UNIVERSIDADE DE BRASÍLIA
Faculdade do Gama

Sistemas de Banco de Dados 2

Tecnologias de Banco de Dados (TI-BD)

Banco de Dados Textuais

Victor Jorge da Silva Gonçalves – 18/0055241

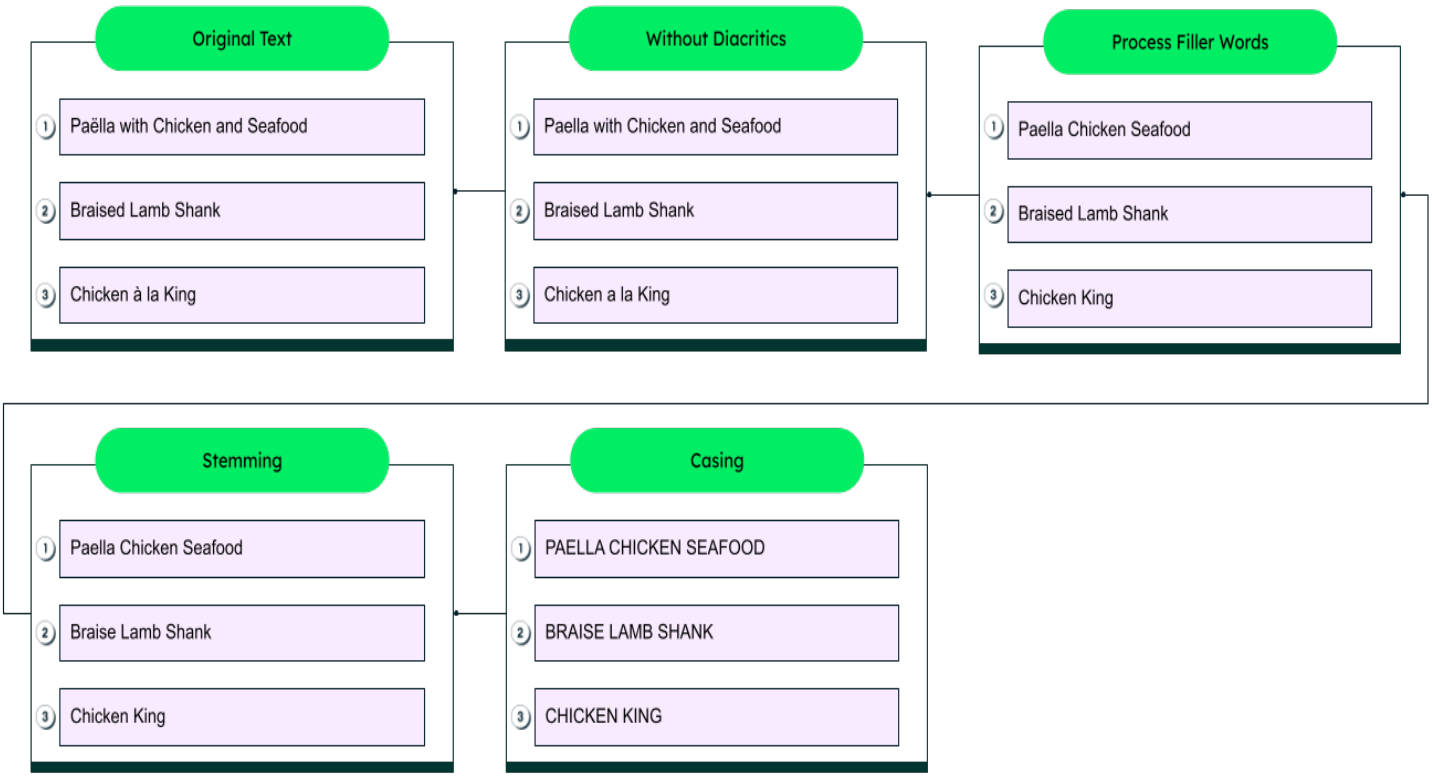
Brasília, DF
2023

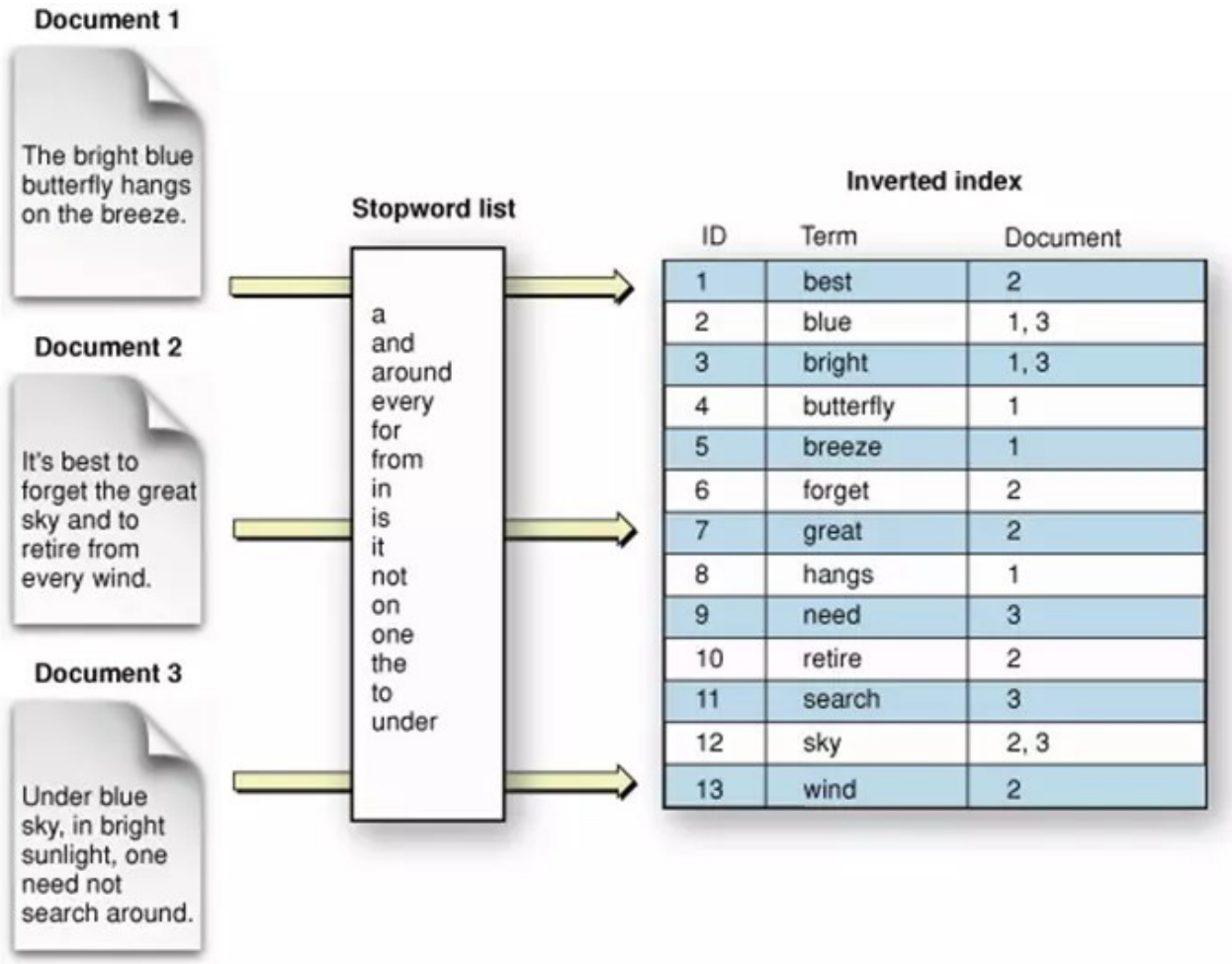
Bancos de Dados Textuais – Definição

No contexto de sistemas gerenciadores de bancos de dados, um banco de dados textual pode ser definido como um sistema que persiste volumes de coleção de textos, normalmente grandes, e que também provê formas adequadas, rápidas e precisas para recuperação desses dados textuais. Os bancos de dados textuais possuem majoritariamente dois grandes objetivos, sendo eles: fácil acesso ao conteúdo armazenado (busca eficiente) e armazenamento de grandes volumes de dados.

Diferentemente de sistemas gerenciadores de bancos de dados convencionais, notavelmente relacionais, que armazenam suas informações em pedaços (***chunks***) de tamanho fixo, e de forma atômica, os bancos de dados textuais possuem a necessidade de persistir dados quem não respeitam regras de atomicidades claras, e que por consequência podem vir a possuir dados de tamanhos variados para cada registro persistido. Um problema também relacionado é o armazenamento de dados multimídia, que possuem características análogas ao armazenamento de texto bruto.

Problema inerente ao projeto de bancos de dados textuais é a busca de dados persistidos. A maneira mais simples de se buscar por texto em uma base de dados seria a utilização de um simples mecanismo de *string pattern match* onde o usuário informa a palavra a ser pesquisada, e o sistema retorna as posições dos textos nos quais aquela palavra foi encontrada. Porém, essa abordagem é pouco efetiva e possui diversas falhas, onde, em casos de busca mais complexos, como a composição de duas ou mais palavras, a busca pode se mostrar não funcional. Existem diversas outras maneiras de se buscar dados em bases de dados textuais, tais como o uso de expressões regulares, também conhecidas como ***RegEx*** e algorítimos similares; porém, o método mais comum, isto é, visto em tecnologias reais, é a busca indexada. Na busca indexada, o texto é decomposto em índices, que são armazenados em estruturas de dados que permitem a utilização de algoritmos eficientes para busca e recuperação de dados.


Original Text
1 Paëlla with Chicken and Seafood
2 Braised Lamb Shank
3 Chicken à la King
Without Diacritics
1 Paella with Chicken and Seafood
2 Braised Lamb Shank
3 Chicken a la King
Process Filler Words
1 Paella Chicken Seafood
2 Braised Lamb Shank
3 Chicken King
Stemming
1 Paella Chicken Seafood
2 Braise Lamb Shank
3 Chicken King
Casing
1 PAELLA CHICKEN SEAFOOD
2 BRAISE LAMB SHANK
3 CHICKEN KING


Document 1
The bright blue butterfly hangs on the breeze.
Document 2
It's best to forget the great sky and to retire from every wind.
Document 3
Under blue sky, in bright sunlight, one need not search around.
Stopword list
a
and
around
every
for
from
in
is
it
not
on
one
the
to
under
Inverted index
ID Term Document
1 best 2
2 blue 1, 3
3 bright 1, 3
4 butterfly 1
5 breeze 1
6 forget 2
7 great 2
8 hangs 1
9 need 3
10 retire 2
11 search 3
12 sky 2, 3
13 wind 2

Objetivos da tecnologia

Como visto na definição, sistemas de bancos de dados textuais possuem como objetivo o armazenamento de grandes volumes de dados de texto, geralmente bruto, e consequentemente a habilidade de recuperação de textos de forma rápida, precisa e adequada. Existem aplicações onde se existe a necessidade de se recuperar um registro do banco de dados utilizando apenas frações de seu conteúdo, isto é, realizando o casamento ou ***pattern matching*** de um trecho específico de seu conteúdo com o registro como um todo. Aplicações como o Google Search Engine, buscadores de um site pessoal, por exemplo, são beneficiários das capacidades de uma tecnologia que gerencia e armazena textos. O usuário não deveria precisar saber digitar o título de um artigo de forma completa e concisa para poder achá-lo em um *site blog*.

Vantagens de bancos de dados textuais

Como já explicado, sistemas de bancos de dados textuais visam resolver um problema de forma coerente e performática onde outros sistemas gerenciadores de bancos de dados não conseguem extrair boa performance do equipamento eletrônico. Para bases de dados onde existe grande volume de texto inserido, e/ou existe a necessidade de se buscar por registros utilizando pattern matching com frações de seu conteúdo, bancos de dados textuais apresentam vantagem em sua performance e uso. Casos onde uma consulta feita em um banco de dados relacional poderia não trazer dado algum, bancos de dados textuais conseguiriam trazer com precisão um conjunto de dados que satisfaçam as condições da busca.

Desvantagens de bancos de dados textuais

A principal desvantagem da utilização de bancos de dados textuais é o ***overhead*** do uso de suas estruturas internas para indexação dos dados. Em casos onde é necessário inserir grandes volumes de dados, e não há a necessidade de se realizar consultas complexas sobre esses dados, um banco de dados textual pode se mostrar ineficiente.

Exemplos de Uso

Um exemplo de uso de tecnologia de banco de dados textual é na empresa na qual trabalho, onde existe um grande fluxo de entrada de documentos de diversos clientes, e por consequência, um grande volume de manipulação desses dados. A tecnologia utilizada é o ***ElasticSearch***, que combina o uso de diversas estratégias de armazenamento, sendo a de armazenamento com índices para consultas indexadas. A utilização dessa tecnologia acontece acompanhado do uso de um banco de dados relacional, no caso, o ***Postgresql***. Tal combinação é feita de forma que o banco de dados relacional se encarregue de manter em disco os dados de forma completa e pouco mutável, enquanto o ***ElasticSearch*** se encarrega de manter o conteúdo atualizado e de fácil acesso ao usuário, para que o mesmo consiga pesquisar com eficiência pelo dado.

Referencia Bibliografica

IAN H. Witten, ALISTAIR Moffat, & TIMOTHY C. Bell. Managing Gigabytes. 2ª Edição. Editora Spring, 1999.

GONZALO, Navaro. Text Databases. Dept. of Computer Science, University of Chile.