



Aprendizado de Máquina Aplicado aos Negócios EAD0759

Prof. Antonio Geraldo **Vidal**

vidal@usp.br

Sala G175



automated data mining survey
responses com ter transcripts
qualatativ root cause
classificati insights
ad-hoc and is product
reviews serv it vor of the
customer dashboards consumer
trends ad-hoc analysis early warning

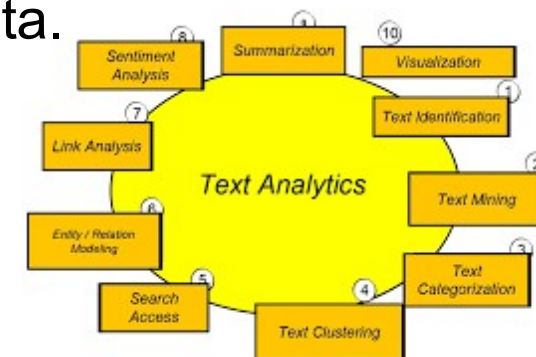


Técnicas para Análise de Textos (dados não-estruturados)



Motivações para Análise de Textos

- Grande disponibilidade de informações interessantes sobre muitos assuntos (pessoas, negócios, tecnologias, etc.) em uma infinidade de textos.
- Acesso rápido e fácil a informações presentes em textos largamente disponíveis na Internet.
- As técnicas utilizadas em análise de dados tem sido úteis no processo de análise de textos e de tomada de decisão.
- Há grande necessidade de se obter conhecimento (de forma objetiva) a partir dos textos disponíveis, sem ter que examiná-los ou lê-los de forma completa.
- Textos são normalmente classificados como dados não-estruturados.





O que é Análise de Textos



- Descoberta de conhecimento em textos, geralmente é relacionado a:
 - ☐ Recuperação de informações
 - ☐ Classificação de conteúdo ou documentos
 - ☐ Extração de resumos de textos
 - ☐ Descoberta de regras associativas
 - ☐ Descoberta de grupos ou padrões contidos nos textos
- Textos livres são denominados “dados não estruturados” em oposição aos dados estruturados, contidos em tabelas ou registros de bancos de dados.
- Na maioria das análises é necessária a combinação e análise de dados textuais e dados estruturados para se obter informações úteis e conhecimento.



Abrangência





Vantagens da Análise de Textos

Organizações e pessoas acumulam grandes volumes de informações textuais e frequentemente não conseguem gerenciá-las de forma eficiente, perdendo tempo e conhecimento.



Sources of Data



Técnicas de análise de dados podem ajudar a melhorar o desempenho dos negócios através da análise de informações textuais, oferecendo conhecimento novo e útil para a tomada de decisões.

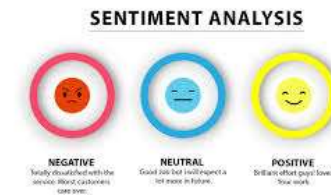


Aplicações da Análise de Textos

- A partir de textos resultantes de pesquisas, a análise de textos pode identificar padrões que sejam úteis para alcançar objetivos de negócios.
- A partir de textos disponibilizados pelas pessoas (clientes, fornecedores, concorrentes etc.), pode-se avaliar oportunidades e riscos.
- Por exemplo, a partir de textos de e-mails de clientes, pode-se utilizar um processo de classificação de textos para identificar os principais motivos de reclamações, elogios ou solicitações.
- Filtragem de mensagens não desejadas (*spam*) ou identificação de textos inadequados (proibidos ou falsos).
- Análise de textos também é útil para processar respostas automáticas para e-mails dos clientes, baseados em casos semelhantes e “*chat-bots*”.



Aplicações



- A partir de documentos sobre avaliação de pessoas, pode-se identificar aquelas que necessitam de cursos de atualização, aquelas que são líderes etc.;
- O processo de análise de textos também pode ser útil para automatizar análise de currículos, para facilitar a identificação do perfil de pessoas.
- Obtenção de informação relevante e estruturada (conhecimento) contida em grandes volumes de textos.
- Análise de sentimento em redes sociais e desenvolvimento de sistemas de recomendação, que envolvem tarefas de classificação.



Principais Categorias

- Análise de textos pode ser classificada em duas categorias principais:
 - **Análise de conteúdo:** “tenho um grande volume de textos e preciso saber algo sobre eles”, isto é **busca do conhecimento** que está contido neles;
 - **Pesquisa de conteúdo:** “há uma grande quantidade de textos e preciso saber o que posso encontrar neles”, isto é **busca de conteúdo**: na web, em documentos legais, em documentos médicos, em documentos técnicos, em estudos e pesquisas científicas etc.



Análise de Textos

- Em bases de dados textuais, também conhecidas como *corpus*, cada exemplar (observação) é tratado como um *documento*.
- Cada documento em um *corpus* pode assumir diferentes características em relação a:
 - ☐ Tamanho do texto (sequência de caracteres);
 - ☐ Tipo de conteúdo (assunto que aborda);
 - ☐ Língua na qual foi escrito;
 - ☐ Linguagem com a qual foi escrito (formal, coloquial, poética, etc.);
 - ☐ Fonte ou origem do texto.
- A **transformação** de um *corpus* em um *conjunto de dados* que possa ser submetido a procedimentos de análise de dados consiste em um processo que gera uma representação capaz de descrever cada documento em termos de suas características relevantes para o objetivo da análise.
- Uma das formas mais comuns e básicas de representar documentos consiste em usar uma lista de ocorrência de palavras ou termos.



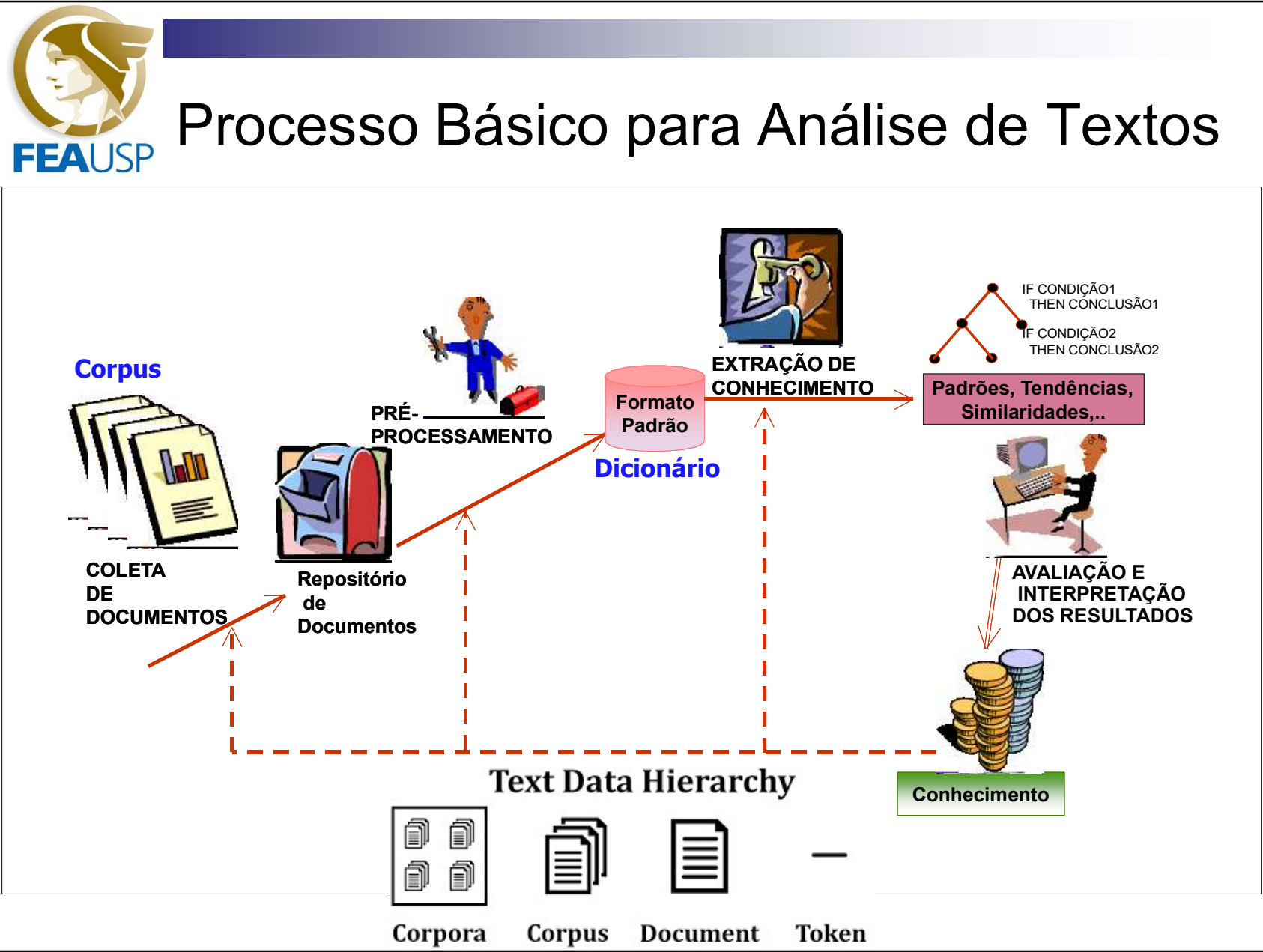
Análise de Textos

- A lista de todas as palavras ou **termos** que ocorrem em todos os **documentos** de um **corpus** (conjunto de documentos) pode ser nomeado de **dicionário**.
- Com base no dicionário e na frequência com que as palavras ou termos do dicionário aparecem nos documentos e no *corpus* é possível construir uma **representação** para o *corpus*.
- Essa representação padrão pode ser tratada como o **conjunto de dados** a ser analisado.
- Um **corpus** é definido por um conjunto de **D Documentos** e cada um dos documentos é definido por um conjunto de **T Termos**.



Análise de Textos

- Qualquer problema de análise de textos está dentro de determinado contexto, domínio ou área de aplicação.
- Informações e conhecimento inerentes a este domínio ou área de aplicação devem sempre ser considerados em todas as tomadas de decisão no processo de pré-processamento do **corpus**.
- A análise de textos **não está** diretamente relacionada à área de **Processamento de Linguagem Natural**, que é bem mais complexa e em geral envolve técnicas de *Deep Learning*.
- Após o pré-processamento o **corpus** assume uma representação que se distancia da linguagem natural, pois perde toda a construção gramatical e semântica complexa.
- Os textos são representados apenas por dados ou **termos** e suas frequências nos **documentos** que compõem o **corpus**.





Técnicas Envolvidas na Análise de Conteúdo

- Análise de textos pode envolver:
 - Técnicas linguísticas
 - Técnicas estatísticas comumente usadas em recuperação e análise de informação
 - Técnicas de aprendizagem de máquina
- As perguntas mais comuns a serem respondidas são:
 - Quais são as palavras mais frequentes?
 - Quais são as palavras mais raras ou inéditas?
 - Quais são as palavras que melhor definem o conteúdo dos documentos?
 - Quais são as frases mais importantes dos textos?
 - Quais são os grupos ou classes de textos existentes?



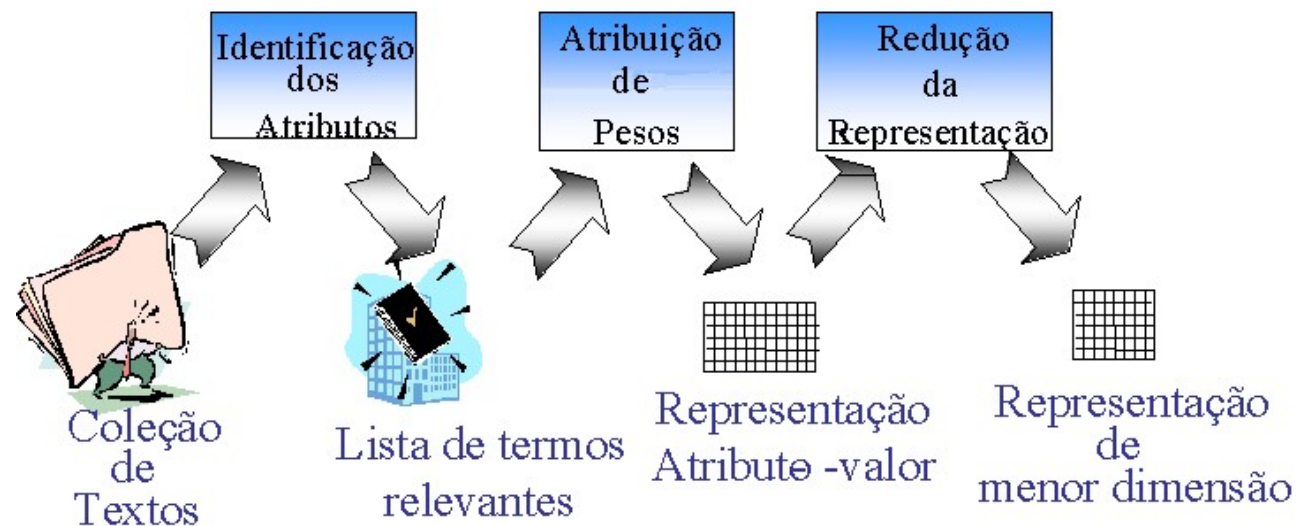
Desafios da Análise de Textos

- Avaliar a semelhança não é algo óbvio – qual é a diferença entre duas frases, termos ou palavras?
- Avaliar o resultado da análise não é fácil: qual é a resposta "certa"? (não há verdade absoluta)
- O texto pode conter termos que você não tenha previsto, por exemplo incorretos, estrangeiros, novos etc.
- O objetivo é diferente do objetivo da classificação: não é necessário modelar todos os dados, mas só considerar os mais relevantes para um determinado objetivo.
- As palavras podem ocultar o conteúdo semântico:
 - **Sinonímia**: uma palavra-chave **T** pode não aparecer em qualquer parte do documento, mesmo que o documento esteja intimamente relacionado com a palavra **T**, por exemplo, **conhecimento**;
 - **Polissemia**: a mesma palavra pode significar coisas diferentes em diferentes contextos, por exemplo, **mineração** na engenharia não tem o mesmo significado que **mineração** em análise de dados.



Preparação de Textos

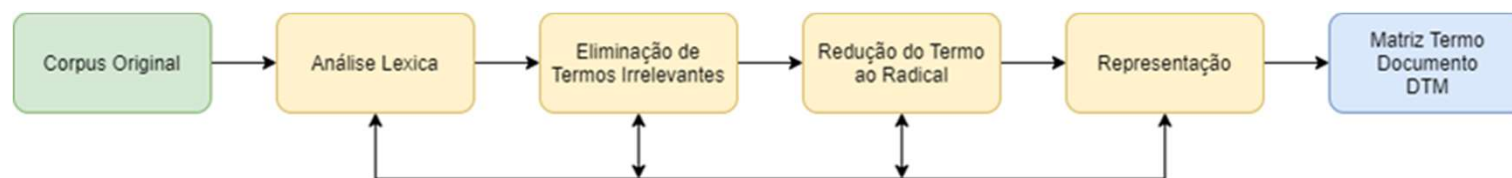
- Fase necessária para estruturação dos textos num formato compatível com as técnicas normalmente utilizadas em análise de dados (classificação, agrupamento e associação).





Matriz Termo Documento (DTM)

- Exige um processo de preparação dos dados.
 1. Análise Léxica (pontuação, capitalização, dígitos, acentos etc.)
 2. Eliminação de Termos Irrelevantes (*stopwords*)
 3. Redução do Termo ao seu Radical (*stemming*)
 4. *Redução de Sinônimos ou Termos Equivalentes no Contexto*
 5. Representação (vetorial ou matricial) do corpus
 6. Obtenção da Matriz Termo Documento ou *Document Term Matrix* - DTM





Identificação de Atributos Relevantes

- Consiste em identificar as palavras ou termos que são relevantes para caracterizar o contexto de cada documento.
- Pode envolver técnicas dependentes do idioma.
- Envolve a remoção de termos pouco significativos;
- Envolve a normalização de palavras para sua respectiva forma canônica (radical);
- Envolve o uso de um dicionário de termos do domínio da aplicação.
- Portanto, envolve muitas decisões do pesquisador ou analista.



Atribuição de Pesos

- Geralmente envolve o uso de medidas estatísticas baseadas na frequência dos termos nos documentos.
 - ☐ Booleano (0 ou 1, se o termo existe ou não existe no documento)
 - ☐ Frequência (de cada termo em cada documento)
 - ☐ tf (*term frequency*) x idf (*inverse document frequency*)
 - ☐ $tf \times idf$ (normalizado)
 - ☐ $\log tf \times idf$ (normalizado)
 - ☐ Baseado em entropia
 - ☐ E diversas outras.
- $tf \times idf$ ou TFIDF, é uma métrica estatística numérica que se destina a refletir a importância de uma palavra para um documento em uma coleção ou corpus. É frequentemente usada como um fator de ponderação em pesquisas de recuperação de informações, mineração de texto e modelagem do usuário.

Document 1		Document 2	
Term	Count	Term	Count
This	1	This	1
is	1	is	1
a	1	a	1
beautiful	2	beautiful	1
day	5	night	2



Frequência Inversa nos Documentos

- A **frequência inversa nos documentos**, ou simplesmente **idf** (do inglês, *inverse document frequency*), fornece essa relação da seguinte maneira:

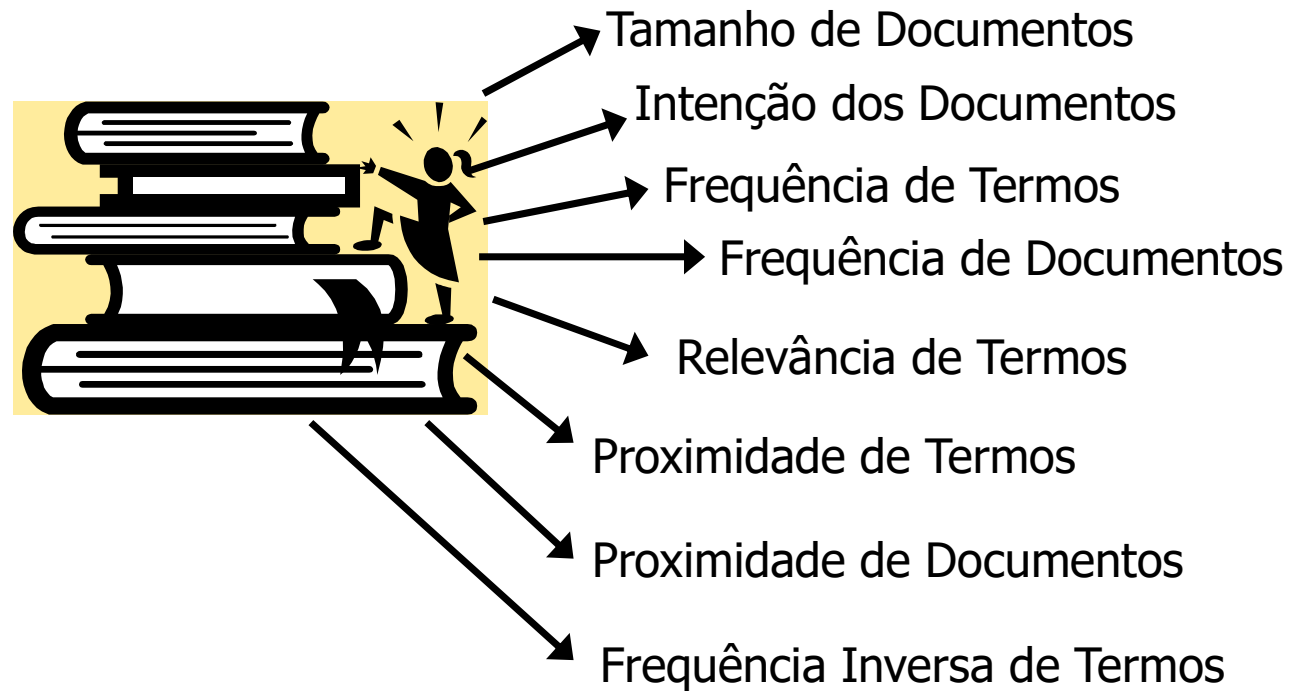
$$idf(te) = \log \left(\frac{n}{nt} \right)$$

- Sendo **n** o número total de documentos no *corpus*, e **nt** o número de documentos em que o termo **te** aparece.
- Essa medida prioriza termos que aparecem em poucos documentos no *corpus*, indicando que existem termos com frequência invertida maior que possuem poder de discriminação mais alto do que termos com frequência invertida menor.
- Ou seja, **termos raros discriminam documentos entre si, termos comuns não.**
- Justifica, portanto, o estabelecimento de uma medida que combine os dois fatores importantes: a frequência do termo dentro de um documento e a frequência inversa dos termos nos documentos do corpus.

$$tf - idf_{norm}(doc_i, te_j) = \frac{tf(doc_i, te_j) * idf(te_j)}{\sum_{j=1}^m tf(doc_i, te_k)}$$



Adicionando Dimensões Numéricas a Textos Não-estruturados





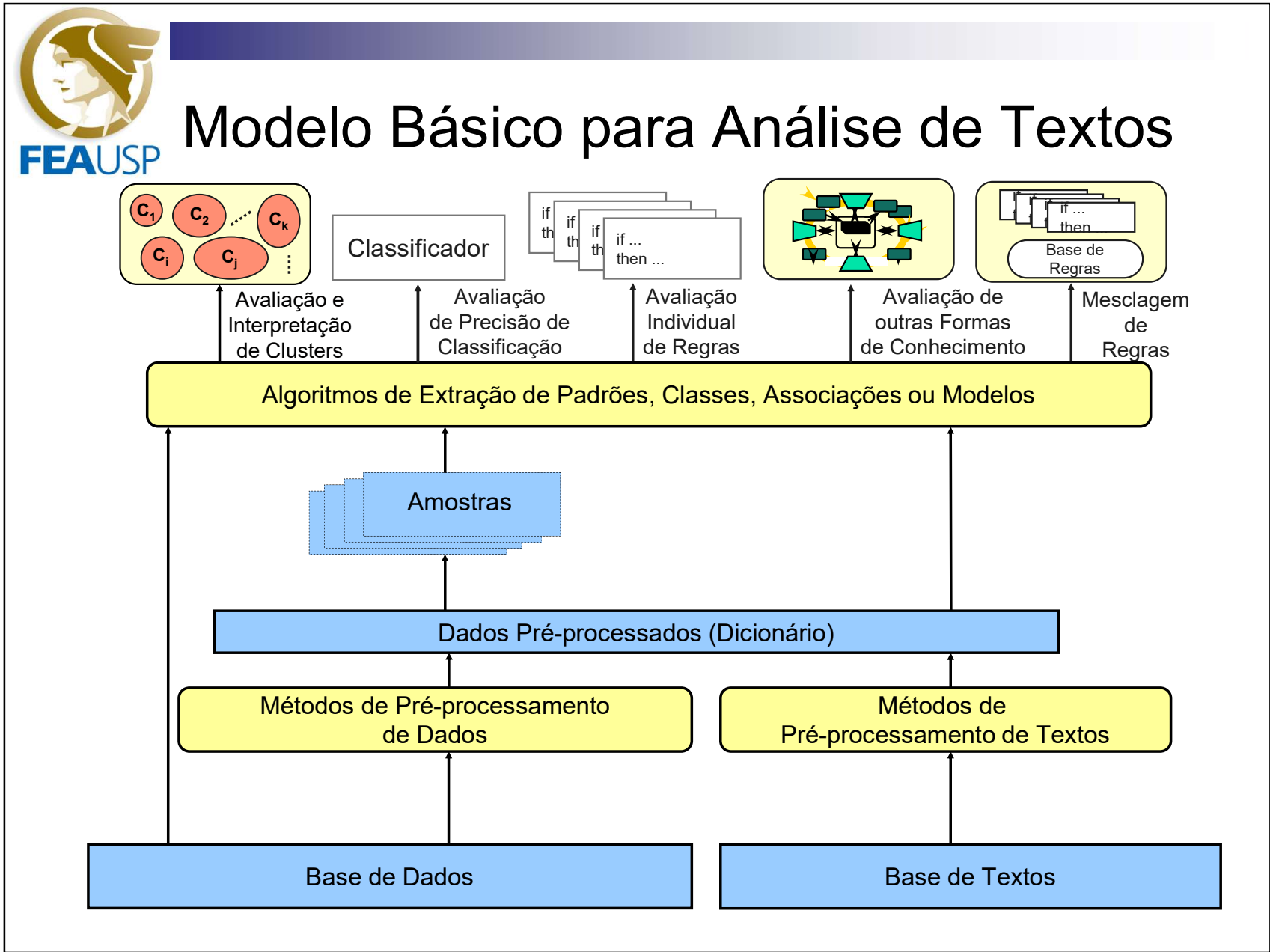
Representação do Corpus


- Os quatro tipos de representação vetorial dos documentos e matricial do *corpus* apresentados, permitem que os algoritmos para classificação e agrupamento que já estudamos possam ser aplicados à análise de textos.
 1. Matriz DTM *tf* binária
 2. Matriz DTM *tf* frequência
 3. Matriz DTM *tf-idf*
 4. Matriz DTM *tf-idf* normalizada



Redução da Representação

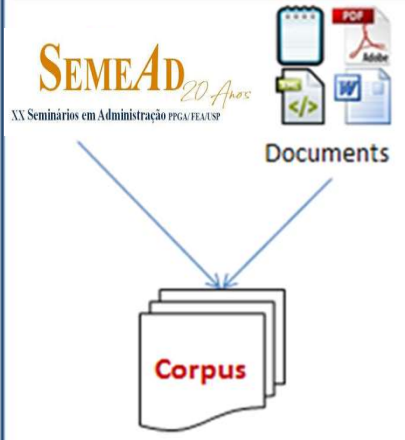
- Atividade que pode ser necessária tendo em vista a grande ou enorme quantidade de termos que podem ser identificados.
- Uma grande quantidade de termos pode exceder a capacidade de processamento dos métodos de descoberta de conhecimento tradicionais.
- Métodos de seleção de atributos
 - Identificam os termos mais importantes para compor uma nova representação e os demais são desconsiderados.
- Métodos de indução construtiva
 - Visam combinar termos para a construção de novos que possuam melhor poder preditivo que os originais.





Análise de Textos no R (Laboratório)

Step 1 – Data Assemble



Documents

Corpus

Step 2 – Data Processing

2A - Explore Corpus
2B - Convert text to lowercase
2C - Remove

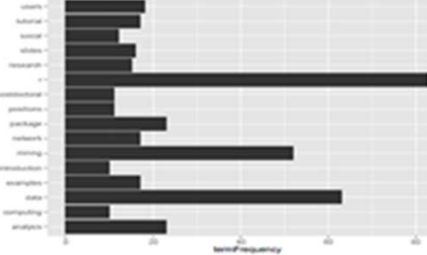
- a) Numbers(if required)
- b) Punctuations
- c) English stop words
- d) Own stop words(if required)
- e) Strip whitespace
- f) Stemming
- g) Sparse terms

2D - Create document term matrix


(Description about each step is on next slide)

Step 3 - Visualization

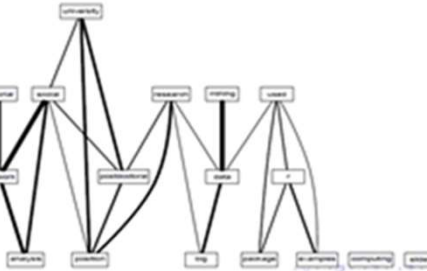
Frequency Items



Word Cloud

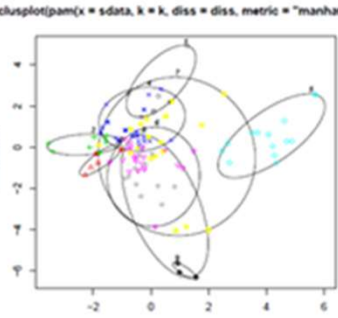



Correlation Plot



Step 4 – Run Model(s)

- ❖ Classification
- ❖ Sentiment Analysis
- ❖ Clustering



clusplot(pam(x = sdata, k = k, diss = diss, metric = "manhattan")

These two components explain 24.81 % of the point variability.



Matriz TD – Termo-Documento

- Forma mais comum de representação em análise de textos é a matriz TD - Termo-Documento:
 - Termo: geralmente uma única palavra, mas pode ser uma palavra-frase como "*data mining*" ou "inteligência artificial".
 - Documento: uma denominação genérica que significa uma coleção de textos a ser recuperada e compõe um corpus.
- O volume pode ser grande:
 - Termos são muitas vezes 50k ou mais.
 - Documentos podem estar na casa dos bilhões (na Web).
 - A matriz TD pode ser binária ou utilizar contagem de frequência de termos nos documentos.



Matriz TD - Exemplo

Documento	Database	SQL	Índice	Regressão	Linear
D1	24	21	9	0	3
D2	32	10	5	0	0
D3	12	16	5	0	0
D4	6	7	2	0	0
D5	43	31	20	0	0
D6	2	0	0	18	6
D7	0	0	1	32	0
D8	3	0	0	22	4
D9	1	0	0	34	25
D10	6	0	0	17	23

Cada documento é um vetor de frequência de termos.



Matriz TD

- Na matriz TD perde-se todo o conteúdo semântico dos textos.
- A lista de termos precisa ser construída com muito critério e cuidado:
 1. Nem todas as palavras são iguais, depende do contexto!
 2. Palavras que possuem o mesmo significado devem ser tratadas da mesma forma!
 3. Quais são as palavras relevantes para o objetivo da análise?
- Remover palavras sem significância (*stopwords*)
- Normalizar palavras decorrentes ou derivadas (*stemming*) e deixá-las em sua forma “canônica”.



Stop Words

- Muitas das palavras mais usadas são inúteis na recuperação e análise do texto - estas palavras são chamadas de palavras de parada ou **stopwords**.
 - Por exemplo, artigos e preposições: o, de, e, para,
 - Tipicamente, há cerca de 400 a 500 de tais palavras em cada idioma.
 - Para uma determinada aplicação, uma lista adicional de domínio específico de stopwords pode ser construída.
- Por que precisamos remover stopwords?
 - Reduzir o tamanho do arquivo de indexação (ou de dados).
 - Stopwords representam 20-30% do total de contagem de palavras.
 - Melhorar a eficiência
 - Stopwords não são úteis para a análise ou mineração de texto.
 - Stopwords têm sempre um grande número de acessos (frequência).



Palavras Derivadas e Similares

- Técnicas usadas para descobrir a raiz de palavras derivadas:
 - Palavra **uso**: usuário, usado, usando, usabilidade, etc.
 - Palavra **engenho**: engenharia, engenheiro, engenhoca, etc.
 - Palavra **análise**: analítico, analisar, analisado etc.
- Técnicas usadas para descobrir palavras similares:
 - Sinônimos (dado, atributo ou campo) (perda ou prejuízo)
 - Remoção de terminações (s, ente, ando, etc.)
 - Transformação de palavras (combinações ou equivalências).
- A combinação de palavras com a mesma raiz e a substituição de palavras similares pode reduzir a necessidade de indexação em 40 a 50%.



Distancia em Matrizes TD

- Para uma dada uma matriz termo-documento, podemos definir distâncias entre os documentos calculando a distância entre seus termos.
- Os elementos da matriz podem ser 0 ou 1 ou a frequência do termo (muitas vezes normalizado).
- Pode-se usar a distância euclidiana ou a distância cosseno:
 - Distância cosseno (dc) é o ângulo entre dois vetores de documentos sendo comparados;
 - Não é intuitiva, mas tem sido utilizada por gerar bons resultados.
 - Se os documentos são os mesmos, $dc = 1$, se nada tiverem em comum $dc = 0$.

$$d_c(D_i, D_j) = \frac{\sum_{k=1}^T d_{ik} d_{jk}}{\sqrt{\sum_{k=1}^T d_{ik}^2 \sum_{k=1}^T d_{jk}^2}}$$

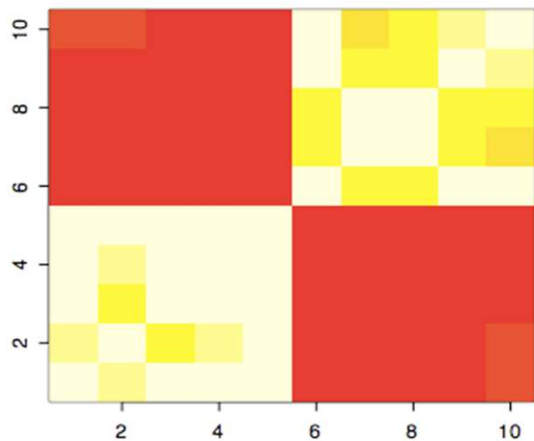


Distância entre Documentos pela Matriz TD

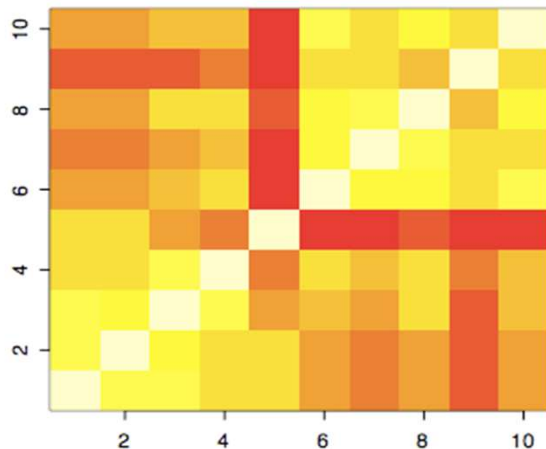
Documento	Banco de Dados	SQL	Índice	Regressão	Linear
D1	24	21	9	0	3
D2	32	10	5	0	0
D3	12	16	5	0	0
D4	6	7	2	0	0
D5	43	31	20	0	0
D6	2	0	0	18	6
D7	0	0	1	32	0
D8	3	0	0	22	4
D9	1	0	0	34	25
D10	6	0	0	17	23



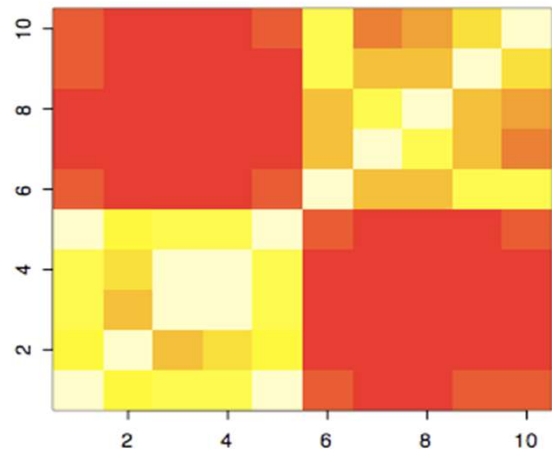
Distância entre Documentos pela Matriz TD



Distância Cosseno



Distância Euclidiana



Distância Euclidiana Escalada



Peso na Matriz TD

- Nem todos os termos são de igual importância para o objetivo da análise:

- Por exemplo, **Capital** pode ser menos importante do que **Social**.
- Se um termo ocorre com frequência em muitos documentos ele tem menor poder discriminatório.
- Uma maneira de corrigir esse problema é a frequência inversa do documento (**IDF**):

$$IDF = \log(N/n_j)$$

- Importância do Termo = Frequência do Termo (TF) x IDF
 - N_j = número de documentos contendo o termo
 - N = número total de documentos
- Um termo é "importante" se ele possuir uma elevada TF e / ou um elevado IDF.
- TF x IDF é uma medida muito utilizada para a importância do termo.

$$tf-idf_{norm}(doc_i, te_j) = \frac{tf(doc_i, te_j) * idf(te_j)}{\sum_{j=1}^m tf(doc_i, te_k)}$$



Consultas

- Uma consulta é uma representação das necessidades de informação do usuário
 - Normalmente uma lista de palavras ou
 - Uma simples pergunta em linguagem natural
- Uma vez obtida a matriz TD, consultas podem ser representadas como um vetor num mesmo espaço:
 - "Índice de Banco de Dados" = (1,0,1,0,0,0)
- Calcular a distância entre uma consulta e a versão TF x IDF TD
 - Retorna um vetor ordenado de documentos (ou respostas).



Análise Textual

- Uma vez que tivermos transformado os textos dos documentos em dados de uma matriz de representação adequada (TD, TDxIDF, ou LSI) podemos então analisá-los utilizando algoritmos específicos:
- Utilizar técnicas de análise de dados para:
 - Classificação de documentos:
 - Se for possível ter dados de treinamento para as classes, **com supervisão**.
 - Agrupamento de documentos:
 - Se não for possível ter dados de treinamento para as classes, **sem supervisão**.
 - Associação de documentos:
 - Regras de associação.



Classificação de Documentos

- Classificação automática de um enorme número de documentos textuais on-line (p.ex: páginas da Web, e-mails, etc.)
- Classificação de textos de clientes (p.ex: pedidos de informações, reclamações, requisições de assistência)
- Um típico problema de classificação de objetos:
 - Conjunto de treinamento: peritos humanos precisam gerar um conjunto de dados de treinamento;
 - Classificação: a ferramenta de TI descobre as regras ou modelo de classificação;
 - Aplicação: as regras descobertas podem ser aplicadas para classificar documentos novos ou desconhecidos.
- Técnicas Utilizadas
 - Regressão linear/logística, Naive Bayes (probabilidade) e Redes Neurais.
 - Árvores de decisão normalmente não são tão aplicáveis neste caso, devido à dimensão enorme e poucas interações.



Agrupamento de Documentos

- Também pode-se aplicar *clustering*, ou aprendizado não supervisionado em documentos.
- Agrupamento automático de documentos com base em seu conteúdo.
- Não necessita de conjuntos de treinamento ou taxonomias pré-determinadas.
- Principais etapas:
 - Pré-processamento: remover stopwords, palavras derivadas, extração de características, análise lexica, ...
 - Agrupamento hierarquico: calcular semelhanças entre os documentos aplicando algoritmos de agrupamento, ...
 - Analisar: revisar a árvore de grupos para o número desejado de níveis.
- Como em todas as aplicações de agrupamento, o sucesso é relativo.



Análise de Sentimento (Classificação)

- **A análise de sentimento** é o uso de processamento de análise de texto (mas não só) para sistematicamente identificar, extrair, quantificar e estudar estados afetivos e informações subjetivas.
- A análise de sentimento é amplamente aplicada à opiniões do cliente, como avaliações e respostas de pesquisa, mídias on-line e sociais, e textos de saúde para aplicações que vão desde marketing e atendimento ao cliente até a medicina clínica.
- Uma tarefa básica na análise do sentimento é classificar a *polaridade* de um determinado texto no documento, sentença, frase ou recurso/aspecto da entidade, isto é, **positiva**, **negativa** ou **neutra**.
- Na classificação de sentimento avançada, "além da polaridade", procura-se, por exemplo, estados emocionais como *prazer*, *raiva*, *nojo*, *tristeza*, *medo* e *surpresa*.



Conclusões

- A análise de textos pode ser muito útil para apoiar processos analíticos de tomada de decisão.
- Porém, a presença de um especialista é fundamental para obtenção de resultados mais expressivos.
- As pesquisas em análise de textos são relativamente recentes, e o interesse em sua realização tem sido cada vez maior.
- Modelos de redes neurais profundas (*Deep Learning*) tem sido criados com muito sucesso para a classificação e análise de conteúdo de textos, interpretação de linguagem natural e tradução.
- A análise de textos está compreendida na Web Análise ou *Web Analytics*.



Laboratório de Análise de Texto



- Exemplos
 - Textos de Artigos do SEMEAD
- Ferramentas
 - R Text Mining
 - R Studio



Referências

- Statistical Analysis & Data Mining Applications
Robert Nisbet, John Elder, Gary Miner – Elsevier, 2009
- Business Intelligence - Um Enfoque Gerencial para a Inteligência do Negócio
Efraim Turban, Ramesh Sharda, Jay E. Aronson, David King - Bookman, 2009
- Tecnologia da Informação para Gestão
Efraim Turban, Dorothy Leidner, Ephraim McLean, James Wetherbe - Bookman, 2010
- Redes Neurais Artificiais – para engenharia e ciências aplicadas
Ivan Nunes da Silva, Danilo Hernane Spatti e Rogério Andrade Flauzino – Editora Artliber, 2010
- Introdução ao Data Mining (Mineração de Dados)
Pang-Ning Tan, Michael Steinbach e Vipin Kumar – Editora Ciência Moderna, 2009
- Data Mining Concepts and Techniques – 3rd Edition
Jiawei Han, Micheline Kamber e Jian Pei – Morgan Kaufmann / Elsevier, 2012
- Data Warehouse - Como Construir o Data Warehouse
W.H. Inmon - Campus, 1997
- Documentação e Tutoriais do banco de dados Microsoft SQL Server 2008 R2
Microsoft
- Chakraborty, D. Introdução ao Processamento de Linguagem Natural (PNL)", PyImageSearch, P. Chugh, A. R. Gosthipaty, J. Haase, S. Huot, K. Kidriavsteva, R. Raha, e A. Thanki, eds., 2022, <https://pyimg.co/60xld>



A cartoon illustration showing three construction workers in hard hats and safety gear. One worker is standing on a large, unstable pile of numbers and symbols (0-9, %, \$, &, etc.), another is at the base, and a third is using a jackhammer on the side. The pile is shaped like a triangle and is precariously balanced. The background is a simple green patch of grass.



The logo for FEAUSP, featuring a circular emblem with a profile of a person's head and the text "FEAUSP" in blue capital letters.

Técnicas de Machine Learning Aplicadas a Negócios

Prof. Antonio Geraldo da Rocha Vidal