Mineração do DHBB

Alexandre Rademaker (FGV/EMAp e IBM Research)
http://arademaker.github.io

http://emap.fgv.br/corpo-docente/alexandre-rademaker
http://researcher.ibm.com/person/br-alexrad

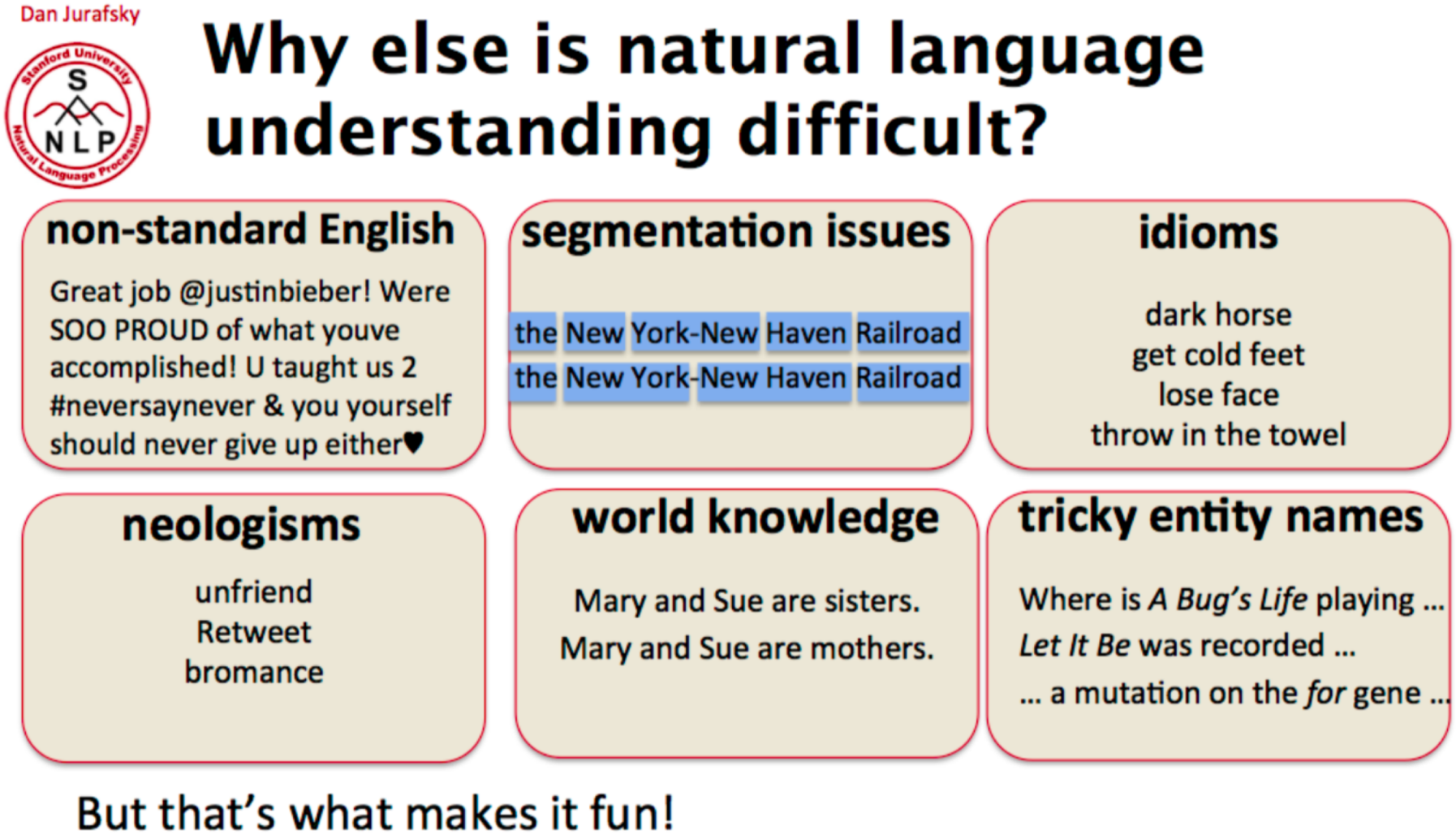
Colaboradores

- Fabricio Chalub (IBM)
- Bruno Cuconato (FGV/EMAp)
- Henrique Muniz (FGV/EMAp e IBM)
- Guilherme Passos (IBM)
- Claudia Freitas (PUC-Rio)
- Suemi Higuchi (CPDOC)
- Flávia M. da R. Pereira da Silva (IBM)
- Valéria de Paiva, Adam Pease, Gerard de Melo etc.

O que é Processamento de Linguagem Natural?

- Resposta à perguntas (IBM Watson ganhou o Jeopardy 2011)
- Extração de Informações (entidades nomeadas, eventos, relações etc)
- Expanção de consultas (via sinômimos)
- Análise de sentimentos (críticas em blogs e em sites online)
- Tradução
- Classificação de textos
- Sumarização
- Linguagens controladas . . .

O que é PLN?

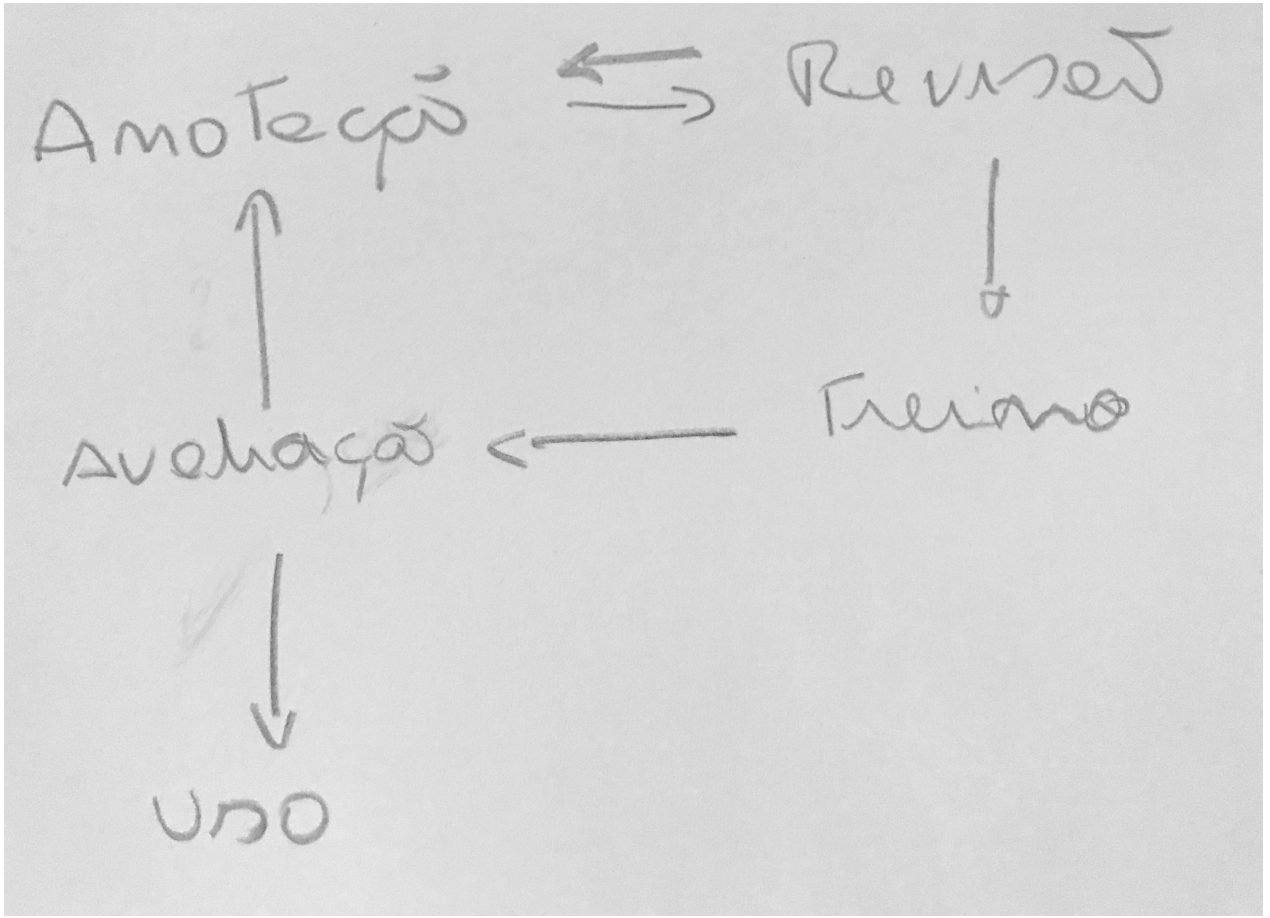

Dan Jurafsky
Why else is natural language understanding difficult?
non-standard English
Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥
segmentation issues
the New York-New Haven Railroad
the New York-New Haven Railroad
idioms
dark horse
get cold feet
lose face
throw in the towel
neologisms
unfriend
Retweet
bromance
world knowledge
Mary and Sue are sisters.
Mary and Sue are mothers.
tricky entity names
Where is A Bug's Life playing ...
Let It Be was recorded ...
... a mutation on the for gene ...
But that's what makes it fun!

Mineração do DHBB

- Extração de Informações ou “reconhecimento de entidades mencionadas” (https://en.wikipedia.org/wiki/Information_extraction)
- Avaliação conjunta HAREM da Linguateca (2008)
- Identificação no texto de menções à:
  - Entidades
  - Relações
  - Co-Referências
- Demo http://184.172.236.221:30000/home

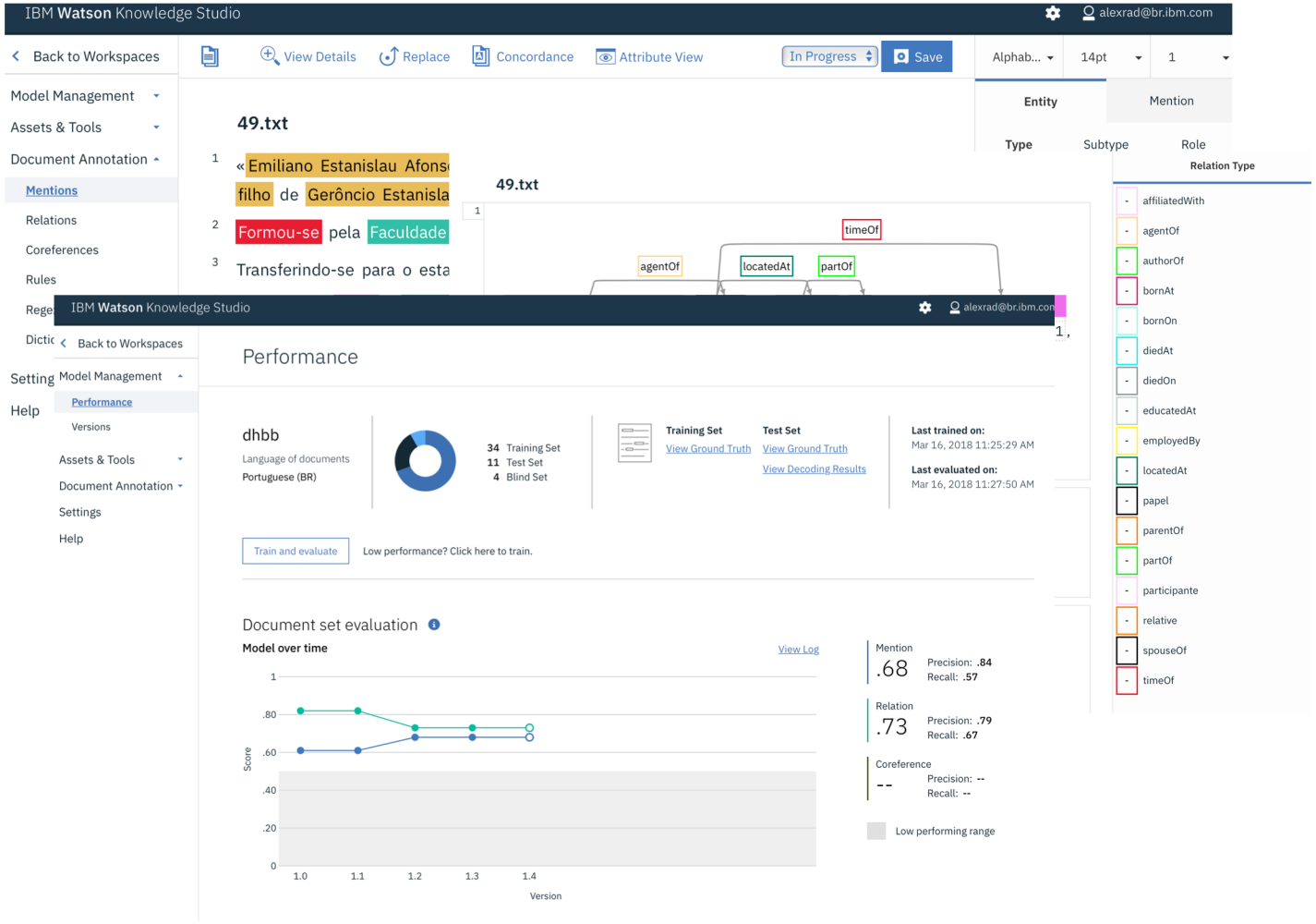
Como?

- Modelagem do domínio: o que extrair? O que é relevante?
- Abordagens: **estatísticas** versus “linguisticamente motivadas”
- Combinação das abordagens.



Ferramentas IBM (Watson)

- Watson Knowledge Studio
- Serviços
  - Natural Language Understanding
  - Watson Discovery Service
  - IBM Cloud

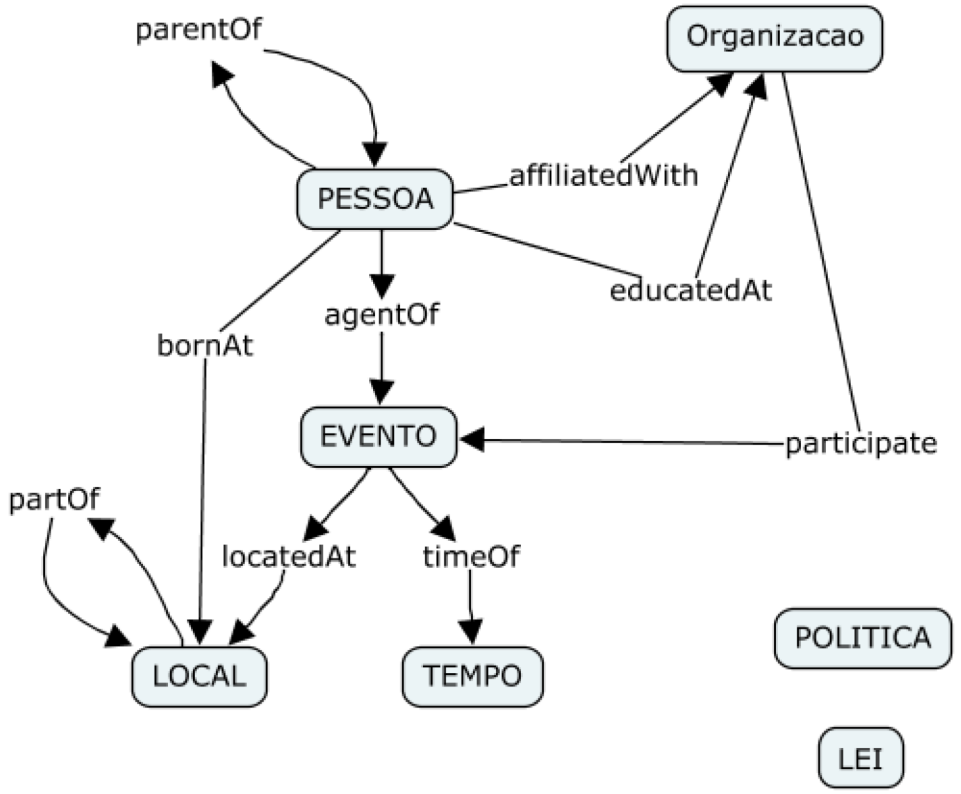


https://youtu.be/r2xYHW0iyZM

Modelagem

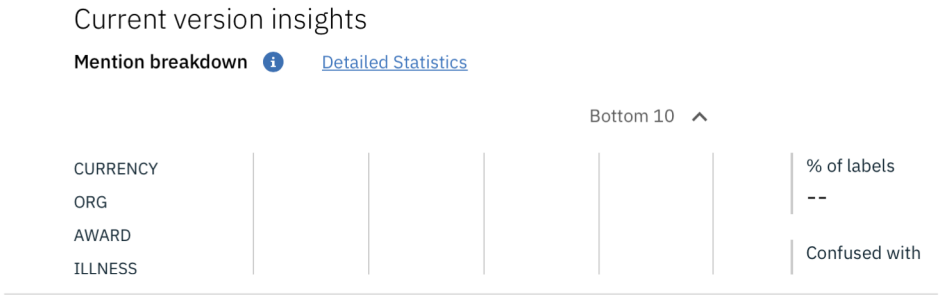
- O que existe e como aparecem nos textos?
- Que informações gostaríamos de ter e com qual estrutura?
- Que consultas queremos poder fazer?
- Qual a capacidade da ferramenta de aprender? Como saber o que não foi aprendido e o motivo?

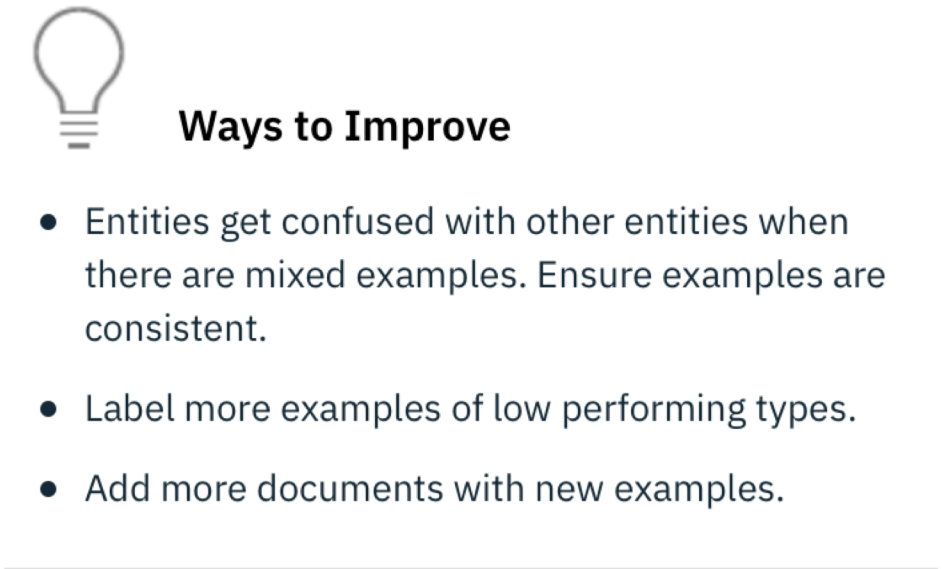
- Caso DHBB: muita regularidade nos textos!

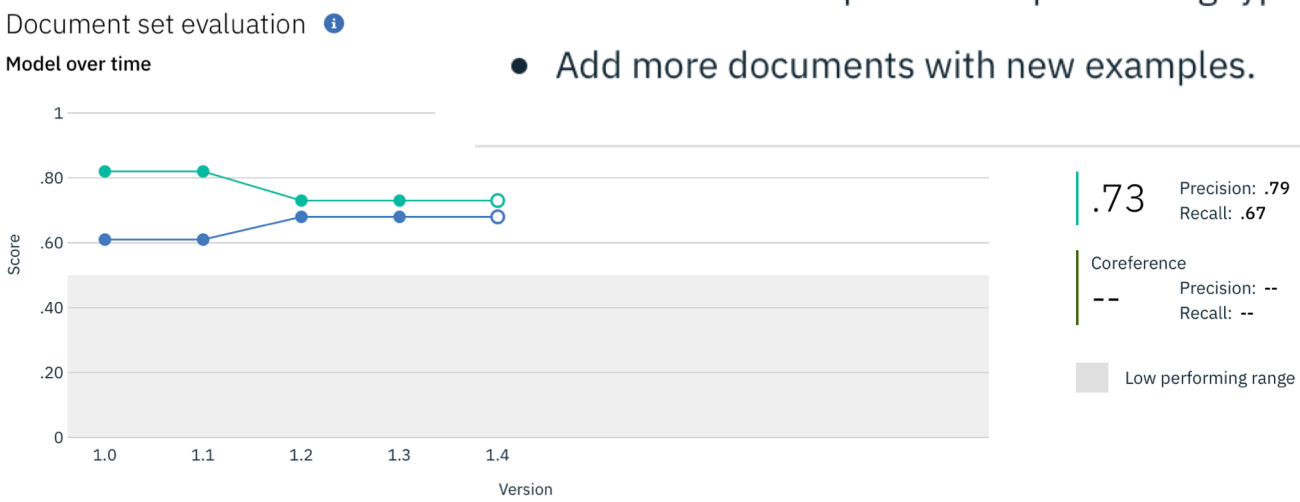


Resultados atuais

- 50 documentos anotados
- Foco inicial parágrafos sobre: filiação, nascimento, educação, conjuge e falecimento.
- Performance de .68 (entidades) e .73 (relações)
- Demo online
- Melhoria Incremental







Próximos passos

- Melhorar o dataset (promover avaliação e competição)
- Melhorar interface de busca
  - Maior flexibilidade
  - Flexibilidade vs facilidade
- Corpus como testes para métodos ‘linguisticamente motivados’ ➔ adaptáveis!

anotacoes.org

```
#+Title: Data Analytics
#+Author: Alexandre Rademaker
#+PROPERTY: session *R*
#+PROPERTY: cache yes

#+BEGIN_SRC R :results output
dados <- read.csv("anotacoes.csv", as.is = TRUE)
table(dados$type)
#+END_SRC

#+RESULTS:
:
:      AGE    AWARD CURRENCY      DOC    EVENT      LAW    LOCAL      ORG
:        6        2        8       75      232       81      789     1849
:   PERSON      POL      TAX     TIME
:     1154       46       27     1464

#+BEGIN_SRC R :results output
dados <- read.csv("anotacoes.csv", as.is = TRUE)
tmp <- as.data.frame(table(dados$mention))
tmp.1 <- subset(tmp, Freq > 28)
tmp.1[order(tmp.1$Freq),]
#+END_SRC

#+RESULTS:
#+begin_example
                    Var1 Freq
2215                 PDS   30
574         Almino Afonso   32
2331                PTB   36
475             Abi-Ackel   39
2323                PSD   39
756  Câmara dos Deputados   40
2390       Rio de Janeiro   41
2256               pleito   42
1988               nasceu   48
733                Brasil   50
1333                filho   52
2260                PMDB   52
2451            São Paulo   67
750                Câmara   73
#+end_example
```

```
-:**-  anotacoes.org   All (21,0)     [(Org)]
Code block evaluation complete.
```

Desafio OAB

Alexandre Rademaker (FGV/EMAp e IBM Research)
http://arademaker.github.io

http://emap.fgv.br/corpo-docente/alexandre-rademaker
http://researcher.ibm.com/person/br-alexrad

Colaboradores

- Fabricio Chalub (IBM)
- Pedro Delfino (FGV/EMAp e FGV/Direito)
- Bruno Cuconato (FGV/EMAp)
- Henrique Muniz (FGV/EMAp e IBM)
- Guilherme Passos (IBM)
- Valéria de Paiva, Adam Pease, Gerard de Melo etc.

O desafio da OAB

- Exame nacional da OAB
- Os exames fornecem uma excelente referência para o desempenho de sistemas de informações jurídicas.
- Interessante problema que requer >> entendimento
- Processamento superficial seguido de processamento profundo.
- No ML approach! GOFAI ("Good Old-Fashioned Artificial Intelligence”)

O Exame

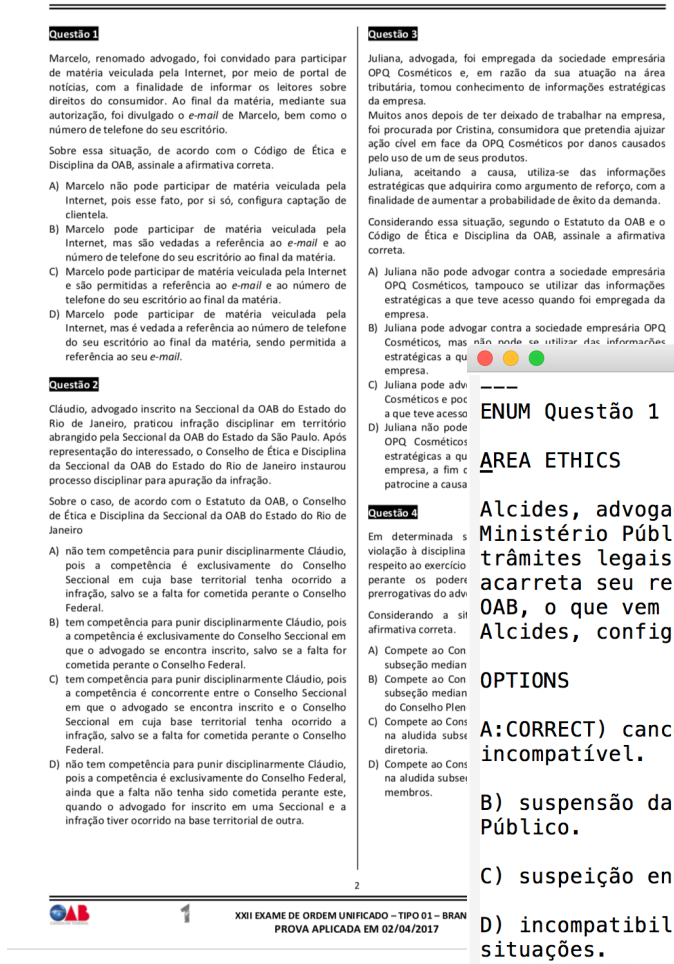
- Apenas em 2010 os exames foram unificados nacionalmente.
- Dois estágios, estamos trabalhando apenas no primeiro, questões multiplas escolhas.
- 80 questões com 4 opções. Performance minima de 50%.
- A cada ano, 3 aplicações do exame.
- O exame tem uma taxa de 80% de reprovação. Em julho de 2017, a reprovação foi de 86%

Questões por assunto

area	#	(%)	area	#	(%)
Ethics	10	65	Constitutional Law	7	42
Consumer's Law	2	56	Civil Procedures	6	40
Children's Law	2	54	Philosophy	2	40
Criminal Procedures	5	47	Labor's Law Proc.	6	40
Regulatory Law	6	47	Criminal Law	6	38
Human Rights	3	47	International Law	2	37
Civil Law	7	44	Business Law	5	33
Environmental	2	43	Taxes	4	42
Labor's Law	5	42			

Preparação dos dados

- Exportação de PDFs
- Testes
- Total de 25 exames e 2061 questões.
- **Reproducible research!**



Questão 1

Marcelo, renomado advogado, foi convidado para participar de matéria veiculada pela Internet, por meio de portal de notícias, com a finalidade de informar os leitores sobre direitos do consumidor. Ao final da matéria, mediante sua autorização, foi divulgado o *e-mail* de Marcelo, bem como o número de telefone do seu escritório.

Sobre essa situação, de acordo com o Código de Ética e Disciplina da OAB, assinale a afirmativa correta.

A) Marcelo não pode participar de matéria veiculada pela Internet, pois esse fato, por si só, configura captação de clientela.
B) Marcelo pode participar de matéria veiculada pela Internet, mas são vedadas a referência ao *e-mail* e ao número de telefone do seu escritório ao final da matéria.
C) Marcelo pode participar de matéria veiculada pela Internet e são permitidas a referência ao *e-mail* e ao número de telefone do seu escritório ao final da matéria.
D) Marcelo pode participar de matéria veiculada pela Internet, mas é vedada a referência ao número de telefone do seu escritório ao final da matéria, sendo permitida a referência ao seu *e-mail*.

Questão 2

Cláudio, advogado inscrito na Seccional da OAB do Estado do Rio de Janeiro, praticou infração disciplinar em território abrangido pela Seccional da OAB do Estado da São Paulo. Após representação do interessado, o Conselho de Ética e Disciplina da Seccional da OAB do Estado do Rio de Janeiro instaurou processo disciplinar para apuração da infração.

Sobre o caso, de acordo com o Estatuto da OAB, o Conselho de Ética e Disciplina da Seccional da OAB do Estado do Rio de Janeiro

A) não tem competência para punir disciplinarmente Cláudio, pois a competência é exclusivamente do Conselho Seccional em cuja base territorial tenha ocorrido a infração, salvo se a falta for cometida perante o Conselho Federal.
B) tem competência para punir disciplinarmente Cláudio, pois a competência é exclusivamente do Conselho Seccional em que o advogado se encontra inscrito, salvo se a falta for cometida perante o Conselho Federal.
C) tem competência para punir disciplinarmente Cláudio, pois a competência é concorrente entre o Conselho Seccional em que o advogado se encontra inscrito e o Conselho Seccional em cuja base territorial tenha ocorrido a infração, salvo se a falta for cometida perante o Conselho Federal.
D) não tem competência para punir disciplinarmente Cláudio, pois a competência é exclusivamente do Conselho Federal, ainda que a falta não tenha sido cometida perante este, quando o advogado for inscrito em uma Seccional e a infração tiver ocorrido na base territorial de outra.

Questão 3

Juliana, advogada, foi empregada da sociedade empresária OPQ Cosméticos e, em razão da sua atuação na área tributária, tomou conhecimento de informações estratégicas da empresa.
Muitos anos depois de ter deixado de trabalhar na empresa, foi procurada por Cristina, consumidora que pretendia ajuizar ação cível em face da OPQ Cosméticos por danos causados pelo uso de um de seus produtos.
Juliana, aceitando a causa, utiliza-se das informações estratégicas que adquirira como argumento de reforço, com a finalidade de aumentar a probabilidade de êxito da demanda.

Considerando essa situação, segundo o Estatuto da OAB e o Código de Ética e Disciplina da OAB, assinale a afirmativa correta.

A) Juliana não pode advogar contra a sociedade empresária OPQ Cosméticos, tampouco se utilizar das informações estratégicas a que teve acesso quando foi empregada da empresa.
B) Juliana pode advogar contra a sociedade empresária OPQ Cosméticos, mas não pode se utilizar das informações estratégicas a qu
empresa.
C) Juliana pode adv
Cosméticos e poc
a que teve acesso
D) Juliana não pode
OPQ Cosméticos
estratégicas a qu
empresa, a fim c
patrocine a causa

Questão 4

Em determinada s
violação à disciplina
respeito ao exercício
perante os poderes
prerrogativas do adv

Considerando a sit
afirmativa correta.

A) Compete ao Con
subseção median
B) Compete ao Con
subseção median
do Conselho Plen
C) Compete ao Cons
na aludida subse
diretoria.
D) Compete ao Cons
na aludida subse
membros.

2

1 XXII EXAME DE ORDEM UNIFICADO – TIPO 01 – BRAN
PROVA APLICADA EM 02/04/2017

2011-05.txt

```
---
ENUM Questão 1

AREA ETHICS

Alcides, advogado de longa data, resolve realizar concurso para o
Ministério Público, vindo a ser aprovado em primeiro lugar. Após os
trâmites legais, é designada data para a sua posse, circunstância que
acarreta seu requerimento para suspender sua inscrição nos quadros da
OAB, o que vem a ser indeferido. No caso em comento, em relação a
Alcides, configura-se situação de

OPTIONS

A:CORRECT) cancelamento da inscrição por assunção de cargo
incompatível.

B) suspensão da inscrição até a aposentadoria do membro do Ministério
Público.

C) suspeição enquanto permanecer no cargo.

D) incompatibilidade, podendo atuar, como advogado, em determinadas
situações.

---
ENUM Questão 2

AREA ETHICS

Na Secretaria Municipal de Fazenda, tramita procedimento
administrativo relacionado à imposição do IPTU em determinada área
urbana. O proprietário do imóvel contrata o advogado Juliano para
solucionar a questão. Portando mandato extrajudicial, o advogado
dirige-se ao local e, em face dos seus conhecimentos pessoais, obtém o
U:---  2011-05.txt    Top (4,0)     Git-master  (Text ARev)
Wrote /Users/arademaker/work/oab-data/OAB/raw/2011-05.txt
```

https://github.com/own-pt/oab-exams

Leis

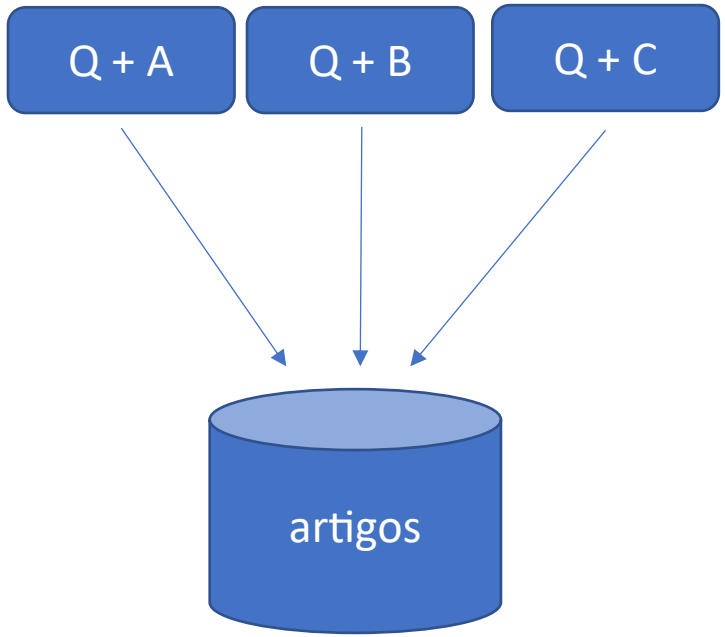
- Como tartar as leis?
- Estrutura das leis é relevante.
- Portal e padrão de codificação de leis (LexML)

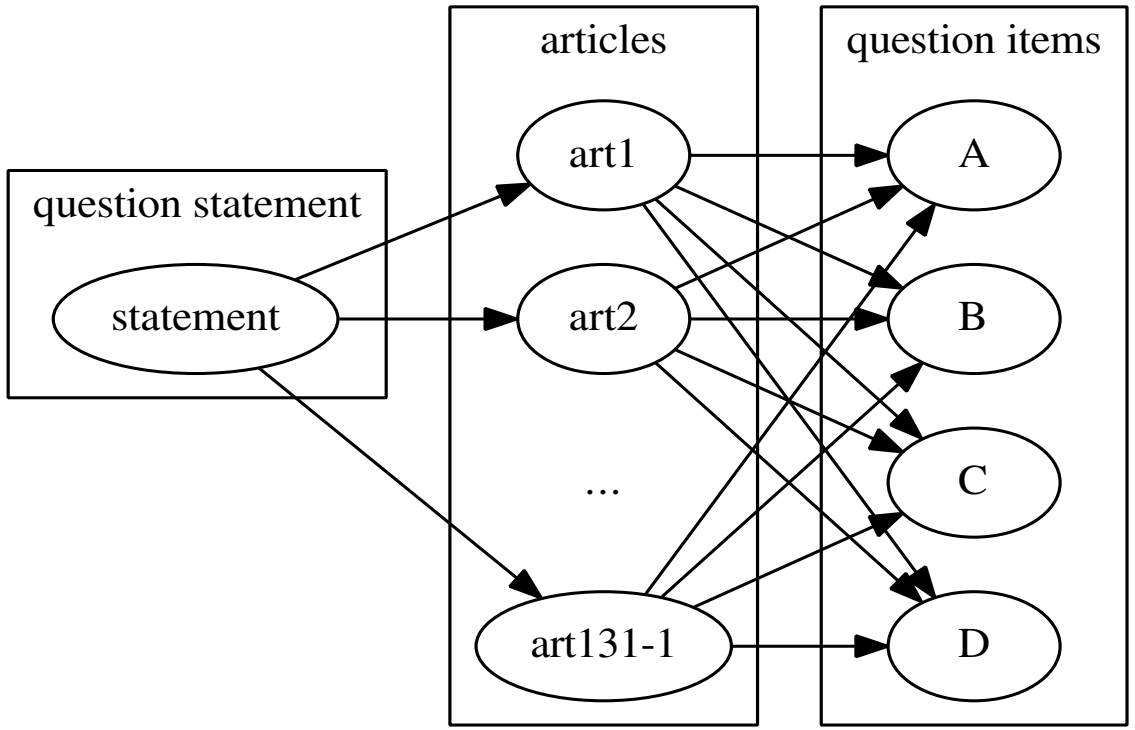
http://www.lexml.gov.br

Conjunto manual de verificação

- Escolhemos 60 questões de ética e constitucional
- Um dos pesquisadores identificou para cada questão qual artigo de qual lei justifica a resposta.
- Em geral um artigo por questão.
- Em Ética, 3 normas (leis).
- Em constitucional, decisões do STJ e outros.
- Como lidar com revisões de leis?

Experimentos: IR e graph





Resultados parciais (baselines)

Os jovens Rodrigo, 30 anos, e Bibiana, 35 anos, devidamente inscritos em certa seccional da OAB, desejam candidatar-se, pela primeira vez, a cargos de diretoria do Conselho Seccional respectivo. Rodrigo está regularmente inscrito na referida seccional da OAB há seis anos, sendo dois anos como estagiário. Bibiana, por sua vez, exerceu regularmente a profissão por três anos, após a conclusão do curso de Direito. Contudo, afastou-se por dois anos e retornou à advocacia há um ano. Ambos não exercem funções incompatíveis com a advocacia, ou cargos exoneráveis ad nutum. Tampouco integram listas para provimento de cargos em tribunais ou ostentam condenação por infração disciplinar. Bibiana e Rodrigo estão em dia com suas anuidades.

Considerando a situação narrada, assinale a afirmativa correta.

OPTIONS

A) Apenas Bibiana preenche as condições de elegibilidade para os cargos.

B) Apenas Rodrigo preenche as condições de elegibilidade para os cargos.

C) Bibiana e Rodrigo preenchem as condições de elegibilidade para os cargos.

D:CORRECT) Nenhum dos dois advogados preenche as condições de elegibilidade para os cargos.

Paulo é contratado por Pedro para promover ação com pedido condenatório em face de Alexandre, por danos causados ao animal de sua propriedade. Em decorrência do processo, houve condenação do réu ao pagamento de indenização ao autor, fixados honorários de sucumbência correspondentes a dez por cento do apurado em cumprimento de sentença. O réu ofertou apelação contra a sentença proferida na fase cognitiva. Ainda pendente o julgamento do recurso, Pedro decide revogar o mandato judicial conferido a Paulo, desobrigando-se de pagar os honorários contratualmente ajustados.

Nos termos do Código de Ética da OAB, a revogação do mandato judicial, por vontade de Pedro,

OPTIONS

A:CORRECT) não o desobriga do pagamento das verbas honorárias contratadas.

B) desobriga-o do pagamento das verbas honorárias contratadas.

C) desobriga-o do pagamento das verbas honorárias contratadas e da verba sucumbencial.

D) não o desobriga do pagamento das verbas honorárias sucumbenciais, mas o desobriga das verbas contratadas.

Tipos de questões

1. Narrativas e respostas longas
2. Factuais
3. Nomes de normas ou entidades
4. Negativa

Mais que extração de informações!

Leonardo, and Bruno. Luana, 35 years old, was already a manager in a bank when she graduated. Leonardo, 30 years, is mayor of the municipality of Pontal. Bruno, 28 years old, is a military policeman in the same municipality. The three want to practice law in the private sector. Considering the incompatibilities and impediments to practice, please select the correct answer.

A) Luana is not prohibited from practicing law because she is an employee of a private institution, so there are no impediments or incompatibilities.

B) Bruno, like all other civil servants, is only prohibited from practicing the law against the government agency that remunerates him.

C) The three graduates, Luana, Leonardo, and Bruno, have functions incompatible with legal practice. They are therefore prohibited from exercising private practice. (CORRECT)

D) Leonardo is banned from practicing law only against or in favor of legal entities from the public sector, public companies, mixed-capital companies, public foundations, parastatal entities or concessionaire corporations or public service licensees.

(2016 OAB exam, 19th edition, question 4)

Art. 28. A advocacia é incompatível, mesmo em causa própria, com as seguintes atividades:

I - chefe do Poder Executivo e membros da Mesa do Poder Legislativo e seus substitutos legais;

II - membros de órgãos do Poder Judiciário, do Ministério Público, dos tribunais e conselhos de contas, dos juizados especiais, da justiça de paz, juízes classistas, bem como de todos os que exerçam função de julgamento em órgãos de deliberação coletiva da administração pública direta e indireta;

III - ocupantes de cargos ou funções de direção em Órgãos da Administração Pública direta ou indireta, em suas fundações e em suas empresas controladas ou concessionárias de serviço público;

IV - ocupantes de cargos ou funções vinculados direta ou indiretamente a qualquer órgão do Poder Judiciário e os que exercem serviços notariais e de registro;

V - ocupantes de cargos ou funções vinculados direta ou indiretamente a atividade policial de qualquer natureza;

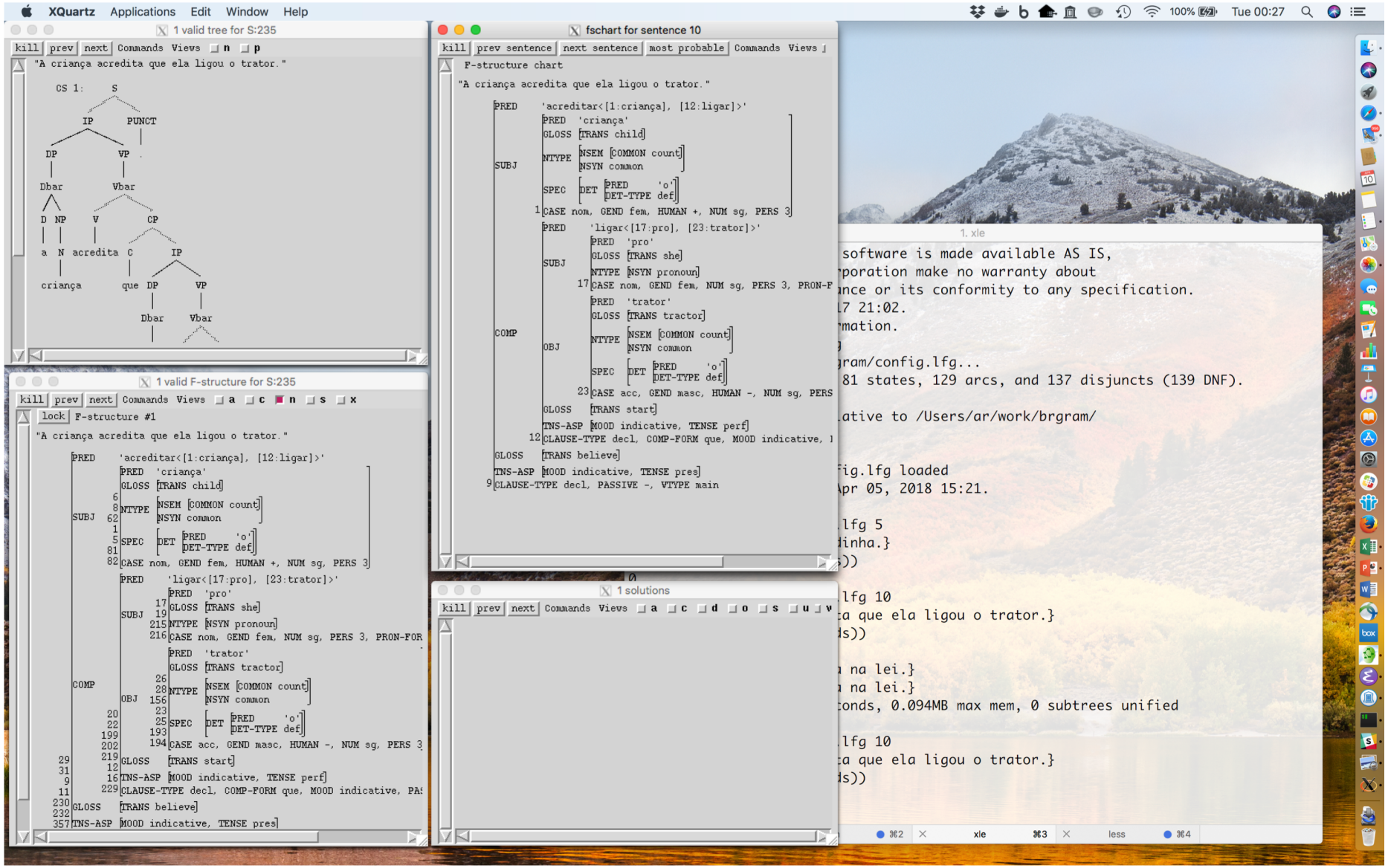
VI - militares de qualquer natureza, na ativa;

VII - ocupantes de cargos ou funções que tenham competência de lançamento, arrecadação ou fiscalização de tributos e contribuições parafiscais;

VIII - ocupantes de funções de direção e gerência em instituições financeiras, inclusive privadas.

§ 1º A incompatibilidade permanece mesmo que o ocupante do cargo ou função deixe de exercê-lo temporariamente.
§ 2º Não se incluem nas hipóteses do inciso III os que não detenham poder de decisão relevante sobre interesses de terceiro, a juízo do conselho competente da OAB, bem como a administração acadêmica diretamente relacionada ao magistério jurídico.

Next


XQuartz Applications Edit Window Help
100% Tue 00:27
1 valid tree for S:235
kill prev next Commands Views n p
"A criança acredita que ela ligou o trator."
CS 1: S
IP PUNCT
DP VP
Dbar Vbar
D NP V CP
a N acredita C IP
criança que DP VP
Dbar Vbar
fschart for sentence 10
kill prev sentence next sentence most probable Commands Views
F-structure chart
"A criança acredita que ela ligou o trator."
PRED 'acreditar<[1:criança], [12:ligar]>'
PRED 'criança'
GLOSS [TRANS child]
NTYPE [NSEM [COMMON count] NSYN common]
SUBJ
SPEC [DET [PRED 'o' DET-TYPE def]]
1 CASE nom, GEND fem, HUMAN +, NUM sg, PERS 3
PRED 'ligar<[17:pro], [23:trator]>'
PRED 'pro'
GLOSS [TRANS she]
SUBJ
NTYPE [NSYN pronoun]
17 CASE nom, GEND fem, NUM sg, PERS 3, PRON-F
PRED 'trator'
GLOSS [TRANS tractor]
COMP
NTYPE [NSEM [COMMON count] NSYN common]
OBJ
SPEC [DET [PRED 'o' DET-TYPE def]]
23 CASE acc, GEND masc, HUMAN -, NUM sg, PERS
GLOSS [TRANS start]
TNS-ASP [MOOD indicative, TENSE perf]
12 CLAUSE-TYPE decl, COMP-FORM que, MOOD indicative,
GLOSS [TRANS believe]
TNS-ASP [MOOD indicative, TENSE pres]
9 CLAUSE-TYPE decl, PASSIVE -, VTYPE main
1 valid F-structure for S:235
kill prev next Commands Views a c n s x
lock F-structure #1
"A criança acredita que ela ligou o trator."
PRED 'acreditar<[1:criança], [12:ligar]>'
PRED 'criança'
GLOSS [TRANS child]
6
8 NTYPE [NSEM [COMMON count] NSYN common]
SUBJ 62
1
5 SPEC [DET [PRED 'o' DET-TYPE def]]
81
82 CASE nom, GEND fem, HUMAN +, NUM sg, PERS 3
PRED 'ligar<[17:pro], [23:trator]>'
PRED 'pro'
17 GLOSS [TRANS she]
SUBJ 19
215 NTYPE [NSYN pronoun]
216 CASE nom, GEND fem, NUM sg, PERS 3, PRON-FOR
PRED 'trator'
GLOSS [TRANS tractor]
26
COMP 28 NTYPE [NSEM [COMMON count] NSYN common]
OBJ 156
20 23
22 25 SPEC [DET [PRED 'o' DET-TYPE def]]
199 193
202 194 CASE acc, GEND masc, HUMAN -, NUM sg, PERS 3
29 219 GLOSS [TRANS start]
31 12
9 16 TNS-ASP [MOOD indicative, TENSE perf]
11 229 CLAUSE-TYPE decl, COMP-FORM que, MOOD indicative, PA
230 GLOSS [TRANS believe]
232
357 TNS-ASP [MOOD indicative, TENSE pres]
1 solutions
kill prev next Commands Views a c d o s u v
1. xle
software is made available AS IS,
poration make no warranty about
nce or its conformity to any specification.
7 21:02.
mation.
gram/config.lfg...
81 states, 129 arcs, and 137 disjuncts (139 DNF).
ative to /Users/ar/work/brgram/
ig.lfg loaded
pr 05, 2018 15:21.
lfg 5
inha.}
))
lfg 10
a que ela ligou o trator.}
s))
na lei.}
na lei.}
conds, 0.094MB max mem, 0 subtrees unified
lfg 10
a que ela ligou o trator.}
s))
⌘2 xle ⌘3 less ⌘4

Obrigado! Thanks!

arademaker@gmail.com

alexrad@br.ibm.com