UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Faculdade de Computação
Av. João Naves de Ávila, nº 2121, Bloco 1A - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4144 - http://www.portal.facom.ufu.br/ facom@ufu.br

PLANO DE ENSINO

1. IDENTIFICAÇÃO

Componente Curricular:	Organização e Recuperação da Informação							
Unidade Ofertante:	Faculdade de Computação							
Código:	GSI024		Período/Série:		5º	Turma:	S	
Carga Horária:						Natureza:		
Teórica:	30	Prática:	30	Total:	60	Obrigatória (X)	Optativa:	( )
Professor(A):	Wendel Alexandre Xavier de Melo					Ano/Semestre:	2023-1	
Observações:								

2. EMENTA

Conceitos de documento, palavra e termo. Indexação de documentos: extração de termos, *stopwords*, *stemming*, criação de índices. *Thesauri*. Modelos de processamento de consultas. Avaliação de Sistemas de Recuperação de Informação (RI). RI em Documentos semi-estruturados, multimídia e documentos na Web. Extração da informação. Classificação de documentos. Redução de dimensionalidade.

3. JUSTIFICATIVA

A Organização e Recuperação de Informação (ORI) é uma disciplina abrangente da Ciência da Computação que se concentra principalmente em prover aos usuários o acesso fácil às informações de seu interesse. Em particular, essa disciplina trata da representação, armazenamento, organização, e acesso a itens de informação, como documentos, páginas da Internet, catálogos online, registros estruturados e semiestruturados, objetos multimídia e etc. Devido ao volume gigantesco de informação gerado pelos sistemas de informação, as máquinas de busca de informação se tornaram ferramentas fundamentais para localizar e recuperar informação. Portanto, os conhecimentos ligados ao funcionamento e construção de máquinas de busca são indispensáveis para permitir uma formação atualizada e multidisciplinar do bacharel em Sistemas de Informação.

4. OBJETIVO

Objetivo Geral:

Propor soluções para o problema de recuperar informações nos documentos de uma determinada coleção (estruturada ou semi-estruturada), a partir de uma consulta formulada pelo próprio usuário.

Objetivos Específicos:

Dominar os conceitos dos três modelos clássicos de RI (booleano, vetorial e probabilístico), assim como as técnicas frequentemente utilizadas para a construção de uma máquina de busca como indexação de documentos usando índices invertidos e o pré-processamento de documentos (análise léxica, eliminação de *stopwords*, *stemming* e seleção de palavras-chave).

5. PROGRAMA

1. Introdução à Recuperação da Informação e modelo booleano

2. Dicionário e lista de *postings*: conceitos de documento, palavra e termo.

3. Indexação de documentos: termos, *stopwords*, *stemming*, Thesauri

4. Compressão de índices

5. Peso de termos

6. Modelo Vetorial

7. Avaliação de sistemas de recuperação de informação

8. Realimentação de relevantes e expansão de consultas

9. Recuperação em documentos semi estruturados (XML)

10. Modelo probabilístico

11. Classificação de documentos

12. Agrupamento de documentos

13. Redução de dimensionalidade

14. Web: busca, *crawling*, indexação, análise de *links*

15. Extração da informação

16. Introdução à recuperação de imagens baseada em conteúdo

6. METODOLOGIA

O curso será ministrado através de aulas expositivas sobre o tema, às segundas-feiras, de 19:00 até 20:40; e quintas-feiras, de 20:50 até 22:30. Para a exposição, serão usados slides, disponibilizados em meio virtual, em conjunto com a exposição oral do professor. A apresentação será complementada, sempre que necessário, com anotações e demonstrações no quadro da sala. Serão ao todo dezessete terças-feiras e dezoito quintas-feiras, totalizando 66 horas-aula presenciais. As 6 horas faltantes serão contabilizadas por meio da realização de atividades extra-classe distribuídas ao longo do semestre.

Para as aulas prática e o desenvolvimento de trabalhos, será utilizada a linguagem Python.

Adicionalmente, será criada uma classe virtual para a disciplina no Microsoft *Teams*, na qual os alunos regularmente matriculados na disciplina serão cadastrados pelo docente responsável. Nessa classe virtual, serão disponibilizados conteúdos da disciplina tais como slides, diagramas, exercícios resolvidos, anotações e código SQL executado nas aulas. Essa classe virtual também servirá como repositório para receber trabalhos dos estudantes e como ferramenta de comunicação extra classe. Estudantes também poderão solicitar ingresso na classe através do enlace : https://teams.microsoft.com/l/team/19%3apWP-tWRfXC3sYHhxYRIYVbsKkVvDfXxbPsS-yr14O7M1%40thread.tacv2/conversations?groupId=4de8c191-1c25-4bdc-a3db-19bc1f6f6727&tenantId=cd5e6d23-cb99-4189-88ab-1a9021a0c451 .

Semana	Módulo	Atividades Presenciais	Carga Horária Presencial	Data e Horário de Atividades Presenciais	Atividades extraclasse	Carga Horária Atividades extraclasse
31/07/2023	Início Semestre	-	-	-	-	

1	Introdução	Apresentação da disciplina	2 horas-aula	01/08/2023 19:00		
	Introdução	Aula expositiva: Introdução à Organização e Recuperação da Informação	2 horas-aula	03/08/2023 20:50		
2	Indexação	Aula prática: Introdução à programação em Python. Entrada e Saída, condicionais	2 horas-aula	08/08/2023 19:00		
	Indexação	Aula expositiva: Problemas gerais de Recuperação da Informação. Indexação	2 horas-aula	10/08/2023 20:50		
3	Modelos Clássicos	Aula expositiva: Modelo Booleano	2 horas-aula	17/08/2023 20:50		
4	Modelos Clássicos	Aula prática: Manipulação de strings	2 horas-aula	22/08/2023 19:00		
	Modelos Clássicos	Aula expositiva: Ponderação de termos: TF, IDF, TF-IDF	2 horas-aula	24/08/2023 20:50		
5	Modelos Clássicos	Aula prática: Dicionários para indexação	2 horas-aula	29/08/2023 19:00		
6	Modelos Clássicos	Aula expositiva: Ponderação de termos: TF, IDF, TF-IDF	2 horas-aula	04/09/2023		
	Modelos Clássicos	Aula prática: Processamento de Língua Natural	2 horas-aula	05/09/2023 19:00		
7	Modelos Clássicos	Aula prática: Processamento de Língua Natural	2 horas-aula	12/09/2023 19:00		
	Modelos Clássicos	Aula expositiva: Modelo vetorial	2 horas-aula	14/09/2023 20:50		

8	Modelos Clássicos	Aula prática: Manipulação de vetores: computação numérica	2 horas-aula	19/09/2023 19:00	Estudo dirigido para desenvolvimento do Trabalho 1	2 hora-aula
	Modelos Clássicos	Aula expositiva: Modelo Probabilístico	2 horas-aula	21/09/2023 20:50		
9	Modelos Clássicos	Aula prática:Manipulação de vetores: computação numérica e geração de gráficos	2 horas-aula	26/09/2023 19:00		
	Modelos Clássicos	Aula expositiva: Modelo Probabilístico	2 horas-aula	28/09/2023 20:50		
10	Modelos Clássicos	Casamento de padrão com expressões regulares	2 horas-aula	03/10/2023 19:00		
	Avaliação da Recuperação	Aula expositiva: Avaliação da Recuperação - fundamentos	2 horas-aula	05/09/2023 20:50		
11		Prova 1	2 horas-aula	10/10/2023 19:00		
12	Modelos Clássicos	Coleta automática de informações na WEB	2 horas-aula	17/10/2023 19:00	Estudo para desenvolvimento do Trabalho 2	2 hora-aula
	Avaliação da Recuperação	Aula expositiva: Avaliação da Recuperação - precisão e revocação	2 horas-aula	19/10/2023 20:50		
13	Web Scraping	Coleta automática de informações na WEB	2 horas-aula	24/10/2023 19:00		
	Realimentação de Relevância	Aula expositiva: Realimentação de relevância	2 horas-aula	26/10/2023 20:50		
14	Web Scraping	Coleta automática de informações na WEB	2 horas-aula	31/10/2023 19:00	Estudo para desenvolvimento do Trabalho 3	2 horas-aula

	Web Scraping	Aula prática: Implementação de coleta automática de informações na WEB	2 horas-aula	03/11/2023		2 horas-aula
15	Classificação de Documentos	Aula prática: técnicas para agrupamento	2 horas-aula	07/11/2023 19:00		
	Realimentação de Relevância	Aula expositiva: Realimentação de relevância - análise local	2 horas-aula	09/11/2023 20:50		
16	Page Rank	Aula expositiva: Ranqueamento de páginas WEB. Page Rank	2 horas-aula	14/11/2023 20:50 (reposição)		
	Classificação de Documentos	Aula expositiva: Classificação de Documentos	2 horas-aula	16/11/2023 20:50		
17	Redução de Dimensionalidade	Aula prática: redução de dimensionalidade	2 horas-aula	21/11/2023 19:00		
	Classificação de Documentos	Aula expositiva: Classificação de documentos - Algoritmo de Rocchio, classificadores de Bayes ingênuos, Ensemble	2 horas-aula	23/11/2023 20:50		
18		Prova 2	2 horas-aula	28/11/2023 19:00		
		Prova de Reposição	2 horas-aula	30/11/2023 20:50		
02/12/2023	Termino do semestre letivo		total de horas-aula presenciais: 66	-	-	total de horas-aula de atividades extraclasse: 6

Carga Horária Total (presencial + atividades extraclasse):	72 horas-aula

B) **Atendimento ao discente:** O atendimento aos discentes será realizado na sala do docente (1B150), às terças-feiras e quintas-feiras de 18:00 às 19:00. Horários adicionais podem ser disponibilizados através de agendamento prévio por email ao docente. Também é possível realizar atendimento remoto por meio da plataforma *Teams*.

7. AVALIAÇÃO

A avaliação da disciplina será feita através de duas provas (P1 e P2) e três trabalhos individuais. As provas serão feitas de forma presencial no horário e sala de aula da disciplina. Cada uma das provas valerá 30 pontos.

ATIVIDADE AVALIATIVA DE RECUPERAÇÃO

De acordo com o Art. 141 das Normas de Graduação (Res. CONDIR Nº 46/2022), Haverá também uma prova de recuperação que poderá recuperar até 50 pontos das provas P1 e P2. A prova de recuperação contemplará todo o conteúdo da disciplina e sua nota substituirá a soma de notas da P1 e P2. Ainda, de acordo com o Art. 141, somente fará jus ao direito de realizar a avaliação de recuperação substitutiva o(a) discente que não obtiver o rendimento mínimo de aprovação (60 pontos) e que possuir no mínimo 75% de frequência na disciplina. A atividade avaliativa de recuperação valerá 50 pontos e substituirá, caso assim se mostre mais vantajoso ao discente, a soma de notas da P1 e P2. Ressalta-se assim que, os discentes que recorrerem à atividade avaliativa de recuperação terão nota final máxima de 90 pontos na disciplina.

CRONOGRAMA DAS ATIVIDADES AVALIATIVAS

Nro	Data	Hora	Descrição	Pontos
1	10/10/2023	19:00 – 20:40	Prova 1	30
2	21/09/2023		Trabalho de implementação 1 - Modelos de Recuperação de Informação - Modelo Booleano. Entrega: 09/10/2023	10
3	10/10/2023		Trabalho de implementação 2 - Modelos de Recuperação de Informação - Modelo Vetorial. Entrega: 24/10/2023	15
4	31/10/2023		Trabalho de implementação 3 - Metodologia de avaliação da Recuperação. Entrega: 15/11/2023	15
5	28/11/2023	19:00 – 20:40	Prova 2	30

Recuperação	30/11/2023	20:50 – 22:30	Atividade Avaliativa de Recuperação (Art. 141 NG)	50 (substitui a soma de notas P1 e P2)
TOTAL:				**100**

8. **BIBLIOGRAFIA**

Básica

BAEZA-YATES, R.; RIBEIRO NETO, B. Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca [S. l]: Bookman, 2013.

BAEZA-YATES, R.; RIBEIRO NETO, B. Modern information retrieval. 2. ed. São Paulo: Addison-Welsey, 2011.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. An introduction to information retrieval. Cambridge: Cambridge University Press, 2009. Disponível em https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Complementar

CRESTANI, F.; PASI, G. Soft computing in information retrieval**:** techniques and_applications. New York: Springer Verlag, 2010.

CROFT, B.; METZLER, D.; STROHMAN, T. Search engines**:** information retrieval in_practice. São Paulo: Addison Wesley, 2009.

FRAKES, W. B.; BAEZA-YATES, R. Information retrieval & data structures. New Jersey: Prentice Hall, 1992.

MOENS, M. F. Information extraction: algorithms and prospects in a retrieval context. New York: Springer Verlag, 2006.

SUMMERFIELD, M. Programação em Python 3. Alta Books, 2013.

MELO, W. Introdução ao Universo da Programação com Python, 2021. Disponível em *https://wendelmelo.net/book*

9. **APROVAÇÃO**

Aprovado em reunião do Colegiado realizada em: ____/____/_____

Coordenação do Curso de Graduação: _____

Documento assinado eletronicamente por **Wendel Alexandre Xavier de Melo**, **Professor(a) do Magistério Superior**, em 27/09/2023, às 21:55, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4729536** e o código CRC **B430A508**.