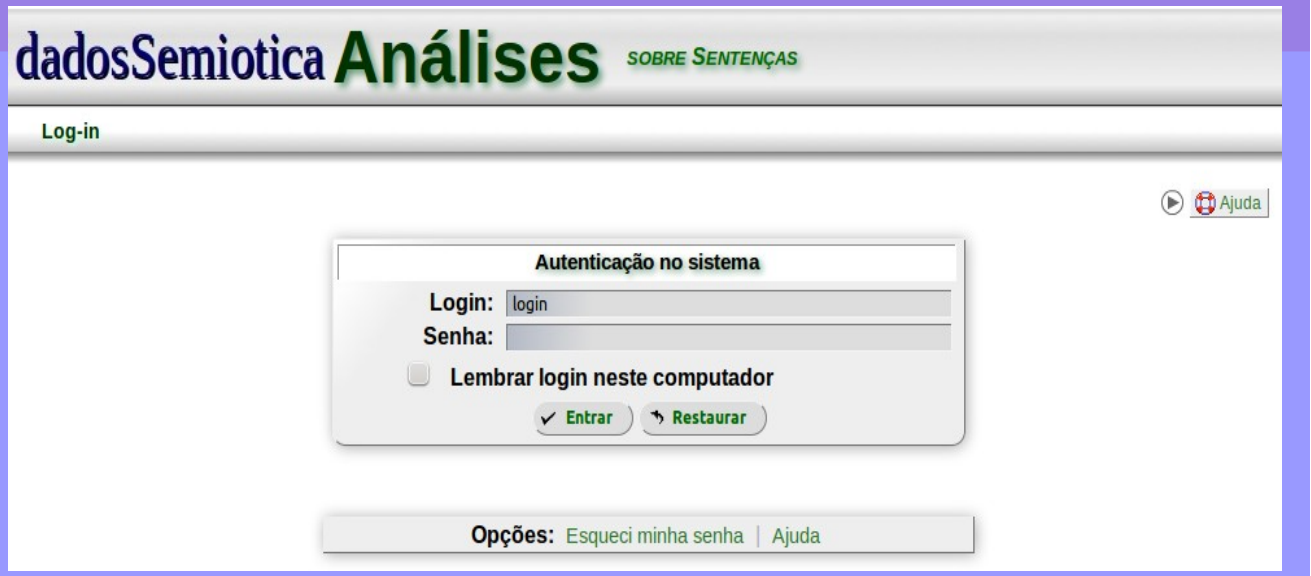
# Manual*

# dadosSemiotica






seguinte

# Sobre - introdução

O programa dadosSemiotica é um coletor e organizador de dados para análise textual.
Quer ser a “mesa de trabalho” do analista de texto.
Seu conceito de texto é proveniente da semiótica:

- *Texto é um todo dotado de sentido.*

A menor unidade para análise, portanto, deve ser também uma unidade dotada de sentido, por isso o dadosSemiotica não trabalha com sílabas nem palavras isoladas, pois palavras isoladas mudam o sentido quando inseridas em uma frase e em um texto.

Ainda dentro deste contexto teórico, a análise consiste em determinar os padrões de cada trecho (que chamaremos, daqui para a frente, de “sentença”) conforme o nível de análise previamente determinado, resultando numa classificação a qual pode ser recuperada por consultas na base de dados do programa.

# Sobre - instalação



1) Checar os requisitos mínimos do servidor:
a - Servidor Web Apache com suporte a PHP e SGBD MySQL ou PostgreSQL;
b - PHP versão 5.2 ou superior;
c - MySQL versão 5.0 ou superior ou PostgreSQL versão 8.0 ou superior.
d - Suporte a Java (Tomcat instalado)
e - R instalado no servidor ($ apt-get install r-base)

2 - Mover o diretório simp para algum local visível na Internet/Intranet;

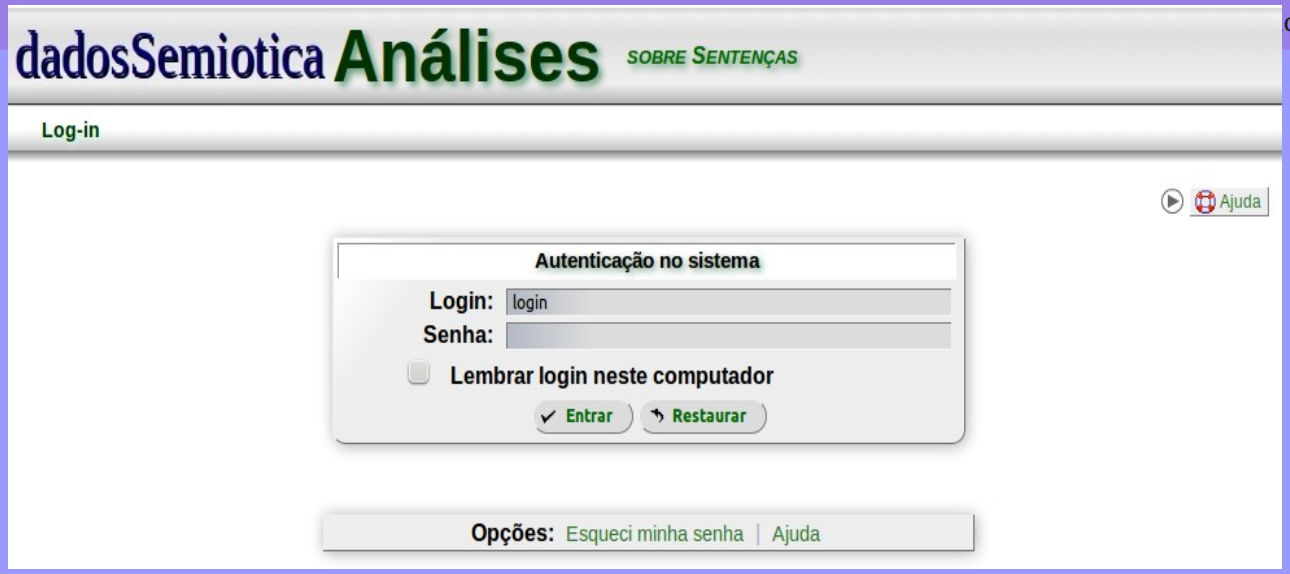
3 - Copiar o arquivo config.bk.php sobre o arquivo config.php ($ cp config.bk.php config.php)
- colocar permissão 777 no arquivo config.php e no diretório "arquivos" ($ chmod 777 config.php arquivos)
- O caminho padrão dos scripts do R é "/usr/bin/Rscript". Dependendo da distribuição Linux do servidor ou configuração específica, o caminho seja outro. Nesse caso, é necessário alterar, no arquivo "constantes.php", a diretiva "CAMINHO_R_SCRIPT".

4 - Acessar o endereço do sistema com visibilidade na Internet/Intranet;

5 - Preencher os formulários de configuração atentando-se à ajuda. Caso você não tenha a senha do root do servidor mysql, deve antes criar um usuário e banco de dados.

6 – Pronto!
Algum problema durante a instalação? Copie o arquivo config.bk.php com o nome config.php, sobrescrevendo-o, apague os cookies do navegador e tente instalar novamente.

# Usos



# Usos - introdução

O dadosSemiótica é uma interface web para análises de sentenças, permitindo a classificação manual, semi-automática e automática de uma grande quantidade de textos.

Sendo um sistema multiusuários, permite a utilização simultânea de diferentes recursos por diferentes pessoas com diferentes ou mesmo nível de acesso (grupos). Para organizar esses trabalhos, o programa baseia-se em projetos.

Um projeto é um conjunto de textos a serem analisados sob um conjunto fechado de categorias. O projeto é dependente do analista, responsável por definir quais textos e quais categorias compõe cada um de seus projetos, bem como responsável por realizar as análises.

Na escolha das categorias reside a orientação teórica das análises e é com base nesta que o analista deve definir um número variável de classificações e, sempre que possível, um código para o registro de cada análise.

# Usos - grupos

Existem 3 grupos de usuários no dadosSemiotica:

- **Administrador**: é o usuário que define as configurações gerais do site, podendo modificar permissões dos grupos, criar e editar usuários e definir quais os recursos que estarão disponíveis naquela instalação. É ele quem cadastra os gerentes
- **Gerente**: é o usuário que cria as categorias, ou seja, define quais os tipos de análises poderão ser feitas naquela instalação, e faz o *upload* dos *corpora*. Ele também gerencia os analistas, cadastrando-os.
- **Analista**: é o usuário que cria os projetos, escolhendo os textos e categorias que os compõe, e realiza as análises.

Para obter cada um desses níveis de acessos é necessário realizar um login diferenciado.

# Usos - projetos

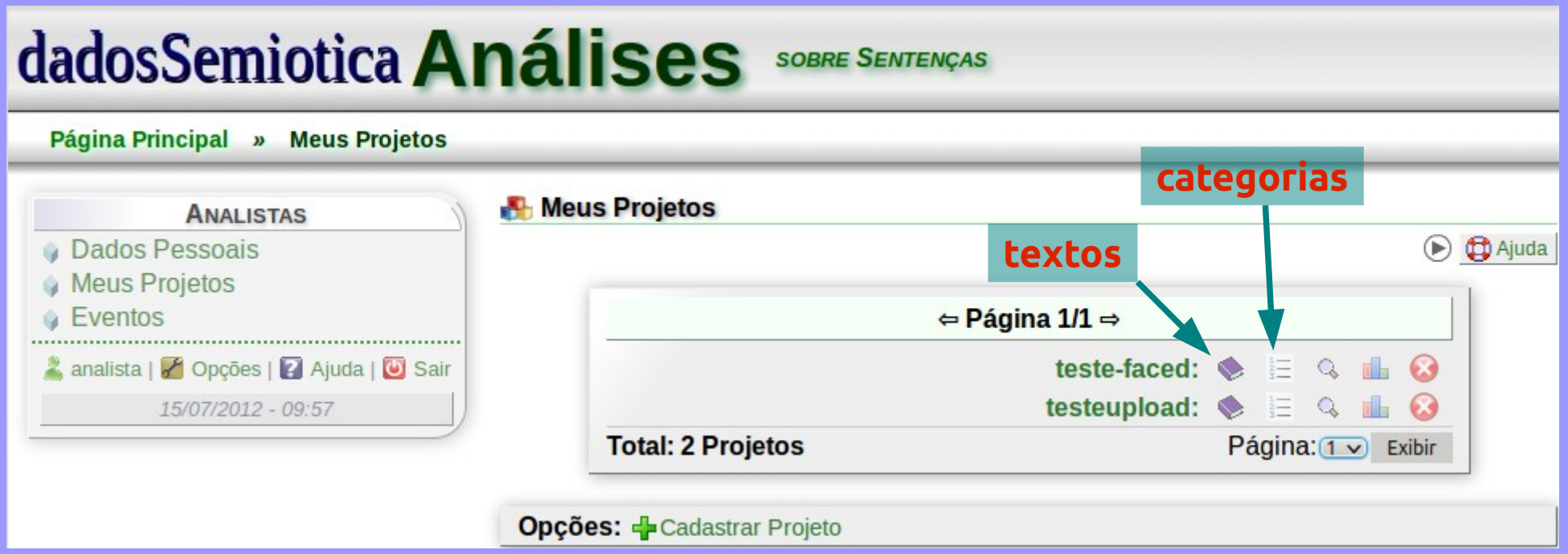



Esta é a página do analista listando seus projetos.

Todo o site trabalha com o acesso aos recursos usando
i) o menu à esquerda,
ii) a lista de navegação no topo, embaixo do logo (onde é possível voltar para a página anterior)
iii) as figuras ao lado dos projetos, frases etc

# Usos - projetos

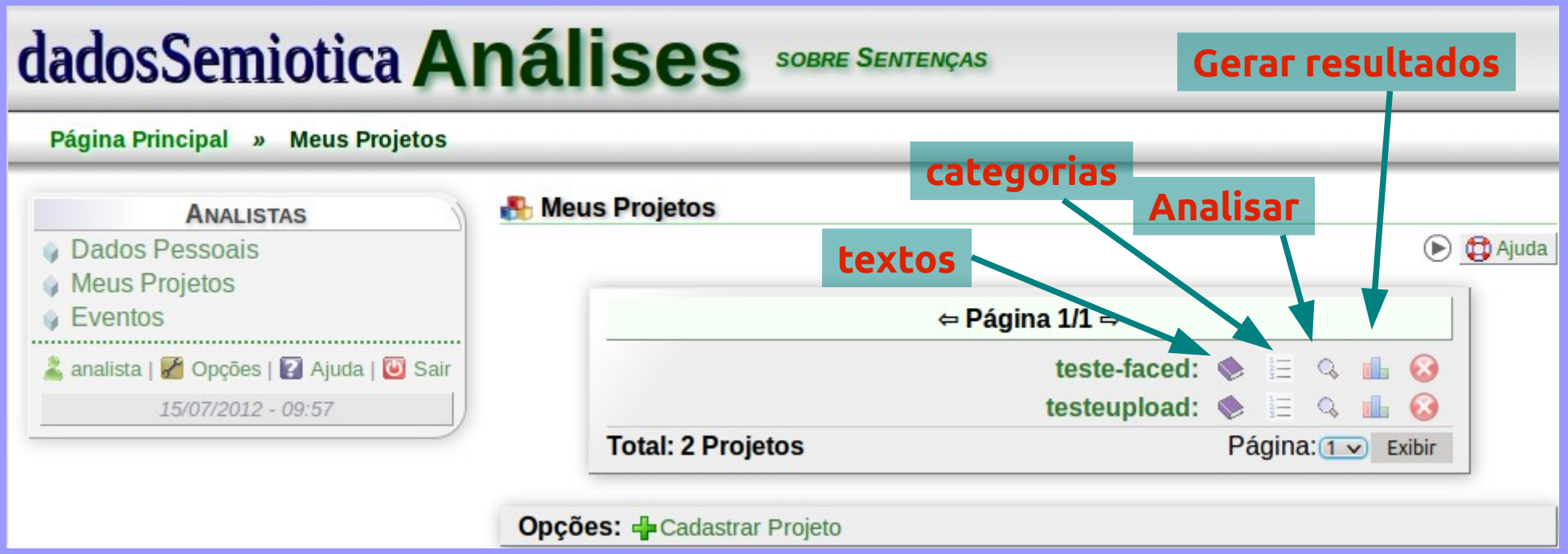





Um projeto é um conjunto de textos e categorias a ser analisado. Para a semiótica, trabalhar com um conjunto de textos significa definir um contexto interno (como o conjunto de obras de um mesmo autor).
Assim, o projeto é um contexto de análise.

# Usos - projetos






Ao lado do nome do projeto eu tenho as seguintes opções:

- *Textos de análise*: permite visualizar, apagar e inserir textos;
- *Categorias de análise*: permite visualizar, apagar e inserir categorias;
- ***Visualizar*: dá acesso à interface de análises.**
- *Gerar estatísticas*: permite gerar resultados e tabelas das análises
- *Excluir*: permite excluir o projeto inteiro.

Abaixo da lista de projetos o analista tem a opção de cadastrar outro projeto.

# Usos - projetos

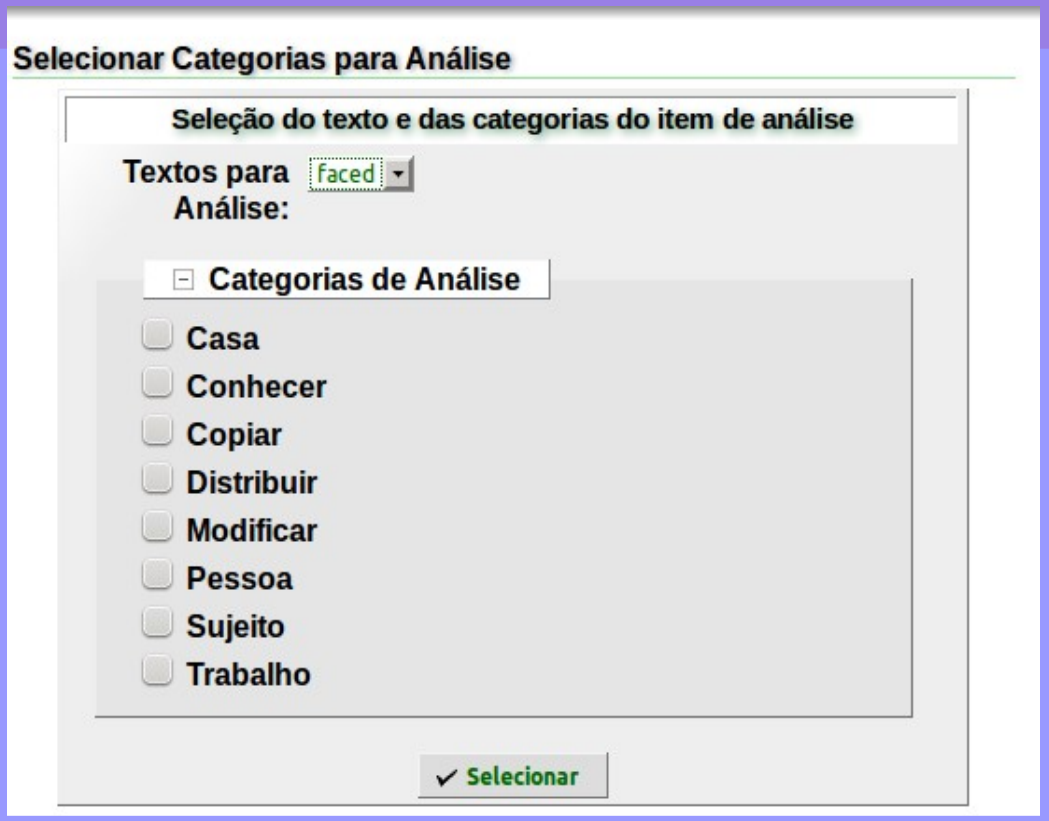
Só o analista pode deletar seus projetos.
Ao deletar um projeto, os textos e as categorias permanecem no sistema, mas as análises manuais e o conjunto de relações entre categorias e textos, que definem o projeto,
são apagados do banco.
Por isso é melhor, se o projeto não foi encerrado, deletar uma categoria, uma análise ou um texto do projeto antes que deletar o projeto inteiro. A deleção do projeto pode ser usada, no entanto, quando, de fato, o projeto está obsoleto, houve má formação das análises ou outro motivo pelo qual o analista acredite que seja melhor começar de novo.

# Usos - análises


Para iniciar uma análise, é necessário escolher um projeto, pois não existe análise sem projeto.
Abrindo a lista de projetos, clique na lupa para iniciar o processo de visualização da interface de análise.

# Usos - análises






No dadosSemiótica, analisa-se um texto por vez. Isso é importante para manter a identidade semiótica do texto. Pode-se, no entanto, escolher até quatro categorias por vez na hora de iniciar as análises. A escolha do número de categorias a ser visualizada deve favorecer a integridade da análise: algumas categorias podem influenciar a análise de outras, o que pode ser desejável ou não, dependendo do caso. A decisão cabe ao analista.

# Usos - análises

Interface das análises manuais.
O sistema foi planejado para receber ilimitado número de categorias diferentes por projeto, mas a tela de análises, para comodidade do analista, permite apresentar até quatro categorias para análises. Note que, dependendo da resolução do monitor, é possível que não caibam 4 categorias ao mesmo tempo na tela, mesmo com o navegador maximizado.

# Usos - análises

Da esquerda para a direita vemos a sentença e os campos de análise, tantos quantas forem as categorias escolhidas.
Nos campos é possível:

- Digitar a análise
- Selecionar e arrastar texto de outra análise para o campo (copia o texto selecionado)
- Selecionar e arrastar texto da sentença para o campo
- Apagar a análise (clique na lixeira)
- Copiar a última análise acima em todos os campos vazios (clique no pincel)

# Usos - análises

A visualização de um campo sempre sobre para o topo quando se digita no campo. Esse recurso permite a visualização das sentenças seguintes, ainda não analisadas, melhorando a contextualização da análise.

# Usos - análises

As análises só são salvas no banco após clicar no botão atualizar, no final da página.
É possível reabrir uma categoria para analisar novamente ou para usar seu conteúdo para fazer uma segunda análise em outra categoria. Nesse caso, todas as análises feitas anteriormente aparecem na tela e, havendo ou não mudanças, todas as categorias abertas são salvas novamente ao clicar no botão atualizar.

# Usos - análises

Ao terminar uma análise, é possível:
1) voltar ao projeto para escolher outras análises.
2) escolher outro projeto para trabalhar

Em qualquer momento é possível salvar uma análise, mesmo incompleta, e voltar depois para continuá-la, revisá-la etc

# Usos - análises

## Análises quantitativas e qualitativas:

- **O sistema aceita qualquer tipo de análise no campo de análises** (numérico ou textual). A maioria dos programas de estatística é capaz de detectar classificações (como s n para sim não) e tratar esses dados de forma adequada, por isso use o campo como for melhor para o tipo de dado desejado.
- **É importante padronizar os textos de análise sempre que se desejar verificar estatisticamente os dados.** Digitação errada pode provocar outliers e afetar os resultados, prefira então evitar acentos e mesmo simplificar (substituindo palavra por letra)sempre que possível.
- **As análises podem ser textualizadas.** Isso permite inserir categorias de comentários, análises descritivas qualitativas etc

# Usos - categorias

As categorias são a base teórico-metodológica das análises. A opção de deixá-las sob responsabilidade do gerente é permite, numa mesma instalação do dadosSemiotica, construir um banco de dados com múltiplas análises e múltiplos analistas alimentando a mesma base.

Se cada analista “inventasse” suas próprias categorias de análise, de pouco valeria que esses resultados todos estivessem no mesmo banco e o sistema funcionaria apenas internamente aos projetos.

Como não cabe ao analista, mas ao gerente, determinar as categorias de análise, é possível gerir o escopo dos resultados de forma a padronizar as classificações e permitir buscas cruzadas nos diferentes corpora.

O Gerente, portanto, é um mentor: ele e somente ele determina que tipo de análises serão realizadas numa dada instalação do dadosSemiotica. Assim, o dadosSemiotica pode ser usado também para fins didáticos, numa situação de professor-aluno.

# Usos - categorias



O sistema possui dois tipos de análises: automáticas e manuais. As categorias automáticas são geradas por módulos de pré e pós processamento, os primeiros acionados pelo Gerente no upload e os últimos acionados pelo analista na fase de geração final de tabelas de resultados. O processo de criação das categorias manuais, pelo Gerente, é um processo muito simples:

# Usos - categorias

A escolha das categorias pelo Gerente deve basear-se nos propósitos gerais da instalação.
Por exemplo, a teoria a ser utilizada pode sugerir categorias relevantes.
Sugerimos a criação hierárquica, por exemplo:
Turma → a que turma pertence?
Turma-código → código da turma no sistema
Disciplina-código → disciplina + código da turma

# Usos - resultados

O trabalho do analista, como sabemos, não termina com as análises pontuais que fazemos do *corpus*: é necessário recuperá-las de forma organizada a fim de, a partir de cruzamentos, comparações e análises gerais, podermos observar padrões (ou a falta deles), tendências e conflitos, a partir do que, finalmente, poderemos chegar a alguma conclusão sobre o projeto desenvolvido.

O dadosSemiotica trabalha essa etapa organizando os resultados em tabelas: para visualizar os resultados das análises, o analista deve solicitar uma tabela para visualização direta ou para análise estatística dos dados. E pode solicitar tantas tabelas quantas desejar para cada projeto.

Todas as tabelas são solicitadas por projeto: elas vem com cada linha correspondendo a uma sentença de cada um dos textos do corpus do projeto e, nas colunas, as informações correspondentes (qual o texto, qual a sentença e outras informações escolhidas pelo analista, dependendo de sua necessidade).

# Usos - resultados


Para acessar a tela de requisição de tabelas, basta clicar no botão “gerar estatísticas” do projeto.

Além dos dados das análises manuais, o analista pode solicitar:

• Dados do módulo de chat, quando se tratar de *corpus* de chat em que o módulo foi ativado no upload.

•Dados de análise morfossintática. Todos os textos são pré-processados pelo módulo de análise morfossintática, baseado no CoGroo, e os resultados podem ser requisitados com buscas específicas no momento da geração das tabelas, no módulo de pós-processamento.

• O texto da sentença analisada e o texto do trecho de texto de onde proveio (quando a sentença faz parte de um parágrafo, por exemplo).

# Usos - resultados

Os dados que podem ser obtidos do módulo de chat são:

- data
- hora
- nick
- histórico
- notificação
- estado (online/offline)
- tempo resposta

Atualmente as análises estão configuradas para receber logs no formato de registro do programa Konversation*, que possuem, em linhas de entrada de textos pelos participantes, a seguinte estrutura:

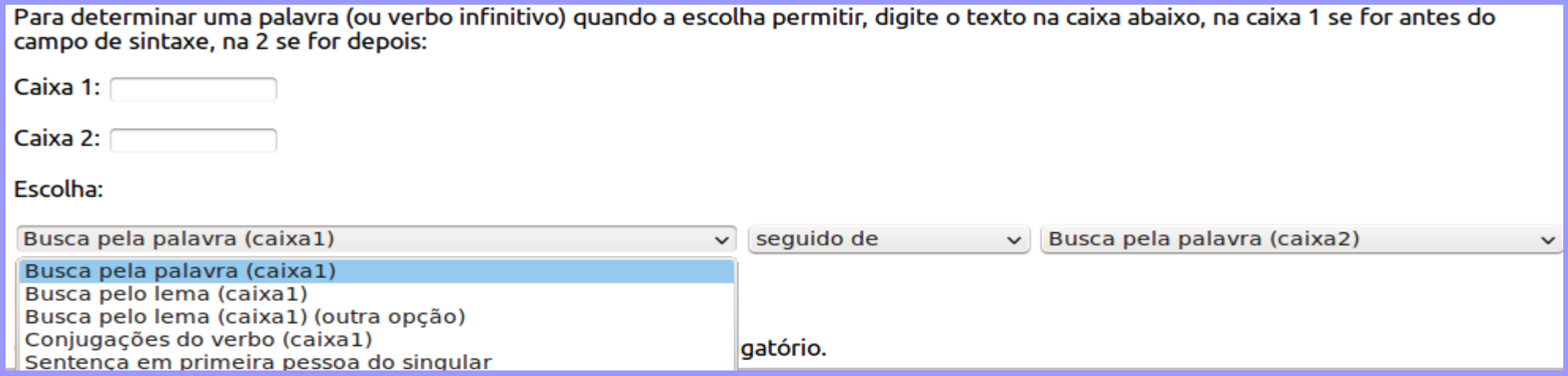
[dia_semana dia_mês mês ano] [hora:minutos:segundos] <nick>  texto

# Usos - resultados



| Symbol | Meaning |
|---|---|
| A << B | A dominates B |
| A >> B | A is dominated by B |
| A < B | A immediately dominates B |
| A > B | A is immediately dominated by B |
| A $ B | A is a sister of B (and not equal to B) |
| A .. B | A precedes B |

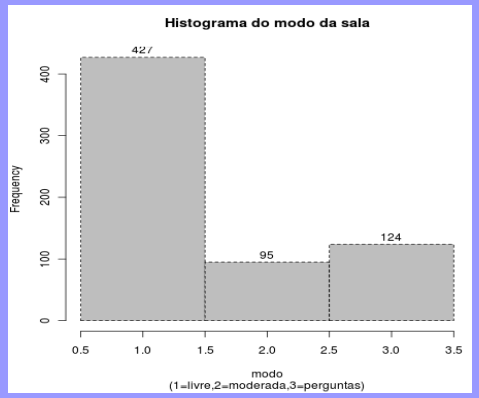
Para incluir resultados da busca morfossintática, o analista deve informar a busca no campo correspondente, seguindo a sintaxe de Class Tregex Pattern*, como na figura acima.
Está em desenvolvimento um formulário para buscas simples, como a figura abaixo.



# Usos - resultados

O dadosSemiotica vai utilizar o R* para obter algumas informações estatísticas sobre os dados indicados pelo analista para inspeção, dentre todos escolhidos para as tabelas:

- Histograma (gráfico)
- Número de observações
- Número de nulos
- Média
- Desvio Pádrão
- Mediana
- Valor Mínimo
- Valor Máximo




Caso não se trate de uma variável independente (numérica), o dadosSemiotica vai transformar os valores em uma sequencia de inteiros.

Estas informações são enviadas por e-mail, junto com a tabela solicitada e permitem ter uma visão global dos dados antes de seu tratamento, de modo que o pesquisador possa verificar a eventual necessidade de ajustes.

# Desenv. - introdução

O dadosSemiotica foi desenvolvido como uma ferramenta auxiliar no processo de análise de dados, especialmente análise de textos, visando possibilitar a manipulação de grande quantidade de dados em teorias que normalmente trabalham com análises qualitativas. A interface da semiótica com a fonética* mostrou que, se por um lado a teoria semiótica, como teoria da área das humanas, é sólida o bastante para interrelacionar-se com teorias das exatas, por outro lado somente com o uso de uma ferramenta para indexar, classificar e organizar os dados com grande facilidade de uso e com suficiente estabilidade e coerência com os pressupostos teóricos envolvidos permitiria a aplicação desta teoria em um corpus de grande extensão.

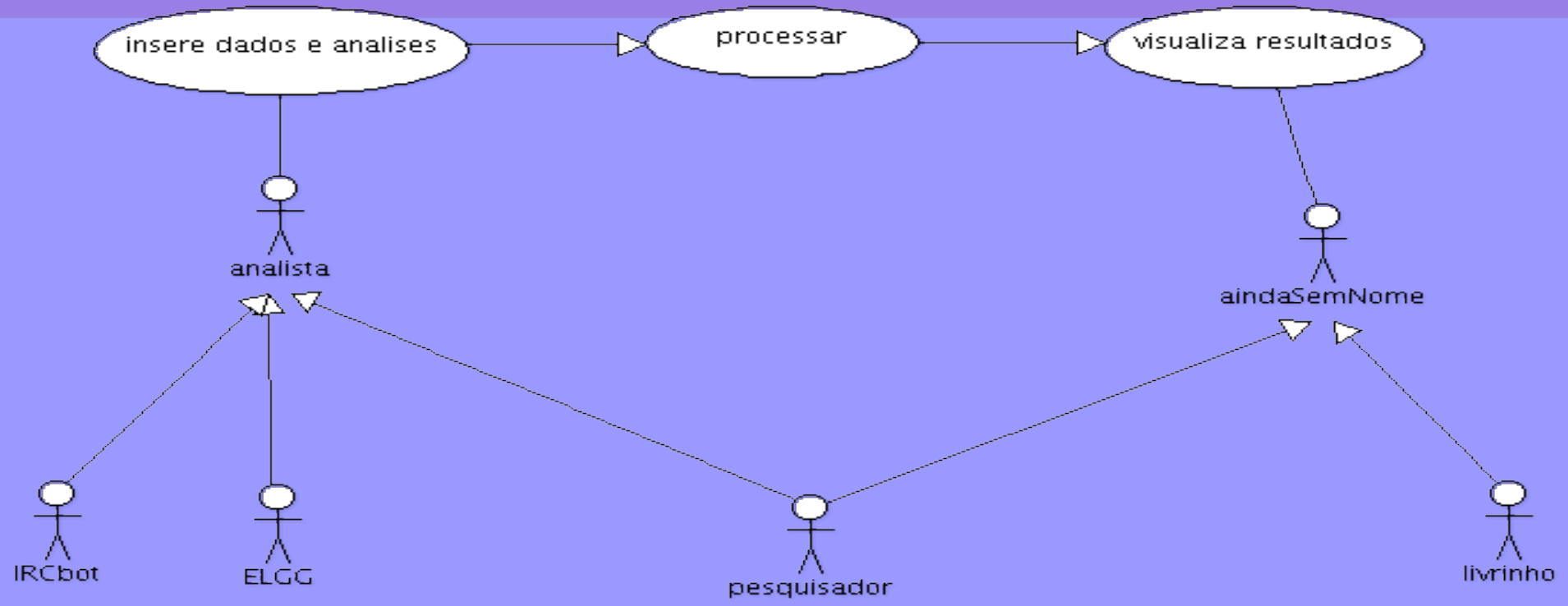
É importante notar que, embora o programa tenha sido criado como auxiliar em pesquisas semióticas, *qualquer análise de texto que tenha o texto como unidade global e a sentença como unidade mínima pode utilizar esta ferramenta* com o mesmo grau de aproveitamento, pois a abordagem teórica não é pré-determinado pelo sistema.


O formato original do programa previa a entrada de dados sob controle exclusivo do usuário Analista, um ser humano que escolhia e preparava os material de entrada. A saída final eram gráficos solicitados pelo próprio analista.

A versão 1.0 possui dois módulos de pré-processamento, o Morfossintático, baseado no Cogroo*, e o Módulo de Chat, especialmente desenvolvido para tratamento de logs de chat do IRC*. Além destes, o dadosSemiotica trabalha com um módulo de buscas morfossintáticas, baseado em Tregex*, e seu módulo de geração de tabelas para download das análises está vinculado ao R*, possibilitando a geração de histogramas e descrição automática para verificação prévia do corpus.

# Desenv. - introdução

dadosSemiotica



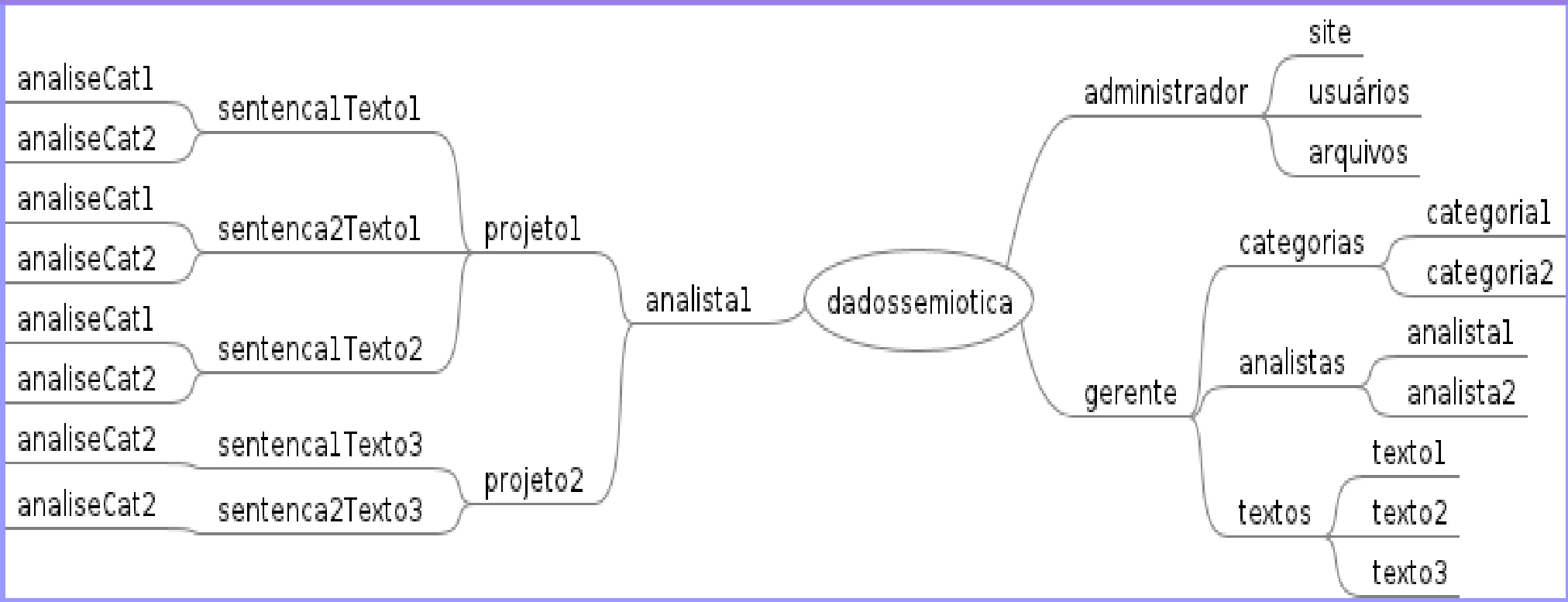

O objetivo das mudanças é duplo: ao mesmo tempo em que se implementam soluções automáticas para coleta dinâmica de dados em interfaces de interação, como o chat do IRC, mantém a possibilidade de intervenção direta do analista/pesquisador em resultados locais ou mesmo acrescentando categorias de análise a qualquer momento do processo. Com isso espera-se manter a produtividade do ambiente sempre aberta ao controle humano sem prejudicar a coleta automatizada.

# Desenv. - download

A versão 1.0 está disponível para download em:
https://sourceforge.net/projects/dadossemiotica

O código em desenvolvimento está disponível no sistema de gerenciamento de versões Subversion e pode ser baixado na mesma página.

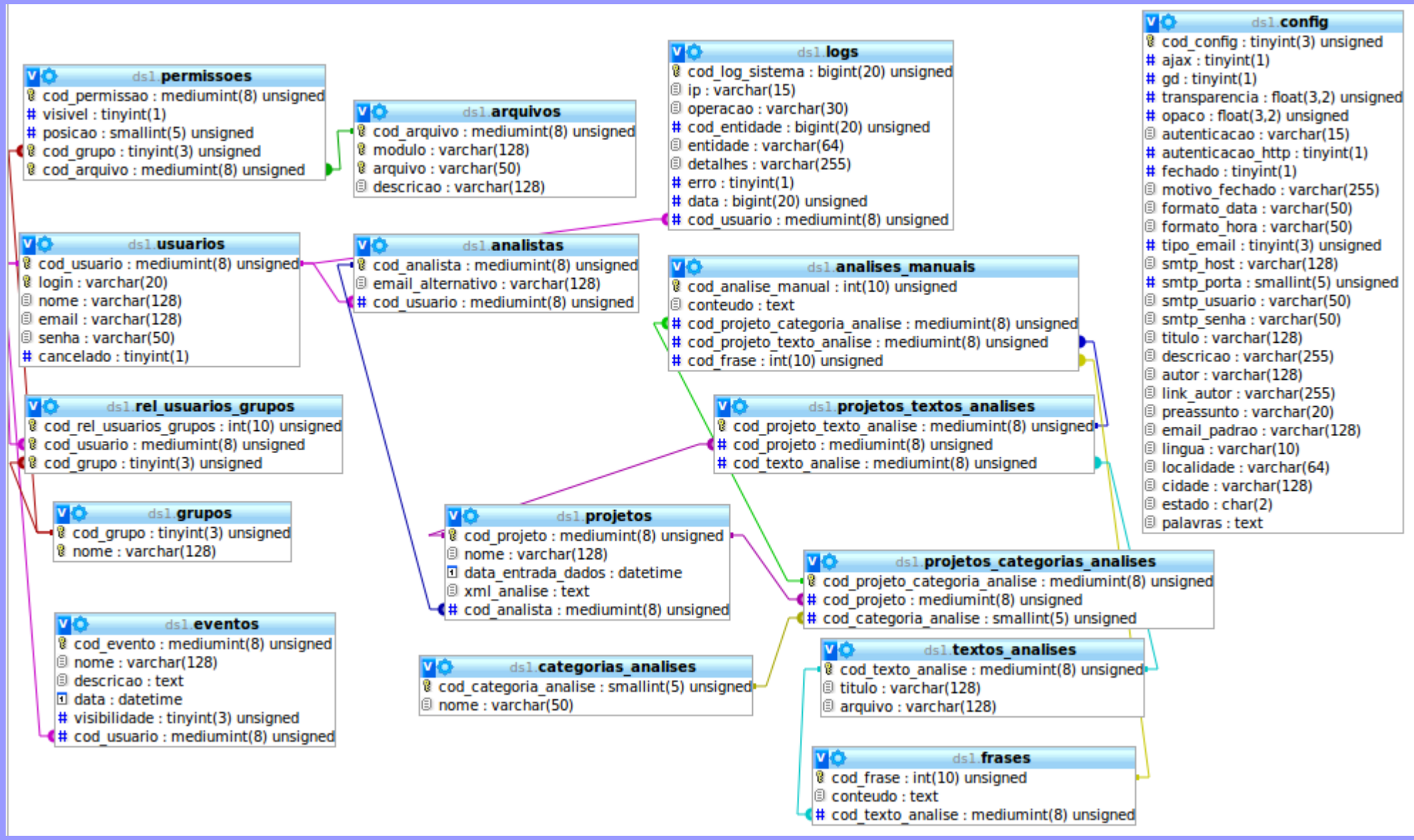
# Desenv. - site e rotas




Os ambientes são diferenciados entre os grupos conforme as atribuições dos usuários.
O SIMP* (framework de desenvolvimento usado no dadosSemiotica) tem um módulo que gera dinamicamente a documentação das entidades/atributos/relacionamentos a partir da área de administração.

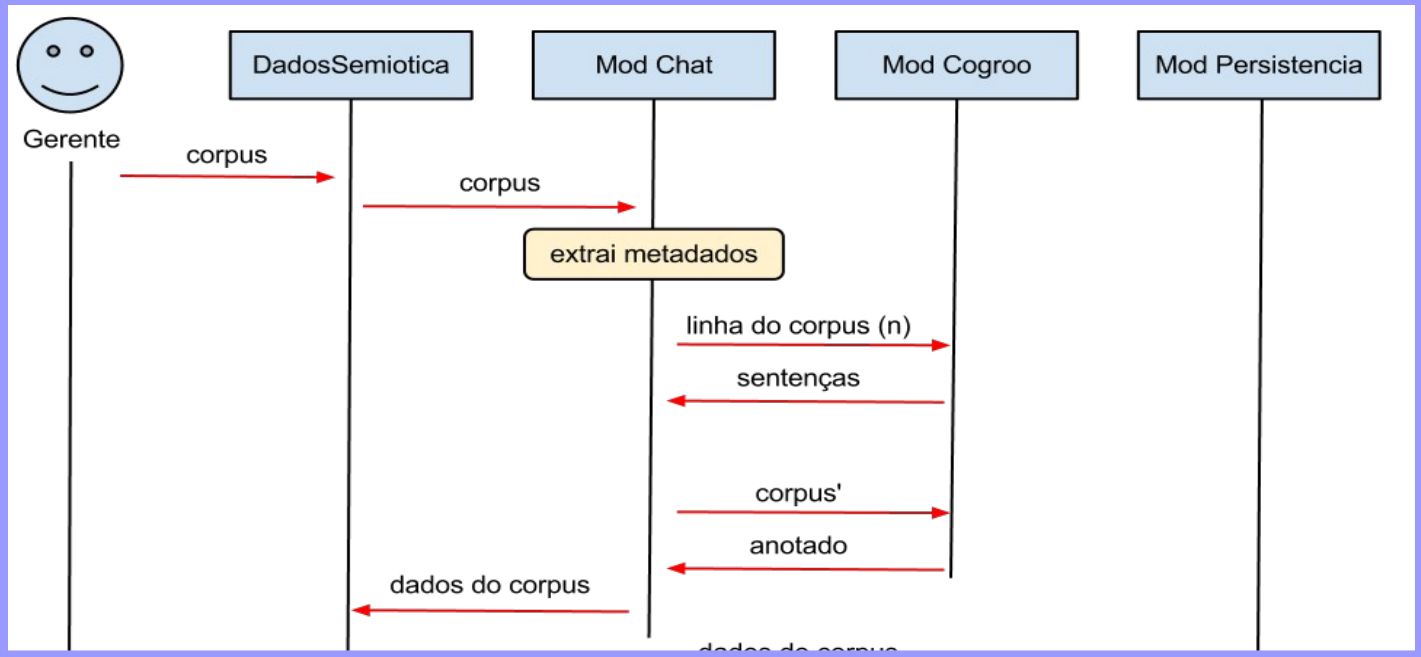
O dadosSemiotica foi desenvolvido em php/mysql/ajax e xhtml com o framework SIMP*, de Rubens Takiguti Ribeiro.

O principal objetivo do SIMP é prover facilidades e padrões para o desenvolvimento de sistemas de informação modulares em plataforma web. Sua estrutura foi preparada para a criação de sistemas com muitas entidades (tabelas de bancos de dados) e relacionamentos entre elas, que exigem muitas ferramentas básicas de cadastro. Para suprir este objetivo, utiliza alguns padrões consagrados da engenharia de software, que foram adequadamente adaptados para o contexto da web para equilibrar agilidade no desenvolvimento e escalabilidade.




Estrutura de Banco de dados da versao alfa 1

# Desenv. - entrada de textos



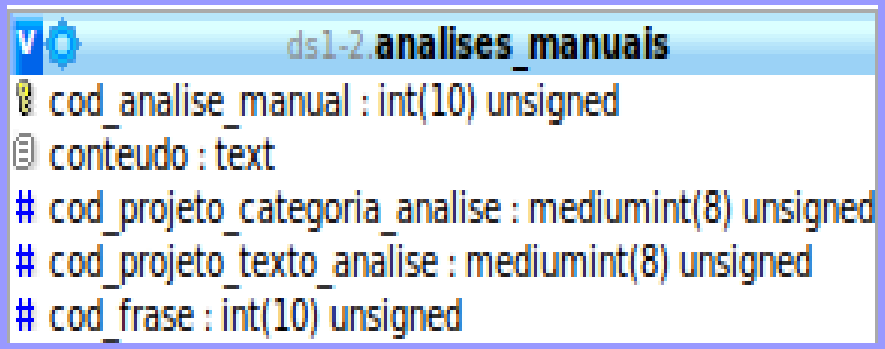

A figura mostra o processo completo de entrada de texto com o módulo de chat ativado.

O sistema possui uma função processamento. O usuário pode escolher entre executá-la no momento da submissão ou enviá-la para background, o que é útil principalmente no caso de textos longos, cujo processamento pode demorar muitos minutos.

Note que, na versão atual do programa, o papel do gerente está restrito a entrada de analistas, textos e categorias no sistema.

# Desenv. - entrada de análises manuais






As análises manuais são salvas na tabela correspondente do banco. Esta tabela não recebe resultados automáticos e, portanto, não possui uma entrada para cada sentença, somente registrando as análises de sentenças cujos campos não foram deixados vazios no momento da atualização.
Ao recuperar os dados desta tabela, o módulo de estatísticas preenche estes campos vazios como nulos.

# Desenv. - Módulos de preprocessamento

O dadosSemiotica possui dois módulos de preprocessamento. O módulo de preprocessamento morfossintático provê os serviços de separação de sentenças e análise morfossintática para todas as sentenças do corpus.
O módulo de chat somente atua se for ativado no momento do upload. Sua função é registrar diversas informações próprias do IRC (tais como entradas, saídas, histórico do nick, notificações) e, antes de acionar a separação de sentenças, retirar da frase elementos estranhos ao texto padrão em língua portuguesa: data, hora e nick. Estes elementos atrapalham a análise morfossintática, mas são resgatados para visualização pelo usuário para análises manuais (exceto data).

O **Módulo de Preprocessamento Morfossintático** possui um serviço de separação de sentenças, detectadas pela pontuação. Para as análises, é fundamental que o analista possa visualizar as sentenças em sequência, para uma visão do todo, mas que seja também capaz de realizar análises pontuais para cada sentença. A única forma de dividir trechos menores é separar esses trechos menores com quebras de linha, já que é enviada ao módulo uma linha por vez.

Texto como registros de chat, por exemplo, possuem esta característica e terão muitas vezes sentenças menores do que o padrão, em virtude do fluxo normal de conversas no chat, em que muitas vezes o usuário quebra a sentença para conseguir envios mais rápidos, mantendo o interesse dos outros participantes na conversa. Mesmo assim, como é também comum haver mais de uma sentença em uma linha de chat, todas as entradas de chat são processadas pelo separador de sentenças.

O serviço de análise do **Módulo de Preprocessamento Morfossintático** automaticamente infere estruturas a partir dos textos coletados pelo sistema. Tais estruturas, então apresentando um modelo de dados computacionalmente tratável, são indexadas e armazenadas em um banco de dados. Tendo como base os anotadores desenvolvidos para o corretor gramatical software livre CoGrOO*, desenvolvido pelo Centro de Competência em Software Livre (CCSL)* para a suíte de escritório LibreOffice*, o módulo segmenta o documento em sentenças, tokens (palavras e símbolos) e sintagmas e classifica, dado o contexto, as palavras de acordo com sua classe gramatical, identifica seu lema e ainda suas características como gênero, número, pessoa e tempo .

# Desenv. - Módulo de posprocessamento

Módulo de poprocessamento estatístico: trata-se do que chamamos de módulo de estatísticas, que é o sistema de geração de tabelas e análises descritivas do *corpus*.

Por este sistema, o analista recupera os dados  na forma de uma tabela com as informações que desejar e no formato adequado à análise estatística – ou outra – que almeja.
O módulo envia por e-mail a tabela com todas as categorias selecionadas e, para algumas, escolhidas pelo analista dentre aquelas, uma análise estatística descritiva básica do corpus (histograma, média, mediana, desvio padrão, mínimo, máximo, número de observações e número de valores nulos). Esta descrição permite ao analista avaliar se existe ou não necessidade de ajustes, por exemplo corrigindo outliers gerados por erros de digitação. Os resultados estatísticos são obtidos pelo software R*

.

É possível incluir resultados das análises do módulo de preprocessamento morfossintático nas tabelas geradas pelo módulo de estatística. Como os resultados não possuem uma estrutura relacional, são armazenados todos em um mesmo campo do banco para cada sentença. A fim de realizar buscas nesse campo, a sintaxe utilizada para registro dos resultados segue o padrão **Tregex***, que é utilizado no módulo de estatística para realizar as buscas. O sistema está instalado na API de análise morfossintática do módulo de preprocessamento correspondente.

Nesta versão do dadosSemiotica é possível incluir uma consulta às análises morfossintáticas em cada requisição de tabela, retornando duas informações: uma apenas indicando a presença (0) ou ausência (1) da estrutura buscada em cada sentença; a outra informando o resultado textual da busca.

# Desenv. - a fazer

i. Separação de palavra longa demais para visualização na tela de análises manuais;

ii. Generalizar as regras do módulo de chat, atualmente só disponível para logs provenientes do programa cliente de IRC Konversation;.

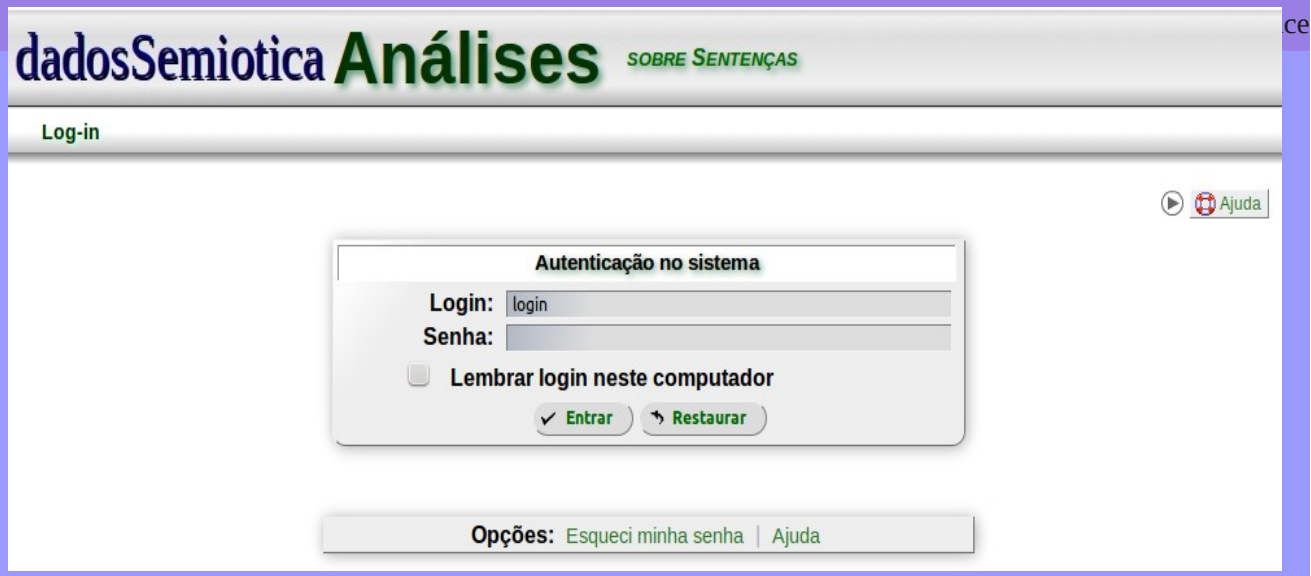
iii. Disponibilização de uma versão diferenciada do módulo de estatísticas para o gerente, com acesso ao banco inteiro em cada consulta;

iv. Ampliação das possibilidades de buscas Tregex via formulário;

v. Possibilidade de recuperar uma análise de uma categoria de um projeto em outro, quando há textos em comum.

# Comunidade





# Comunidade - licença

O dadosSemiotica foi criado sob uma licença GPL-v.2, a qual pode ser encontrada em:
http://www.gnu.org/licenses/gpl-2.0.html
Trata-se de um software livre de código aberto que está registrado no Sourceforge:
Http://dadossemiotica.sf.net

# Comunidade – documentação e suporte

Página do projeto:
http://sourceforge.net/projects/dadossemiotica/

Para assinar a lista de suporte, envie uma mensagem com o assunto
Subscribe
para:
dadossemiotica-ajuda-request@lists.sourceforge.net

Equipe de desenvolvedores do dadosSemiotica:
Ana Cristina Fricke Matte
Rubens Takiguti Ribeiro
Willian Daniel Colen de Moura Silva
Hugo Leonardo Canalli

Além da equipe de desenvolvedores, o dadossemiotica conta com a equipe de betatesters do Grupo Texto Livre do Laboratório SEMIOTEC de Semiótica e Tecnologia FALE/UFMG.
http://semiotec.textolivre.org

# REFERÊNCIAS

Matte, A., Ribeiro, R., Colen, W., Canalli, H. - *Dadossemiotica: análise e processamento de texto escrito*. Anais do WSL2012 Workshop Internacional de Software Livre. Porto Alegre, 2012.

Matte, A. - *Análise quantitativa da tensividade do conteúdo verbal tendo em vista o estudo da expressão da emoção na fala e o modelamento prosódico*. Cadernos de Estudos Linguísticos 46(1), Campinas, 2004.

**Texto Livre:** http://www.textolivre.org
**SemioTec:** http://semiotec.textolivre.org
**CoGroo**: http://www.cogroo.org
**SIMP:**
http://sourceforge.net/projects/simpframework/
**Tregex:**
http://nlp.stanford.edu/software/tregex.shtml
**Tregex Pattern:**
http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/trees/tregex/TregexPattern.html
**Centro de Competência em Software Livre (CCSL):**
http://ccsl.ime.usp.br/
**LibreOffice:**
http://pt-br.libreoffice.org/
**Konversation:** http://konversation.kde.org/
**R:** http://cran.r-project.org/
**IRC:**
http://pt.wikipedia.org/wiki/Internet_Relay_Chat

Apoio
FAPEMIG 2010-2012 processo PPM-00206-10.

Faculdade de Letras

Textolivre

UFMG

Semiotec

CoGrOO
Corretor Gramatical para OpenOffice.org

Este manual sobre o dadosSemiotica
foi desenvolvido por Ana Cristina Fricke Matte
e está disponível sob uma Licença Creative Commons.

voltar