Projeto Deep Learning UFG / EasySearch

Pedro Vítor Quinta de Castro



Sobre o EasySearch

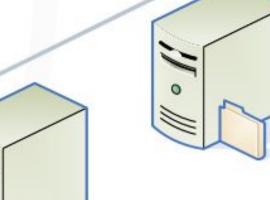
EasySearch

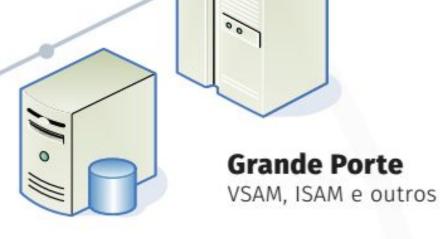
- Indexação de múltiplas plataformas
- Acesso a diversas fontes de dados
- 'Syndication' dos resultados
- Alta disponibilidade e balanceamento de carga
- Navegação pelos resultados
- Categorização
- Linguística





Arquivos MS Word, MS Excel, PDF entre outros





Bases de Dados Relacionais

MS SQL Server, ORACLE, Sybase, DB2 e outros

Bases de Gestão de Documentos

MS CMS, Documenntun

Base de e-mails

MS Exchange, Lotus Notes, Novell Groupwise entre outros

Weblogic, Websphere entre outros



Problema

Correspondência entre termo de busca e resultados é puramente textual Sem conceitos de semântica . Categorização dos documentos é feita manualmente por

- taxonomia ou facetas obtidas a partir de metadados dos documentos
- Similaridade entre documentos somente por TF-IDF



Soluções Propostas

- LDA (Latent Dirichlet Allocation)
 Modelagem e Extração de Tópicos
 Clusterização de Documentos
 Doc2vec
 Similaridade entre documentos
 Substituição do TF-IDF para recurso de MoreLikeThis

