
DEPARTAMENTO: DEINFO **ÁREA:** Inteligência Artificial
CARGA HORÁRIA TOTAL : 60 h
NÚMERO DE CRÉDITOS: 04
CARGA HORÁRIA SEMANAL: 4 h **TEÓRICAS:** 2h **PRÁTICAS**: 2h
PRÉ-REQUISITOS: Nenhum
CO-REQUISITOS: Nenhum

EMENTA

Processamento de Linguagem Natural (subtarefas). Modelos de Linguagens (*n-gram*, probabilísticos). Representação de documentos (esparsa e densa). Semântica Distribucional (*Word Embedding*). Classificação de documentos. Agrupamento de documentos. Extração de Informação. Recuperação de Informação. Introdução à Aprendizagem de Máquina Simbólica. Sumarização Automática de Documentos. Análise de Sentimentos. Repositórios Semânticos. Estudo/discussão de Artigos

CONTEÚDO

UNIDADES E ASSUNTOS

(1) Processamento de Linguagem Natural (PLN)
- . Introdução e Histórico
- . Subtarefas em PLN
- . Pré-processamento em Aplicações
  - . Lemmatização
  - . Categorização gramatical
  - . Parsing
    - . constituinte
    - . dependências
  - . Resolução de Correferência
  - . Frames semânticos
- . Prática com toolkits para PLN

(2) Representação de Documentos e dados textuais
- . Modelo baseado em N-gramas
- . Modelo de representação densa
- . Semântica Distribucional

(3) Classificação de documentos
- . Abordagens baseadas em Aprendizagem de Máquina
- . Aplicações

(4) Agrupamento de documentos
. Abordagens e Aplicações

(5) Recuperação de Informação
. Modelos de Representação: Vetorial e Probabilísticos
. Indexação
. Busca
. Métricas de Avaliação

(6) Extração de Informação (EI)
. Histórico da área de EI
. Abordagens existentes
. Extração de Informação
. baseada em Ontologias
. Aprendizagem e Povoamento de Ontologias
. Métricas de avaliação
. *Shared-tasks* em EI
. Tarefas específicas em EI:
. Reconhecimento de Entidades Nomeadas
. Extração de Relações
. Extração de Eventos
. Extração de Informação Temporal

(7) Introdução à aprendizagem de máquina simbólica
. Programação em Lógica Indutiva

(8) Sumarização Automática de Documentos (SAD)
. Abordagens: Extrativa, Semi-Extrativa e Abstrativa
. Métodos: estatísticos, baseados em PLN superficial e PLN profundo, baseados em grafos
. *Shared tasks* em SAD
. Métricas de avaliação

(9) Análise de Sentimentos (AS)
. Abordagens baseadas em dicionários
. Abordagens baseadas em Aprendizagem de Máquina
. Abordagens baseadas em Ontologias
. Shared tasks em AS
. AS baseada em aspectos

(10) Repositório Semânticos
. tesauros
. baseados em ontologias

(11) Estudos/Discussão de Periódicos

BIBLIOGRAFIA

BÁSICA

1. Dan Jurafsky e James H. Martin. Speech and Language Processing, 2nd Edition. Prentice Hall, 2008.

2. Christopher D. Manning, Hinrich Schütze. Foundations of Statistical Natural Language Processing (1st Edition). MIT Press, 1999.

3. Ricardo Baeza-Yates e Berthier Ribeiro-Neto. Recuperação de Informação. Conceitos e Tecnologia das Máquinas de Busca (2a. Edição). Editora Bookman, 2013.

COMPLEMENTAR

1. Steven Bird, Ewan Klein, e Edward Loper. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, 2009.

2. Peter Jackson e Isabelle Moulinier. Natural Language Processing for Online Applications: Text retrieval, extraction and categorization (2nd Edition), John Benjamins Publishing Company, 2007.

3. Torres-Moreno, Juan-Manuel. Automatic Text Summarization, John Wiley & Sons, 2014.

4. Michael W. Berry, Jacob Kogan. Text Mining: Applications and Theory. Wiley, 2010.

5. Ashok N. Srivastava, Mehran Sahami. Text Mining: Classification, Clustering, and Application. Chapman & Hall. CRC Data Mining and Knowledge Discovery Series, 2009.