Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação

PROGRAMA DE DISCIPLINA

ANO: 2020
DISCIPLINA: TECC: RECUPERAÇÃO. DE INFORMAÇÃO – MÁQUINAS DE BUSCA NA WEB
PROFESSOR: BERTHIER RIBEIRO DE ARAÚJO NETO
CURSO:

CÓDIGO: DCC831
CLASSIFICAÇÃO: OP
CRÉDITOS: 04
CARGA HORÁRIA: **TEÓRICA:** 060 **horas**
PRÁTICA: 000 **horas**
TOTAL: 060 **horas**
PRÉ-REQUISITO: O estudante deve estar familiarizado com programação, familiaridade com C++ é desejável.
PERÍODO: optativa

EMENTA: Ementa variável, focalizando tópicos em Sistemas de Informação.

A – OBJETIVO

O objetivo principal da disciplina é estudar o projeto e a implementação de uma máquina de busca básica para a Web. Ao final do curso o aluno deverá conhecer e ter implementado uma versão básica dos principais componentes de uma máquina de busca, a saber: (a) o coletor de páginas Web, (b) o indexador de páginas Web e (c) o processador de consultas responsável por ordenar os documentos da coleção com relação a uma consulta dos usuários.

Cada aluno terá que construir sua própria máquina de busca. Adicionalmente, cada aluno deverá avaliar os resultados retornados por esta máquina de busca. Que seja, ao final do curso o alundo terá construído uma máquina de busca e terá realizado uma avaliação preliminar da mesma, tendo uma compreensão geral de como uma máquina de busca funciona.

B – PROGRAMA

1. Coleta: arquitetura de um coletor de páginas web, principais problemas envolvidos.
2. Indexação: arquivos invertidos e listas invertidas, compressão de textos, compressão de listas invertidas.
3. Modelagem: modelos de RI, modelos clássicos de RI (booleano, vetorial e probabilístico), redes de inferência, aprendizado de máquina, "clicks" dos usuários.
4. Avaliação dos resultados: precisão e revocação, coleções de referência, DCG, teste A/B.

C – AVALIAÇÃO

A avaliação será como se segue:

1. Prova I – 30 pontos
2. Prova II – 30 pontos
3. Trabalhos práticos (40 pontos):
    a. Coletor --- 10 pontos
    b. Indexador –-- 10 pontos
    c. Processador de Consultas --- 10 pontos
    d. Avaliação dos resultados --- 10 pontos

D – BIBLIOGRAFIA

RICARDO BAEZA-YATES, BERTHIER RIBEIRO-NETO, **Modern Information Retrieval**, Pearson, 2011.

IAN H. WITTEN, ALISTAIR MOFFAT, THIMOTHY C. BELL, **Managing Gigabytes: Compressing and Indexing Documents and Images**, Morgan Kaufmann Publishers, 1999, 2nd. Ed.

CHRISTOPHER MANNING, PRABHAKAR RAGHAVAN, HINRICH SCHUTZE, **Introduction to Information Retrieval,** Cambridge Press, 2008.

Artigos técnicos especializados que serão fornecidos e discutidos em sala.