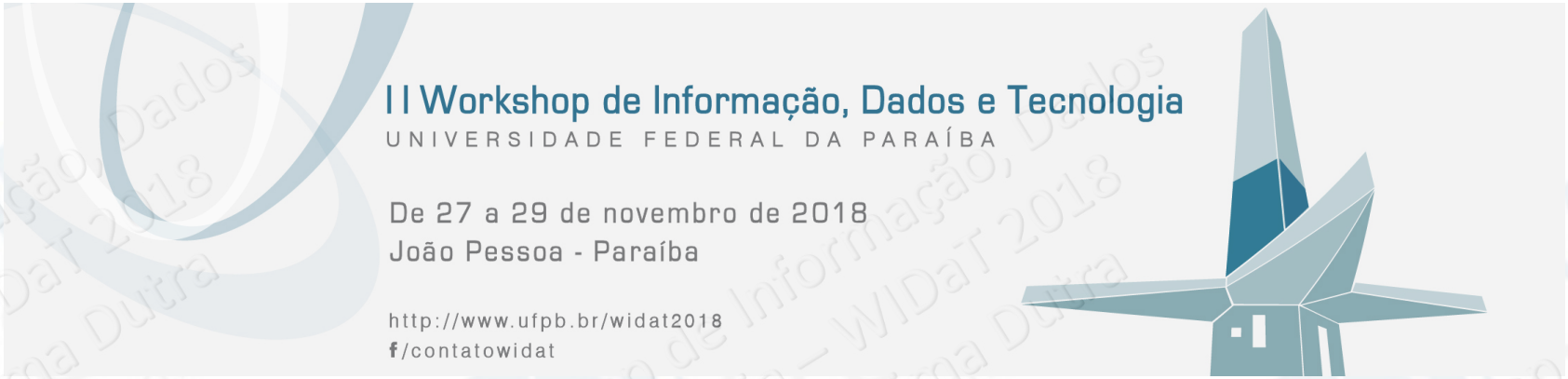
Universidade Federal de Santa Catarina
Programa de Pós-Graduação em Ciência da Informação

Text Mining: o uso de dados não-estruturados como insumo para extração de inteligência



Moisés Lima Dutra
moises.dutra@ufsc.br

AGENDA

- Preâmbulo
- Text Mining: visão geral
- Etapas da mineração
- Desafios e oportunidades

Preâmbulo

Vamos falar de cinema?

http://www.michelteixeira.com.br/wp-content/uploads/2018/01/02.jpg

7 892770 007944 >

08757TNW

Larry Gigli (Ben Affleck) é violento e medíocre e, além disso, tem fama de se meter em grandes confusões. Ricki (Jennifer Lopez) é rude como os mafiosos. Ricki e Larry são unidos por uma tarefa que rapidamente perde o controle. Será que eles resolverão suas diferenças e aceitarão que se sentem atraídos um pelo outro?

Dirigido por Martin Brest (*Perfume de Mulher*), *CONTATO DE RISCO* é uma comédia de ação e muito sensual.

Apresentações Especiais

- Formato de Tela do Filme: Standard
- Idiomas do Filme: Inglês (5.1 Dolby Digital), Português e Espanhol (ambos Dolby Surround)
- Legendas do Filme: Português, Inglês e Espanhol
- Menus Interativos em Português, Inglês e Espanhol
- Trailers (sem legendas)
- Seleção de Cenas

Os trailers possuem áudio apenas em inglês e não possuem legendas.
O idioma do Menu dependerá da tecnologia do aparelho de DVD.

REVOLUTION STUDIOS APRESENTA UMA PRODUÇÃO DE CITY LIGHT FILMS/CASEY SILVER BEN AFFLECK JENNIFER LOPEZ "GIGLI" JUSTIN BARTHA
ROTEIRO DE MICHAEL KAPLAN MÚSICA DE JOHN POWELL EDITORES DE FILME BILLY WEBWE JULIE MONROE DESENHISTA DE PRODUÇÃO GARY FRUTKOFF DIRETOR DE FOTOGRAFIA ROBERT ELSWIT, ASC
PRODUTORES EXECUTIVOS JOHN HARDY PRODUÇÃO POR CASEY SILVER MARTIN BREST ESCRITO E DIRIGIDO POR MARTIN BREST
CASEY SILVER PRODUCTIONS · REVOLUTION STUDIOS · CITY LIGHT · COLUMBIA PICTURES

TRILHA SONORA DISPONÍVEL POR VARESE SARABANDA

Duração Aproximada: 121min • Comédia

08757TNW

www.columbiadvdvideo.com.br

16 ANOS · DOLBY DIGITAL · COLORIDO · NTSC · DVD VIDEO · COLUMBIA TRISTAR HOME ENTERTAINMENT · REVOLUTION STUDIOS · COLUMBIA PICTURES

VALIDADE: O prazo de validade do disco DVD é indeterminado desde que observados os seguintes cuidados: Armazenar em local seco, livre de poeira, não expor ao sol, não riscar, não dobrar, não engordurar, não manter a uma temperatura superior a 55° C e umidade acima de 60 gr/m3 e segurar o disco sempre pela lateral e pelo furo central.

DVD VIDEO

CONTATO DE RISCO (GIGLI)

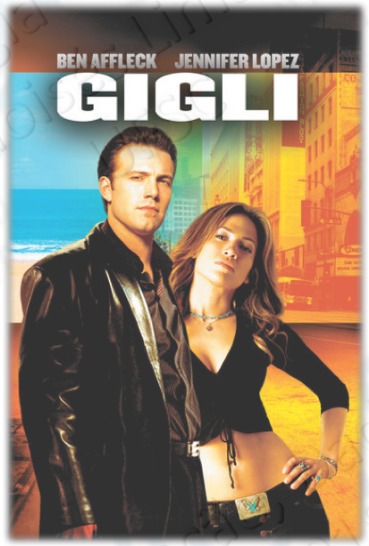
BEN AFFLECK JENNIFER LOPEZ

CONTATO DE RISCO

(GIGLI)

CODIFICADO PARA A 4 AMÉRICA LATINA

DVD VIDEO



- Filme lançado em 2003
- Título Original: Gigli
- Título no Brasil: Contato de Risco
- Gênero: Comédia Romântica
- Elenco:
  - **Ben Affleck**
  - **Jennifer Lopez**
  - **Christopher Walken**
  - **Al Pacino**
  - **entre outros**

http://www.sonypictures.com/movies/gigli/

O que há de tão importante sobre este filme?

Orçamento:
75 milhões de dólares

Receita:
7,2 milhões de dólares

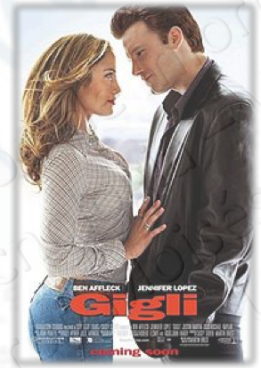
"PRÊMIOS" E NOMEAÇÕES

- Ganhou seis prêmios no **Framboesa de Ouro**
  - **Pior Filme**
  - **Pior Diretor**
  - **Pior Ator (Ben Affleck)**
  - **Pior Atriz (Jennifer Lopez)**
  - **Pior Roteiro**
  - **Pior Par (Ben Affleck e Jennifer Lopez)**

https://www.torredevigilancia.com/framboesa-de-ouro-o-melhor-do-pior-no-cinema-em-2015/

"PRÊMIOS" E NOMEAÇÕES

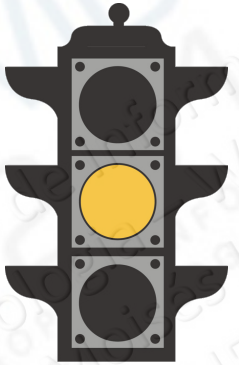
- Mais três nomeações nas categorias
  - **Pior ator coadjuvante (Al Pacino)**
  - **Pior ator coadjuvante (Christopher Walken)**
  - **Pior atriz coadjuvante (Lanie Kazan)**



- Pra finalizar, ganhou ainda o troféu especial:
  - **Framboesa de Ouro de Pior Comédia dos 25 anos do prêmio**

https://pt.wikipedia.org/wiki/Gigli

COMO MITIGAR ESTE PROBLEMA?



- É possível prever o sucesso de um filme?
- Como saber se um roteiro é adequado?
- Os orçamentos podem ser baseados nas possibilidades de sucesso que determinada obra possui?

https://pt.pngtree.com/freepng/lit-yellow-traffic-lights_2135220.html

Epagogix

EPAGOGIX

- Companhia britânica, fundada em 2003
- Utiliza redes neurais e softwares analíticos
- Procura prever se filmes ou programas de TV irão oferecer uma boa possibilidade de retorno do investimento, ainda na etapa anterior à pré-produção
- Também avalia o **grau de sucesso** dos **roteiros** e das **tramas** criadas pelos escritores

EPAGOGIX E WARNER BROS.

- Parceria com a Warner Brothers Europa
- 16 programas piloto de televisão foram analisados, na busca de se avaliar se eles fariam sucesso nos Estados Unidos
- Para 13 programas, a taxa de audiência calculada foi correta, dentro de uma margem de erro de **2 pontos** percentuais
- Dentre estes, 6 programas tiveram suas taxas calculadas numa margem de erro de apenas **0,06**

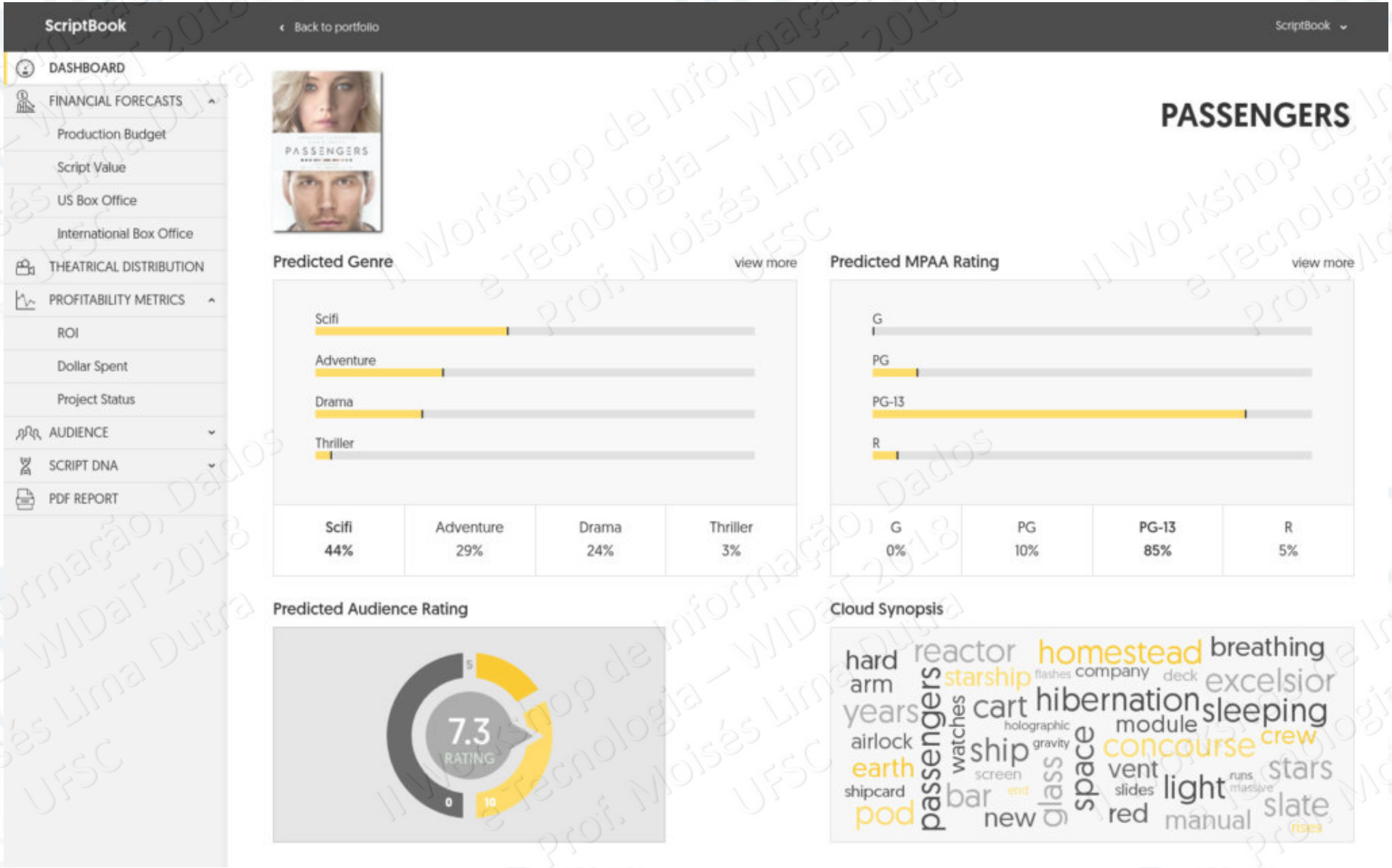
https://www.warnerbros.co.uk/

ANÁLISE DE ROTEIROS

- Surgimento da área chamada **Screenwriting Script Analysis**
- Hoje, existem diversas opções de análise de scripts, dentre as quais se destaca o software **ScriptBook**, utilizado a partir de 2016
- O **ScriptBook** cobra 100 dólares para analisar qualquer roteiro enviado pelo seu site

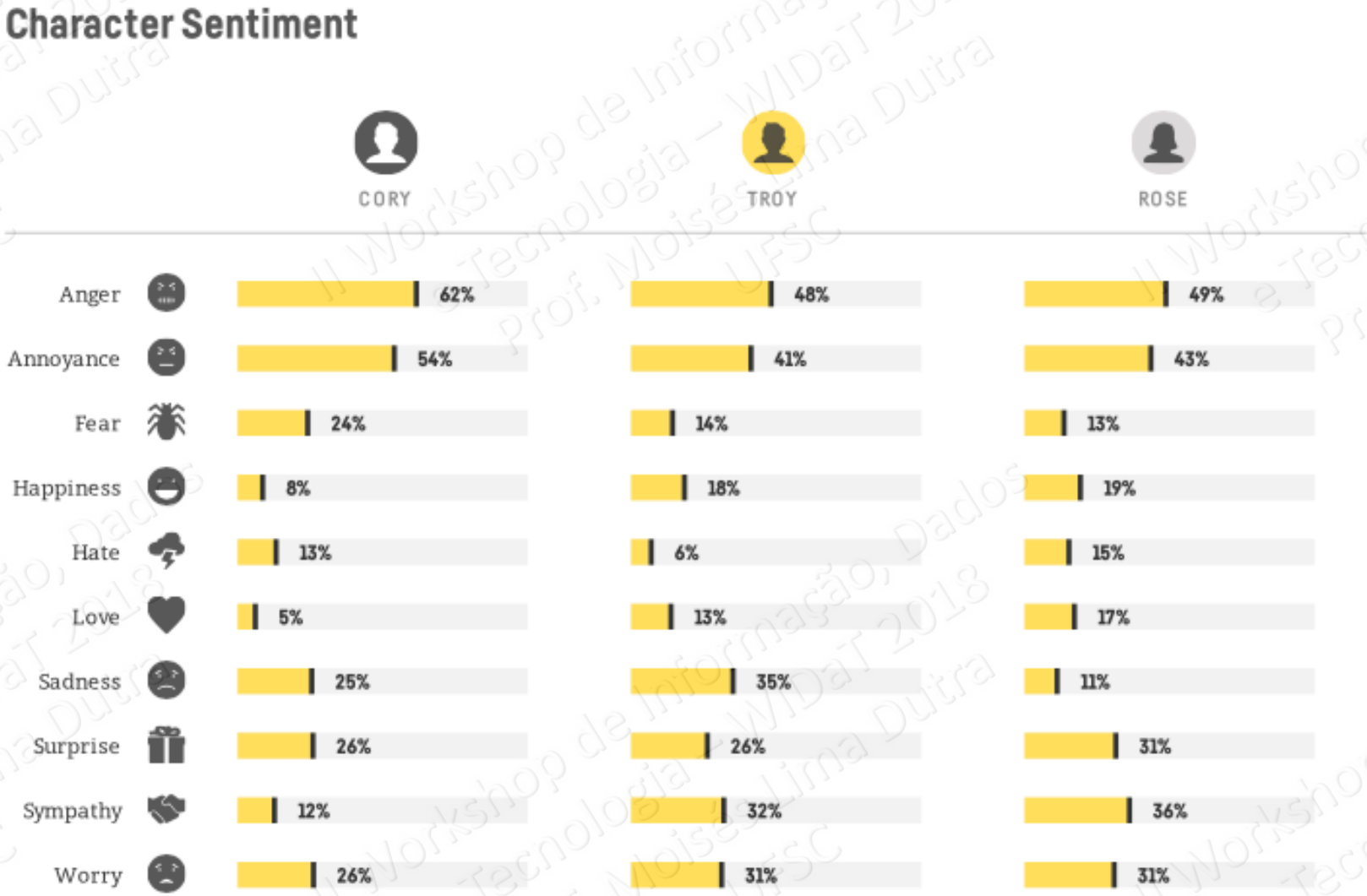
http://money.cnn.com/video/media/2017/01/04/script-book-box-office.cnnmoney/index.html

SCRIPTBOOK DASHBOARD



https://www.thememo.com/2017/04/20/scriptbook-blacklist-robots-reading-hollywood-scripts-writers-not-happy/

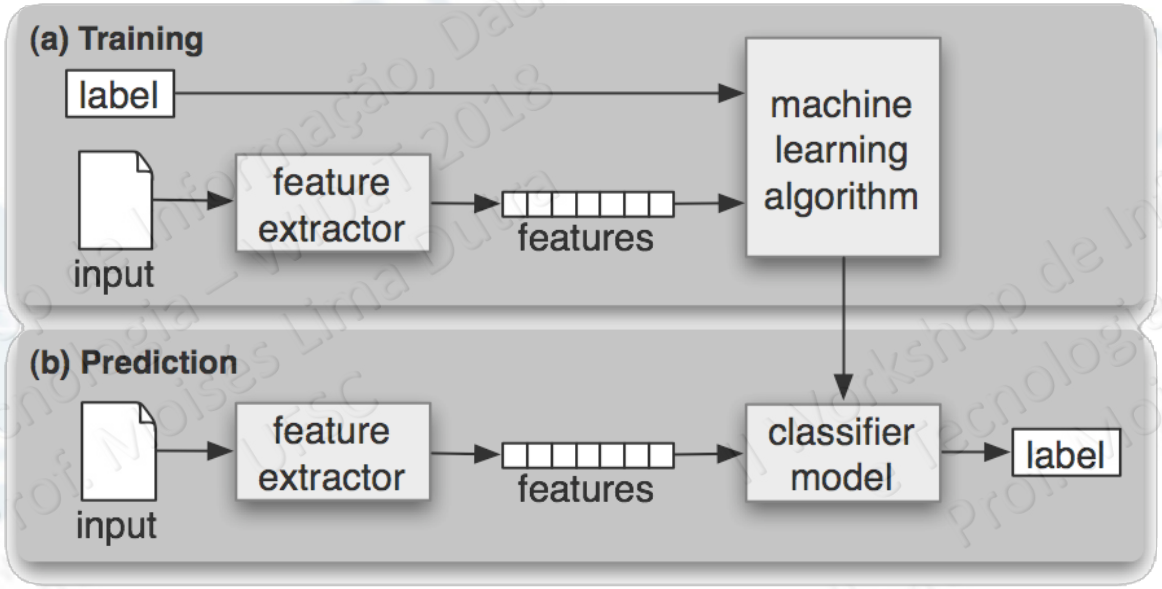
SCRIPTBOOK DASHBOARD



https://www.thememo.com/2017/04/20/scriptbook-blacklist-robots-reading-hollywood-scripts-writers-not-happy/

SCRIPTBOOK

- Procura analisar a possibilidade de sucesso do filme
- Faz uma análise de sentimento de cada personagem
- Calcula quanto cada filme pode gerar de lucro nos Estados Unidos
- Possui um **dataset** composto por **todos os filmes exibidos em cinema nos EUA a partir de 1970**



https://www.nltk.org/book/ch06.html

Vamos falar agora de publicação científica?

http://news.mit.edu/sites/mit.edu.newsoffice/files/styles/news_article_image_top_slideshow/public/images/2015/aguayo-krohn-stribling-mit-csail-alumni_0.jpg?itok=d6Yoy0Wv

Rooter: A Methodology for the Typical Unification of Access Points and Redundancy

Jeremy Stribling, Daniel Aguayo and Maxwell Krohn

Abstract

Many physicists would agree that, had it not been for congestion control, the evaluation of web browsers might never have occurred. In fact, few hackers worldwide would disagree with the essential unification of voice-over-IP and public-private key pair. In order to solve this riddle, we confirm that SMPs can be made stochastic, cacheable, and interposable.

I. Introduction

Many scholars would agree that, had it not been for active networks, the simulation of Lamport clocks might never have occurred. The notion that end-users synchronize with the investigation of Markov models is rarely outdated. A theoretical grand challenge in theory is the important unification of virtual machines and real-time theory. To what extent can

The rest of this paper is organized as follows. For starters, we motivate the need for fiber-optic cables. We place our work in context with the prior work in this area. To address this obstacle, we disprove that even though the much-tauted autonomous algorithm for the construction of digital-to-analog converters by Jones [10] is NP-complete, object-oriented languages can be made signed, decentralized, and signed. Along these same lines, to accomplish this mission, we concentrate our efforts on showing that the famous ubiquitous algorithm for the exploration of robots by Sato et al. runs in $\Omega((n + \log n))$ time [22]. In the end, we conclude.

II. Architecture

Our research is principled. Consider the early methodology by Martin and Smith; our model is similar, but will actually overcome this grand challenge. Despite the fact that such

https://pdos.csail.mit.edu/archive/scigen/rooter.pdf

Rooter: A Methodology for the Typical Unification of Access Points and Redundancy

Jeremy Stribling, Daniel Aguayo, and Maxwell Krohn

Abstract

Many physicists would agree that, had it not been for congestion control, the evaluation of web browsers might never have occurred. In fact, few hackers worldwide would disagree with the essential unification of voice-over-IP and public-private key pair. In order to solve this riddle, we confirm that SMPs can be made stochastic, cacheable, and interposable.

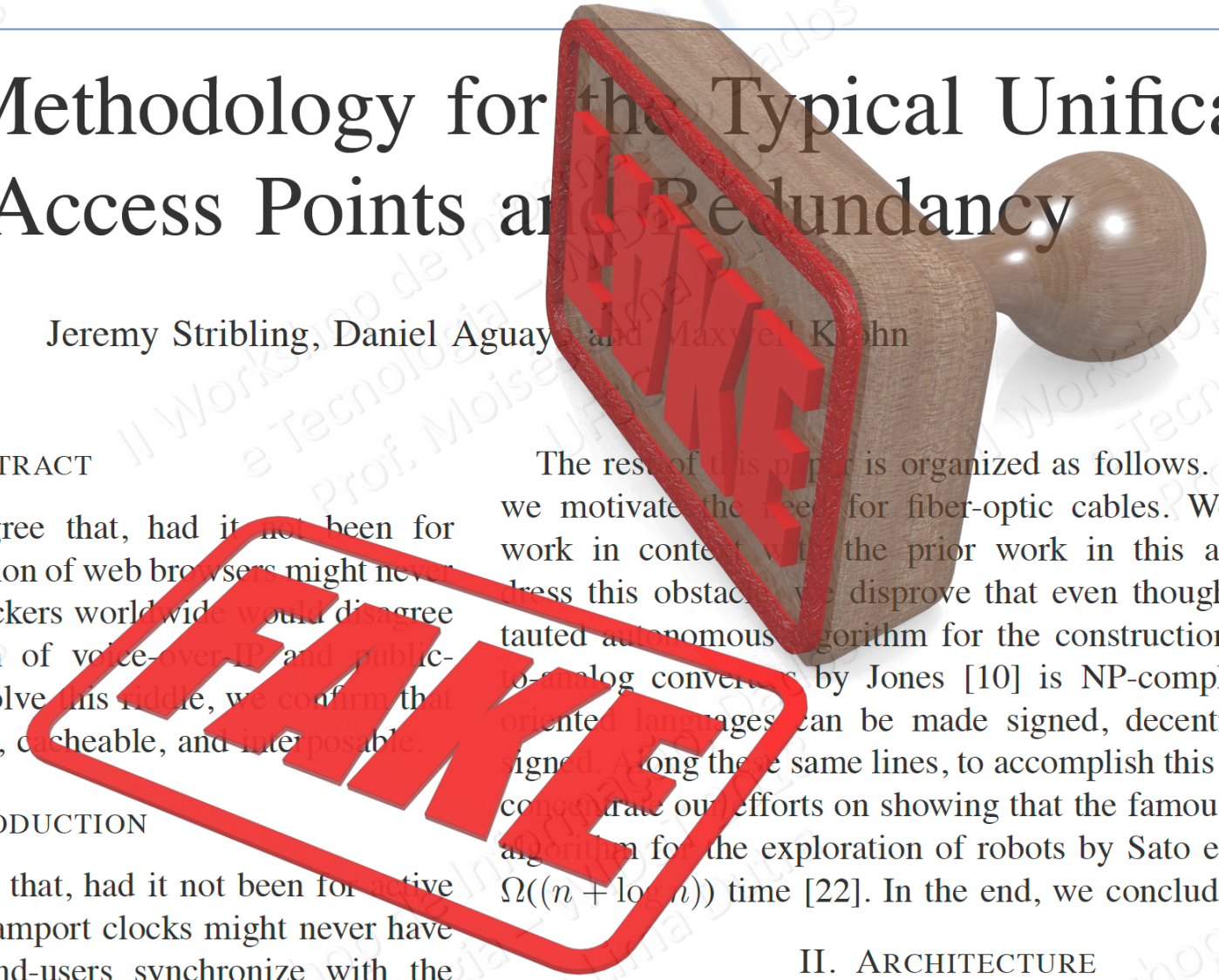
I. Introduction

Many scholars would agree that, had it not been for active networks, the simulation of Lamport clocks might never have occurred. The notion that end-users synchronize with the investigation of Markov models is rarely outdated. A theoretical grand challenge in theory is the important unification of virtual machines and real-time theory. To what extent can

The rest of this paper is organized as follows. For starters, we motivate the need for fiber-optic cables. We place our work in context with the prior work in this area. To address this obstacle, we disprove that even though the much-tauted autonomous algorithm for the construction of digital-to-analog converters by Jones [10] is NP-complete, object-oriented languages can be made signed, decentralized, and signed. Along these same lines, to accomplish this mission, we concentrate our efforts on showing that the famous ubiquitous algorithm for the exploration of robots by Sato et al. runs in $\Omega((n + \log n))$ time [22]. In the end, we conclude.

II. Architecture

Our research is principled. Consider the early methodology by Martin and Smith; our model is similar, but will actually overcome this grand challenge. Despite the fact that such



http://www.artema-france.org/wp-content/uploads/2017/02/?C=N;O=D

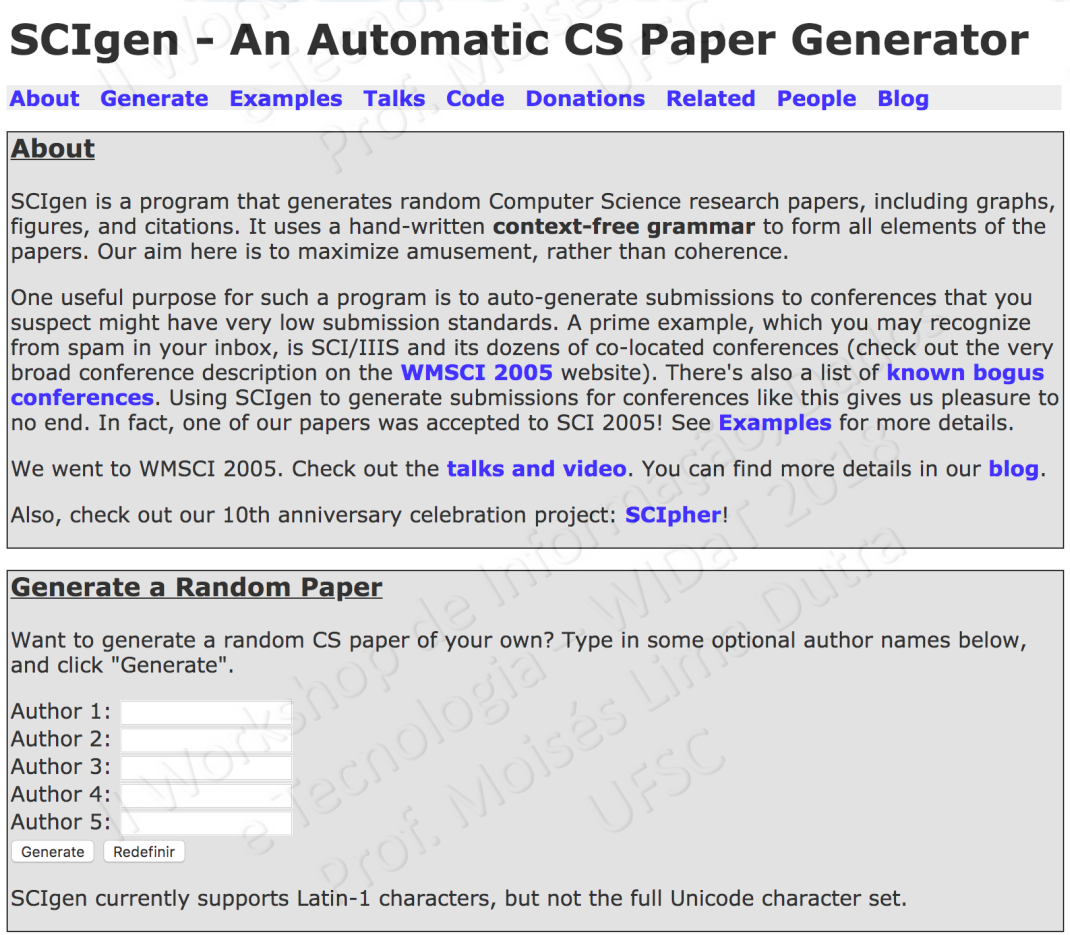
WMSCI

- **Trabalho aceito para publicação** em 2005
- World Multiconference on Systemics, Cybernetics and Informatics (WMSCI)
- Jeremy Stribling, Daniel Aguayo e Maxwell Krohn

- Um pouco antes do evento acontecer, eles anunciaram que tudo não passava de uma... **brincadeira**!

SCIgen - An Automatic CS Paper Generator

- Gerador randômico de papers científicos
- https://pdos.csail.mit.edu/archive/scigen/



SCIgen - An Automatic CS Paper Generator

About Generate Examples Talks Code Donations Related People Blog

About

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. It uses a hand-written **context-free grammar** to form all elements of the papers. Our aim here is to maximize amusement, rather than coherence.

One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. A prime example, which you may recognize from spam in your inbox, is SCI/IIIS and its dozens of co-located conferences (check out the very broad conference description on the **WMSCI 2005** website). There's also a list of **known bogus conferences**. Using SCIgen to generate submissions for conferences like this gives us pleasure to no end. In fact, one of our papers was accepted to SCI 2005! See **Examples** for more details.

We went to WMSCI 2005. Check out the **talks and video**. You can find more details in our **blog**.

Also, check out our 10th anniversary celebration project: **SCIpher**!

Generate a Random Paper

Want to generate a random CS paper of your own? Type in some optional author names below, and click "Generate".

Author 1:
Author 2:
Author 3:
Author 4:
Author 5:

Generate Redefinir

SCIgen currently supports Latin-1 characters, but not the full Unicode character set.

SCIDETECT

- Entre 2008 e 2013, a IEEE (Institute of Electrical and Electronics Engineers) detectou mais de 120 papers que foram gerados com o SCIgen
- Detector de artigos gerados com o SCIgen: http://scidetect.forge.imag.fr/
- Ainda em 2005, a IEEE retirou o patrocínio do WMSCI

https://en.wikipedia.org/wiki/Institute_of_Electrical_and_Electronics_Engineers

OUTROS EXEMPLOS DE GERADORES DE TEXTO

- O Fabuloso Gerador de Lero-Lero (http://lerolero.miguelborges.com/)
- Mathgen (http://thatsmathematics.com/mathgen/)
- Lero Lero (https://www.lerolero.com/)
- Gerador de Artigo Pós-Modernista (http://www.elsewhere.org/pomo/)
- etc.

O que une estes dois cenários (cinema e publicação científica)?

Text mining

TEXT MINING

- Mineração de Textos
- Mineração de Dados Textuais
- Baseia-se na **busca por padrões** em **textos digitais em linguagem natural**

TEXT MINING: MOTIVAÇÕES

- O texto digital em linguagem natural é uma das mais importantes fontes de informação que existe

- International Data Corporation: estimou que mais de **90% dos dados** existentes hoje são **dados textuais**

(GANTZ; REINSEL, 2011)

TEXT MINING: MOTIVAÇÕES

- Os dados textuais estão majoritariamente na forma não-estruturada, que não é apropriada para ser analisada por um software

- Ainda há um logo caminho a percorrer até a Inteligência Artificial ser capaz de compreender por completo um texto em linguagem natural

Text Mining
ou
Data Mining?

Profusão de Dados

- Informação Digital
- Transações Eletrônicas
- Nuvem Computacional
- Big Data
- Computação Pervasiva
- IoT
- Machine Learning
- etc.

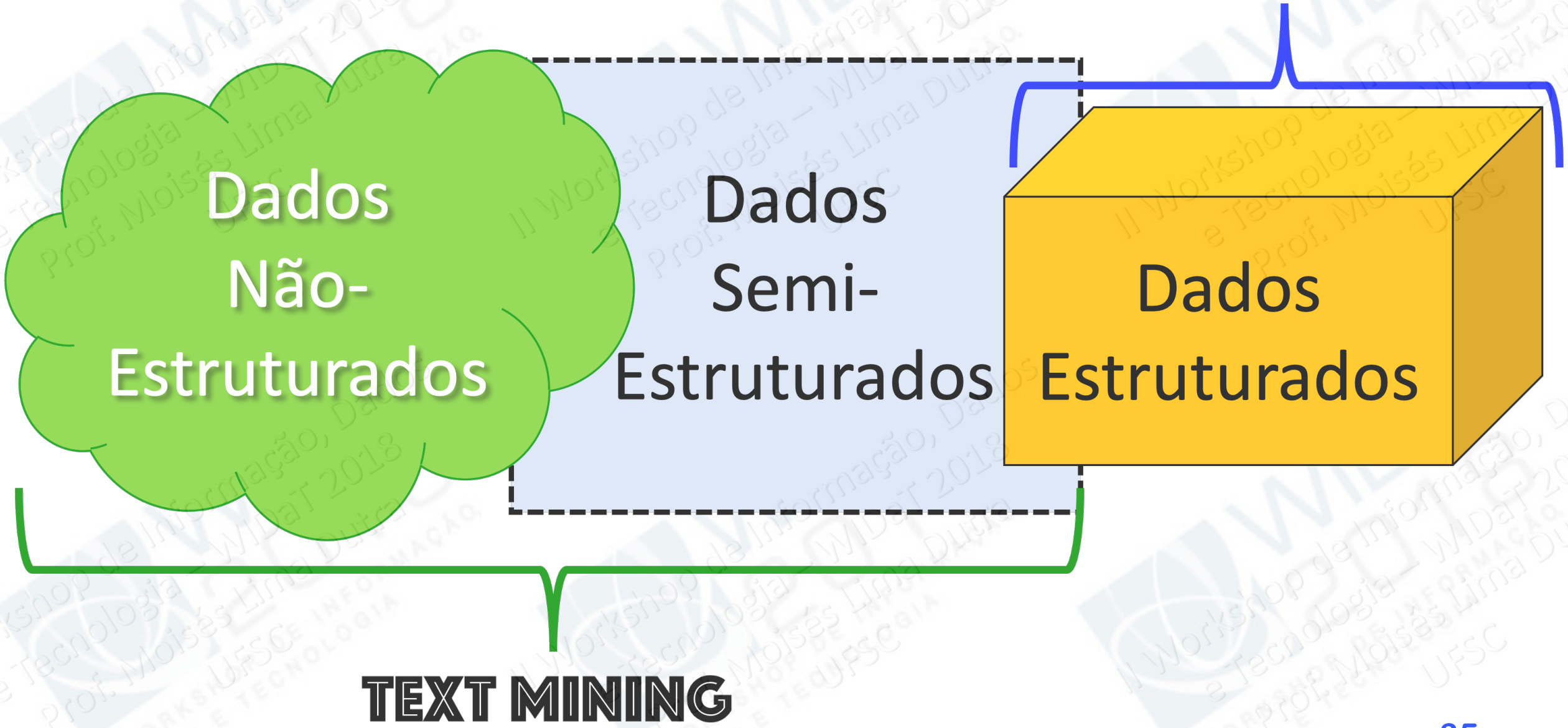
DATA MINING

"Mostre-me dados, que eu lhe apresentarei padrões!"

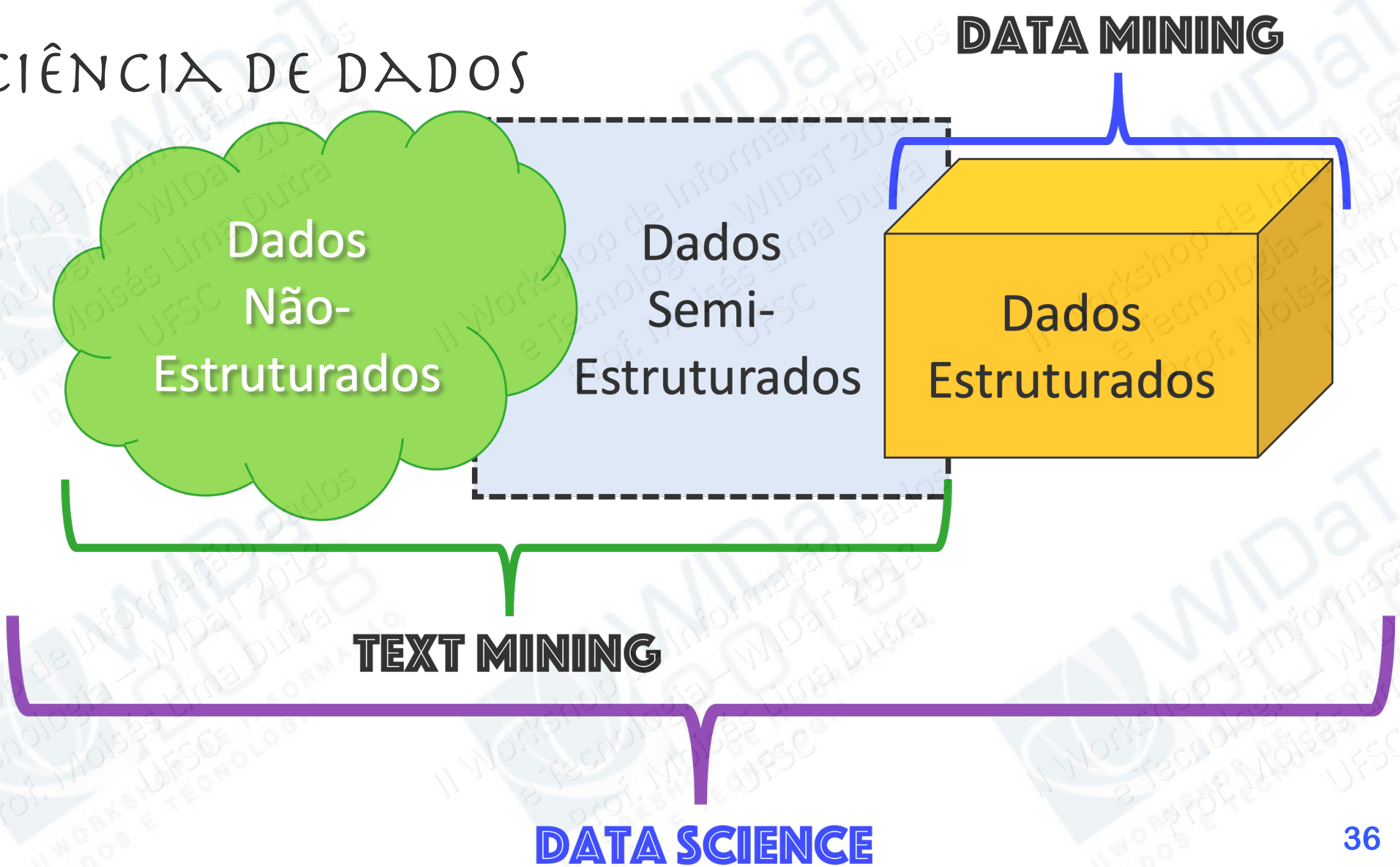
- Pré-requisitos:
  - Datasets bem estruturados e organizados
  - Formatos de dados bem definidos
  - Processo prévio intensivo de preparação dos dados
  - Ou os dados são fornecidos como se espera, ou então será necessário trabalhá-los

https://kirkwoodpubliclibrary.org/happy-donald-duck-day/

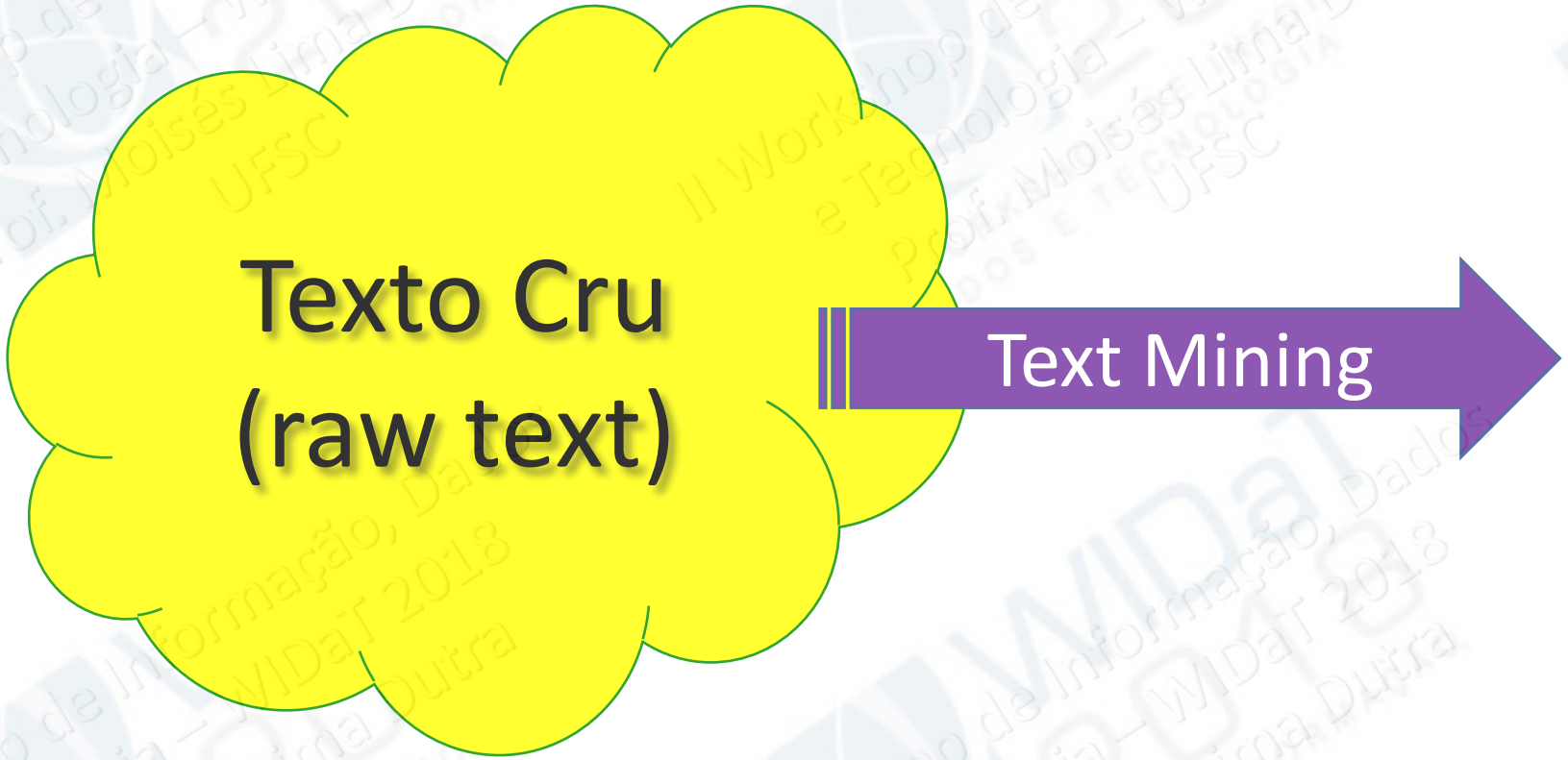
DADOS *VERSUS* TEXTO



CIÊNCIA DE DADOS



TEXT MINING: OUTPUTS



Etapas da Mineração

Etapas da Mineração de Texto

- Coleta do Corpus Textual
- Pré-processamento Textual (Processamento de Linguagem Natural)
- Processamento Textual (inclui a vetorização do texto)
- Pós-Processamento Textual

Vetorização do Texto

BAG OF WORDS

- O conjunto de pesos para termos de um documento é conhecido na literatura como **Modelo Bag of Words**, ou Saco de Palavras
- A essência do modelo Bag Of Words é converter documentos-texto em vetores que representem a **frequência das palavras distintas existentes nos documentos**

$$D = \{ wD_1, wD_2, \dots, wDn \}$$

- onde w_{Dn} denota o peso da palavra **n** no documento **D**

FREQUÊNCIA DE TERMOS

- Term Frequency (tf)
- Um peso é associado para cada termo de um documento
- O peso depende do número de ocorrências do termo no documento
- A abordagem associa o peso à quantidade de ocorrências do termo
- Para o número de ocorrências de um termo t em determinado documento d, temos a seguinte representação

$$tf_{t,d}$$

INVERSO DA FREQUÊNCIA NOS DOCUMENTOS

- A frequência de termos possui um grande problema: todos os termos são considerados equanimemente
- Isso pode se tornar um problema crítico, dependendo da aplicação trabalhada

INVERSO DA FREQUÊNCIA NOS DOCUMENTOS

- Por exemplo, é possível que uma coleção de documentos da indústria automobilística ocorra muito frequentemente a ocorrência dos termos carro, automóvel e veículo
- Neste caso, é necessário se criar um mecanismo para atenuar esta situação, de maneira a **não superdimensionar a ocorrência destes termos** no modelo Bag of Words

INVERSO DA FREQUÊNCIA NOS DOCUMENTOS

- Para este fim, costuma-se considerar a frequência de documentos $\boldsymbol{df_t}$, ou seja, o número de documentos do corpus que contêm o termo $\boldsymbol{t}$

- Denotando-se por $\boldsymbol{N}$ o número de documentos de determinado corpus, o inverso da frequência de documentos ($\boldsymbol{idf}$) para um termo $\boldsymbol{t}$ é definido por

$$\boldsymbol{idf_t = \log \frac{N}{df_t}}$$

INVERSO DA FREQUÊNCIA NOS DOCUMENTOS

- Exemplo:

term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

Manning, Raghavan e Schütze (2008)

TF-IDF

- O $\boldsymbol{tfidf}$ é o produto das duas métricas anteriores, e pode ser definido por

$$\boldsymbol{tfidf = tf\ x\ idf}$$

- Este cálculo é capaz de estabelecer uma **ponderação aos termos presentes em um documento**, considerando **tanto o próprio documento quanto seu relacionamento com o corpus textual analisado**

TF-IDF

- É importante ressaltar, no entanto, que embora o cálculo do $tfidf$ crie as representações para os textos, ainda é preciso considerar que os documentos de um mesmo corpus podem possuir tamanhos diferentes

- Neste caso, para corrigir esta distorção, é comum a realização de um procedimento de normalização, dando origem à medida $tfidf$ **normalizada**

- O cálculo do TF-IDF é utilizado no processo de vetorização para preencher a Bag of Words

Exemplo de Vetorização

EXEMPLO

- Seja $\Psi = \{ doc_1, doc_2, \ldots, docn \}$ um determinado corpus de documentos

- Vamos considerar a situação na qual $n = 3$, portanto,

$$\Psi = \{ doc_1, doc_2, doc_3 \}$$

doc1	Ambiente agradável e tranquilo. Comida e música com qualidade. Adoramos o Filé à Parmegiana.
doc2	O Filé à Parmegiana da cidade. Ambiente agradável e qualidade no atendimento.
doc3	O Filé à Parmegiana com fritas é uma delícia.

Silva, Peres e Boscarioli (2016)

EXEMPLO

- **Lista de tokens** obtida após a execução da análise lexical sobre o corpus

doc1	ambiente | agradável | e | tranquilo | comida | e | música | com | qualidade | adoramos | o | filé | à | parmegiana
doc2	o | filé | à | parmegiana | da | cidade | ambiente | agradável | e | qualidade | no | atendimento
doc3	o | filé | à | parmegiana | com | fritas | é | uma | delícia

Silva, Peres e Boscarioli (2016)

EXEMPLO

- Lista de tokens obtida **após a remoção das stopwords**

doc1	ambiente | agradável | tranquilo | comida | música | qualidade | adoramos | filé | parmegiana
doc2	filé | parmegiana | cidade | ambiente | agradável | qualidade | atendimento
doc3	filé | parmegiana | fritas | delícia

Silva, Peres e Boscarioli (2016)

EXEMPLO

- Lista resultante após aplicação da **radicalização**

doc1	ambient | agrad | tranquil | com | músic | qualidad | ador | fil | parmegian
doc2	fil | parmegiana | cidad | ambient | agrad | qualidad | atend
doc3	fil | parmegian | frit | delíci

Silva, Peres e Boscarioli (2016)

EXEMPLO

- **Dicionário de termos** obtidos

termo 1	ador	termo 8	fil
termo 2	agrad	termo 9	frit
termo 3	ambient	termo 10	músic
termo 4	atend	termo 11	parmegian
termo 5	cidad	termo 12	qualidad
termo 6	com	termo 13	tranquil
termo 7	delíc		

Silva, Peres e Boscarioli (2016)

EXEMPLO

- Conjunto de dados Ψ considerando uma **representação binária**

	ador	agrad	ambient	atend	cidad	com	delíc	fil	frit	músic	parmegian	qualidad	tranquil
Ψ	wte_1	wte_2	wte_3	wte_4	wte_5	wte_6	wte_7	wte_8	wte_9	wte_{10}	wte_{11}	wte_{12}	wte_{13}
doc_1	1	1	1	0	0	1	0	1	0	1	1	1	1
doc_2	0	1	1	1	1	0	0	1	0	0	1	1	0
doc_3	0	0	0	0	0	0	1	1	1	0	1	0	0

Silva, Peres e Boscarioli (2016)

EXEMPLO

- Conjunto de dados Ψ considerando uma **representação $tfidf$**

	ador	agrad	ambient	atend	cidad	com	delíc	fil	frit	músic	parmegian	qualidad	tranquil
Ψ	wte_1	wte_2	wte_3	wte_4	wte_5	wte_6	wte_7	wte_8	wte_9	wte_{10}	wte_{11}	wte_{12}	wte_{13}
doc_1	1,58	0,58	0,58	0,00	0,00	1,58	0,00	0,00	0,00	1,58	0,00	0,58	1,58
doc_2	0,00	0,58	0,58	1,58	1,58	0,00	0,00	0,00	0,00	0,00	0,00	0,58	0,00
doc_3	0,00	0,00	0,00	0,00	0,00	0,00	1,58	0,00	1,58	0,00	0,00	0,00	0,00

Silva, Peres e Boscarioli (2016)

EXEMPLO

- Conjunto de dados Ψ considerando uma **representação $tfidf$ normalizada**

	ador	agrad	ambient	atend	cidad	com	delíc	fil	frit	músic	parmegian	qualidad	tranquil
Ψ	wte_1	wte_2	wte_3	wte_4	wte_5	wte_6	wte_7	wte_8	wte_9	wte_{10}	wte_{11}	wte_{12}	wte_{13}
doc_1	0,18	0,06	0,06	0,00	0,00	0,18	0,00	0,00	0,00	0,18	0,00	0,06	0,18
doc_2	0,00	0,08	0,08	0,23	0,23	0,00	0,00	0,00	0,00	0,00	0,00	0,08	0,00
doc_3	0,00	0,00	0,00	0,00	0,00	0,00	0,40	0,00	0,40	0,00	0,00	0,00	0,00

Silva, Peres e Boscarioli (2016)

Text Mining: Aplicações

TEXT MINING: APLICAÇÕES

- Classificação de Documentos
- Clusterização (ou Categorização) de Documentos
- Sumarização de Documentos
- Similaridade de Documentos
- Reconhecimento de Entidades Nomeadas (ou Mencionadas)
- Análise de Sentimentos (Mineração de Opiniões)
- Sistemas de Pergunta e Resposta
- Chatbots
- entre outras

Desafios e Oportunidades

DESAFIOS

- Mineração de texto em outras línguas que não o inglês
- Padrões para os mineradores em língua portuguesa
- Relações léxico-semânticas

OPORTUNIDADES

- Muitos softwares disponíveis
- Pacotes, "bibliotecas", APIs (Application Programming Interfaces)
- Poucas linhas de código em Python, Java ou R, por exemplo
- Percorrimento da curva de aprendizagem é bastante rápido
- Cinema, publicação científica e inúmeras outras áreas de aplicação

MATERIAL CONSULTADO

- BUDER, Emily. **Can AI Predict a Movie's Success? Algorithmic Screenplay Service 'Scriptbook' Causes Major Backlash.** Disponível em: <https://nofilmschool.com/2017/04/scriptbook-black-list-screenwriting-ai-algorithm>. Acesso em: 01 maio 2018.
- GLADWELL, Malcolm. **The Formula:** What if you built a machine to predict hit movies?. Disponível em: <https://www.newyorker.com/magazine/2006/10/16/the-formula>. Acesso em: 01 maio 2018.
- HALL, Jacob. **WTF:** ScriptBook Will Predict If a Screenplay Will Make Money; Will Also Possibly, Maybe Trigger Nuclear Armageddon. Disponível em: <http://www.slashfilm.com/scriptbook-software/>. Acesso em: 01 maio 2018.
- WIKIPÉDIA. **Epagogix.** Disponível em: <https://en.wikipedia.org/wiki/Epagogix>. Acesso em: 01 maio 2018.
- WESTBROOK, Adam. **Robots are reading Hollywood scripts – and writers aren't happy.** Disponível em: <https://www.thememo.com/2017/04/20/scriptbook-blacklist-robots-reading-hollywood-scripts-writers-not-happy/>. Acesso em: 01 maio 2018.
- WIKIPÉDIA. **Gigli.** Disponível em: <https://pt.wikipedia.org/wiki/Gigli>. Acesso em: 01 maio 2018.
- WIKIPÉDIA. **The Hero with a Thousand Faces.** Disponível em: <https://pt.wikipedia.org/wiki/The_Hero_with_a_Thousand_Faces>. Acesso em: 01 maio 2018.
- WILCOCK, David. **The Synchronicity Key:** The Hidden Intelligence Guiding the Universe and You. New York: Dutton - Est. 1852, 2013. 528 p.
- PRESSBERG, Matt. **Swimming With Algorithms:** Can Software Predict The Next Hollywood Hit Based On The Script Alone?. Disponível em: <http://www.ibtimes.com/swimming-algorithms-can-software-predict-next-hollywood-hit-based-script-alone-2266229>. Acesso em: 01 maio 2018.

MATERIAL CONSULTADO

- APPLE INC. **Dicionário**. 2. ed. Cupertino (CA), EUA: macOS High Sierra, 2017.
- DIAS, Maria; MALHEIROS Marcelo. (2005). **Estudo de técnicas de radicalização para a Língua Portuguesa**. Disponível em: <https://www.researchgate.net/publication/242193490_Estudo_de_tecnicas_de_radicalizacao_para_a_Lingua_Portuguesa>. Acesso em 20 mar. 2018.
- GANTZ, John; REINSEL, David. Extracting value from chaos. IDC iview, v. 1142, n. 2011, p. 1-12, 2011. Disponível em: <http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf>. Acesso em: 27 nov. 2018.
- INGERSOLL, Grant S.; MORTON, Thomas S.; FARRIS, Andrew L.. **Taming Text:** How to find, organize and manipulate it. Shelter Island, NY (USA): Manning Publications Co., 2013. 298 p.
- MANNING, Christopher D.; RAGHAVAN**,** Prabhakar; SCHÜTZE, Hinrich. **Introduction to Information Retrieval**, Cambridge (UK): Cambridge University Press. 2008.
- NEVES, P. I. ; CORREA, D. A. ; GOLDSCHMIDT, R. R. ; MOURA, Ana Maria Carvalho ; CAVALCANTI, M. C. . UMA ANÁLISE SOBRE ABORDAGENS E FERRAMENTAS PARA EXTRAÇÃO DE INFORMAÇÃO. **Revista Militar de Ciência e Tecnologia** , v. 30, p. 32-58, 2013.
- SARKAR, Dipanjar. **Text Analytics with Python:** A Practical Real-World Approach to Gaining Actionable Insights from Your Data. Middletown, DE (USA): Apress, 2016. 385 p.
- SOUSA, Afonso F. (Org.). **Máximas e Mínimas do Barão de Itararé**. Rio de janeiro: Record, 1985.
- WEISS, Sholom M.; INDURKHYA, Nitin; ZHANG, Tong. **Fundamentals of Predictive Text Mining**. New York: Springer, 2010. 226 p. (Texts in Computer Science).

OBRIGADO!