



Universidade de Brasília
Faculdade de Ciência da Informação - FCI

Text Mining na prática

Aplicação a um Serviço de
Atendimento ao Cliente

Dr. Marcelo Schiessl



Workshop de Informação, Dados e Tecnologia - WIDaT 2019

Agenda



- ↵ Introdução
- ↵ Dado Convencional X Textual
- ↵ Conceitos
- ↵ Aplicações
- ↵ Problema REAL – na prática
- ↵ Conclusão

Introdução



- ⇒ 80% das informações estão no formato textual e não podem ser tratados com ferramentas padrão de base de dados (TAN,1999).
- ⇒ Também chamado de não-estruturado

O texto expressa uma fonte de informação tão vasta, quanto rica, porém codificada de maneira que é difícil de ser decifrada automaticamente (HEARST,1999).

Tipo de Dado Textual

- 📄 Cartas de clientes
- 📄 E-Mail
- 📄 Gravações de chamadas telefônicas
- 📄 Contratos
- 📄 Mídias sociais (WhatsApp, Facebook, Twitter etc)

- 📄 Documentação técnica
- 📄 Patentes
- 📄 Notícias e artigos
- 📄 Páginas de internet

BD Tradicionais x Textuais

Tradicionais

Nome	Endereço	Telefone	Curso
José	Q 06	3487-3245	Francês
Lisa	Q 03	3228-9932	Francês
Márcio	Q 06	3487-5631	Francês

↳ Dados estruturados e organizados

Textuais

José mora na quadra 6. Seu amigo Márcio, telefone 3487-5631, também mora na mesma quadra. Ambos estudam com Lisa.

Lisa também é aluna do curso de Francês e mora na Q 03. Seu telefone é 3228-9932. O de José, 3487-3245.

- ↳ Não possuem “estrutura” pré-definida
- ↳ A informação pode estar disposta em qualquer parte ou mesmo não existir explicitamente

Conceitos

“É o processo **não-trivial** de identificação, em **grandes bases de dados**, de **padrões válidos**, previamente **desconhecidos**, potencialmente **úteis** e, finalmente, **compreensíveis** (Fayyad et al, 1997)”.



Data
Mining?



Mineração
de Dados?

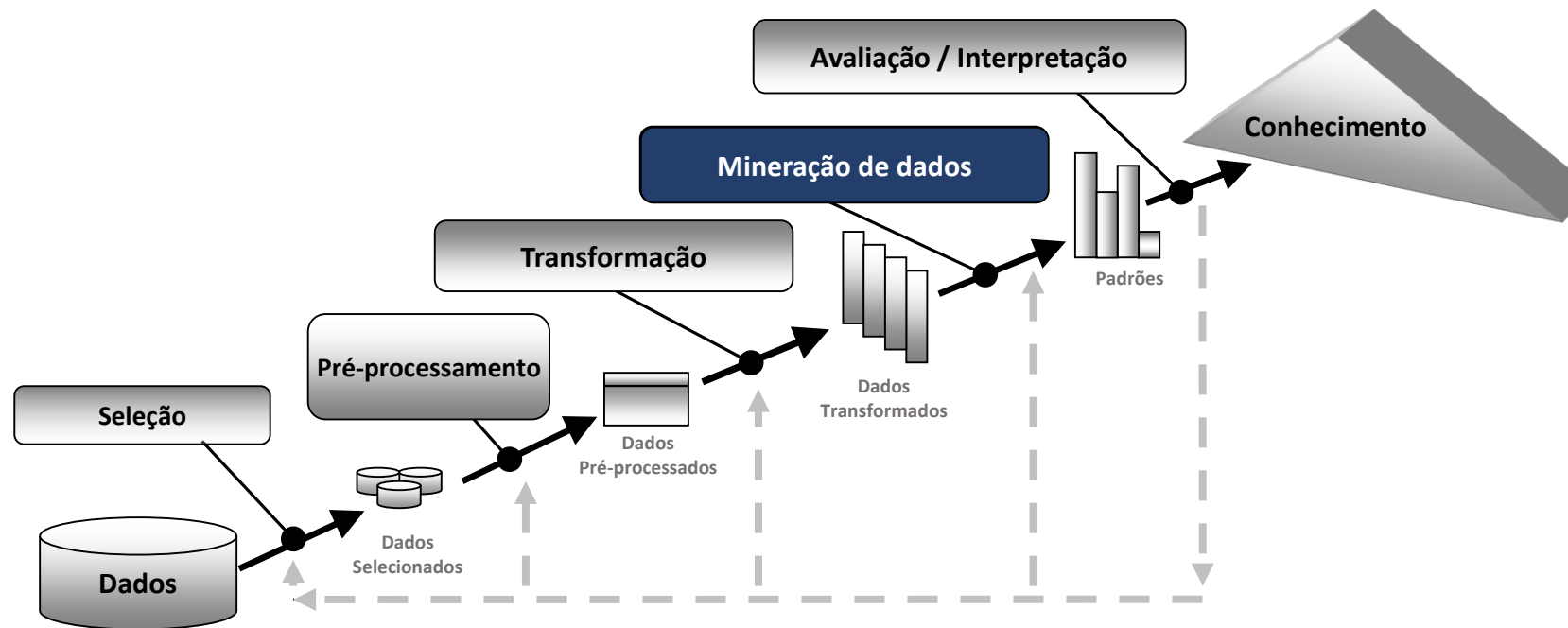


Knowledge
Discovery
in Data?



Descoberta de
Conhecimento
em Dados?

Processo DCD – Passo a passo



Fonte: Fayyad et al, 1997, pág. 41

Conceitos



Text
Mining?

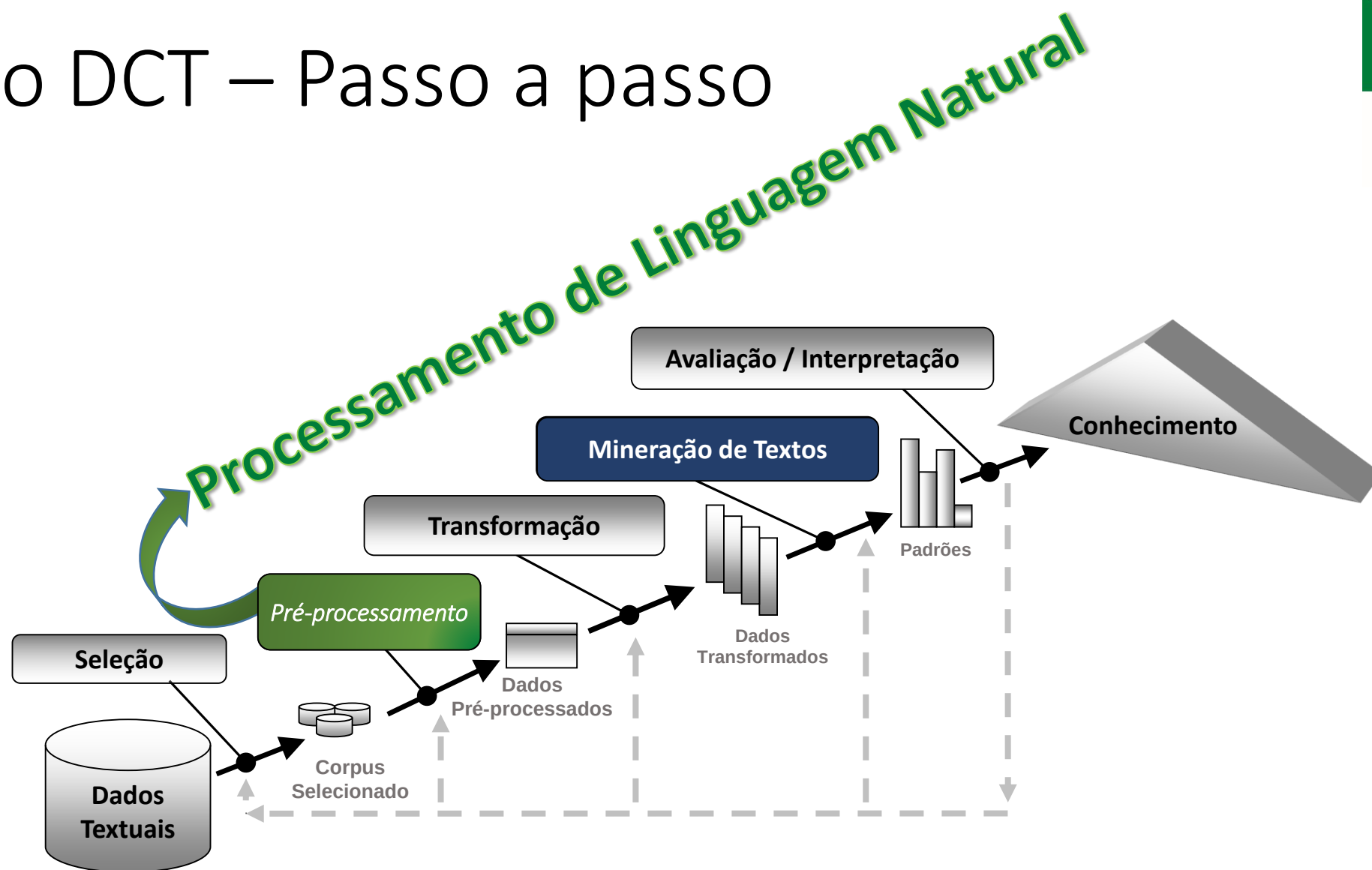
Mineração
de Texto?

“É o processo que emprega um grupo de algoritmos para converter o dado textual **não-estruturado** em dado **estruturado** e métodos quantitativos **similares** aos utilizados na **DCD** para descoberta de conhecimento (Weiss et al, 2005).”

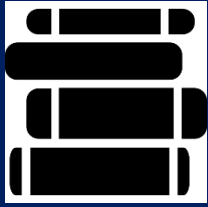
Knowledge
Discovery
in Text?

Descoberta de
Conhecimento
em Texto?

Processo DCT – Passo a passo



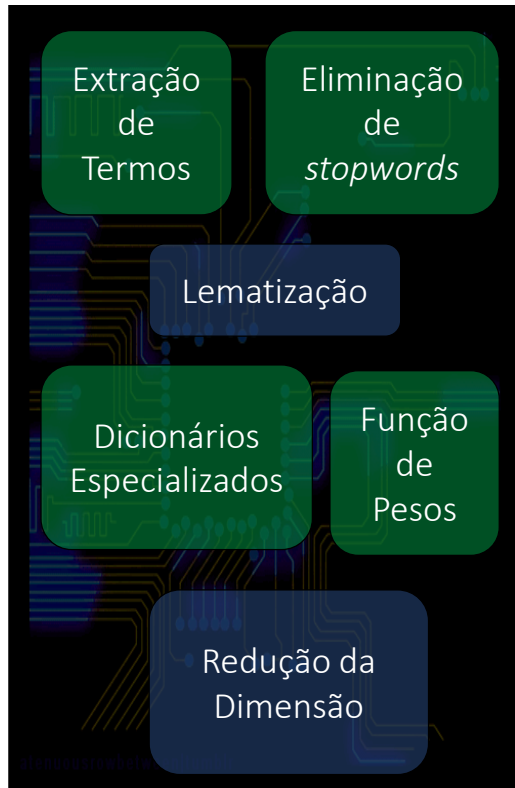
Pré-processamento



Doc_1 – Política é o fim.

Doc_2 – O caminho da @& política é informacao.

Doc_3 – A\$ segurança da informação é a política.



Documento	Termos				
	política	fim	caminho	informação	segurança
Doc_1	1	1	0	0	0
Doc_2	1	0	1	1	0
Doc_3	1	0	0	1	1

Aplicações da DCT

- 📄 Classificação de Documentos
 - 📄 Categorização de textos
 - 📄 Filtro de e-mail, Inteligência Competitiva na Web
- 📄 Grupamento de textos
 - 📄 Criação novas categorias a partir de uma coleção
- 📄 Recuperação da Informação
 - 📄 Motores de busca com técnicas da DCT
- 📄 Similaridade, link analysis

Aplicações da DCT



↳ Extração da Informação

↳ Notícias sob demanda, temas científicos associados a autores ou instituições

↳ Trending topics

↳ Acompanhamento de redes sociais

Exemplo: 27out2018 – <http://blablabra.net/>



Aplicações da DCT



A evolução anual
dos termos mais
frequentes na ATA
do Copom
De 2007 - 2017



2007



2008



2009



2010



2011



2012



2013



2014



2015



2016



2017



📄 Predição e Avaliação

📄 Identificação de tendências

📄 Validação de hipóteses

O problema REAL – na prática

- ↵ SAC – mensagens textuais
- ↵ SAC – fonte de informação estratégica
- ↵ Leitura de mensagens – compreensão do tema
- ↵ Velocidade de atendimento – mutirões de leitura
- ↵ Classificação manual X experiência do analista
- ↵ Informação não-estruturada X acompanhamento



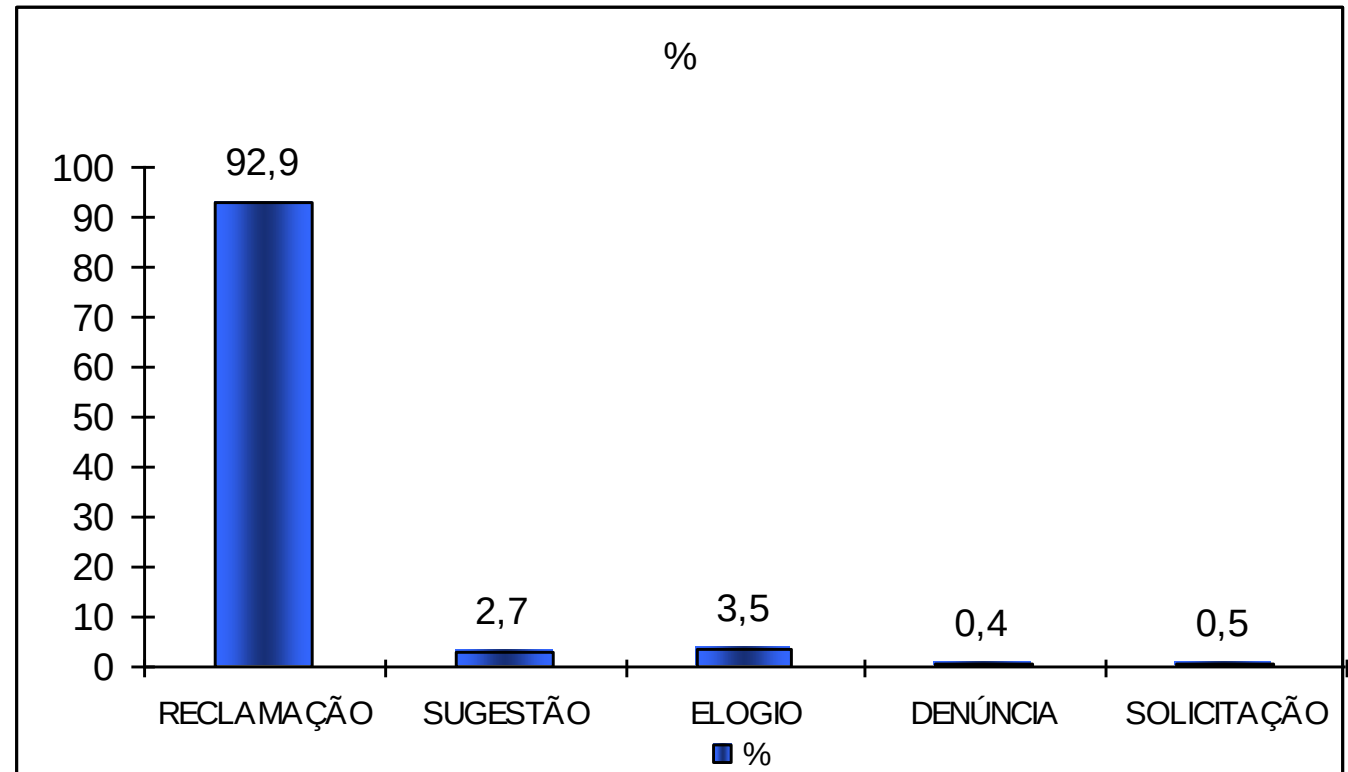
Fonte dos dados

📄 A Base de Dados

- 📄 499 mil registros
- 📄 Histórico de 6 anos

📄 O Escopo

- 📄 Janeiro – Junho – 6 meses
- 📄 52 mil registros – Reclamação
- 📄 2.413.586 termos



Exemplo de registro

GRUPO ASSUNTO: AGENCIA

ORIGEM: TELEFONE

NATUREZA: RECLAMAÇÃO

PRODUTO: AGÊNCIA

MOTIVO: FALTA DE CORTESIA DO GERENTE

DESCRIÇÃO DA OCORRÊNCIA: Cliente reclama que dia 07/02 esteve nesta agência para solicitar um empréstimo e a Gerente Sra. X informou os documentos necessários. Dia 08/02 retornou a agência e ao entregar o RG a Sra.X ela dirigiu-se ao Gerente Sr. Y com irônia e disse que o documento não seria aceito pelo caixa. A cliente foi falar com o Gerente Sr. Y que não aceitou o RG, mesmo a cliente dizendo que não houve tempo disponível para retirar outro, mas o gerente mencionou em voz alta que o RG estava um "lixo" e a cliente era incompetente até mesmo para retirar outro documento. Pede providências urgentes quanto ao atendimento prestado.

Exemplo de registro



GRUPO ASSUNTO: AGENCIA

ORIGEM: TELEFONE

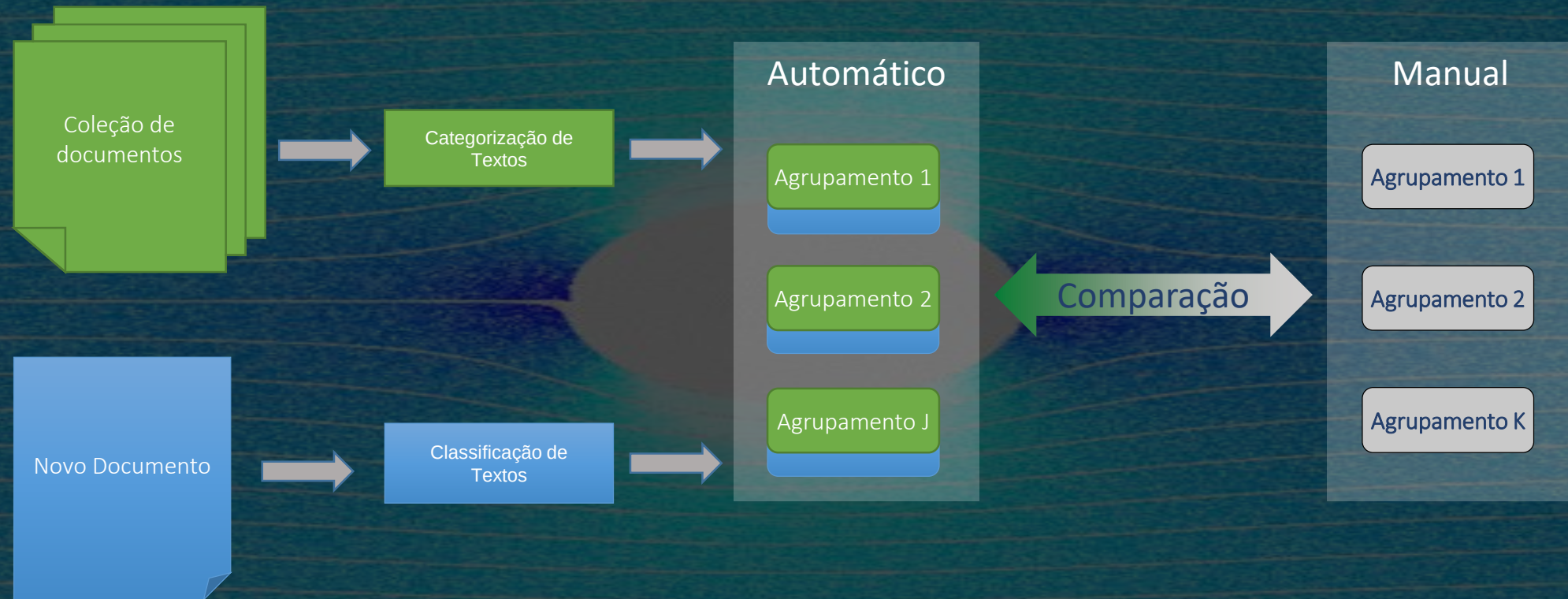
NATUREZA: RECLAMAÇÃO

PRODUTO: AGÊNCIA

MOTIVO: FALTA DE CORTESIA DO GERENTE

DESCRIÇÃO DA OCORRÊNCIA: Cliente reclama que dia 07/02 esteve nesta agência para solicitar um empréstimo e a Gerente Sra. X informou os documentos necessários. Dia 08/02 retornou a agência e ao entregar o RG a Sra. X ela dirigiu-se ao Gerente Sr. Y com **irônia** e disse que o documento não seria aceito pelo caixa. A cliente foi falar com o Gerente Sr. Y que não aceitou o RG, mesmo a cliente dizendo que não houve tempo **disponível** para retirar outro, mas o gerente mencionou em voz alta que o RG estava um "lixo" e a cliente era incompetente até mesmo para

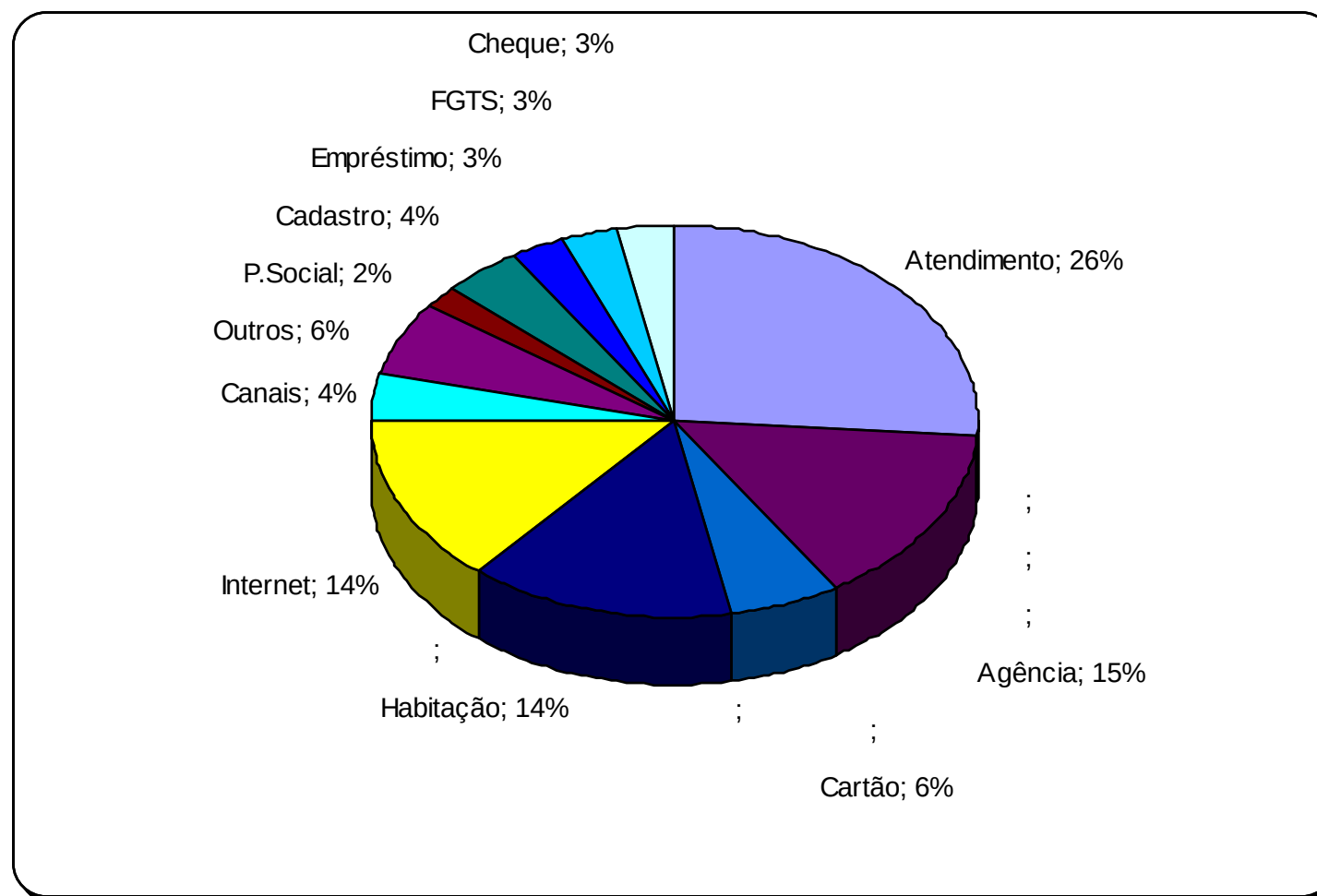
Fluxo Operacional



Nomeando dos agrupamentos

Grupo Assunto X Agrupamento											
Código do Grupo Assunto		Agrupamento (Cluster ID)									Total
		...	11	12	13	14	15	16	17	18	
Empréstimo	Qtde	...	27	1	5	0	14	188	4	5	303
	% linha	...	9	0	2	0	5	62	1	2	
	% coluna	...	3	1	2	0	1	71	2	2	
Cartões	Qtde	...	65	156	26	250	31	1	0	3	1087
	% linha	...	6	14	2	23	3	0	0	0	
	% coluna	...	7	82	8	89	3	0	0	1	

Classificação Automática Final



Conclusão



- ⇒ O SAC é fonte rica de informação
- ⇒ A preparação da base é difícil
 - ⇒ Erros ortográficos, pontuação e padronização
- ⇒ Dicionário especializado
 - ⇒ atualizado continuamente pela equipe
- ⇒ Dicionário em Português ou corpus de treinamento está restrito a alguns grupos acadêmicos
- ⇒ Há muito espaço para o profissional da Ciência da Informação

Referências



- ↵ FAYYAD, U., et al. **From Data Mining to Knowledge Discovery in Databases**. AAAI/MIT Press, 1997.
- ↵ HEARST, M.A. 1999. Untangling text data mining. In **Proceedings of ACL'99**: the 37th annual meeting of the association for computational linguistics, University of Maryland, June 20–26 (invited paper). Disponível em: <<http://www.ai.mit.edu/people/jimmylin/papers/Hearst99a.pdf>>.
- ↵ TAN, A.-H. Text mining: The state of the art and the challenges. In **Proceedings of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99** workshop on Knowledge Discovery from Advanced Databases, p. 65–70, 1999.
- ↵ WEISS, S. M., et al. **Text Mining**: Predictive Methods for Analyzing Unstructured Information. Springer, New York, 2005.



UnB

Obrigado!

Dr. Marcelo Schiessl

marcelo.schiessl@gmail.com