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 3, 2018

Roteiro

Contexto

As atas registram assuntos discutidos em reuniões e podem ser utilizadas como base de dados.

- Utilizadas como referência e apoio a decisões;
- Um assunto pode ser discutido diversas vezes em reuniões diferentes;
- É desejável recuperar um histórico desses assuntos ao longo do tempo;
- Necessidade de ferramentas automáticas.

Contexto

Recuperação de Informação em documentos textuais:

- Informações contidas em grandes quantidades de texto;
- Inerentemente não estruturados;
- Documentos com múltiplos assuntos;
- Assuntos dispersos pela base de documentos.

Contexto

Nesse cenário, o desafio é encontrar trechos de texto que tratem somente do assunto pesquisado.

Essa tarefa consiste em 2 passos principais:

- Encontrar pontos onde há transição de assuntos;
- Identificar o teor desses assuntos;

Contexto

Os algoritmos de segmentação textual são utilizados para dividir um texto em segmentos contendo um assunto completo e relativamente independente.

- Úteis em aplicações com textos sem indicações de quebras de assunto, como transcrições de áudio, e diálogos em chats.
- Não dão indicações sobre o conteúdo dos segmentos.

Contexto

Os modelos de Extração de Tópicos podem estimar o assunto de cada documento de uma coleção.

- Agrupam documentos por tópico.
- Identificam palavras para descrever o tópico do documento.
- Incorporam conhecimento de domínio aos dados.

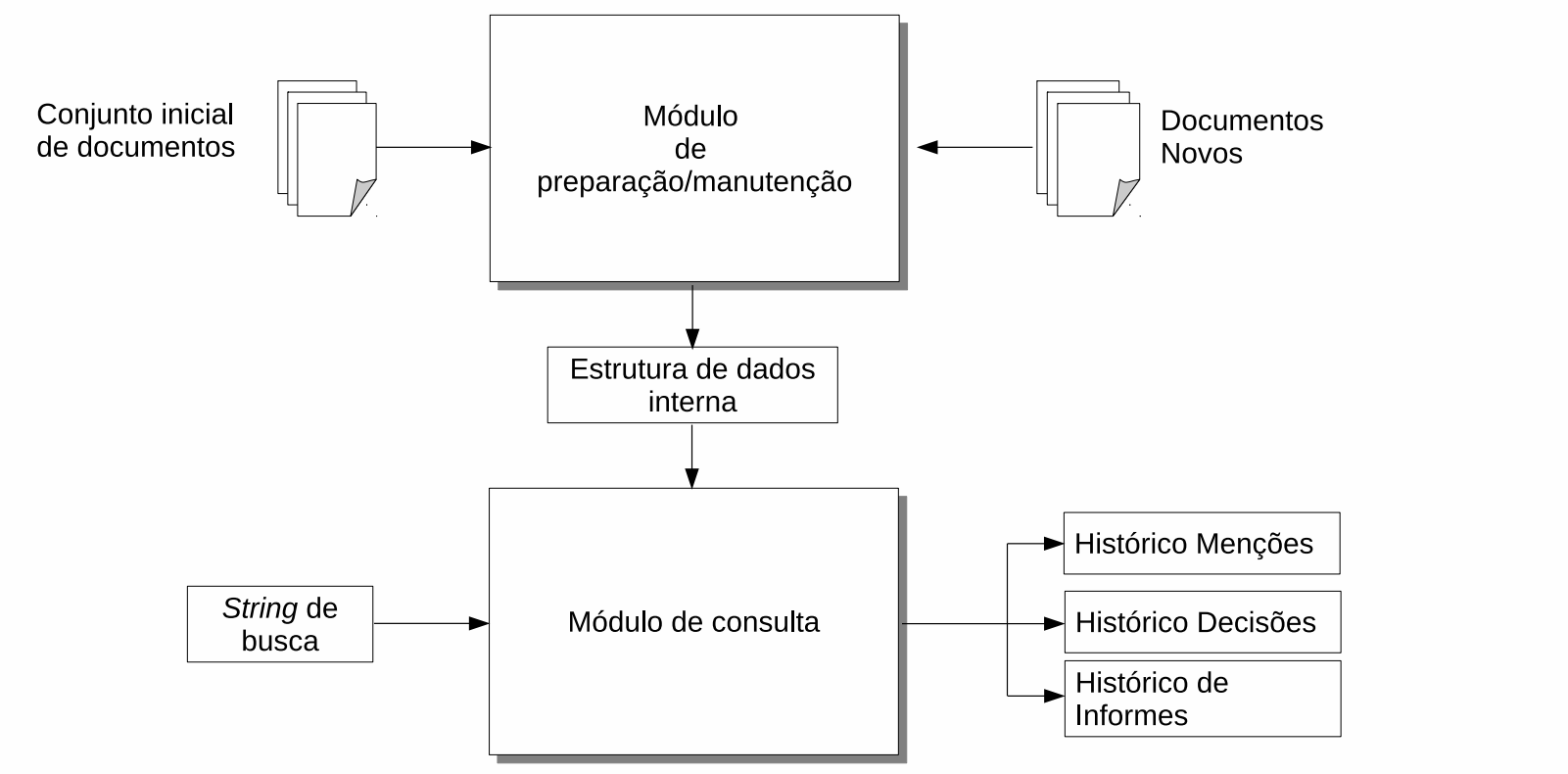
Objetivos

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:

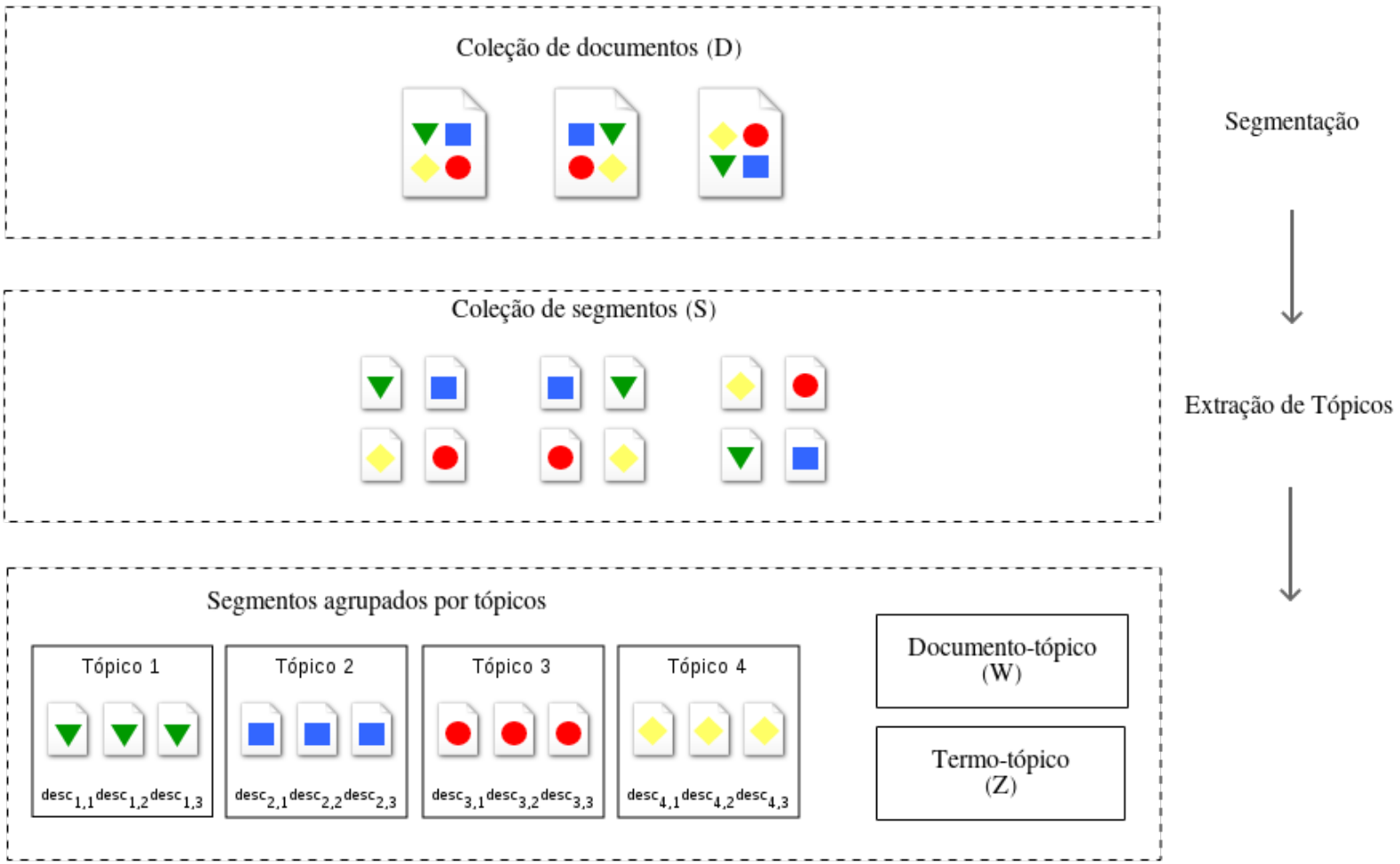
- Gerar uma estrutura mais organizada que a coleção original.
- Utilizar a estrutura latente dos segmentos para Recuperação de Informação.

Visão Geral do Sistema.



Proposta

Estrutura de dados interna e seu processo de geração.



Proposta

Essa abordagem 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.

Distribuição de tópicos em uma ata real.

UNIVERSIDADE FEDERAL DE SAO 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 SAO 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;

docentes; técnica; administrativo; presidente; dia;

computação; conselho; aprovado; acordo; ficou;

cursadas; conselho; coordenação; computação; presidente;

disciplinas; cursadas; libras; conselho; aprovado;

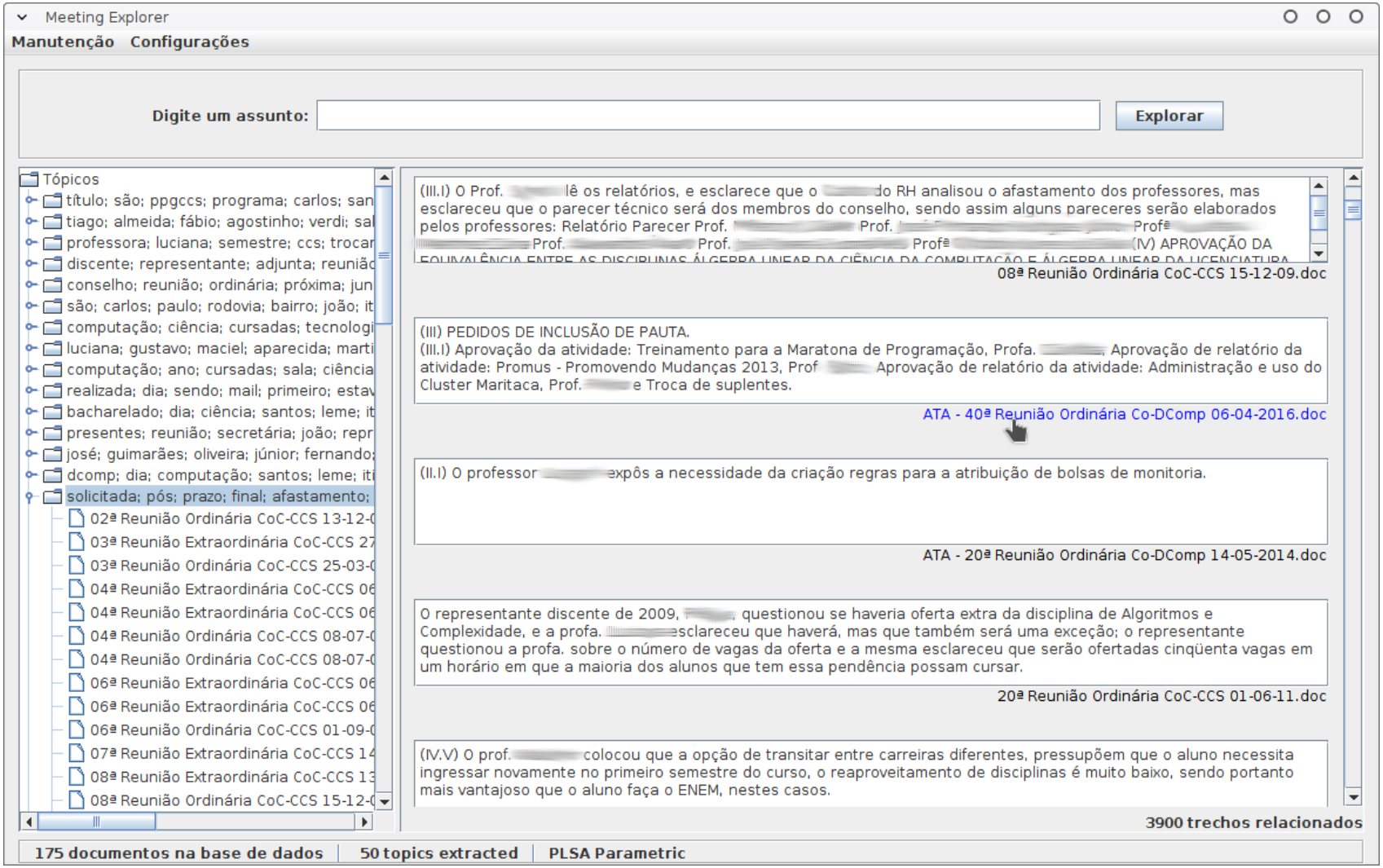
representante; discente; presidente; secretária; turma;

Proposta

A Estrutura de Dados Interna é aproveitada para Recuperação de Informação.

- Os tópicos são representados por seus descritores;
- Usa-se o modelo de espaço vetorial para ranquear os tópicos;
- Exibe-se os segmentos atribuídos ao primeiro tópico do ranking;
- Busca exploratória pelos tópicos mais similares à consulta.

Proposta


Meeting Explorer
Manutenção Configurações
Digite um assunto:
Explorar
Tópicos
título; são; ppgccs; programa; carlos; san
tiago; almeida; fábio; agostinho; verdi; sal
professora; luciana; semestre; ccs; trocar
discente; representante; adjunta; reuniã
conselho; reunião; ordinária; próxima; jun
são; carlos; paulo; rodovia; bairro; joão; it
computação; ciência; cursadas; tecnologi
luciana; gustavo; maciel; aparecida; marti
computação; ano; cursadas; sala; ciência
realizada; dia; sendo; mail; primeiro; estav
bacharelado; dia; ciência; santos; leme; it
presentes; reunião; secretária; joão; repr
josé; guimarães; oliveira; júnior; fernando;
dcomp; dia; computação; santos; leme; iti
solicitada; pós; prazo; final; afastamento;
02ª Reunião Ordinária CoC-CCS 13-12-
03ª Reunião Extraordinária CoC-CCS 27
03ª Reunião Ordinária CoC-CCS 25-03-
04ª Reunião Extraordinária CoC-CCS 06
04ª Reunião Extraordinária CoC-CCS 06
04ª Reunião Ordinária CoC-CCS 08-07-
04ª Reunião Ordinária CoC-CCS 08-07-
06ª Reunião Extraordinária CoC-CCS 06
06ª Reunião Extraordinária CoC-CCS 06
06ª Reunião Ordinária CoC-CCS 01-09-
07ª Reunião Extraordinária CoC-CCS 14
08ª Reunião Extraordinária CoC-CCS 13
08ª Reunião Ordinária CoC-CCS 15-12-
(III.I) O Prof. lê os relatórios, e esclarece que o do RH analisou o afastamento dos professores, mas esclareceu que o parecer técnico será dos membros do conselho, sendo assim alguns pareceres serão elaborados pelos professores: Relatório Parecer Prof. Prof. Profª Prof. Prof. Profª (IV) APROVAÇÃO DA
08ª Reunião Ordinária CoC-CCS 15-12-09.doc
(III) PEDIDOS DE INCLUSÃO DE PAUTA.
(III.I) Aprovação da atividade: Treinamento para a Maratona de Programação, Profa. Aprovação de relatório da atividade: Promus - Promovendo Mudanças 2013, Prof Aprovação de relatório da atividade: Administração e uso do Cluster Maritaca, Prof. e Troca de suplentes.
ATA - 40ª Reunião Ordinária Co-DComp 06-04-2016.doc
(II.I) O professor expôs a necessidade da criação regras para a atribuição de bolsas de monitoria.
ATA - 20ª Reunião Ordinária Co-DComp 14-05-2014.doc
O representante discente de 2009, questionou se haveria oferta extra da disciplina de Algoritmos e Complexidade, e a profa. esclareceu que haverá, mas que também será uma exceção; o representante questionou a profa. sobre o número de vagas da oferta e a mesma esclareceu que serão ofertadas cinqüenta vagas em um horário em que a maioria dos alunos que tem essa pendência possam cursar.
20ª Reunião Ordinária CoC-CCS 01-06-11.doc
(IV.V) O prof. colocou que a opção de transitar entre carreiras diferentes, pressupõem que o aluno necessita ingressar novamente no primeiro semestre do curso, o reaproveitamento de disciplinas é muito baixo, sendo portanto mais vantajoso que o aluno faça o ENEM, nestes casos.
3900 trechos relacionados
175 documentos na base de dados
50 topics extracted
PLSA Parametric

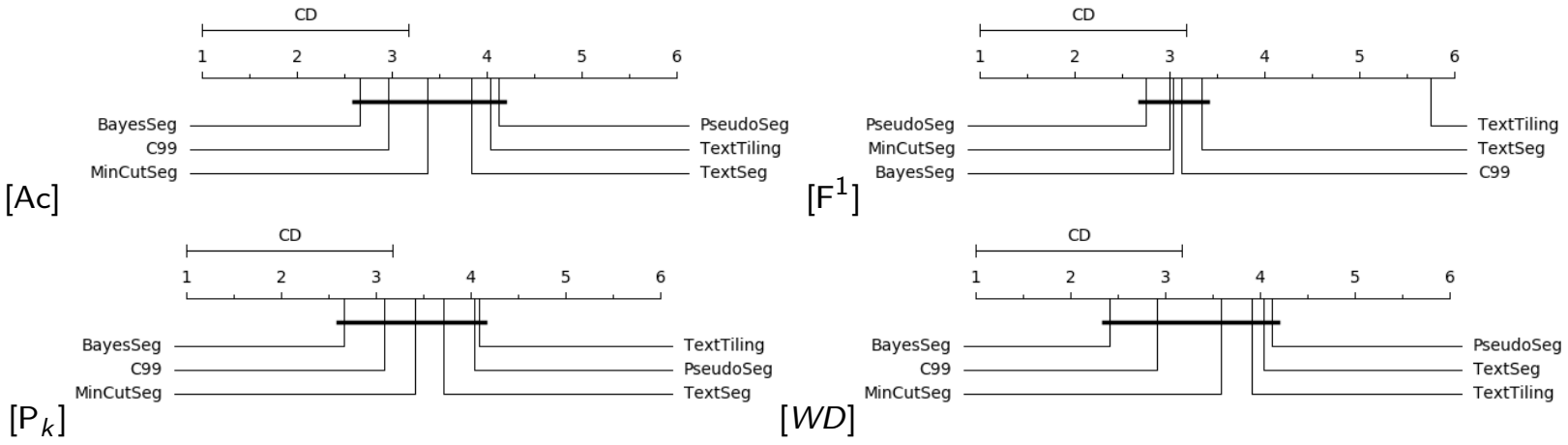
Análise dos Resultados

Os segmentadores foram avaliados objetivamente.

- Criação de segmentações de referência.
- Medidas de similaridade entre a referência e os resultados.
- Testes estatísticos para comparação (Friedman e Nemenyi);

Análise dos Resultados

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



Não há diferença significativa entre os métodos.

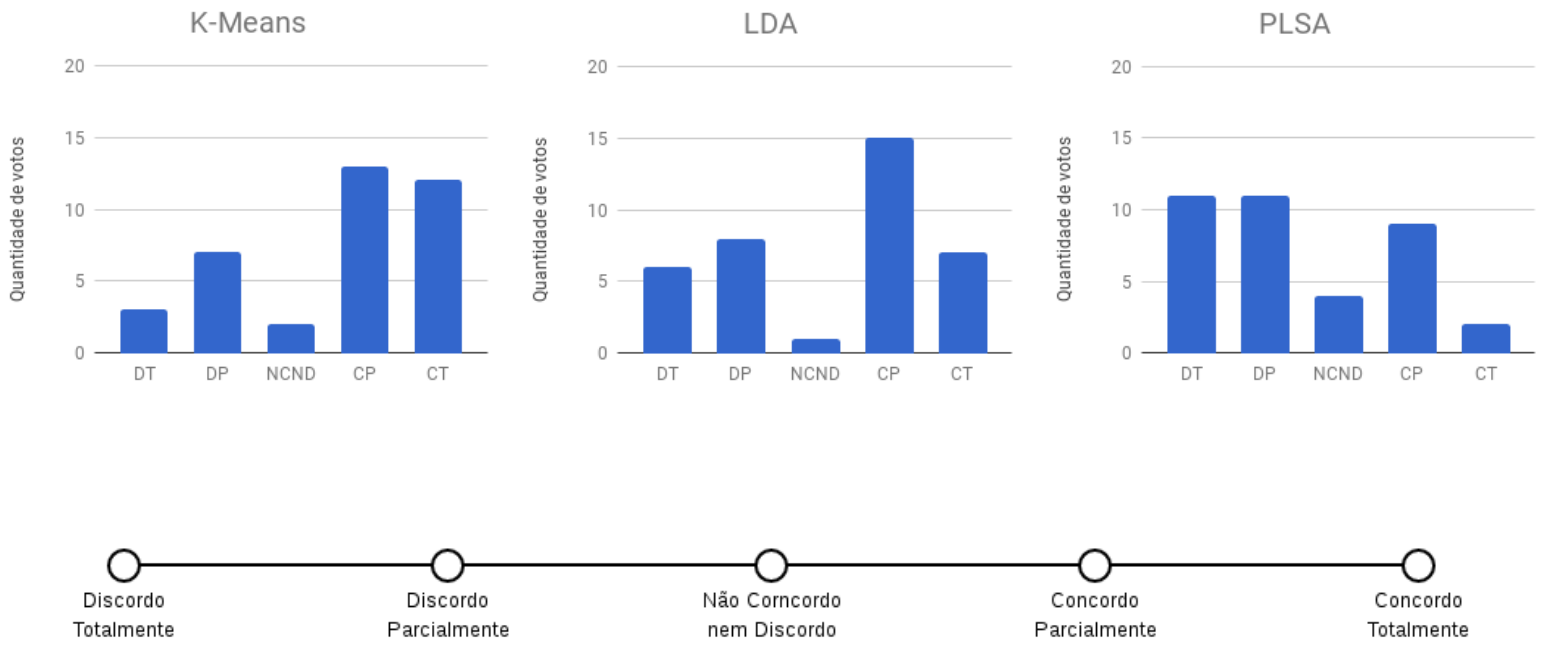
Análise dos Resultados

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.

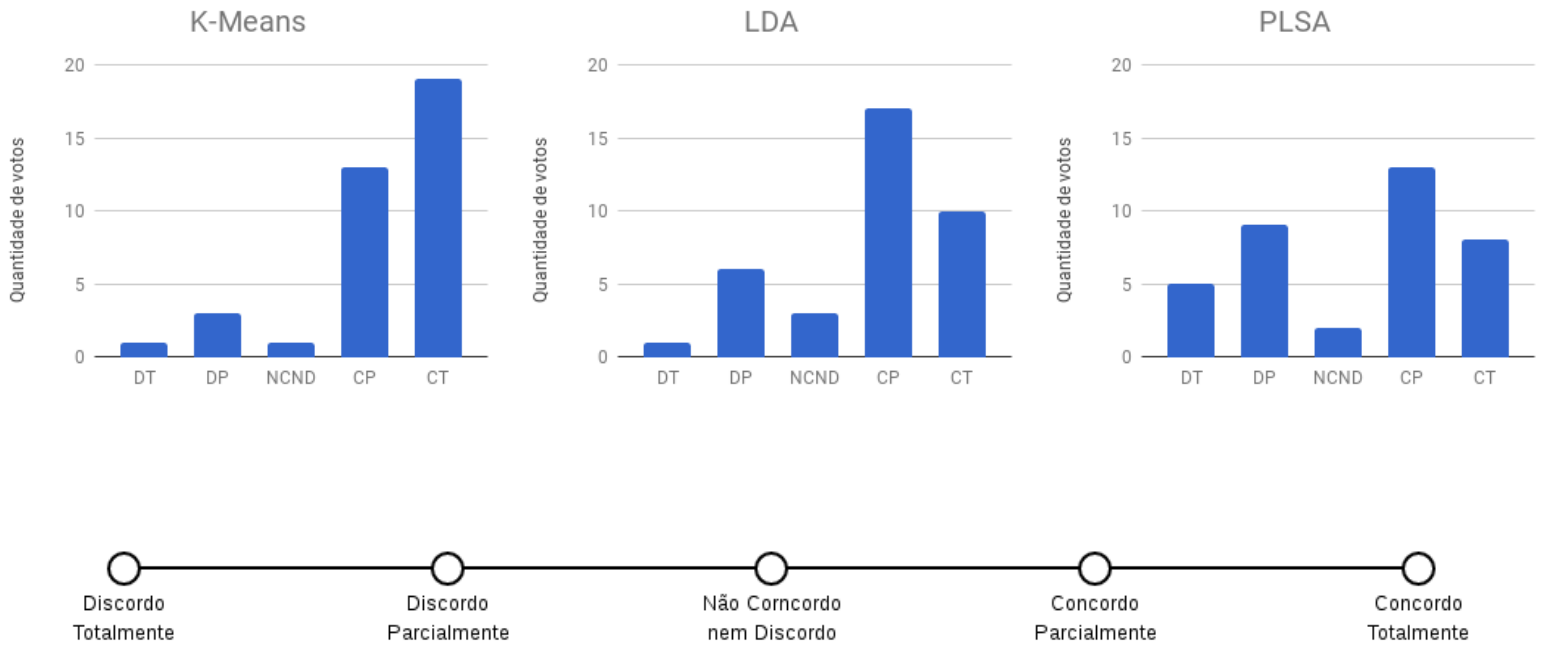
Análise dos Resultados

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



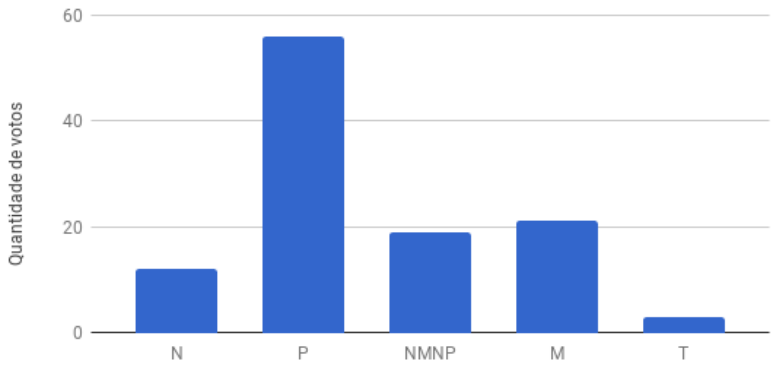
Análise dos Resultados

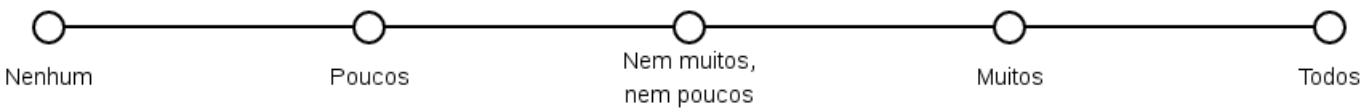
Representatividade dos descritores.
Segunda questão: *"As palavras <descritores> resumem bem o assunto tratado nos trechos."*.



Análise dos Resultados

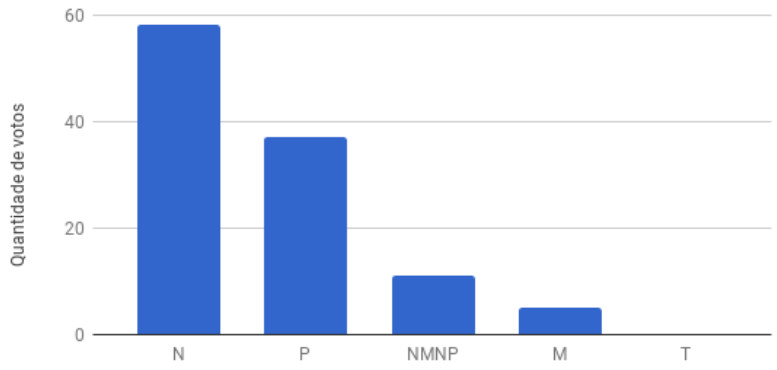
Coesão dos segmentos.
Terceira questão: "*Existem trechos que não tratam de um único assunto?*".

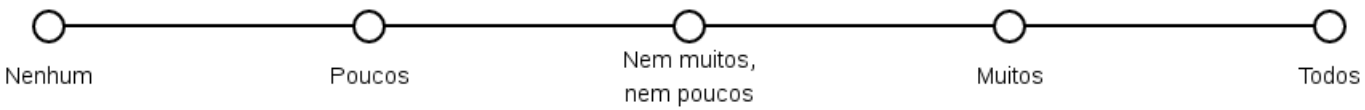




Análise dos Resultados

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





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

- Resultados abaixo do esperado para os segmentadores;
- Impressões satisfatórias dos usuários;
  - Completude;
  - Coesão;
- Segmentação de referência com mais anotadores;
- Melhor treinamento dos anotadores;
- Maior concordância entre anotadores;
- Segmentação de referência mais confiável e representativa.

Conclusão

- Extraem padrões úteis do *corpus*;
  - Capacidade Representativa do descritores;
  - Coesão dos grupos;
- Melhores resultados com o K-Means;
  - Inconstância entre as consultas;

Conclusão

Contribuições

- O método para extração de conhecimento em documentos multi-temáticos;
- 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.

Trabalhos Futuros

- 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*);
- Algoritmos de agrupamento incremental
- Classificação dos segmentos em relação ao tipo de menção ao assunto
- Testes voltados a experiência do usuário

Obrigado!