# Avaliação de Técnicas de Recuperação de Informação para Organização e Extração de Conhecimento de Documentos de Reunião

Ovídio José Francisco
Orientadora: Prof.ª Dr. Katti Faceli
Coorientador: Prof. Dr. Rafael Geraldeli Rossi


August 20, 2018

# Roteiro

# Roteiro

# Contexto

- As atas registram assuntos discutidos em reuniões;
- Uma ata pode conter diversos assuntos discutidos.

- Um assunto pode ser discutido em diversas reuniões ao longo do tempo;
- Recuperar o histórico de um assunto pode apoiar tomadas decisão.

- Informações contidas em grandes quantidades de texto;
- Inerentemente não estruturados;
- Documentos com múltiplos assuntos;
- Necessidade de ferramentas automáticas.

# Contexto

Desafios na recuperação de informações em atas:

- Apresentar uma ata integral pode dificultar o usuário encontrar determinado assunto;
- É desejável agrupar trechos de atas com assuntos semelhantes.

# Contexto

Algoritmos de Segmentação Textual:

- Dividem um texto em trechos com um único assunto completo;
- Úteis em aplicações com textos sem indicações de quebras de assunto, como transcrições de áudio, e diálogos em chats;
- Podem ser uma etapa de pré-processamento para outras aplicações;
- Não dão indicações sobre o conteúdo dos segmentos.

# Contexto

Modelos de Extração de Tópicos:

- Estimam o assunto de cada documento de uma coleção;
- Agrupam documentos por tópico;
- Identificam palavras para descrever os tópicos;
- Incorporam conhecimento de domínio aos dados.

# Trabalhos Relacionados

| | |
|---|---|
| (ZAMIR; ETZIONI, 1998) | Web document clustering (1998) |
| (MASAO; KôITI, 2000) | Multi-topic multi-document summarization (2000) |
| (JEONG; TITOV, 2010) | Multi-document topic segmentation (2010) |
| (NGUYEN, 2011) | A Study on Statistical Generation of a Hierarchical Structure of Topic-information for Multi-documents (2011) |
| (TAGARELLI; KARYPIS, 2013) | A segment-based approach to clustering multi-topic documents (2013) |

| **Trabalho** | **Divisão** | **Método de inferência** | **Representação** | **Idioma** |
|---|---|---|---|---|
| (ZAMIR; ETZIONI, 1998) | Sentenças | Agrupamento | | Inglês |
| (MASAO; KôITI, 2000) | Sentenças | Agrupamento | Rotulação manual | Inglês |
| (JEONG; TITOV, 2010) | Segmentação | Modelo Bayesiano | | Inglês |
| (NGUYEN, 2011) | Segmentação | Agrupamento hierárquico | Sumarização e palavras-chave | Inglês |
| (TAGARELLI; KARYPIS, 2013) | Parágrafos | Agrupamento e classificação | | Inglês |

# Roteiro

# Objetivos

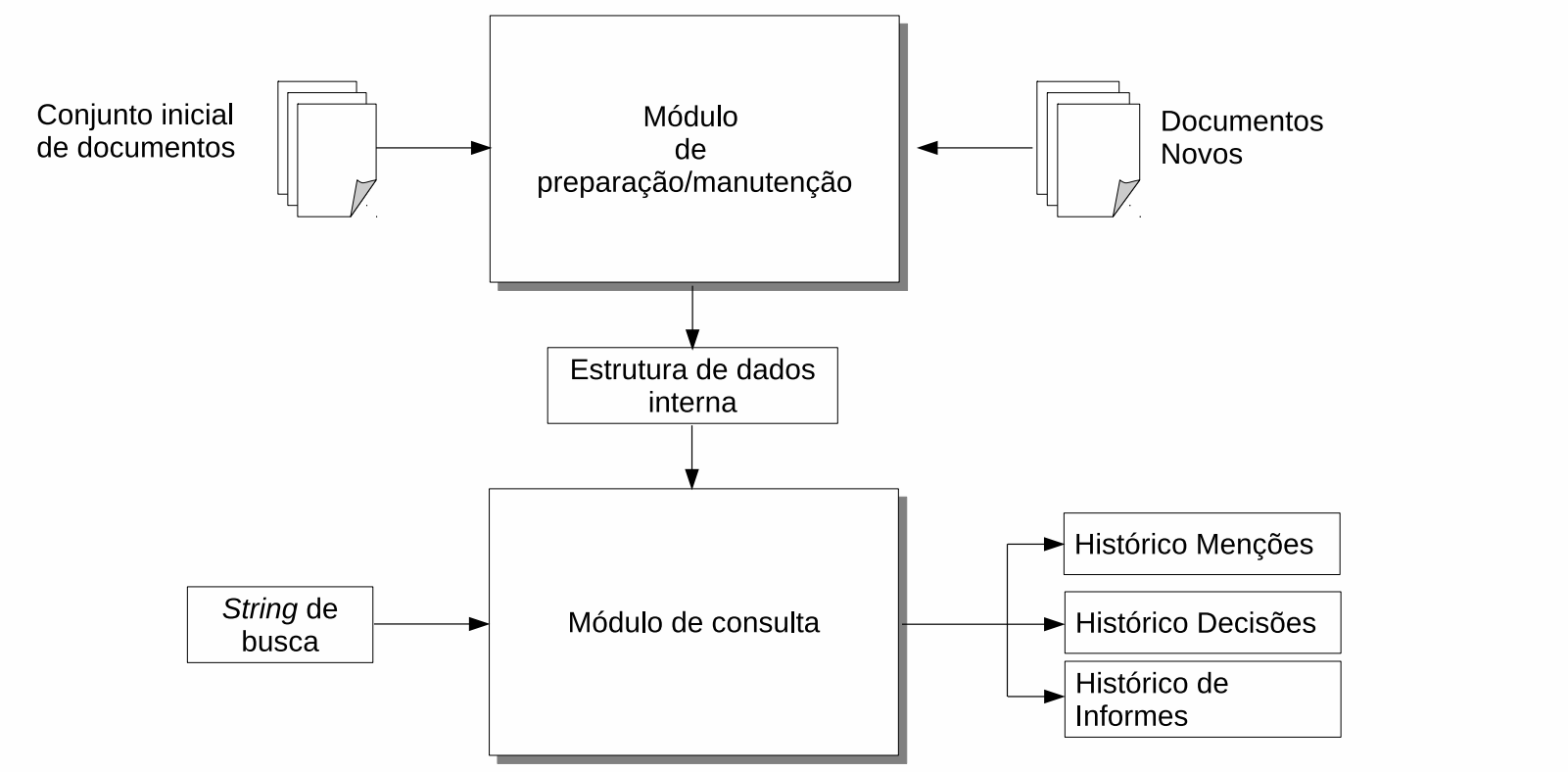
Propor uma solução para identificar, organizar e consultar assuntos registrados em atas de reunião.

Utilizar técnicas de Segmentação Textual em conjunto com modelos de Extração de Tópicos para:

- Gerar uma estrutura mais organizada que a coleção original;
- Utilizar a estrutura latente dos segmentos para Recuperação de Informação.

# Roteiro

# Visão Geral do Sistema



# Módulo de Preparação

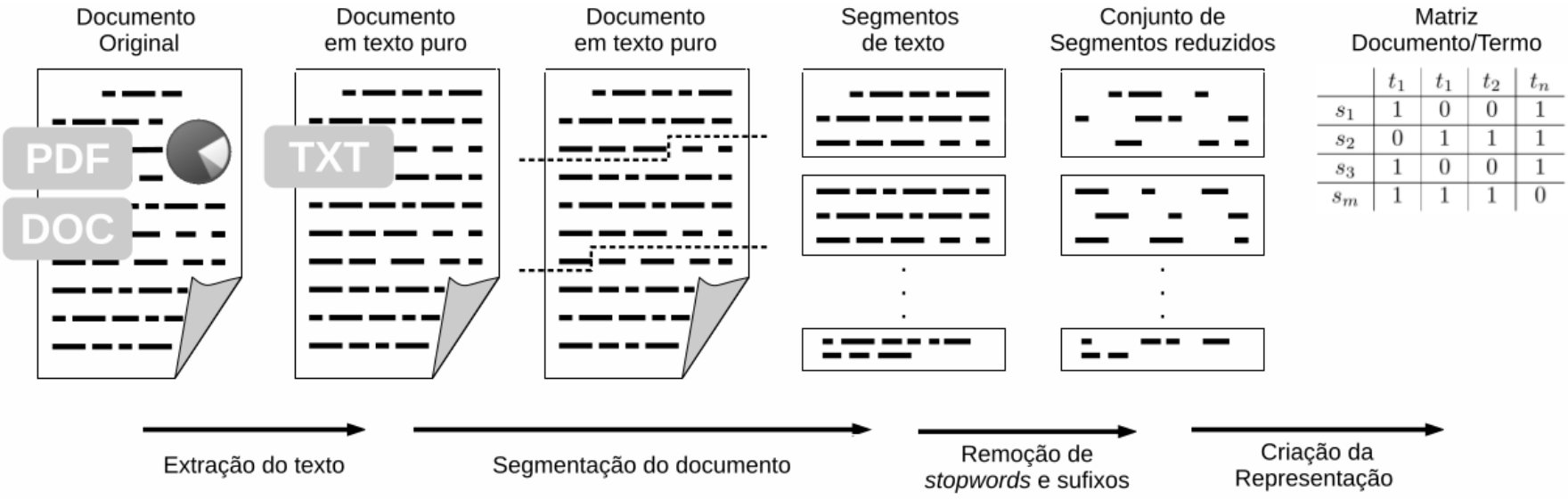
## Preparação

- Extração de texto plano;
- Segmentação;
- Remoção de termos;
- Representação dos Segmentos;
- Extração de Conhecimento.
  - Extração de tópicos;
  - Classificação;

## Manutenção

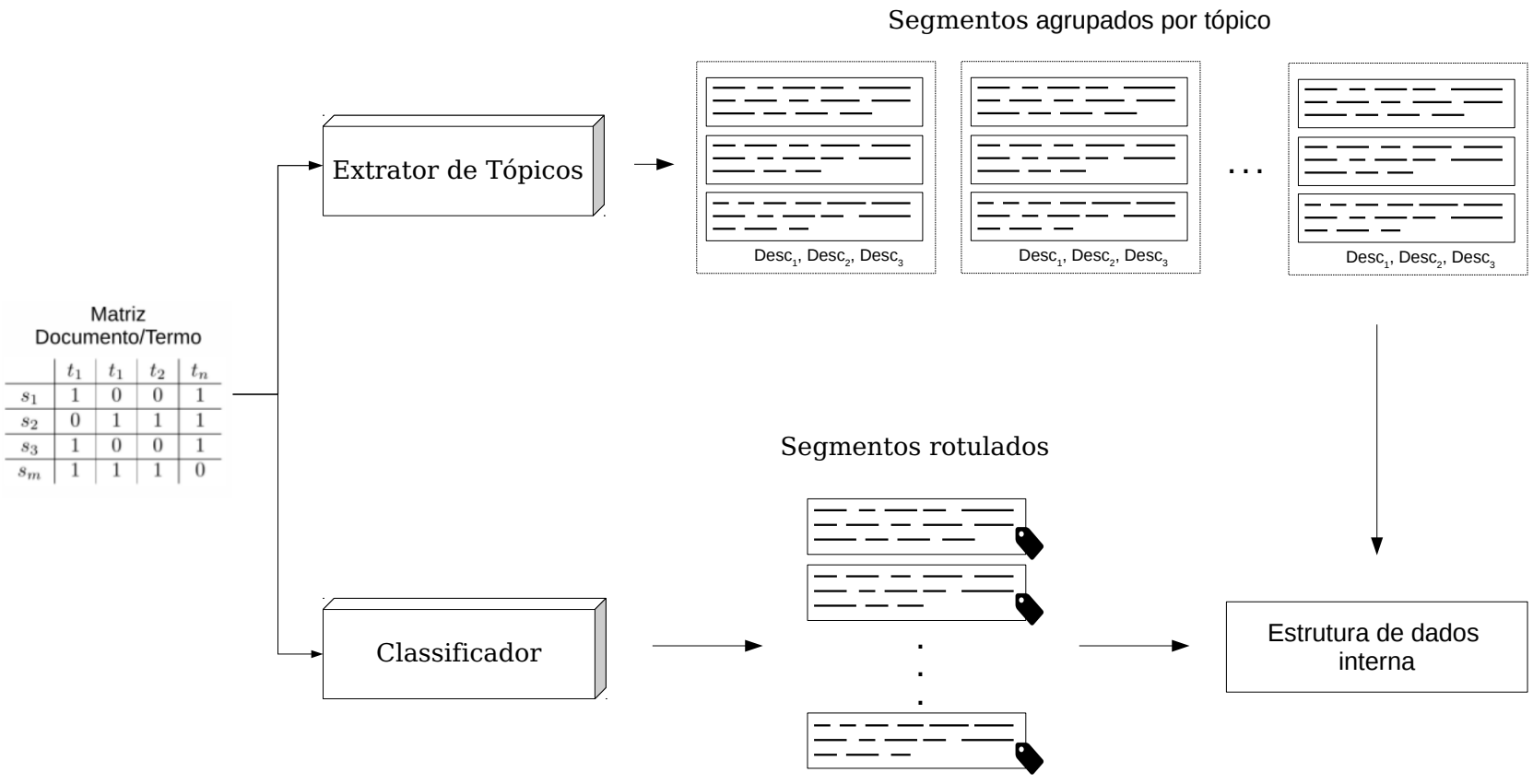
Realimentação do sistema

# Preparação

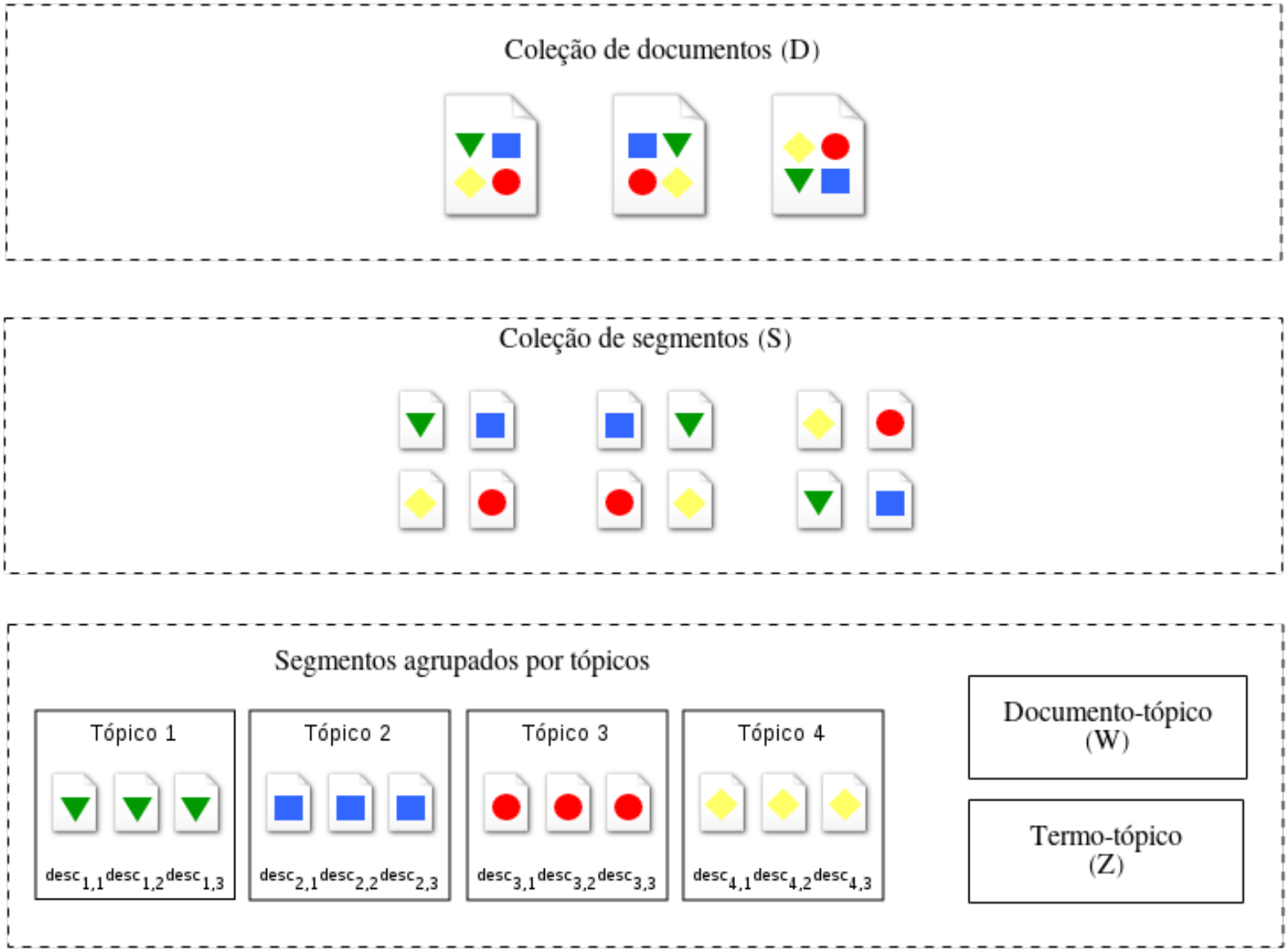



| | $t_1$ | $t_1$ | $t_2$ | $t_n$ |
|---|---|---|---|---|
| $s_1$ | 1 | 0 | 0 | 1 |
| $s_2$ | 0 | 1 | 1 | 1 |
| $s_3$ | 1 | 0 | 0 | 1 |
| $s_m$ | 1 | 1 | 1 | 0 |

# Extração de Conhecimento



# Estrutura de dados interna



# Estrutura de dados interna

Obtém-se uma estrutura:

- Mais organizada que a coleção original;
- Assuntos concentrados em grupos;
- Acrescida de novos atributos;
- Distribuição dos tópicos conhecida.

# Distribuição de tópicos em uma ata real

UNIVERSIDADE FEDERAL DE SÃO CARLOS – Campus Sorocab
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO
Rodovia João Leme dos Santos, Km 110 (SP-264)
Bairro do Itinga - Sorocaba-SP - CEP 18052-780
Telefone: (15) 3202-2022 / www. ufscar. br

**Ata da 17ª Reunião Extraordinária do Conselho do Curso de Bacharelado em Ciência da Computação, UFSCar – Sorocaba**

**Local, Dia e Hora:** No laboratório de Pedagogia, situada nas dependências da Universidade Federal de São Carlos - campus Sorocaba, à Rodovia João Leme dos Santos, quilômetro 110, Bairro do Itinga, na cidade de Sorocaba, Estado de São Paulo e realizada aos vinte e cinco dias do mês de agosto de dois mil e dez, às 14h00. **Convocação e Presença:** A convocação foi realizada sete dias antes da data de sua realização, estando presentes os membros do Conselho do Curso de Bacharelado em Ciência da Computação - Sorocaba, sendo eles os professores

representantes discentes Sr.

(2010). **Comunicação da Presidência:** A prof.ª informou que na última reunião do ConCam foram indicados os nomes do prof. , titular e suplente respectivamente, como representantes do campus Sorocaba no ConsUni (Conselho Universitário). A prof.ª esclareceu também, que o campus já possui um assento ocupado pelo prof. ; e que agora teremos dois assentos no respectivo conselho. A prof.ª também informou que na mesma reunião do ConCam foram indicados os nomes do prof. , titular e suplente, para o assento do campus Sorocaba no CoAd (Conselho de Administração da UFSCar), destacando que agora o campus Sorocaba não está mais no plano de implantação do campus, e irá concorrer com os outros campi da universidade pela distribuição de verbas. A prof.ª também informou que no próximo dia trinta de agosto haverá uma reunião com a ProGrad em Sorocaba para retorno sobre o Projeto Político Pedagógico do curso no qual a coordenação do curso irá participar. **Comunicação dos membros:** O prof. comunicou que houveram diversos problemas com o lançamento do edital para o concurso público da vaga docente para Banco de Dados. O prof. comunicou que já solicitou uma retificação do edital, mas que até o momento tal solicitação está em trâmite no setor de recursos humanos. O Sr. comunicou que foi criado uma lista de e-mails com os e-mails de todos os alunos do curso e docentes do curso, sendo que esta foi uma solicitação da coordenação do curso para que houvesse maior facilidade na comunicação entre alunos e docentes. Foi destacado que os docentes, por padrão, não receberão os e-mails da lista, estando autorizados apenas ao envio de mensagens. Caso desejem receber e-mails da lista o docente deve se comunicar com o Sr. Rubens. **Ordem do Dia:** (I) APROVAÇÃO DAS FICHAS DE CARACTERIZAÇÃO. (I.I) A prof.ª apresentou as fichas de caracterização que seriam analisadas, esclarecendo que todas elas foram criadas considerando fielmente o Projeto Político Pedagógico do curso. As fichas contemplam as disciplinas obrigatórias do curso que ainda não foram oferecidas e as disciplinas optativas que poderão ser oferecidas no primeiro semestre de 2011, sendo respectivamente as obrigatórias: Física para Computação, Algoritmos e Complexidade, Trabalho de Graduação 1, Estágio Supervisionado 1, Trabalho de Graduação 2, Estágio Supervisionado 2 e Seminários de Computação; e as optativas: Tópicos Avançados em Ciência da Computação, Tópicos Avançados em Redes de Computadores e Sistemas Distribuídos, Segurança e Auditoria de Sistemas e Mineração de Dados. Além das fichas foram apresentados os checklists das respectivas disciplinas, que foi uma solicitação da Coordenação Acadêmica para este semestre. DISCIPLINAS OPTATIVAS. (I.II) A prof.ª destacou que haviam sido elaboradas fichas de quatro disciplinas optativas, porém somente três disciplinas serão oferecidas. Desta maneira haverá uma maior flexibilidade para decidir dentre as quatro disciplinas quais seriam as três oferecidas no próximo semestre. DISCIPLINAS DE FÍSICA PARA COMPUTAÇÃO E ALGORTIMOS E COMPLEXIDADE. (I.III) A prof.ª também colocou pra análise as fichas de caracterização das disciplinas de Física para Computação e Algoritmos e

UNIVERSIDADE FEDERAL DE SÃO CARLOS – Campus Sorocaba
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO
Rodovia João Leme dos Santos, Km 110 (SP-264)
Bairro do Itinga - Sorocaba-SP - CEP 18052-780
Telefone: (15) 3202-2022 / www. ufscar. br

Complexidade, explicando que estas disciplinas foram readequadas segundo discussões anteriores sobre a ementa e densidade das mesmas. Juntamente com as fichas de caracterização foi colocada um ofício justificando a alteração das disciplinas. AVALIAÇÃO DAS FICHAS. (I.IV) A prof.ª colocou as fichas para que fosse avaliadas pelos membros do conselho. O prof. questionou se quando o aluno vai se matricular na disciplina de Estágio ou de Trabalho de Graduação o sistema verifica se o pré-requisito de créditos mínimos cursados é checado. O prof. respondeu que acreditava que esta verificação era realizada. A prof.ª colocou que no checklist da disciplina de Algoritmos e Complexidade estava que a disciplina era pré-requisito para outras disciplinas do curso, perguntando se realmente havia alguma disciplina que utilizava essa como pré-requisito. A prof.ª consultou o Projeto Político Pedagógico e verificou que existe uma disciplina optativa, Tópicos Avançados em Teoria da Computação que tem como pré-requisito a disciplina de Algoritmos e Complexidade. (I.V) As fichas foram aprovadas pelo Conselho. **Encerramento:** Estando todos de acordo e nada mais havendo a deliberar, lavra-se, leia-se, aprova-se e assina-se esta Ata por todos os membros do Conselho do Curso de Bacharelado em Ciência da Computação, Campus Sorocaba, participantes desta reunião que, em 02 (duas) vias, será levada a registro e arquivamento junto à Coordenação do Conselho do Curso de Bacharelado em Ciência da Computação - Sorocaba, ficando ali à disposição para consulta restrita aos professores da UFSCar - Sorocaba. Nada mais.

**Profª Drª**
Presidente desta Reunião

**Prof. Dr.**
Professor Associado

**Prof. Dr.**
Professor Adjunto

**Profa. Dra**
Professora Adjunta

**Profa. Dra.**
Professora Adjunta

**Prof. Dr.**
Professor Adjunto

**Prof. Dr.**
Professor Adjunto

**Prof. Dr.**
Professor Adjunto

Representante Discente - 2008

**Prof. Dr.**
Professor Adjunto

Representante Discente - 2009

Representante Discente - 2010

dia; realizada; chamada; estado; conselho;

cursadas; conselho; coordenação; computação; presidente;

docentes; técnica; administrativo; presidente; dia;

disciplinas; cursadas; libras; conselho; aprovado;

computação; conselho; aprovado; acordo; ficou;

representante; discente; presidente; secretária; turma;

# Distribuição de tópicos em uma ata real

**UNIVERSIDADE FEDERAL DE SÃO CARLOS – Campus Sorocaba**

**BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

Rodovia João Leme dos Santos, Km 110 (SP-264)
Bairro do Itinga - Sorocaba-SP - CEP 18052-780
Telefone: (15) 3202-2022 / www. ufscar. br

**Ata da 17ª Reunião Extraordinária do Conselho do Curso de Bacharelado em Ciência da Computação, UFSCar – Sorocaba**

**Local, Dia e Hora:** No laboratório de Pedagogia, situada nas dependências da Universidade Federal de São Carlos - campus Sorocaba, à Rodovia João Leme dos Santos, quilômetro 110, Bairro do Itinga, na cidade de Sorocaba, Estado de São Paulo e realizada aos vinte e cinco dias do mês de agosto de dois mil e dez, às 14h00. **Convocação e Presença:** A convocação foi realizada sete dias antes da data de sua realização, estando presentes os membros do Conselho do Curso de Bacharelado em Ciência da Computação - Sorocaba, sendo eles os professores

representantes discentes Sr.

(2010). **Comunicação da Presidência:** A prof.ª informou que na última reunião do ConCam foram indicados os nomes do prof. , titular e suplente respectivamente, como representantes do campus Sorocaba no ConsUni (Conselho Universitário). A prof.ª esclareceu também, que o campus já possui um assento

dia; realizada; chamada; estado; conselho;

# Distribuição de tópicos em uma ata real

data de sua realização, estando presentes os membros do Conselho do Curso de Bacharelado em Ciência da Computação - Sorocaba, sendo eles os professores

representantes discentes Sr.

(2010). **Comunicação da Presidência:** A prof.ª informou que na última reunião do ConCam foram indicados os nomes do prof. , titular e suplente respectivamente, como representantes do campus Sorocaba no ConsUni (Conselho Universitário). A prof.ª esclareceu também, que o campus já possui um assento ocupado pelo prof. e que agora teremos dois assentos no respectivo conselho. A prof.ª também informou que na mesma reunião do ConCam foram indicados os nomes do prof. , titular e suplente, para o assento do campus Sorocaba no CoAd (Conselho de Administração da UFSCar), destacando que agora o campus Sorocaba não está mais no plano de implantação do campus, e irá concorrer com os outros campi da universidade pela distribuição de verbas. A prof.ª também informou que no próximo dia trinta de agosto haverá uma reunião com a ProGrad em Sorocaba para retorno sobre o Projeto Político Pedagógico do curso no qual a coordenação do curso irá participar. **Comunicação dos membros:** O prof. comunicou que houveram diversos problemas com o lançamento do edital para o concurso público da vaga docente para Banco de Dados. O prof. comunicou que já solicitou uma retificação do edital, mas que até o momento tal solicitação está em trâmite no setor de recursos humanos. O Sr. comunicou que foi criado uma lista de e-mails

cursadas; conselho; coordenação; computação; presidente;

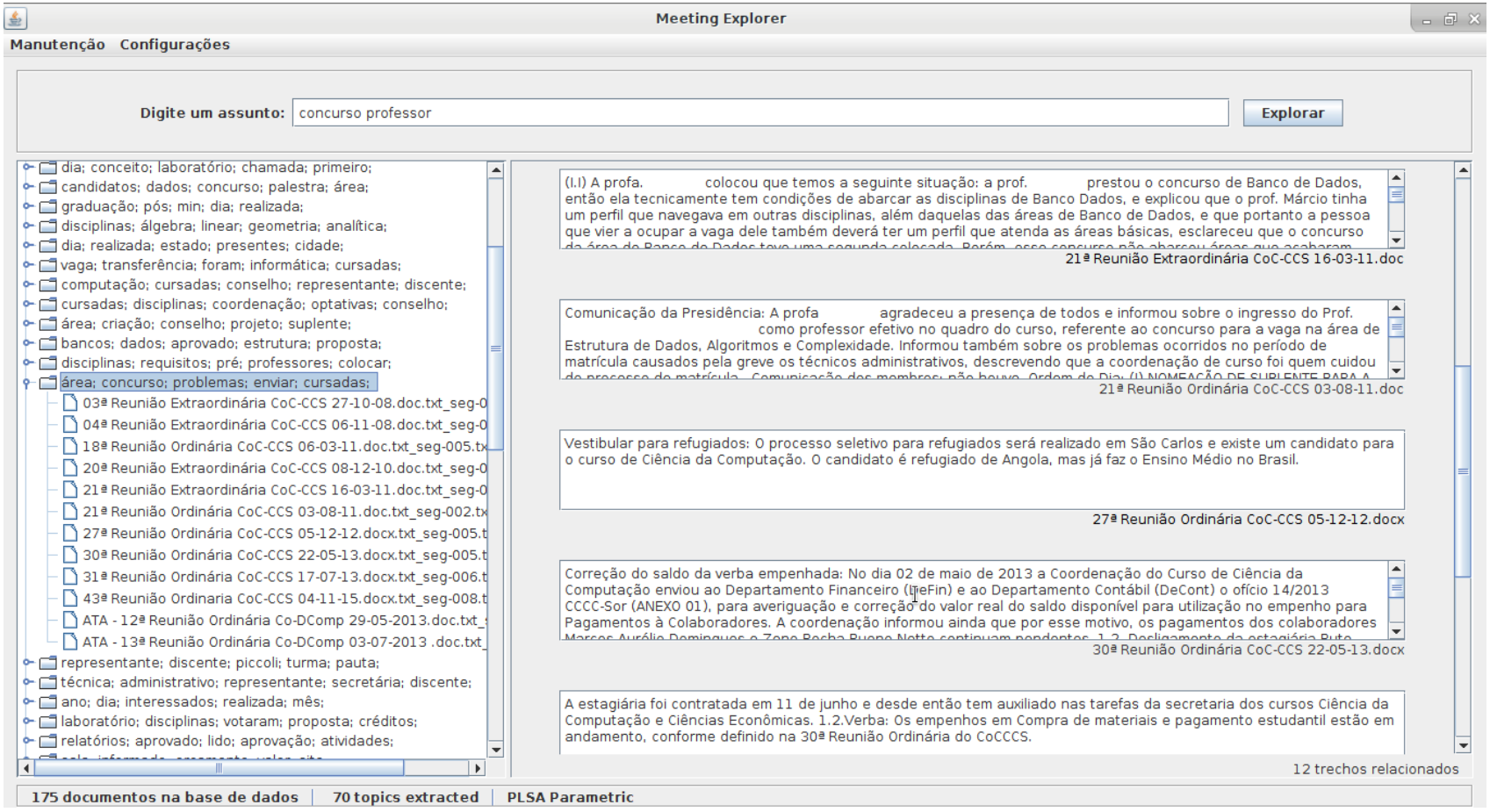
# Módulo de Consulta

Módulo de Consulta:

- Utiliza a Estrutura de Dados Interna como base;
- Os tópicos são representados por seus descritores;
- Os segmentos de atas são agrupados em tópicos;
- Usa o Modelo de Espaço Vetorial para ranquear os tópicos;
- Exibe os segmentos atribuídos ao primeiro tópico do ranking;

# Módulo de Consulta

## Interface do Sistema após uma consulta



# Sistema

O sistema permite:

- Receber uma base de dados não estruturada;
- Identificar os assuntos tratados em cada ata;
- Agrupar segmentos por tópico;
- Adicionar novos atributos (descritores) aos segmentos;
- Expandir o espaço de busca;
- Retornar trechos relevantes à consulta.

# Roteiro

# Avaliação Experimental

A Avaliação Experimental foi dividida em duas etapas:

- Segmentação Textual;
- Extração de Tópicos.

# Avaliação Experimental

Os segmentadores foram avaliados objetivamente.

- Processo de anotação manual em segmentos;
- Criação de segmentações de referência;
- Configuração dos algoritmos;
- Medidas de desempenho;
  - Acurácia e $F^1$;
  - WindowDiff e $P_k$.

- *Testes de significância estatística (Friedman e Nemenyi)*

# Processo de anotação manual em segmentos

A tarefa dos anotadores consistiu em:

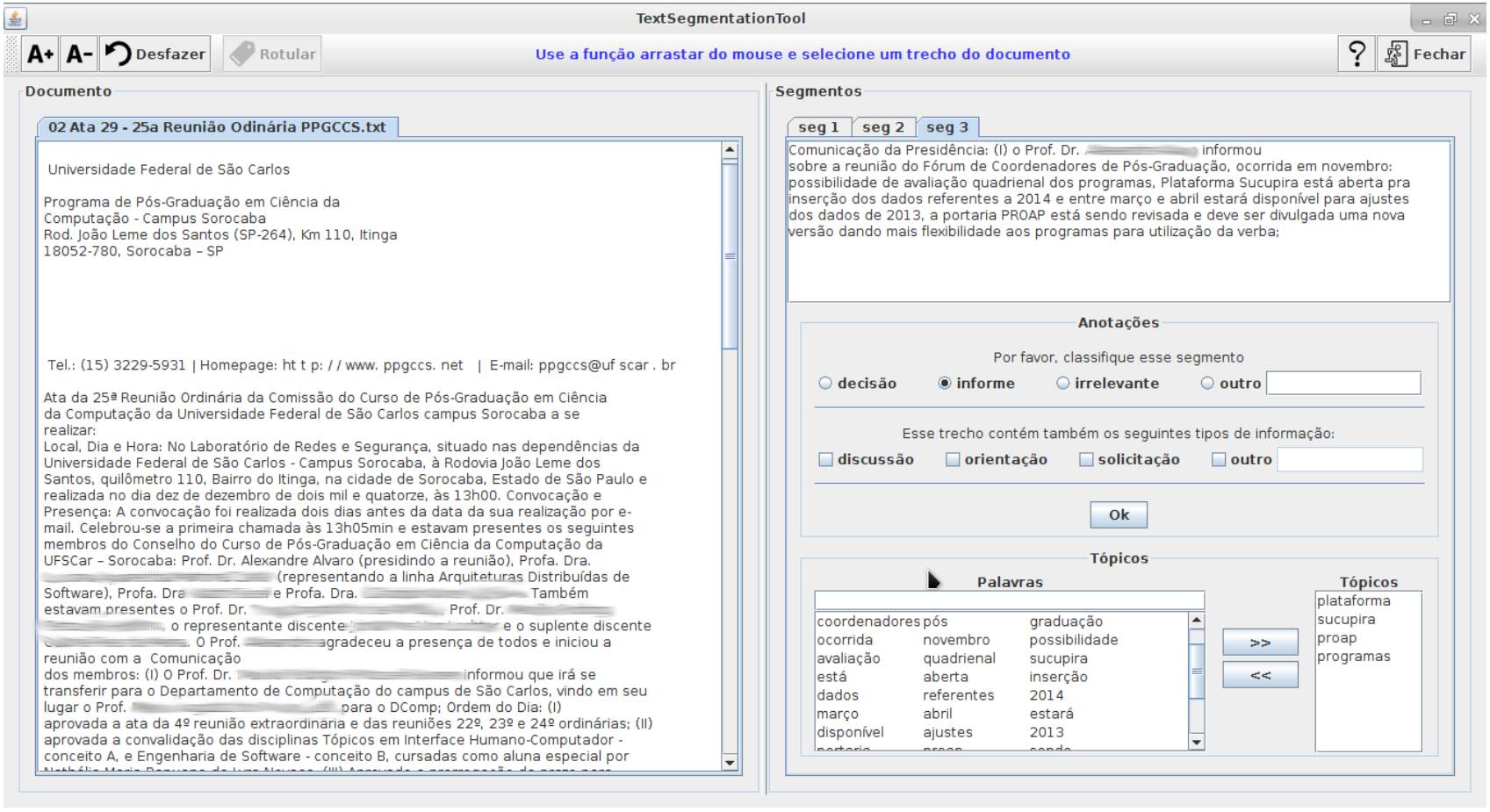
- Selecionar trechos com um único assunto;
- Rotular os trechos selecionados;
  - Tipo comunicação;
  - Contexto onde se gerou o assunto;
  - Descrição do assunto.

Utilizou-se:

- 12 atas da UFSCar;
- 09 anotadores;
- Ferramenta para anotações em segmentos.

# Processo de anotação manual em segmentos

Ferramenta desenvolvida para anotação em segmentos

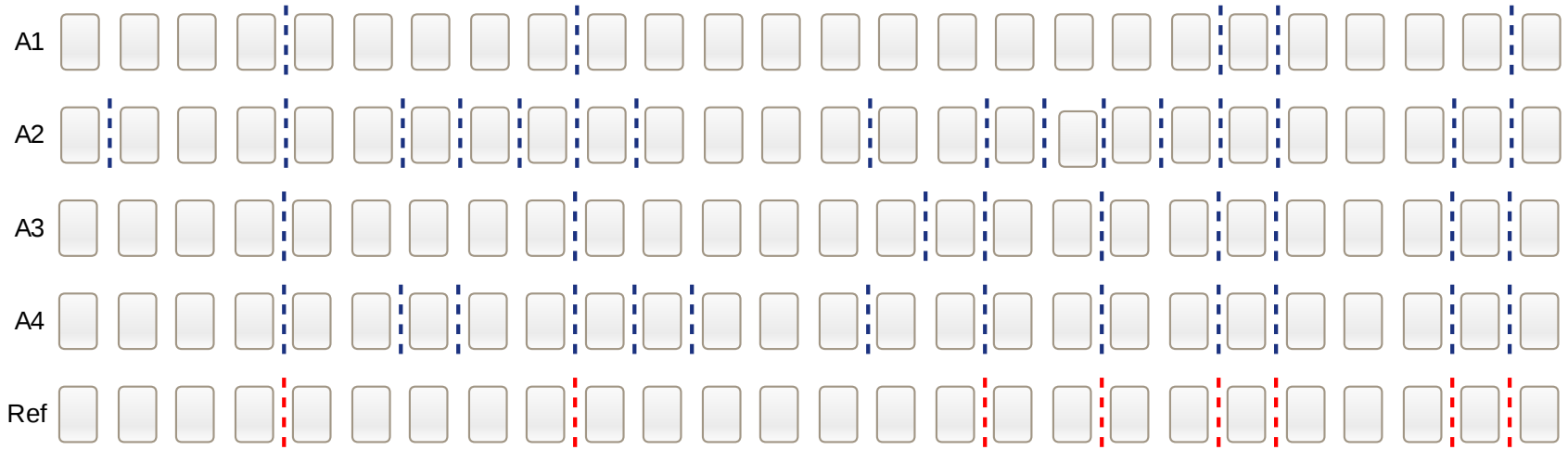


## Descrição dos resultados obtidos com anotadores

| **Ata** | **#Sent.** | **A1** | **A2** | **A3** | **A4** | **A5** | **A6** | **A7** | **A8** | **A9** |
|---|---|---|---|---|---|---|---|---|---|---|
| Ata 1 | 25 | 7 | 4 | 11 | 6 | 16 | 8 | 8 | 15 | 16 |
| Ata 2 | 17 | 4 | 4 | 8 | 6 | 11 | 6 | 6 | 15 | 14 |
| Ata 3 | 26 | 6 | 6 | 8 | 4 | 15 | 9 | 10 | 18 | 14 |
| Ata 4 | 26 | 5 | 5 | 10 | 6 | 14 | 17 | 7 | 11 | 12 |
| Ata 5 | 33 | 4 | 4 | 6 | 5 | 17 | 22 | 9 | 18 | 16 |
| Ata 6 | 11 | 3 | 4 | 6 | 4 | 9 | 9 | 4 | 7 | 5 |
| Ata 7 | 20 | 3 | 7 | 5 | 4 | 11 | 14 | 5 | 5 | 4 |
| Ata 8 | 35 | 4 | 8 | 3 | 8 | 12 | 17 | 5 | 11 | 9 |
| Ata 9 | 24 | 3 | 5 | 3 | 6 | 11 | 11 | 3 | 9 | 9 |
| Ata 10 | 50 | 4 | 5 | 4 | 7 | 31 | 29 | 5 | 9 | 8 |
| Ata 11 | 43 | 4 | 7 | 5 | 7 | 29 | 19 | 5 | 9 | 12 |
| Ata 12 | 56 | 3 | 10 | 4 | 16 | 33 | 25 | 4 | 13 | 11 |
| **Total** | **366** | **50** | **69** | **73** | **79** | **209** | **186** | **71** | **140** | **130** |

# Exemplo de criação de uma segmentação de referência

Referência criada a partir da concordância entre segmentações manuais.



# Configuração Experimental - Segmentadores

| **Algoritmo** | **Parâmetros** (Configuração) | | | | | |
|---|---|---|---|---|---|---|
| *TextTiling* | **Win** | (20-60) | **Step** | (30-55) | | |
| *C99* | **SR** | (.2-.7) | **W** | (sim/não) | **RS** | (3-7) |
| *BayesSeg* | **SR** | (auto, .3-.9) | **Prior** | (.08-.11) | **Disp** | (.1-.7) |
| *MinCut* | **SR** | (.2-.7) | **LenCut** | (5-15) | | |
| *TextSeg* | **SR** | (auto, .1-.9) | | | | |
| *PseudoSeg* | | | | | | |

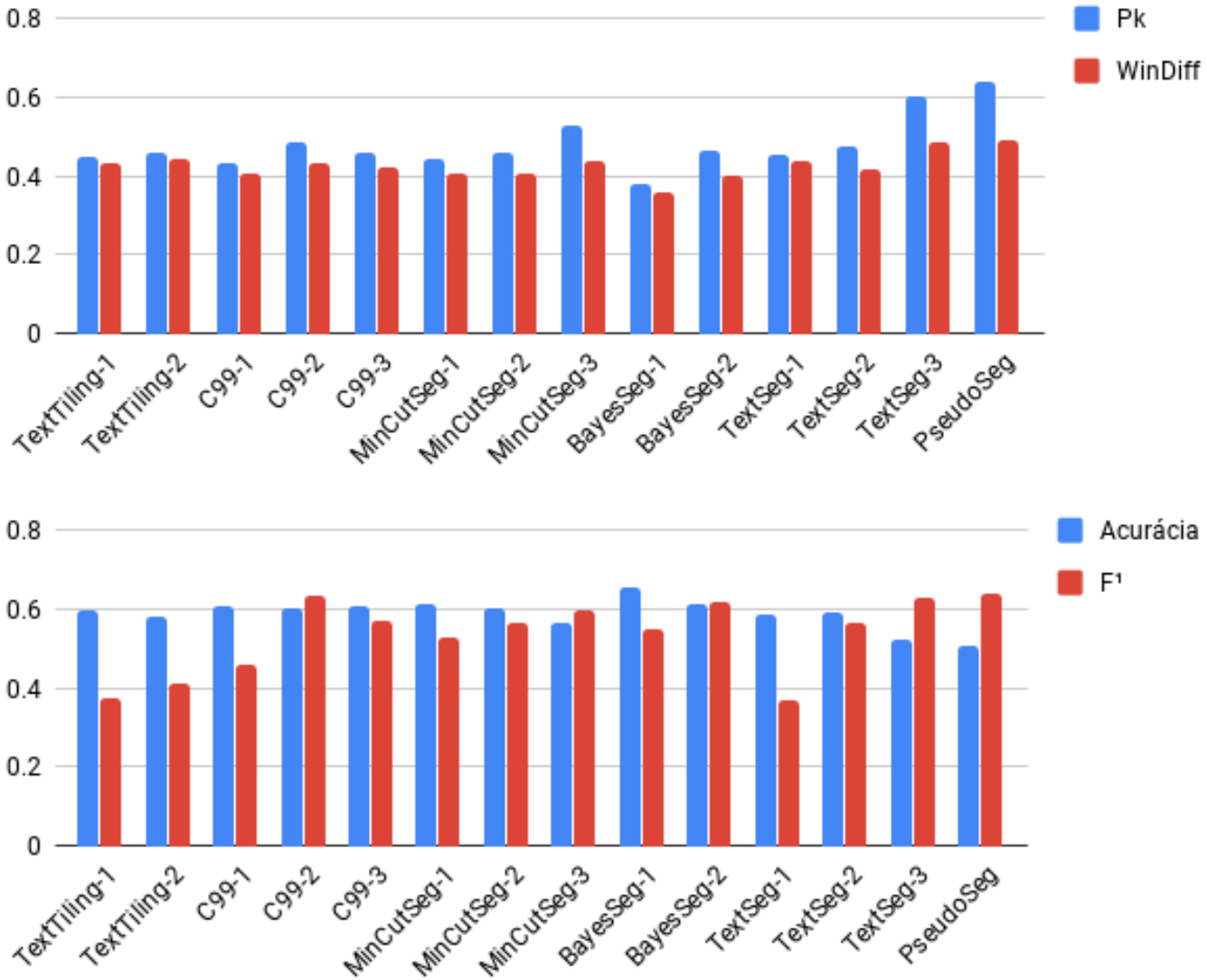
# Configuração Experimental

Resumo dos melhores resultados obtidos por cada configuração

| **Algoritmo** | | **Step** | **Win** | $P_k$ | **WD** | **Ac** | $F^1$ | **#Segs** |
|---|---|---|---|---|---|---|---|---|
| TextTiling-1 | | 20 | 30 | 0.461 | 0.444 | 0.581 | **0.411** | 8.833 |
| TextTiling-2 | | 30 | 45 | **0.450** | **0.435** | **0.596** | 0.373 | 6.417 |
| **Algoritmo** | **RS** | **W** | **SRate** | $P_k$ | **WD** | **Ac** | $F^1$ | **#Segs** |
| C99-1 | 3 | true | 0.300 | **0.434** | **0.407** | 0.607 | 0.457 | 9.250 |
| C99-2 | 3 | true | 0.700 | 0.485 | 0.431 | 0.602 | **0.633** | 21.417 |
| C99-3 | 5 | true | 0.500 | 0.460 | 0.421 | **0.609** | 0.571 | 15.500 |
| **Algoritmo** | | **Cut** | **SRate** | $P_k$ | **WD** | **Ac** | $F^1$ | **#Segs** |
| MinCutSeg-1 | | 9 | 0.400 | **0.444** | 0.408 | **0.614** | 0.526 | 11.917 |
| MinCutSeg-2 | | 11 | 0.500 | 0.459 | **0.407** | 0.603 | 0.563 | 15.000 |
| MinCutSeg-3 | | 5 | 0.700 | 0.528 | 0.438 | 0.567 | **0.599** | 21.000 |
| **Algoritmo** | **Prior** | **Disp.** | **SRate** | $P_k$ | **WD** | **Ac** | $F^1$ | **#Segs** |
| BayesSeg-1 | 0.080 | 0.500 | Auto | **0.380** | **0.361** | **0.655** | 0.551 | 10.000 |
| BayesSeg-2 | 0.110 | 0.100 | 0.600 | 0.462 | 0.399 | 0.615 | **0.619** | 18.417 |
| **Algoritmo** | | | **SRate** | $P_k$ | **WD** | **Ac** | $F^1$ | **#Segs** |
| TextSeg-1 | | | Auto | **0.455** | 0.439 | 0.585 | 0.368 | 6.417 |
| TextSeg-2 | | | 0.500 | 0.475 | **0.417** | **0.594** | 0.566 | 15.500 |
| TextSeg-3 | | | 0.900 | 0.604 | 0.484 | 0.524 | **0.627** | 27.500 |
| **Algoritmo** | | | **SRate** | $P_k$ | **WD** | **Ac** | $F^1$ | **#Segs** |
| PseudoSeg | | | 1.000 | **0.640** | **0.490** | **0.506** | **0.638** | 30.500 |

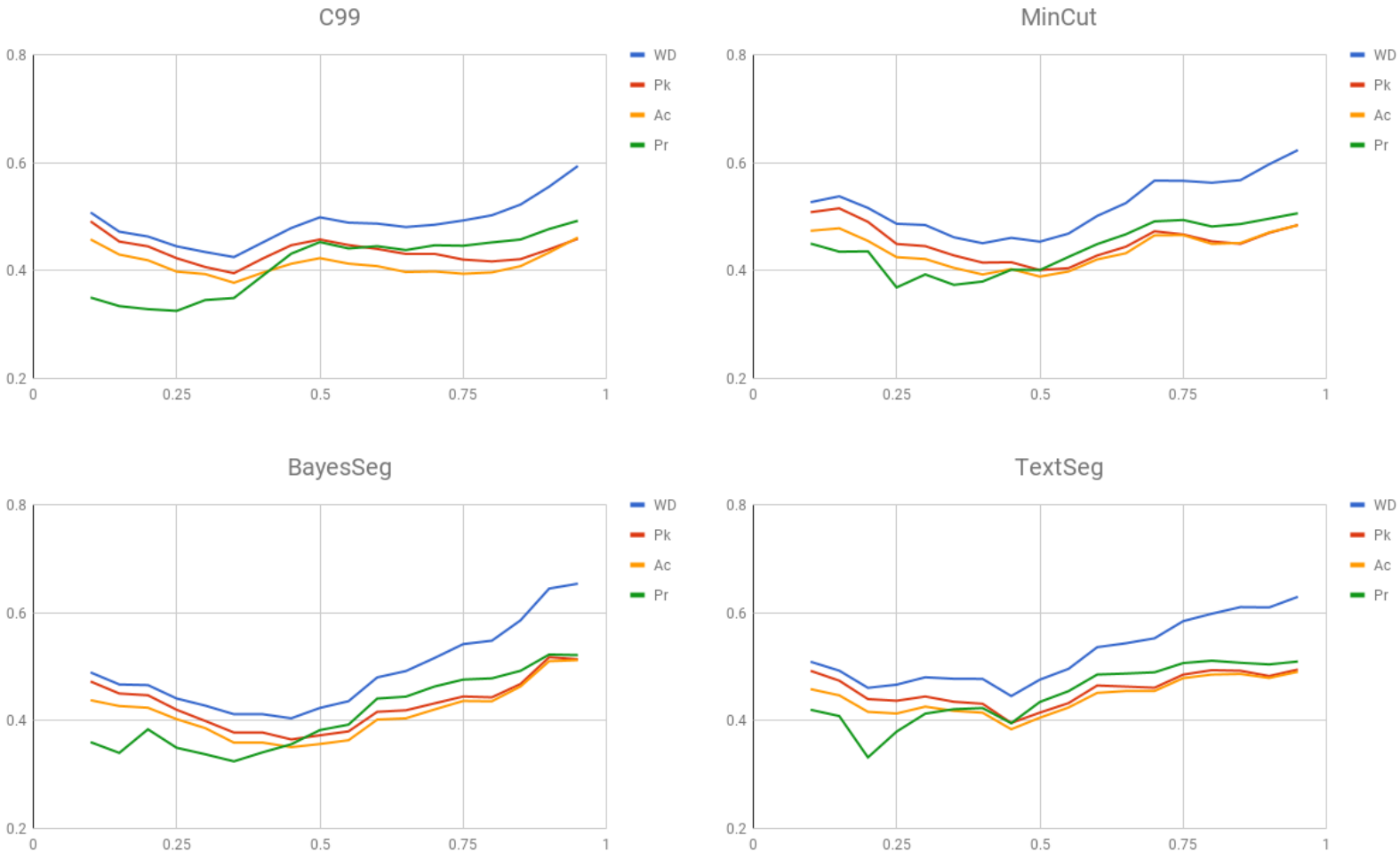
# Avaliação Experimental

Resumo dos melhores resultados obtidos por cada configuração



# Avaliação Experimental

Influência da taxa de segmentos na eficiência dos algoritmos



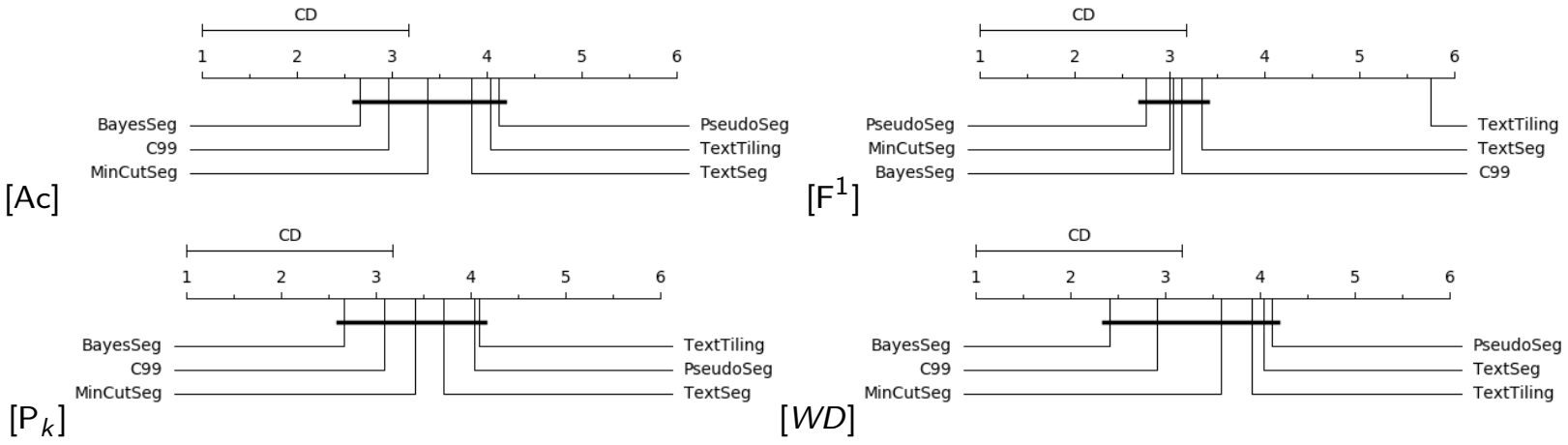
# Avaliação Experimental

Testes de significância estatística de Friedman com pós teste de Nemenyi

- Inicialmente com as configurações de cada algoritmo.
- Novamente com as melhores configurações de cada algoritmo.

# Avaliação Experimental

Diagramas de Diferença Crítica sobre *ranking* considerando valores de Acurácia, $F^1$, *WindowDiff*, e $P_k$.




Não há diferença significativa entre os métodos.

# Avaliação Experimental

Os modelos de Extração de Tópicos foram avaliados junto aos usuários.

- Resultados de 2 consultas ao Sistema usando
- 3 Extratores (K-Means, LDA, PLSA);
- Impressões dos usuários coletadas via questionários.

# Consultas ao Sistema

Entrada:

1. *"defesa de dissertação"*;
2. *"compra de equipamentos"*.

*Corpus*

- Formado por 175 atas;
- Segmentadas com o *BayesSeg*;
- Conjunto de 1276 segmentos.

Resultados utilizando 3 modelos para Extração de Tópicos:

1. K-Means;
2. LDA;
3. PLSA.

# Avaliação Subjetiva

Os resultados foram apresentados à um grupo de avaliadores:

- 24 profissionais da UFSCar;
- 13 profissionais de escolas técnicas;
- 03 profissionais de escolas do Ensino Fundamental.

Perfil:

- 17 membros de conselhos;
- 12 gestores;
- 05 administrativos;
- 03 professores;
- 03 sem afinidade com atas (descartados).

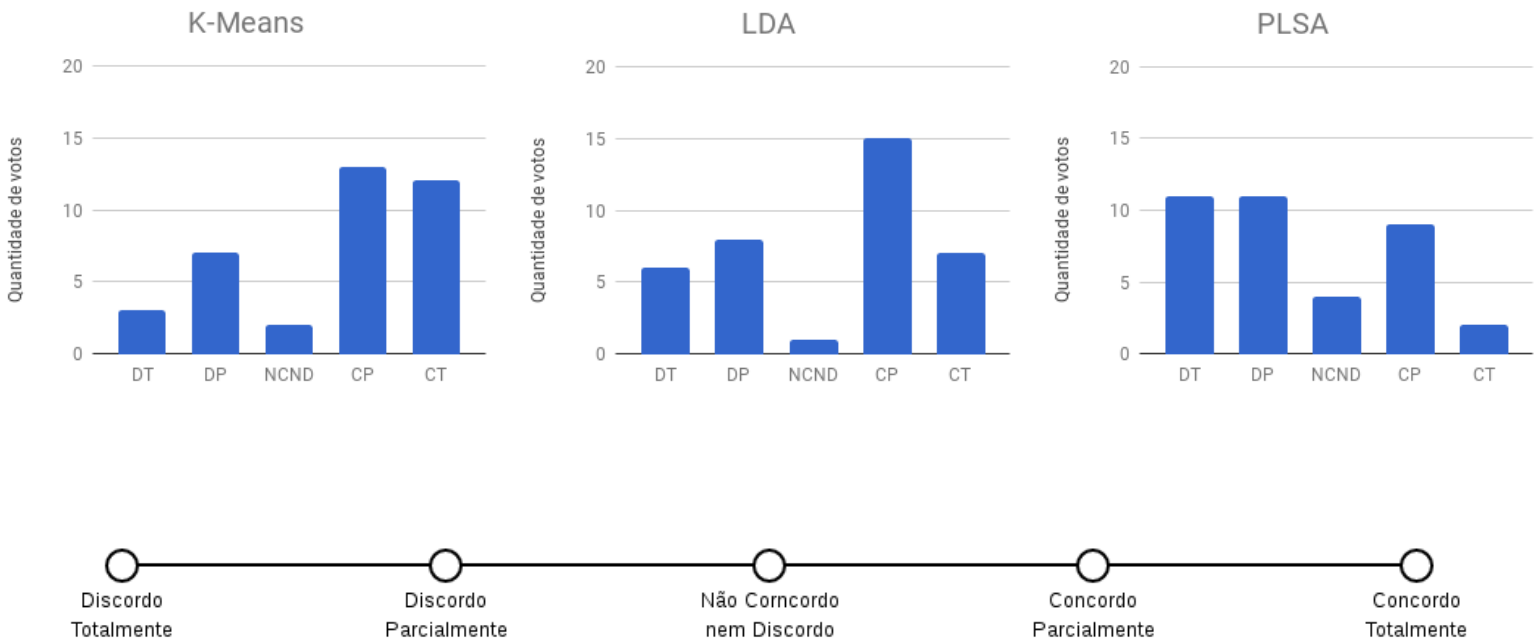
# Questionário

Coletar respostas referentes à:

1. Coesão dos tópicos;
2. Representatividade dos descritores;
3. Coesão dos segmentos;
4. Completude dos segmentos.

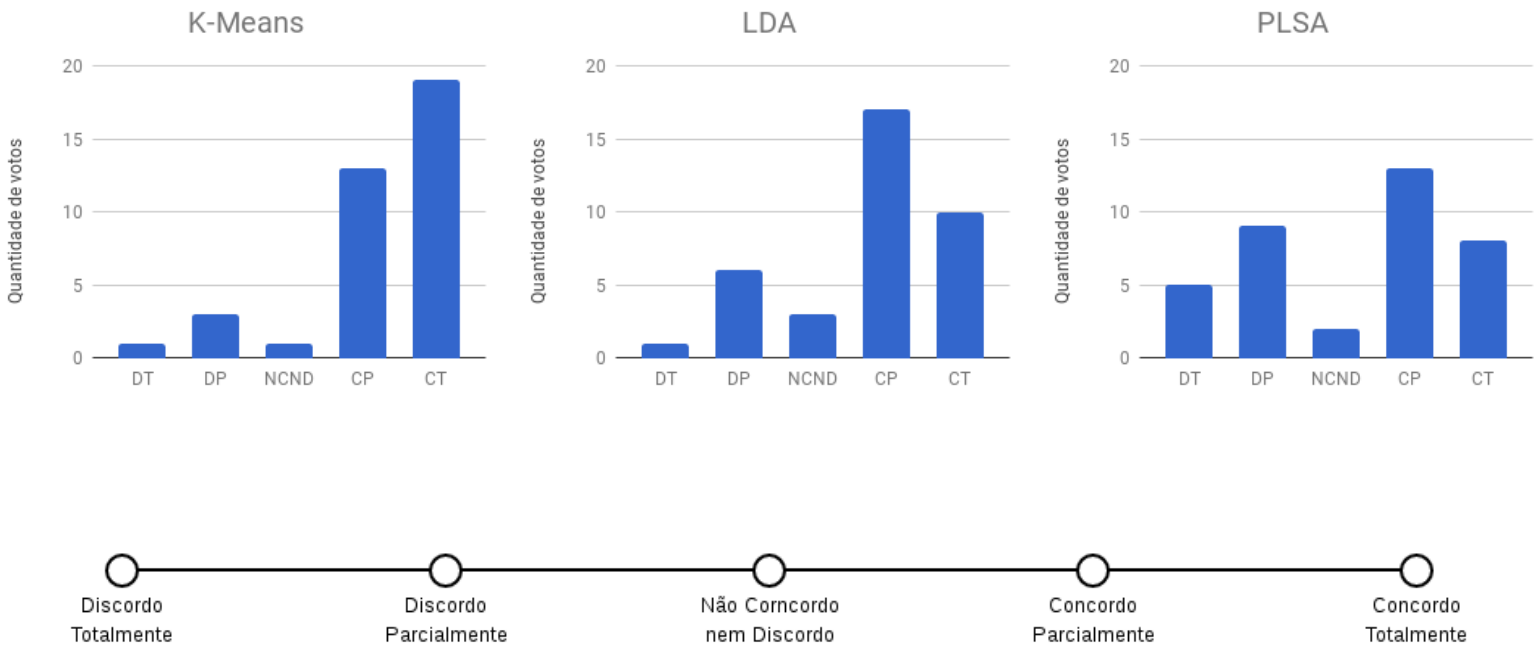
# Coesão dos tópicos

Primeira questão: "*Todos os trechos apresentados compartilham um mesmo assunto.*".



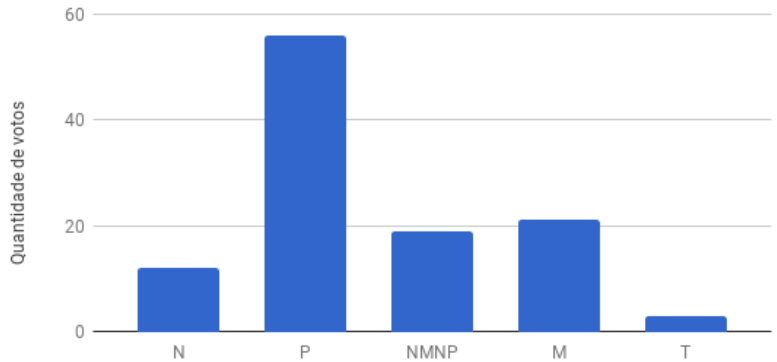
# Representatividade dos descritores

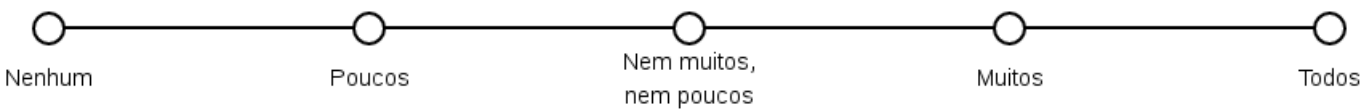
Segunda questão: *"As palavras <descritores> resumem bem o assunto tratado nos trechos."*.



# Coesão dos segmentos

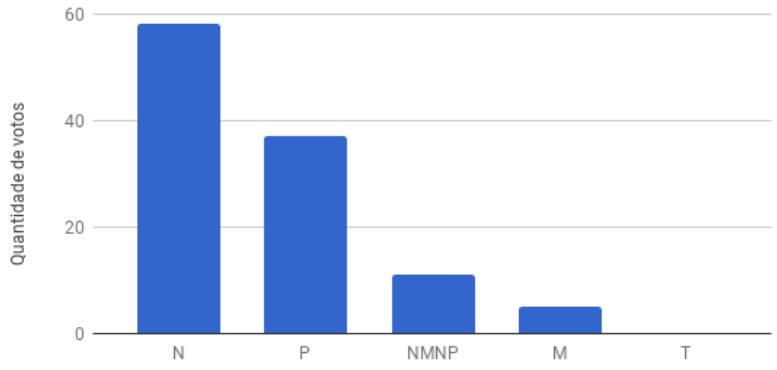
Terceira questão: "*Existem trechos que não tratam de um único assunto?*".

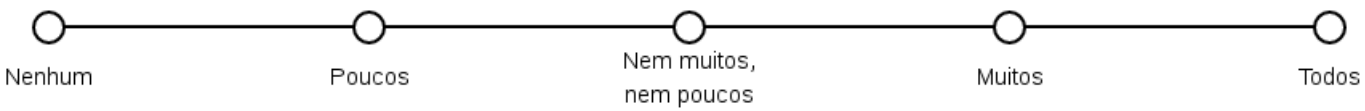


# Completude dos Segmentos

Quarta questão: "*Existem trechos incompletos e insuficientes para compreensão do assunto do trecho?*".



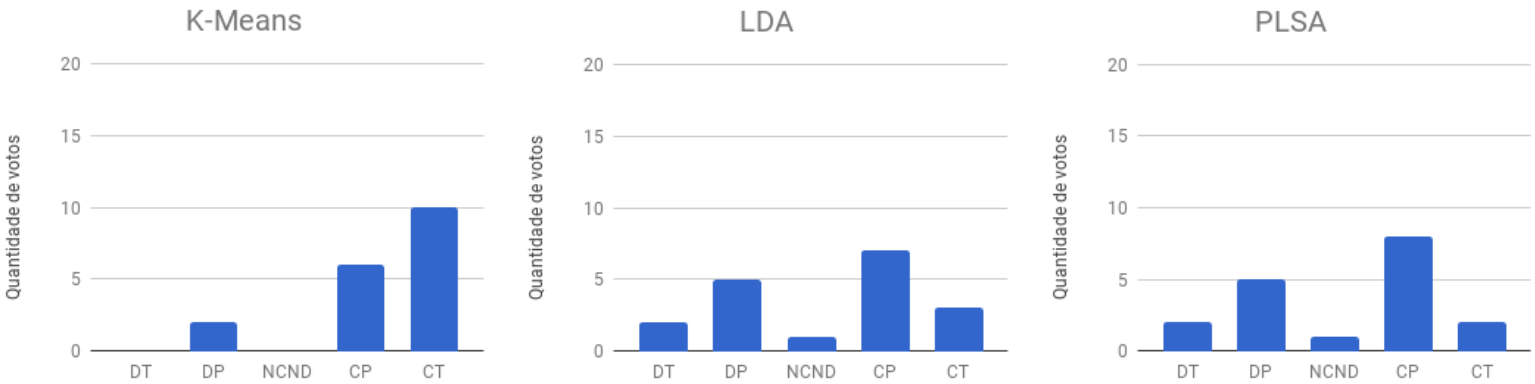



# Comportamento em consultas diferentes

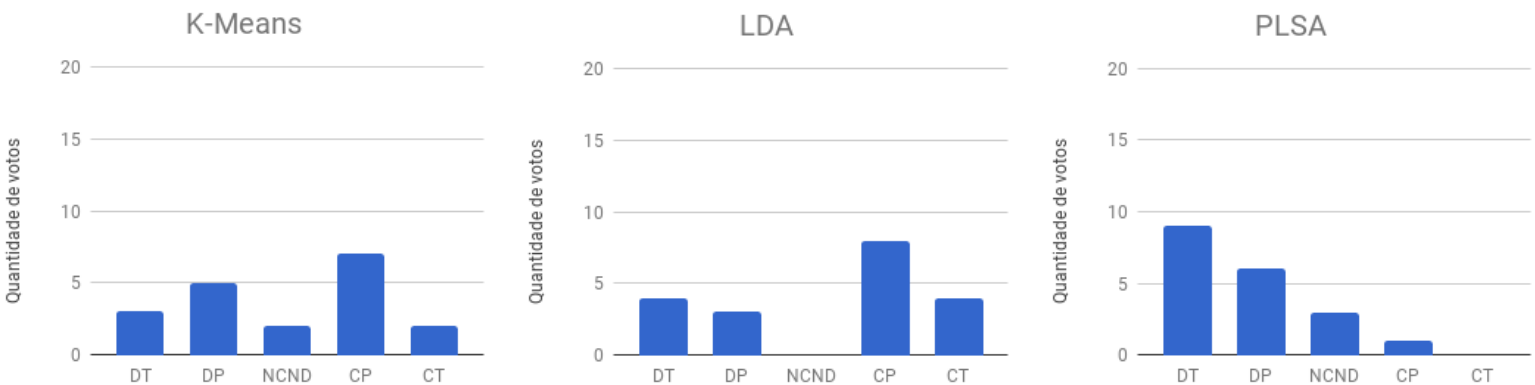
## Coesão dos tópicos

Primeira questão: *"Todos os trechos apresentados compartilham um mesmo assunto."*.

Primeira Consulta: *"defesa de dissertação"*




Segunda Consulta: *"compra de equipamentos"*

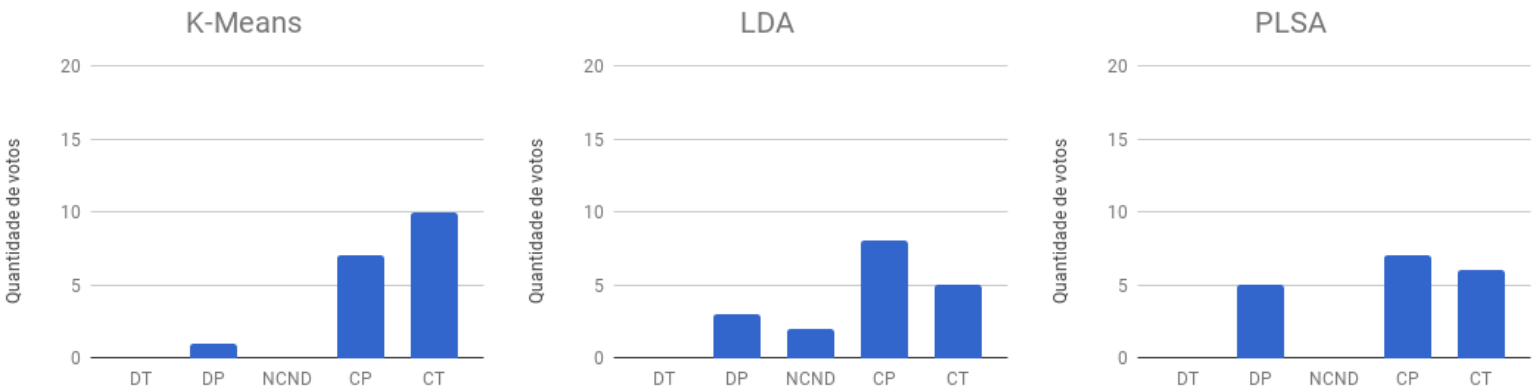


# Comportamento dos extratores em consultas diferentes

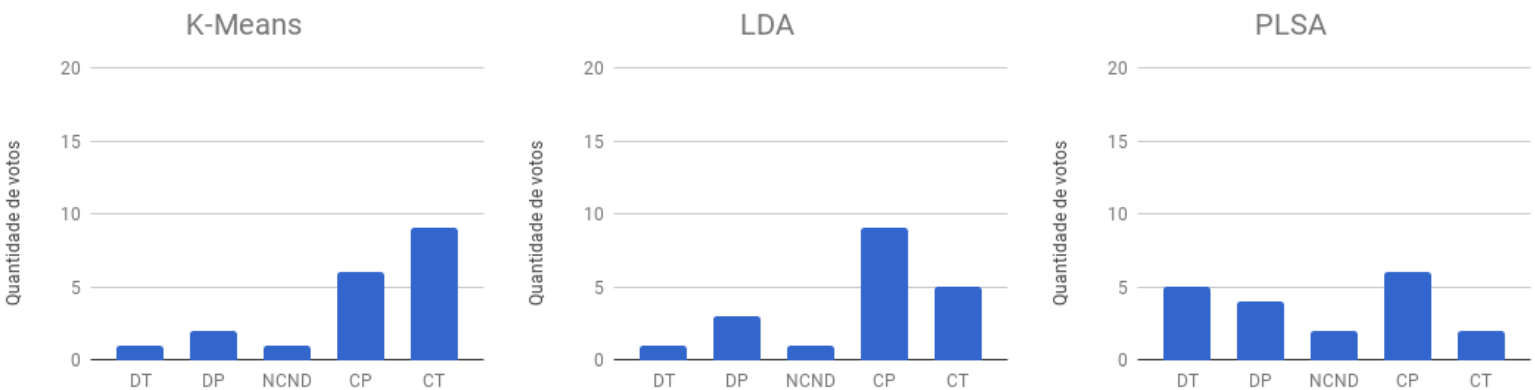
## Representatividade dos descritores

Segunda questão: *"As palavras <descritores> resumem bem o assunto tratado nos trechos."*.

Primeira Consulta: *"defesa de dissertação"*



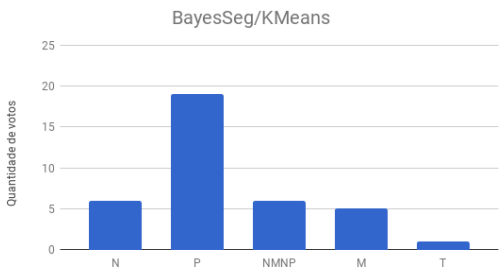

Segunda Consulta: *"compra de equipamentos"*

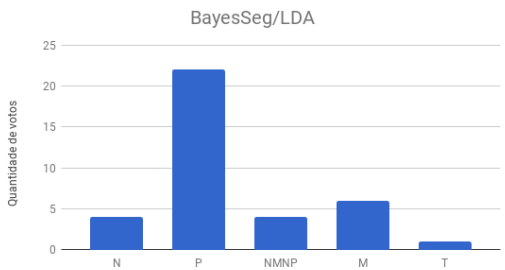


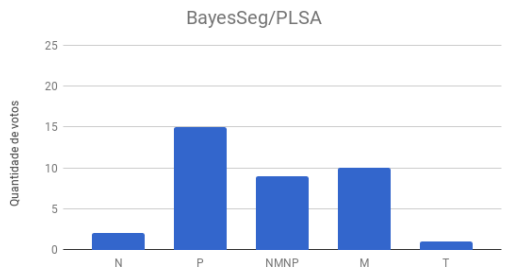
# Qualidade dos segmentos apresentados em diferentes técnicas

## Coesão dos segmentos

Terceira questão: *"Existem trechos que não tratam de um único assunto?"*.



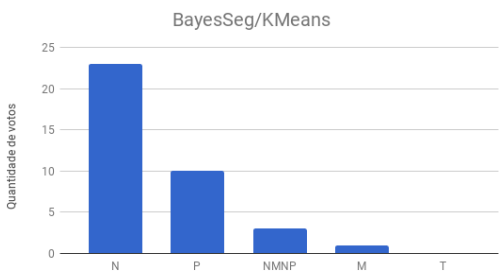



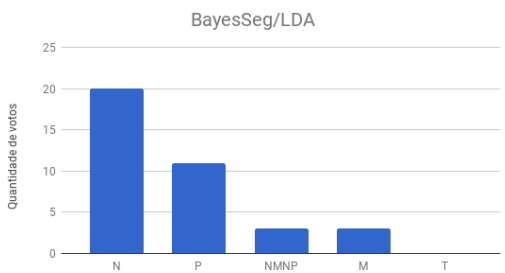



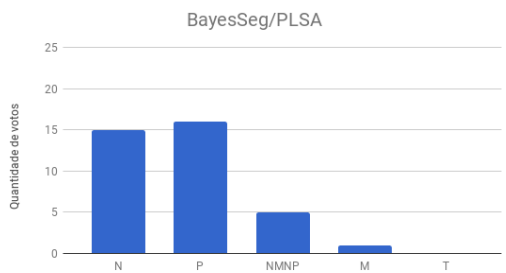
# Qualidade dos segmentos apresentados em diferentes técnicas

## Completude dos Segmentos

Quarta questão: *"Existem trechos incompletos e insuficientes para compreensão do assunto do trecho?"*.



# Roteiro

# Conclusão

A metodologia utilizada nesse trabalho:

- Conecta as técnicas de segmentação textual aos modelos de Extração de Tópicos;
- Gera um estrutura derivada de um *corpus* não estruturado;
- Utiliza variáveis latentes em conjunto com técnicas de Recuperação de Informação.

# Conclusão

## Segmentação

Resultados:

- Medidas abaixo do esperado;
- Impressões satisfatórias dos usuários;
  - Completude;
  - Coesão;

Possíveis melhorias:

- Segmentação de referência com mais anotadores;
- Treinamento dos anotadores;
- Maior concordância entre anotadores;
- Segmentação de referência mais confiável e representativa.

# Conclusão

## Extração de Tópicos

- Melhores resultados com o K-Means;
  - Coesão dos grupos;
  - Capacidade representativa do descritores;

# Conclusão

Contribuições:

- O método para extração de conhecimento em documentos multi-temáticos na língua portuguesa;
- O corpus de atas anotadas;
- A ferramenta para segmentação e anotação manual;
- O sistema proposto e sua implementação;
- As avaliações dos Segmentadores e Extratores de Tópicos.

# Roteiro

# Trabalhos Futuros

Adição de novas técnicas ao sistema:

- Algoritmos de agrupamento incremental
- Categorização dos segmentos (decisão, informe, irrelevante);

Melhorias:

- Inclusão de novos corpora (transcrições de conversas, diálogos em chats, discursos e atas de outras organizações)
- Fontes externas para melhorar os métodos de segmentação textual (*thesaurus* e *clue words*);
- Testes voltados a experiência do usuário.

- Submissão de artigo.

Obrigado!