Apresentação

MINERAÇÃO DE DADOS E PROCESSAMENTO DE LINGUAGEM NATURAL

ADAUTO BENEVIDES, DURVAL JUNIOR, HIGO ALVES E LUIS FERNANDO

Mineração de texto

Figura 11 – Etapas do processo de extração de conhecimento em textos.

Coleta

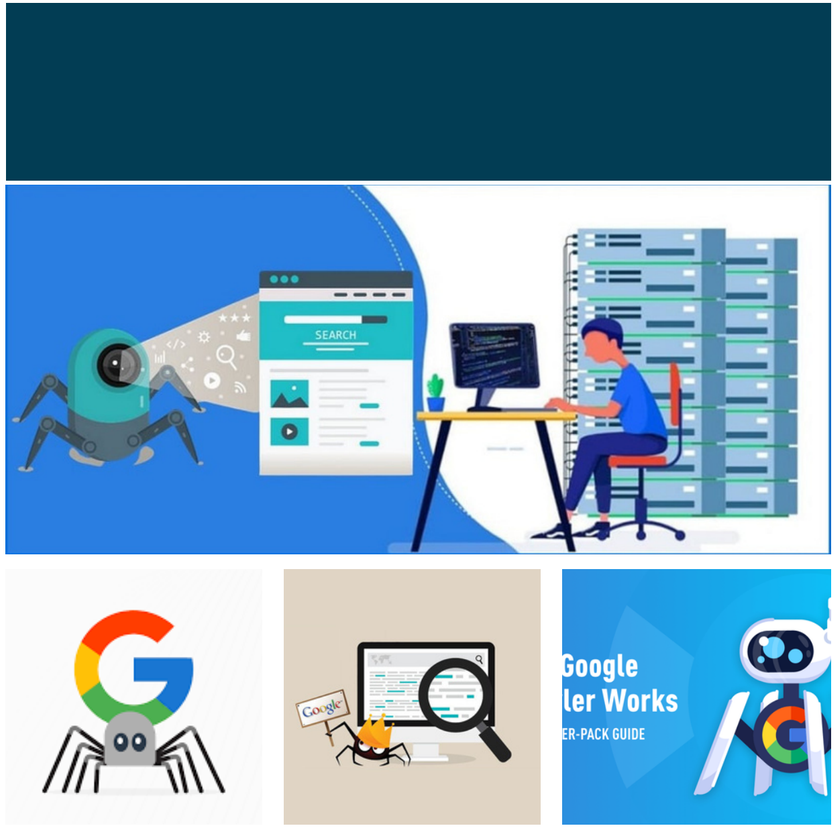
A coleta é a tarefa responsável por adquirir os elementos sob os quais serão utilizados no restante do trabalho, onde normalmente são coletados documentos. Na Web, a coleta pode ser realizada de forma automatizada através de crawlers.

Crawlers

Um crawler é um robô que visita páginas na Web e repassa as informações coletadas para outro componente responsável pela indexação dessas páginas.

Uma forma variante, e que pode ser mais interessante de se coletar documentos na Web, é através do uso de crawlers focados, que dispensam a utilização de grandes recursos de hardware

Esses crawlers costumam implementar alguma forma de classificação durante a varredura na Web de maneira a armazenar apenas os documentos considerados relevantes.



Coleta de Texto com a Turma

- Entrega de Papel e Caneta.
- Escrita de frase com uma palavra chave.
- Recolhimento dos papeis e entrega para a proxima etapa.

Pré-Processamento

- O pré-processamento é necessário para representar o texto numa forma mais estruturada, capaz de alimentar algoritmos de aprendizado de máquinas (GONÇALVES, 2006)

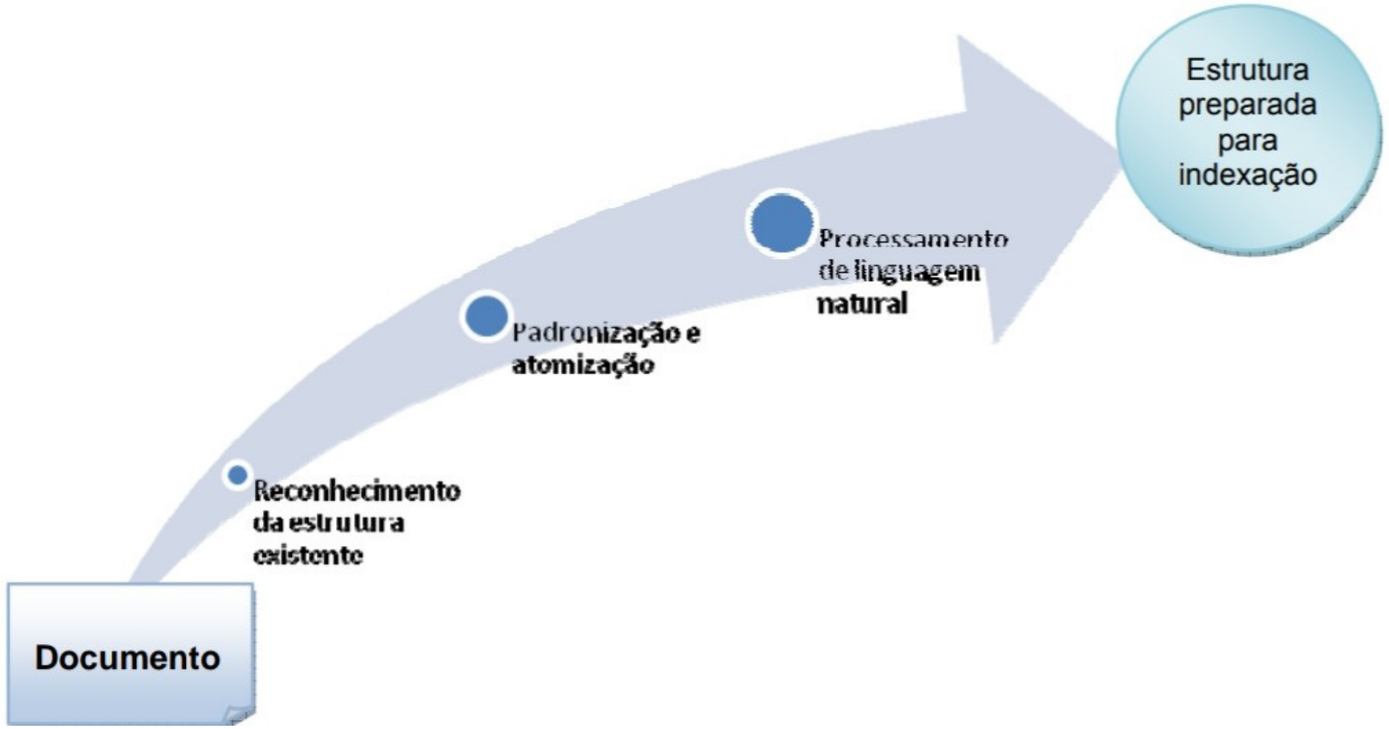
Pré-Processamento

- Nessa etapa são extraídas palavras de documentos, desconsiderando algumas stopwords cuja utilização está mais relacionada com a organização estrutural das sentenças e não têm poder discriminatório.

Pré-Processamento

- Um exemplo para a não utilização de stopwords, é a busca da frase “to be or not to be”, onde todas as palavras poderiam ser consideradas stopwords e nesse caso essa consulta não retornaria nenhum resultado, pois nenhuma das palavras seria indexada.

Pré-Processamento



Pré-Processamento

Tokenização

- O primeiro passo para o processamento de texto escrito é a atomização, amplamente conhecida na literatura como tokenizaçãO. Token é o nome que se dá aos termos extraídos dos textos, sejam eles palavras ou expressões compostas por mais de uma palavra.

Pré-Processamento

Tokenização

- Um texto é uma sequência de letras e delimitadores e, após o processo de tokenização, uma seqüência de tokens.

Pré-Processamento

Processamento de linguagem natural

- O texto é dividido sintaticamente usando informações de uma gramática formal e um léxico.
- faz uso de conceitos linguísticos como classes de palavras (substantivos verbos, adjetivos, etc.)
- estrutura das palavras (radical, vogal temática, desinências)
- formação das palavras (derivação prefixal, sufixal, parassintética, etc.)

Pré-Processamento

Identificação de colocações

- dentro de processamento de linguagem natural, há a Identificação de colocações onde, podemos encontrar nos textos conjuntos de palavras que trazem um significado adicional em relação às palavras que o compõem, analisadas em separado

- um conjunto de palavras pode até mesmo ter um significado completamente novo e adaptado, como no caso de expressões idiomáticas. Esses conjuntos especiais de palavras são conhecidos como colocações

Pré-Processamento

Classes Gramaticais

- Classes de palavras em Português:
  - Substantivo: usado para nomear entidades.
  - Adjetivo: usado para qualificar o substantivo.
  - Verbo: usado para expressar uma ação na sentença.

Pré-Processamento

Classes Gramaticais

- TBL (Tranformation Based Learning):
    - Aprendizado simbólico supervisionado e construção automática de regras lógicas.
    - Processo iterativo de incorporação de regras baseadas em template pré-apresentado.
    - Critérios de pontuação para melhorar as classificações a cada iteração.

Pré-Processamento

Lematização

- Tem como objetivo reduzir uma palavra à sua forma base e agrupar diferentes formas da mesma palavra. Por exemplo, os verbos no tempo passado são alterados para presente (por exemplo, “foi” é alterado para “vai”) e os sinônimos são unificados (por exemplo, “melhor” é alterado para “bom”), padronizando palavras com significado semelhante à sua raiz.

Pré-Processamento

Lematização

- A lematização resolve as palavras em sua forma de dicionário (conhecida como lema), para a qual requer dicionários detalhados nos quais o algoritmo pode pesquisar e vincular palavras aos lemas correspondentes.

- Por exemplo, as palavras “correr", "corre" e "correu" são todas formas da palavra "correr", portanto "correr" é o lema de todas as palavras anteriores.

Pré-Processamento

Análise de Discurso

- Elucidar relacionamentos entre sentenças em um texto.
- Dificuldades: identificação de anáforas, termos diferentes referindo-se à mesma entidade.
- Resolução de relações anafóricas é complexa e requer conhecimento semântico.

Pré-Processamento

Análise de Discurso

“Durante uma viagem, D. Pedro recebeu uma nova carta de Portugal que anulava a Assembléia Constituinte e exigia a volta imediata dele para a metrópole. Então, o Príncipe, próximo ao riacho do Ipiranga, levantou a espada e gritou: "Independência ou Morte!". Este fato ocorreu no dia 7 de setembro de 1822 e marcou a Independência do Brasil. No mês de dezembro de 1822, D. Pedro foi declarado imperador do Brasil."

Indexação

- Técnicas de indexação e recuperação de informações:
  - Necessidade de localizar e recuperar informações armazenadas.
  - Armazenamento conveniente das informações para facilitar a busca.
- Abordagens na representação de documentos:
  - Modelo "saco de palavras" - abordagem estatística, sem conhecimento linguístico.
  - Abordagem com conhecimento semântico dos textos.

Indexação

- Pré-processamento para indexação:
  - Processamento de linguagem natural ou não é usado para obter os termos.
  - Etapa de pré-processamento fornece os termos para a estrutura de dados de indexação.
- Estrutura de dados para indexação:
  - Índice invertido é amplamente utilizado.
  - Funciona como um índice remissivo de um livro, onde as palavras referenciam as páginas em que aparecem.

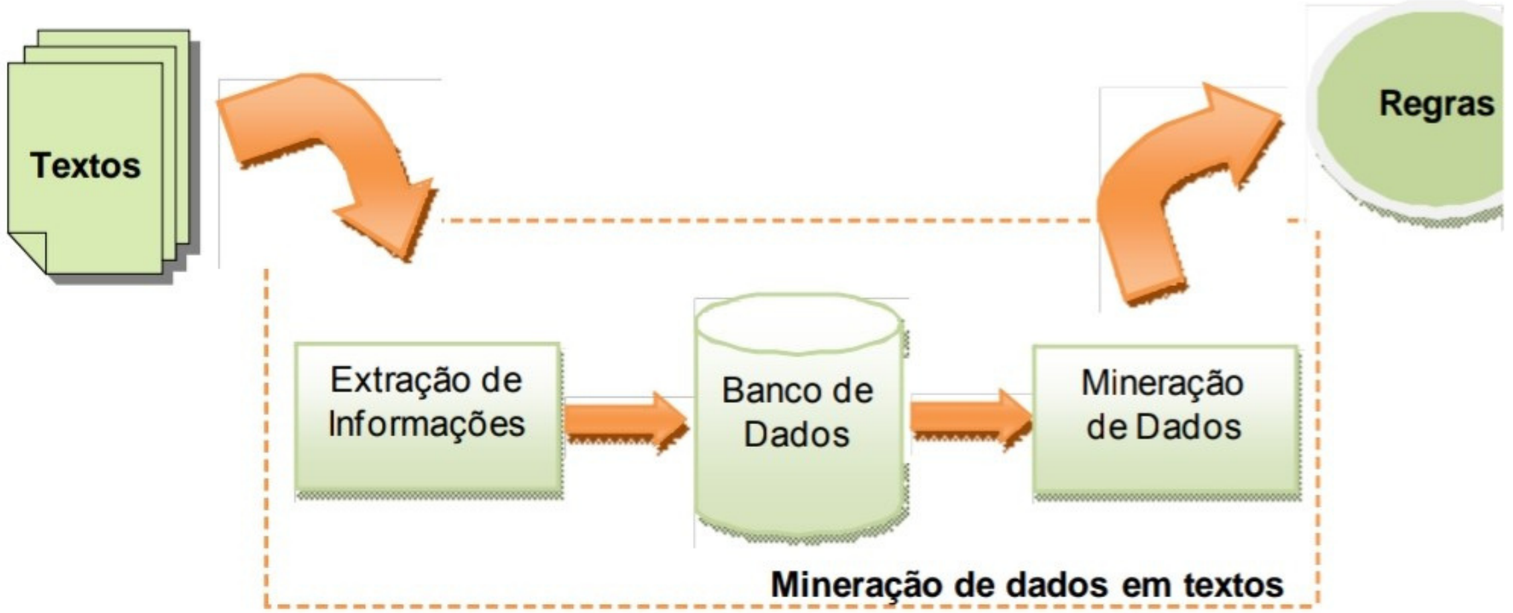
Indexação

- Procedimentos para construção de um índice invertido:
    - Colecionar os documentos a serem indexados.
    - Estabelecer as unidades dos documentos, geralmente palavras (tokens).
    - Tokenização: extrair as unidades (tokens) dos documentos.
    - Análise léxica dos documentos: eliminar caracteres inválidos, filtrar sequências de controle, correções ortográficas.
    - Restrições para entrada no índice: algumas palavras não significativas (preposições) ou muito frequentes (stopwords) são excluídas.
    - Uso de stoplist: lista de stopwords para evitar indexação.
    - Seleção criteriosa de stopwords reduz a dimensão do léxico.

Mineração

- A etapa de mineração visa à obtenção de algum tipo de conhecimento útil oriundo da coleção de textos.

Mineração


Textos
Regras
Extração de Informações
Banco de Dados
Mineração de Dados
Mineração de dados em textos

Análise

- Análise é a etapa em que o ser humano interpreta as informações obtidas pela fase de mineração. Pode tirar proveito de alguma forma de pós processamento que facilite a apresentação dos dados e visualização dos resultados com a possibilidade de navegação sobre as informações fornecidas.

Análise

- Essa fase é mais bem fundamentada sobre interfaces gráficas amigáveis, ferramentas para geração de relatórios, gráficos e ferramentas configuráveis de consulta.

Análise

- Sistemas de mineração de texto precisam prover os usuários com um grande leque de ferramentas para interação com os dados e interpretação dos resultados.

OBRIGADO!