Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Trabalho 1 - Resultados

Profa. Giseli Rabello Lopes

2019 / 1

Roteiro

- Sistemas usados pelos participantes
- Resultados por consultas
  - Descrição das consultas
  - Melhores desempenhos em termos de MAP
- Avaliação geral (para as 10 consultas)
  - Curvas de Recall x Precision Interpolada
  - MAP, R-precision, relevantes retornados

Sistemas usados pelos participantes

Sistemas de RI

elasticsearch

(Whoosh)

Xapian

Resultados por consulta

Descrição das consultas
Melhores desempenhos em termos de MAP

Consulta 1

Número de relevantes: **18**

<top>
<num> 1 </num>
<PT-title> Boicotes de consumidores </PT-title>
<PT-desc> Encontrar documentos que descrevam ou discutam o impacto de boicotes por consumidores. </PT-desc>
<PT-narr> Documentos relevantes devem relatar discussões ou pontos de vista sobre a eficácia de boicotes pelos consumidores. São também relevantes as questões morais envolvidas nesses boicotes. Apenas boicotes feitos por consumidores são relevantes, boicotes políticos são ignorados. </PT-narr>
</top>

Consulta 1

- Thiago Paixão
  - Sistema: *Whoosh*
  - MAP: **52,17%**

Remoção Stopwords, Stemmer Português, caixa baixa

BM25 (padrão b=0.75, k1=1.2)

```
boicotes de consumidores NOT politicos
```

```
num_ret                 1       21
num_rel                 1       18
num_rel_ret             1       13
```

Consulta 2

Número de relevantes: **11**

```
<top>
<num> 2 </num>
<PT-title> Actividades da ETA em França </PT-
title>
<PT-desc> Encontrar documentos sobre as
actividades do grupo terrorista basco ETA em
França. </PT-desc>
<PT-narr> Apenas documentos descrevendo as
actividades do grupo terrorista basco ETA em
território francês, sejam armadas, políticas,
financeiras, ou outras, são relevantes. As
actividades da ETA em Espanha não são
interessantes. </PT-narr>
</top>
```

Consulta 2

- Gustavo Monteiro e Letícia
  - Sistema: *Apache Solr*
  - MAP: **45,69%**

LowerCase, remoçao de stopwords, processamento de stem (leve)
BM25 (k=1.2 e b=0.75)

```
text:(terrorista~ +ETA +França~)
```

num_ret	2	28
num_rel	2	11
num_rel_ret	2	11

Consulta 3

Número de relevantes: **11**

```
<top>
<num> 3 </num>
<PT-title> Filmes passados na Escócia </PT-
title>
<PT-desc> Encontrar informação sobre filmes cuja
acção decorre na Escócia. </PT-desc>
<PT-narr> Os documentos relevantes devem
mencionar filmes cuja acção se passa na Escócia.
Filmes filmados na Escócia, mas cuja acção
decorre noutro local, não são relevantes. Peças,
tais como «Macbeth», não são relevantes, mas as
suas versões filmadas são. Igualmente de
interesse são todos os documentários que incluam
cenas escocesas. </PT-narr>
</top>
```

Consulta 3

- Thiago Paixão
  - Sistema: *Whoosh*
  - MAP: **31,06%**

Remoção Stopwords, Stemmer Português, caixa baixa

BM25 (padrão b=0.75, k1=1.2)

```
filmes escócia
```

```
num_ret                 3       15
num_rel                 3       11
num_rel_ret             3       4
```

Consulta 4

Número de relevantes: **8**

```
<top>
<num> 4 </num>
<PT-title> Tratamento de resíduos industriais </
PT-title>
<PT-desc> Encontrar documentos descrevendo
métodos usados para o tratamento ou remoção de
resíduos industriais. </PT-desc>
<PT-narr> Documentos devem mencionar não apenas
os diferentes meios de tratar os resíduos
industriais, tais como incineração ou
reciclagem, mas também indicar o país, cidade,
ou região onde o método é empregado. </PT-narr>
</top>
```

Consulta 4

- Natasha e Guilherme
  - Sistema: *Tantivy*
  - MAP: **49,33%**

LowerCase, stemmer RSLP (NLTK), remoção de stopwords (NLTK)
BM25 (*default*)

```
trat residu industr
```

```
num_ret                 4       100
num_rel                 4       8
num_rel_ret             4       6
```

Consulta 5

Número de relevantes: **31**

```
<top>
<num> 5 </num>
<PT-title> Desemprego na Europa </PT-
title>
<PT-desc> Encontrar informação e números
sobre o nível de desemprego em países
europeus. </PT-desc>
<PT-narr> Documentos relevantes devem
discutir o problema do desemprego em um ou
mais países europeus, fornecendo números
relativos às taxas de desemprego em causa.
</PT-narr>
</top>
```

Consulta 5

- Natasha e Guilherme
  - Sistema: *Tantivy*
  - MAP: **46,03%**

LowerCase, stemmer RSLP (NLTK), remoção de stopwords (NLTK)
BM25 (*default*)

```
desempreg europ
```

```
num_ret                 5       100
num_rel                 5       31
num_rel_ret             5       24
```

Consulta 6

Número de relevantes: **111**

```
<top>
<num> 6 </num>
<PT-title> Celebrações de centenários </PT-
title>
<PT-desc> Encontrar documentos relatando a
celebração do 100º aniversário de qualquer
evento. </PT-desc>
<PT-narr> Documentos relevantes devem conter
informação sobre o centésimo aniversário do
nascimento ou da morte de uma pessoa famosa ou a
passagem de cem anos após um acontecimento
histórico significativo. </PT-narr>
</top>
```

Consulta 6

- Thiago Coelho
  - Sistema: *Terrier*
  - MAP: **45,42%**

Com Remoção de Stopwords (lista externa - GitHub), Portuguese Snowball Stemmer
DPH (derivado do *framework* DFR - *Divergence from Randomness* - probabilístico)

```
+centenario +celebracao
```

```
num_ret                 6       100
num_rel                 6       111
num_rel_ret             6       58
```

Consulta 7

Número de relevantes: **2**

```
<top>
<num> 7 </num>
<PT-title> Espécies em vias de extinção </
PT-title>
<PT-desc> Que medidas estão a ser tomadas
na Europa para proteger animais ou plantas
em vias de extinção? </PT-desc>
<PT-narr> Documentos relevantes devem
discutir medidas tomadas para salvaguardar
espécies de animais ou plantas em vias de
extinção em países europeus. </PT-narr>
</top>
```

Consulta 7

- André e Ingrid
  - Sistema: *Lucene*
  - MAP: **55%**

Com Remoção Stopwords (*default*), Com Stemmer para português
BM25 (k1 = 1.2 e b = 0.75)

Espécies em vias de extinção

num_ret	7	100
num_rel	7	2
num_rel_ret	7	2

Melhor Resultado (para uma consulta)

1. FSP950224-081: 6.1609006 ✔
2. FSP950808-070: 6.045853 ✖

...

19. FSP950723-046: 5.2278304 ✔

...

VALMIR DENARDIN
Da Agência Folha, em Sorocaba

FSP950224-081 ✔

O trabalho de preservação do lobo-guará desenvolvido pelo zoológico municipal Quinzinho de Barros, de Sorocaba (87 km de SP), será destaque em um programa da rede de TV inglesa BBC.
O programa "Nature" da emissora de Londres está elaborando um documentário sobre iniciativas realizadas em todo o mundo para a preservação de **espécies** ameaçadas de **extinção.** Só mais um projeto brasileiro fará parte do documentário: o trabalho do pesquisador Cláudio Pádua, de Piracicaba (SP), com o mico-leão-preto.
Segundo a bióloga do zôo de Sorocaba, Cecília Pessutti, 29, as pesquisas com o lobo-guará animal típico do cerrado brasileiro e sob risco de **extinção** começaram há cinco anos.
"Estamos estudando animais em cativeiro e também os que vivem na natureza", disse Pessutti, coordenadora do Grupo de Estudos de Canídeos Brasileiros, mantido pela Sociedade de Zoológicos do Brasil.
Quinze zoológicos de todo o país integram o grupo que pesquisa o lobo-guará. Entre as atividades estão a reprodução da espécie em cativeiro, a contagem da população de lobos-guarás e a localização de áreas públicas e particulares que possam abrigar os animais que forem devolvidos à natureza.
O zoológico de Sorocaba mantém dois casais de lobo-guará, que em abril devem entrar no período de reprodução. O zôo é um dos mais importantes do país. Possui 1.400 animais, de 350 **espécies**, e é reconhecido internacionalmente.
O programa da BBC é patrocinado pela Fundação Jersey, um centro inglês de reprodução de **espécies** ameaçadas de **extinção**. As gravações foram realizadas na semana passada no parque ecológico de São Carlos (SP). "Esse parque possui lobos-guarás e ocupa uma área de cerrado, que é o habitat natural da espécie", afirmou Pessutti. "A BBC quis mostrar o animal como ele vive na natureza."

Procriação de animais em extinção em Sorocaba ganha destaque na imprensa internacional

VALMIR DENARDIN

Da Agência Folha, em Sorocaba

O zoológico municipal Quinzinho de Barros, de Sorocaba (87 km a oeste de São Paulo), se transformou em um dos principais centros brasileiros de reprodução de animais ameaçados de **extinção**.

Em cinco anos, o zôo conseguiu a procriação de sete **espécies** de mamíferos e aves da fauna brasileira sob risco de desaparecimento: lobo-guará, ararajuba, mico-leão-de-cara-dourada, veado-campeiro, macaco-aranha-de-testa-branca, macaco-barrigudo e tamanduá-bandeira.

Segundo o Ibama (Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis), há no país 207 **espécies** ameaçadas de **extinção**. A principal causa é o avanço de cidades e áreas de agricultura sobre as regiões de mata.

O Ibama não possui um ranking dos zoológicos do país que mais conseguiram reproduzir animais em **extinção**.

``Mas o zôo de Sorocaba é um dos mais bem sucedidos nesta área, principalmente porque conseguiu a procriação de um grande número de **espécies**", diz a chefe do Departamento de Vida Silvestre do Ibama, Iolita Bampi, 35.

O Quinzinho de Barros foi o primeiro zoológico do país a reproduzir o lobo-guará e a ararajuba em cativeiro.

Esses dois trabalhos alcançaram repercussão internacional. O nascimento de um filhote de ararajuba, em março, mereceu um artigo na revista inglesa ``New Scientist", uma das principais publicações científicas do mundo.

Em fevereiro, uma equipe da rede de TV inglesa BBC visitou Sorocaba para registrar o trabalho de reprodução do lobo-guará.

Para o veterinário Adauto Veloso Nunes, 40, diretor do zoológico, o sucesso na procriação de animais em cativeiro depende de diversos fatores. ``Você precisa achar o casal certo, colocá-lo no ambiente certo, oferecer uma alimentação correta e fazer um bom controle de doenças", diz Nunes.

O príncipe Charles, herdeiro do trono do Reino Unido, anunciou apoio ao programa de adoção de vegetais para ajudar a salvar **espécies** em **extinção**. Ele deve apoiar financeiramente **espécies** raras de batata, feijões e outros vegetais.

Jornalista pode ter de dizer salário nos EUA
O Senado dos EUA anunciou que considera a possibilidade de obrigar repórteres que fazem cobertura do Congresso a revelar seus salários e rendimentos, incluindo pagamentos adicionais que recebem por palestras e cursos. A maioria dos repórteres em Washington seria atingida.

Nigéria executa 43 condenados por roubo
Um pelotão de fuzilamento nigeriano executou ontem pelo menos 43 condenados por roubo em Lagos (Nigéria). Vistas pela população, as execuções ocorreram na prisão de segurança máxima de Kirikiri. O governo havia anunciado anteontem que executaria 53 condenados.

Consulta 8

Número de relevantes: **185**

```
<top>
<num> 8 </num>
<PT-title> Greves </PT-title>
<PT-desc> Quem está em greve e porquê? </PT-
desc>
<PT-narr> Encontrar documentos que relatem casos
específicos de greves em qualquer parte do
mundo. Os documentos devem fornecer informação
específica
sobre as causas da greve e os objectivos que os
grevistas esperam alcançar através dessa greve.
Apenas documentos que discutam greves concretas
são relevantes; informações sobre greves
potenciais ou planeadas não o são. </PT-narr>
</top>
```

Consulta 8

- Natasha e Guilherme
  - Sistema: *Tantivy*
  - MAP: **19,58%**

LowerCase, stemmer RSLP (NLTK), remoção de stopwords (NLTK)
BM25 (*default*)

```
grev
```

```
num_ret                 8       100
num_rel                 8       185
num_rel_ret             8       56
```

Consulta 9

Número de relevantes: **11**

```
<top>
<num> 9 </num>
<PT-title> Produção global de ópio </PT-
title>
<PT-desc> Encontrar informação sobre o
cultivo ou produção mundial de ópio. </PT-
desc>
<PT-narr> Documentos relevantes darão
informação sobre o cultivo de papoilas ou
a produção de ópio e seus derivados, a
nível mundial. Tanto a produção legal como
a ilegal são de interesse. </PT-narr>
</top>
```

Consulta 9

- Gustavo Monteiro e Leticia
  - Sistema: *Apache Solr*
  - MAP: **47,61%**

LowerCase, remoçao de stopwords, processamento de stem (leve)
BM25 (k=1.2 e b=0.75)

```
text:(+ópio produção global~ mundial~ cultivo papoilas)
```

```
num_ret          9       70
num_rel          9       11
num_rel_ret      9       11
```

Consulta 10

Número de relevantes: **22**

```
<top>
<num> 10 </num>
<PT-title> Crises energéticas </PT-
title>
<PT-desc> Encontrar informação sobre
faltas de energia ou de combustível. </
PT-desc>
<PT-narr> Documentos relevantes devem
mencionar onde ocorreu a crise
energética e mencionar suas causas. </
PT-narr>
</top>
```

Consulta 10

- Thiago Paixão
  - Sistema: *Whoosh* Remoção Stopwords, Stemmer Português, caixa baixa
  - MAP: **4,55%** BM25 (padrão b=0.75, k1=1.2)

```
(crise energia) OR (crise combustivel) OR (crise energia país) OR (falta energia
país) OR (crise combustível país) OR (crise combustível país) OR' + '(crise energia
    cidade) OR (falta energia cidade) OR (crise combustível cidade) OR (crise
combustível cidade) OR'+'(crise energia região) OR (falta energia região) OR (crise
combustível região) OR (crise combustível região) OR' +'(crise energia país causa)
    OR (falta energia país causa) OR (crise combustível país causa) OR (crise
combustível país causa) OR'+'(crise energia cidade causa) OR (falta energia cidade
causa) OR (crise combustível cidade causa) OR (crise combustível cidade causa) OR'
    +'(crise energia região causa) OR (falta energia região causa) OR (crise
        combustível região causa) OR (crise combustível região causa)
```

num_ret	10	4
num_rel	10	22
num_rel_ret	10	1

Pior resultado (para uma consulta)

1. FSP941021-070: 111.119191326744467 ✔
2. FSP951222-030: 95.879142653785666 ✖
3. FSP950401-024: 47.369230250110675 ✖
4. FSP940206-064: 29.909105115611111 ✖

Da Sucursal do Rio

O calor que fez ontem na região Sudeste do **país** levou as hidrelétricas que compõem o complexo de Furnas a pedir **energia** de outras duas usinas para evitar uma pane no sistema de distribuição de eletricidade nos Estados do Rio de Janeiro e Espírito Santo.

Durante a tarde de ontem, faltou **energia** em diversos pontos da **cidade**.

A direção de Furnas alegou um excesso de consumo por **causa** dos aparelhos de ar-condicionado ligados entre 13h20 e 13h40, o que teria provocado a elevação da demanda por **energia** de Furnas de 4.100 para 4.500 megawatts.

Isto equivale a meia unidade geradora da hidrelétrica de Itaipu. Meia unidade gera **energia** suficiente para abastecer uma **cidade** de um milhão de habitantes ou metade de uma **cidade** do tamanho de Curitiba (PR).

Para evitar a pane, Furnas teve que captar **energia** das Usinas de Santa Cruz e do Funil.

O assessor da diretoria de produção de Furnas, Carlos Garnier da Silva, 52, disse que os termômetros de Furnas acusaram ontem a temperatura média de 33 graus Celsius.

``Nós obtivemos informações de que na **região** de São Cristóvão (zona norte do Rio) os termômetros chegaram a registrar 44 graus Celsius'', disse Silva.

De acordo com o Instituto de Meteorologia do Rio, a temperatura máxima registrada ontem foi de 41,2 graus Celsius. A máxima foi registrada em Bangu (zona oeste do Rio).

Da Folha Vale

Os prefeitos das pequenas **cidades** do Vale do Paraíba se declararam em "estado de guerra" contra o Estado e o governo federal por **causa** da **crise** financeira que atinge os municípios.

Segundo o presidente do Codivap (Consórcio de Desenvolvimento Integrado do Vale do Paraíba) e prefeito de Campos do Jordão (85 km de São José), João Paulo Ismael (PMDB), "o Estado e o governo federal querem municipalizar todos os serviços e ainda cortar os recursos das prefeituras".

Em reunião realizada na quarta-feira, os prefeitos decidiram programar um locaute (paralisação promovida por empresas ou órgãos do setor público) de três dias para janeiro e a suspensão do pagamento pela iluminação pública das cidades, a partir de fevereiro.

A suspensão de pagamentos também deve atingir os serviços de segurança. A maior parte das prefeituras paga o aluguel e o **combustível** utilizados pelas polícias Militar e Civil.

No início de janeiro, os prefeitos das 22 **cidades** com até 20 mil habitantes da **região** querem que o Codivap organize uma caravana até Brasília. O objetivo é pressionar os deputados e senadores para que eles rejeitem o FEF (Fundo de Estabilização Fiscal).

Segundo os prefeitos, a nova versão do FSE (Fundo Social de Emergência) vai proporcionar uma queda adicional de 30% na arrecadação das pequenas **cidades**, com a retenção de parte do Fundo de Participação dos Municípios.

Ismael disse que já está mantendo contatos com as associações de municípios estaduais e federais para ampliar a adesão à caravana.

"Queremos os 5.000 prefeitos das pequenas **cidades** do país acampados em frente ao Congresso e, com isso, forçar o governo federal e os congressistas a rever o FSE", afirmou.

O vice-governador do Estado, Geraldo Alckmin (PSDB), vem tentando obter a liberação de créditos relativos a convênios firmados com as prefeituras das pequenas **cidades** da **região**.

Alckmin não foi localizado ontem pela Folha para comentar o calendário de protestos programado pelas prefeituras.

Corte de **energia** elétrica atingiu 27 ruas; alagamentos provocaram congestionamentos em diversos pontos da **cidade**

FSP940114-118

Rel. Não Rec.

Da Reportagem Local e da FT

A chuva forte que atingiu a **cidade** ontem arrebentou os fios elétricos de um poste no Alto da Boa Vista (zona sul) e deixou sem **energia**, entre as 16h59 e 18h20, parte dos bairros do Brooklin e Santo Amaro. De acordo com a CET (Companhia de Engenharia de Tráfego), nesse período formou-se um congestionamento de 2 km na avenida Santo Amaro.

Houve pontos de alagamento com lentidão no trânsito na marginal Tietê (Lapa-Penha), junto à ponte do rio Tamanduateí e à ponte da Nova Fepasa (Penha-Lapa), no Vale do Anhangabaú, próximo ao viaduto Eusébio Stevaux, no viaduto Bresser, junto à ligação Leste-Oeste e em pontos da avenida Pacaembu.

Segundo a CET, em janeiro, com as férias escolares, não é comum nem morosidade no trânsito da Santo Amaro. Ao todo, 20 semáforos de 20 cruzamentos ficaram sem **energia** elétrica.

Balanço da Eletropaulo dá conta que 27 ruas e avenidas foram alvo do corte. A Eletropaulo não soube informar quantas pessoas foram atingidas pela **falta** de luz.

O problema, segundo a Eletropaulo, começou quando uma cruzeta de poste quebrou por causa do vento. A assessoria da empresa afirma que os dois bairros ficaram sem luz por 45 minutos.

Pelo menos 20 equipes da CET tentaram restabelecer o trânsito no local. Segundo a CET, só por volta das 20h os veículos começaram a trafegar na Santo Amaro com mais fluidez.

A chuva foi provocada por uma área de nebulosidade que se estende da **região** Norte até a Sudeste. Essa área de nuvens, associada a uma frente fria, deve permanecer em São Paulo até a próxima terça-feira, segundo o Instituto Nacional de Meteorologia, provocando um fim-de-semana chuvoso.

As chuvas provocaram o desabamento de um muro sobre uma casa na rua Florestal, em Heliópolis (zona sul), sem feridos.

Avaliação Geral (para as 10 consultas)

Resultados (em termos de MAP)

Curvas de Recall x Precision Interpolada

MAP, R-precision, relevantes retornados

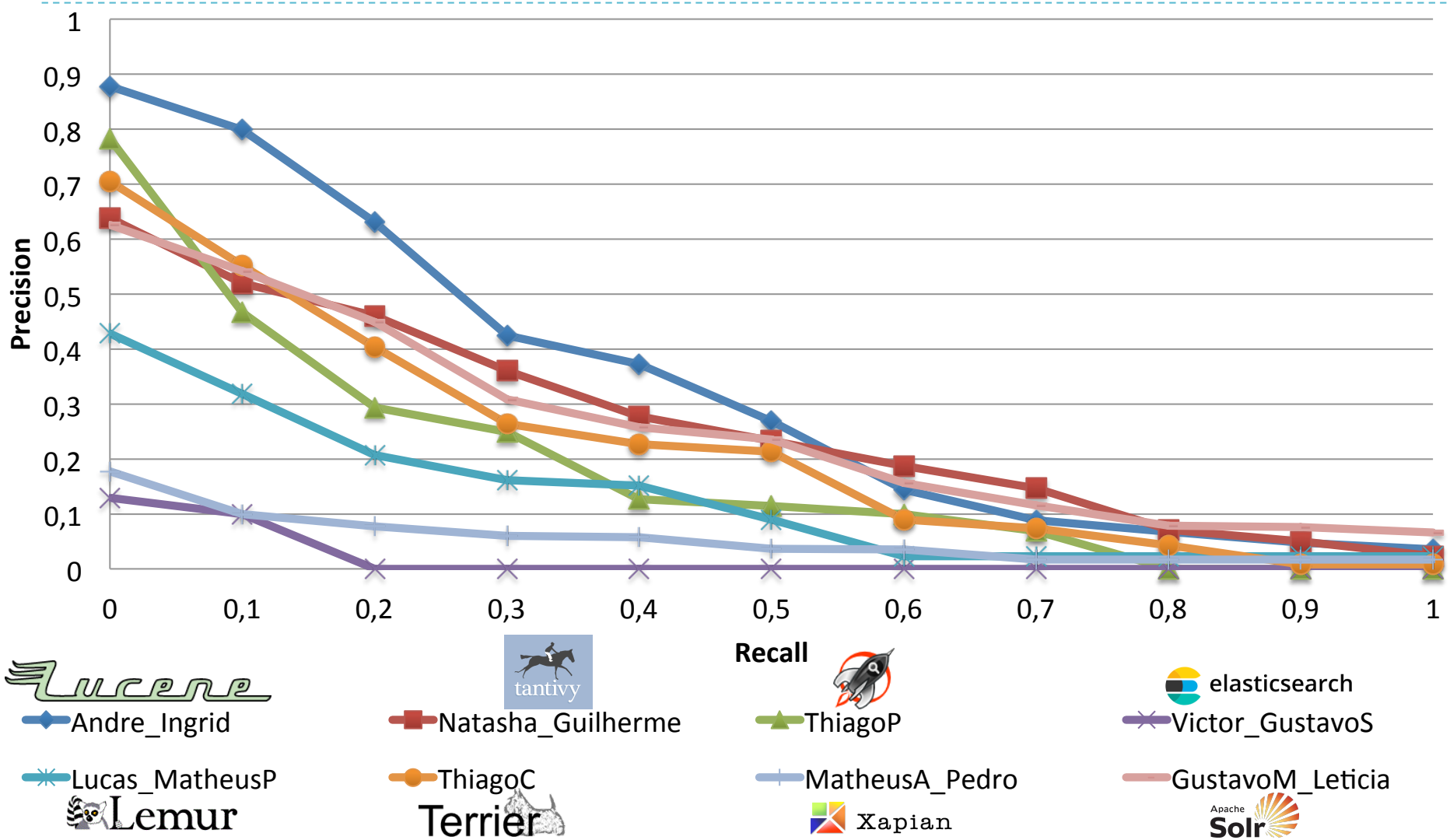
- Gustavo Monteiro
- Leticia Freire

- Natasha Rocha
- Guilherme Franco

tantivy

- André Queiroz
- Ingrid Canaane

Recall x Precision Interpolada



MAP, *R*-precision, Relevantes Retornados

Aluno/Dupla	Sistema	Melhor	MAP*	Rprec	RelRet
André Queiroz e Ingrid Canaane	*Lucene*		**32,60%**	**34,06%**	**169**
Natasha Rocha e Guilherme Franco	*Tantivy*	arquivo 1	*25,55%*	*26,17%*	*131*
Gustavo Monteiro e Leticia Freire	*Apache Solr*		23,77%	22,86%	124
Thiago Coelho	*Terrier*		21,72%	24,07%	130
Thiago Paixão	*Whoosh*		17,32%	21,74%	77
Lucas Rampazzo e Matheus Panno	*Lemur*	com stemming	12,16%	16,66%	83
Matheus Andrade e Pedro Paulo Ferreira	*Xapian*		4,97%	6,25%	50
Victor Lisboa e Gustavo Soares	*elasticsearch*	arquivo 2	1,27%	1,39%	3
Total relevantes					***410***

Resultados detalhados

- Julgamentos de relevância:
  - http://dcc.ufrj.br/~giseli/2019-1/ri/trabalho1/relevantes.txt

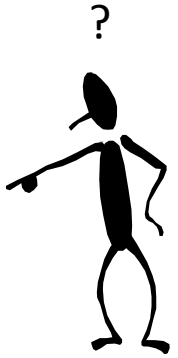
- Ferramenta **trec_eval**
  - Disponível em: http://trec.nist.gov/trec_eval/
  - Informações básicas sobre instalação e uso: http://faculty.washington.edu/levow/courses/ling573_SPR2011/hw/trec_eval_desc.htm
  - Exemplo de comandos para avaliação:
    - `trec_eval relevantes.txt resultado.txt`
      (avaliação geral)
    - `trec_eval -q relevantes.txt resultado.txt`
      (avaliação detalhada por consulta)

Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Dúvidas?



Profa. Giseli Rabello Lopes
giseli@dcc.ufrj.br
CCMN - DCC - Sala E-2012

2019 / 1