

TEXT ANALYTICS EM DOCUMENTOS HISTÓRICOS SENSÍVEIS: CONFIANÇA E ESCALABILIDADE

Flávio Codeço Coelho Bruno Cuconato

Do ponto de vista da matemática...

Corpus Conjunto de documentos: $\{d_1, d_2, \dots, d_n\}$.

Documento Conjunto de Palavras, ou frases, ou parágrafos.

Morfologia elementos únicos de um documento, Tokens: $t_i \in d_j$

Sintaxe Classificação dos tokens de acordo com as suas classes gramaticais e funções sintáticas. $d_i = \{(t_1, \textit{artigo}), (t_2, \textit{substantivo}), \dots\}$

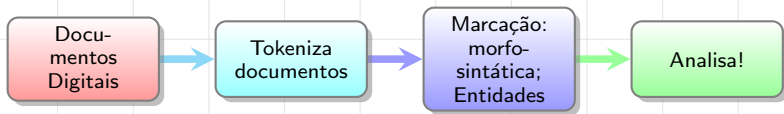
Semântica Significado de cada token.

... níveis mais abstratos.

Cada um destes elementos recebe interpretações probabilísticas, para permitir um tratamento estatístico adequado.

- Análise automatizada de textos digitais
- Classificação de Documentos
- Recuperação de informações
- Modelagem de assuntos
- Escalabilidade é componente chave.





- Algumas destas etapas requerem supervisão.
- etapas intermediárias dependem do Domínio do Corpus
- Pre-processamento para permitir escalabilidade da etapa de análise.

Envolvimento Humano na tokenização

- Tokenização é sensível à língua, terminologia de domínio, abreviações, etc.
- A tokenização é a base da construção do vocabulário que servirá de base às análises.
- Especialista garante a construção de um vocabulário conciso e preciso.
- Lematização pode ser realizada nesta etapa

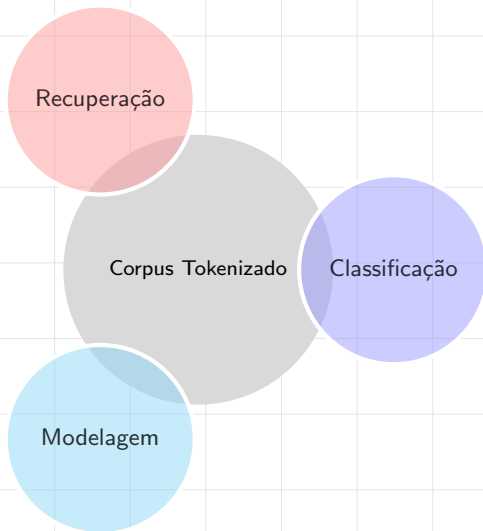
trabalho = { *trabalha*
Trabalho
trabalhar
trabalhando
trabalhador
:
:

- Identificação de Entidades Nomeadas: Pessoas, instituições, Locais, Leis, etc.
- Também requer supervisão humana.

sportsteam sportsteam geo-loc
India vs Australia 2014-15 , 4th Test in Sydney

company product
Samsung to launch Galaxy S6 in March

tvshow tvshow
New Suits and Brooklyn Nine-Nine tomorrow ... Happy days

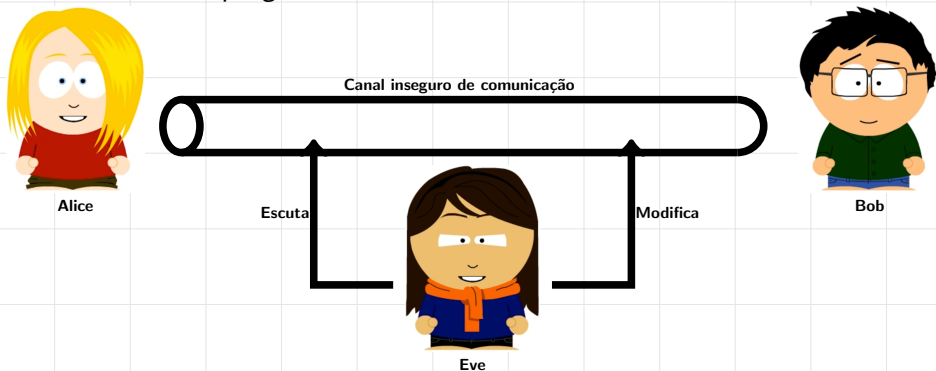


Principais desafios:

- Acesso restrito.
- Pessoas com acesso raramente são analistas técnicos.
- Alta densidade informacional.
- Extremamente relevantes para fins acadêmicos, políticos, jurídicos, etc.
- Arquivos digitais são gerados em velocidade crescente.
- Análise manual torna-se inviável.
- Terceirização de serviços analíticos ajudaria a dar escala ao processo.

É possível terceirizar a análise sem violar a confidencialidade?

Podemos Usar Criptografia?



É possível analisar um texto encriptado?

O empresário inicia explicando como e quando os políticos começaram a agir como "organizações criminosas". Segundo Joesley Batista, tudo começou há cerca de 10, 15 anos, quando surgiram grupos com divisão de tarefas: um chefe, um operador e um tesoureiro.

De acordo com o empresário, são organizações criminosas que existem para ganhar dinheiro cometendo crimes.

Na entrevista, Joesley afirma que esses esquemas organizados começaram no governo do PT e diz que "Lula e o PT" institucionalizaram a corrupção com a criação de núcleos, divisão de tarefas entre integrantes, em estados, ministérios, fundos de pensão e bancos, entre os quais o Banco Nacional de Desenvolvimento Econômico e Social (BNDES).

Fonte: G1, 18 de junho de 2017

gAAAAABZRplEJ-19w1CcbXFfaBZ1wErzXjNyAoMK
dwhxr2my9M0Ck3HFGBTsvFaJnkSdg0sumi5CCtYX
DdVcmZ-xBbcfHgRq2clZrY9XK_VcR6g_tUMWnhmK
CwbJEa5v2HiCIe0r38twfOGcwXuT014TrBRcLrYR
GLNAD6vjVsuBjRgK-TH0ShWtZe1idwglA3jpAL5w
e21b0fpWORQY3woHQQkIUslctdqT6V6g_WzUzWQ-
sdaB78rTJON_FodMPAMzLe4-BfkPcXLX_iPvaigN
OR7wFyFIC83jhN_5yFQw66aLo6-8KHw1ZYgGyDwH
UjJpia_j-z6j45Y3wcAnyEUFbtVIqKVvFyiBL3bp
ipg4eDvUeOf42K0vQdNimRfUULEkoW5SWKk7heYg
M3EIysbEeMQZAgaciaA1DNNkRqh9dVW8YLoRbA4sv
ZZ2t5PFBa4ZpY_Eg98CvvYjzTtwJmojaD7JONR0g
AoW2VaNUhRu3nhQAm-Z6FXM69n2w4hcj05-111qN
vs_WZ4Q2_uVB4csB805d9SEoR9xwgmeHbrs8_GIf
s3v_KDQt28ACZu53r7vGLGPkIC7wpanvM5E7vdSy
Fgh3rIZAFpW94HkpVpT2gx8iQD1jd8CH6IcgHqr6CC

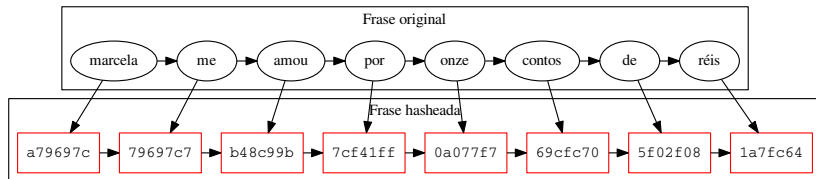
Solução:

- Análise matemática de textos, depende de suas propriedades estatísticas: (frequências de palavras, probabilidades de co-ocorrência, etc.)
- Funções de Hash aplicadas ao nível das palavras, são a solução: $H(p) = h_p$
 - ▶ Determinísticas
 - ▶ Não invertíveis (na prática)
 - ▶ Valores não correlacionados ao dado original

Palavra	Hash value – sha256
<i>word</i>	98c1eb4ee93476743763878fcb96a25fbc9a175074d64004779ecb5242f645e6
<i>words</i>	dba36bffa5cab0f922d087a3aeb179f9d4e745df40b323e1b1471402848c8a3e

Coelho FC, Cuconato B. (2017) Secure trustless text processing of sensitive documents. PeerJ Preprints 5:e2994v1 <https://doi.org/10.7287/peerj.preprints.2994v1>

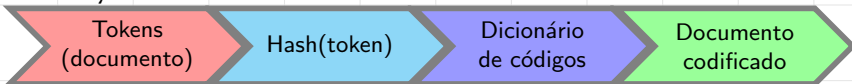
Codificando um documento com funções de Hash



Cada palavra é "salgada"

Na prática, uma sequência de caracteres aleatória é adicionada a cada palavra para aumentar a segurança.

Codificação – Gestor:



Análise e Decodificação – Analista:



Homomorfismo

Documentos codificados são homomórficos, nos parâmetros exigidos pela maioria dos algoritmos de Análise de textos.

Propriedades preservadas: contagem de palavras, ordem das palavras, estrutura de frases e parágrafos.

Análises Possíveis:

- Recuperação de informações
- Classificação de documentos (Machine Learning)
- Modelagem de assuntos (LDA, LSI, etc.)
- Sumarização
- Análise de estilos (Identificação de autores)
- etc.

Vantagens:

- Open-source software¹
- Código pequeno e auditável.
- Facilmente executável em ambiente controlado sem relaxamento das regras de acesso.
- Apenas o corpus codificado é enviado ao analista.
- Corpus só pode ser decodificado de posse do dicionário de decodificação.
- Algoritmo Eficiente: $O(n)$ em relação ao número de palavras.
- implementações em Python e D.

Limitações:

- Requer um corpus tokenizado a priori
- Eliminação de stop-words, lematização e identificação de entidades, devem ser feitas a priori.
- Requer conhecimentos básicos de programação.

¹<https://github.com/NAMD/corpushash>

Obrigado pela atenção!