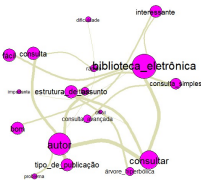


Curso de Análise de Dados Textuais por meio do Software Iramuteq

Período: 13 a 16/03/2017 (13:00-16:30 hs)

Carga horária: 14 horas/aula

Instrutora: Maria Elisabeth Salviati



Introdução

Análise textual

A análise textual é um tipo de análise de dados, que trata especificamente de material verbal transcrito, ou seja, de textos produzidos em diferentes contextos.

Ela é aplicada nos estudos de pensamentos, crenças e opiniões produzidas em relação a determinado fenômeno, tema de investigação, permitindo a quantificação de variáveis essencialmente qualitativas originadas de textos, a fim de descrever o material produzido por determinado sujeito ou sujeitos (CAMARGO & JUSTO, 2013).

Emprepa

Análise textual

Para se analisar grande volume de textos têm sido utilizados softwares específicos de análise textual tais como Alceste e Iramuteq.

O uso de novas técnicas para manipular e apresentar grandes volumes de dados leva a novas possibilidades de análise – pois construir uma representação, naturalmente, é propor uma interpretação.

Emprepa

Análise textual

Esses softwares possibilitam identificar o contexto em que as palavras ocorrem.

Eles executam análise lexical do texto e o particionam em classes hierárquicas, identificadas a partir dos segmentos de textos que compartilham o mesmo vocabulário, facilitando, assim, o pesquisador conhecer seu teor.

Emprepa

Iramuteq

O software Iramuteq - Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires foi criado em 2009 por Pierre Ratinaud.

É um software gratuito de código fonte aberto, licenciado por GNU GPL (v2), que utiliza o ambiente estatístico do software R. Assim como os outros softwares de fonte aberta, ele pode ser alterado e expandido por meio da linguagem Python.

Embrapa

Iramuteq

Ele é utilizado no estudo das Ciências Humanas e Sociais e utiliza o mesmo algoritmo do software Alceste para realizar análises estatísticas de textos, porém, incorpora, além da CHD - Classificação Hierárquica Descendente, outras análises lexicais que auxiliam na análise e interpretação de textos.

Embrapa

Aplicações na Embrapa

Análises qualitativas: Francisco Eduardo de Castro Rocha

Embrapa

Programação

Empresa

Programa do curso

Dia 13/03

Introdução
Nomenclatura
Análises realizadas pelo Iramuteq
Instrução para instalação dos aplicativos utilizados

Dia 14/03

Construção do corpus
Realização de exemplo real (importação do corpus; estatísticas textuais e correção do corpus)
Dicionário de termos

Dia 15/03

Realização de exemplo real (dendrograma; análises de especificidades e AFC)

Dia 16/03

Realização de exemplo real (análise de similitudes; nuvem de palavras.)

Empresa

Nomenclatura

Noção de *corpus*, texto e segmentos de texto:

Empresa

Nomenclatura

- **UTF-8 (8-bit Unicode Transformation Format):** é um tipo de codificação Unicode de comprimento variável criado por Ken Thompson e Rob Pike. Pode representar qualquer carácter universal padrão do Unicode, sendo também compatível com o ASCII.
- **CP1252:** Windows-1252 ou CP1252 é uma codificação de caracteres do alfabeto latino, usado por padrão nos componentes herdados do Microsoft Windows em Inglês e algumas outras línguas ocidentais.

Embrapa

Nomenclatura

- **Dicionário de termos:** dicionário que pode ser atualizado pelo usuário e que contém termos e suas variantes. É utilizado pelo sistema para classificação das palavras conforme o tipo gramatical, a fim de identificar as palavras ativas e suplementares do corpus.
- **Dicionário de lemmes:** criado a partir de um *corpus* submetido à análise estatística e contém o lema (palavra sem flexão), suas variantes e sua frequência de ocorrência.

Embrapa

Nomenclatura

- **Lematização:** é o processo de deflexionar uma palavra para determinar o seu lema. Por exemplo, as palavras *gato*, *gata*, *gatos*, *gatas* são todas formas do mesmo lema: *gato*.

No Iramuteq existem regras próprias de lematização. Os verbos são convertidos ao infinitivo, os substantivos ao singular e os adjetivos ao masculino singular.

O Iramuteq realiza a lematização a partir dos dicionários, sem realizar a desambiguação.

Embrapa

Análises do Iramuteq

- ✓ Estatísticas textuais
Esta análise executa estatísticas simples sobre o "*corpus*" e fornece: o número de textos e segmentos de textos; frequência média e total das palavras; e classificação gramatical das palavras, de acordo com o dicionário. ▶
- ✓ Classificação Hierárquica Descendente (CHD)
Nesta análise o sistema procura obter classes formadas por palavras que são significativamente associadas com aquela classe (a significância começa com o qui-quadrado = 2). Ele apresenta um esquema hierárquico de classes, tornando possível inferir quais ideias o *corpus* textual deseja transmitir. ▶

Emprepa

Análises do Iramuteq

✓ Análises de especificidades e AFC

A análise de Especificidades associa textos com variáveis, ou seja, possibilita a análise dos textos em função das variáveis de caracterização. Todo *corpus* possui variáveis associadas estabelecidas pelo pesquisador. Por exemplo: sexo, escolaridade, Estado, etc. Essa análise permite, então, comparar os resultados por variável (por exemplo entre homens e mulheres, etc). ▶

A Análise Fatorial por Correspondência (AFC) é uma representação gráfica dos dados para ajudar a visualização da proximidade entre classes ou palavras. ▶

Emprepa

Análises do Iramuteq

✓ Análises de similitudes

Mostra um grafo que representa a ligação entre palavras do *corpus* textual. A partir desta análise é possível inferir a estrutura de construção do texto e os temas de relativa importância, a partir da coocorrência entre as palavras.


Ela auxilia o pesquisador na identificação da estrutura da base de dados (*corpus*), distinguindo as partes comuns e as especificidades, além de permitir verificá-las em função das variáveis descritivas existentes. ▶

Emprepa


Análises do Iramuteq

✓ Nuvem de palavras

Mostra um conjunto de palavras organizadas em forma de nuvem. As palavras são apresentadas com tamanhos diferentes, ou seja, as palavras maiores são aquelas que detêm maior importância no *corpus* textual, a partir do indicador de frequência ou outro escore estatístico escolhido. É uma análise lexical mais simples, porém, bastante interessante, na medida em que possibilita rápida identificação das palavras-chaves de um *corpus*, isto é, a rápida visualização de seu conteúdo, pois as palavras mais importantes estão mais perto do centro e graficamente são escritas com fonte maiores. ▶



Instrução para instalação de aplicativos



Instalação de Aplicativos


✓ **Instalação do Open Office**

O Open Office é o editor de texto e planilhas padrão para ser utilizado em conjunto com o Iramuteq.

Do pacote do Open Office serão utilizados o Writer para digitação de texto, leitura dos relatórios e resultados das análises efetuadas e o Calc para digitação de planilhas, leitura e exportação de resultados.

Passos para instalação:

- ✓ Fazer download do arquivo:
- ✓ Apache_OpenOffice_4.1.3_Win_x86_install_pt-BR no endereço:
- ✓ https://sourceforge.net/projects/openofficeorg.mirror/files/4.1.3/binaries/pt-BR/Apache_OpenOffice_4.1.3_Win_x86_install_pt-BR.exe/download
- ✓ Instalar o pacote do Open Office, aplicativos Writer e Calc.



Instalação de Aplicativos


✓ **Instalação do software estatístico R**

O R é um software gratuito para elaboração de gráficos e computação estatística.

O Iramuteq executa as análises utilizando as Bibliotecas do R. Por isso, antes de instalar o Iramuteq é necessário instalar o R e as bibliotecas.

Passos para instalação:


- ✓ Fazer o download do software R versão 3.2.3 para Windows em <https://cran.r-project.org/bin/windows/base/old/3.2.3/>
- ✓ Instálá-lo (arquivo: "R-3.2.3-win.exe") de preferência na pasta de Arquivos de Programas.
- ✓ Durante a instalação, escolha corretamente se 32 ou 64 bits.
- ✓ Aguarde finalizar a instalação.



Instalação de Aplicativos

Instalar as bibliotecas (pacotes) do R: (o computador tem que estar conectado na Internet):

- ✓ Executar o R;
- ✓ Escolher no menu principal Pacotes/Instalar pacotes;
- ✓ Escolher o país (França/Paris2), outros países podem não possuir todas as bibliotecas necessárias;
- ✓ Escolher na lista apresentada em ordem alfabética, o primeiro pacote a ser instalado (ape) e clicar Ok. Se o sistema perguntar para criar uma pasta para armazenar a biblioteca, optar por criá-la e deixar o R escolher a pasta padrão. O sistema realiza o download da biblioteca e a instala.
- ✓ Depois de terminada a instalação do primeiro, instalar todos os demais (ca, gee, igrph, iriba, proxy, rgl, textometry, wordcloud) repetindo os passos d e e;
- ✓ Fechar o R.




Instalação de Aplicativos

✓ **Instalação do Iramuteq**

Passos para instalação:

- ✓ Fazer o download do software IRAMUTEQ em <https://sourceforge.net/projects/iramuteq/files/iramuteq-0.6-alpha3/>
- ✓ Instalá-lo (arquivo: "setup_iramuteq-0.6-alpha3.exe") na mesma pasta onde foi instalado o R. Exemplo: Arquivos de programas.
Se não for instalado na mesma pasta, o Iramuteq não reconhece as bibliotecas do R, mesmo se o caminho for informado em Preferências do Iramuteq.
- ✓ Entrar no Iramuteq e aguardar a instalação das bibliotecas do R automaticamente.

Quando a versão do R for a 3.2.3 as bibliotecas do R são carregadas automaticamente. Porém, se o usuário instalar outra versão pode ser necessário solicitar a instalação das Bibliotecas.



Instalação de Aplicativos

Nota: Se a instalação do R foi correta, bem como das bibliotecas, o Iramuteq encontra todas as bibliotecas e fica pronto para proceder às análises. Caso tenha algum problema na instalação do R ou das bibliotecas, o Iramuteq não conseguirá trazê-las e continuará solicitando a sua atualização toda vez que se entrar no Iramuteq ou, ainda, não executará as análises porque as bibliotecas não existem. Pode ser, ainda, que ele não encontre o R e informe que o R não está instalado. Nesse caso, não adianta acrescentar o caminho do R em **Edição/Preferências** porque ele continuará não encontrando as bibliotecas. A solução desse problema normalmente envolve a desinstalação do Iramuteq e do R e nova instalação nas pastas corretas.

Referências

CAMARGO, B.V.; JUSTO, A.M. IRAMUTEQ: um software gratuito para análise de dados textuais. *Temas em Psicologia*, v. 21, n. 2, p.513-518, 2013.

SALVIATI, M. E. **Introdução ao Gephi 0.9.1**. Planaltina, DF: Embrapa Cerrados, 2016. 21p.

SALVIATI, M. E. (comp.). **Manual do Aplicativo Iramuteq (versão 0.7 Alpha 2 e R Versão 3.2.3)**. Planaltina, DF: Embrapa Cerrados, 2016. 93p.

Estatísticas textuais

Forma	Freq.	Tipos
achar	169	ver
estar	159	ver
saber	82	ver
coisa	63	nom
gente	61	nom
autor	58	nom
postar	57	nr
biblioteca_eletrônica	56	nr
consultar	49	ver
so	45	adv
vez	43	nom
informação	42	nom
substitui	39	ver
entrar	39	ver
exemplo	39	nom
tipo_publicação	37	nr
estrutura_de_assunto	35	ver
usar	35	ver
ficar	34	ver
querer	34	ver
procurar	33	ver
consulta	32	nom
uso	32	nom
bons	31	adv
dar	31	ver
fácil	31	adv

Classificação Hierárquica Descendente (CHD)

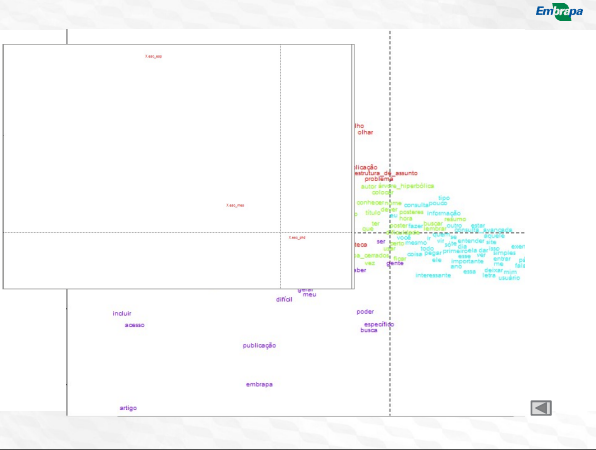
Arquivo Editar Formatar Exibir Ajuda

```

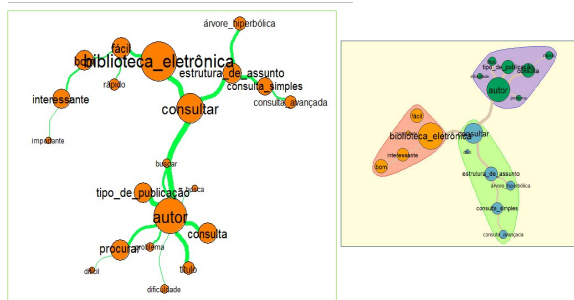
**** $suj_01 $uni_cpac $car_peso $sex_mas $ida_46
mas outras vezes e diária não eu não uso todas as
**** $suj_01 $uni_cpac $car_peso $sex_mas $ida_46
não tenho nenhuma dificuldade não porque se eu cc
**** $suj_01 $uni_cpac $car_peso $sex_mas $ida_46
erro de digitação erro de não saber exatamente o
**** $suj_01 $uni_cpac $car_peso $sex_mas $ida_46
eu acho interessante para quando é seu mesmo sabe
**** $suj_01 $uni_cpac $car_peso $sex_mas $ida_40-49 $esc
eu usei assim quando eu vim para cá há três anos então eu
    
```

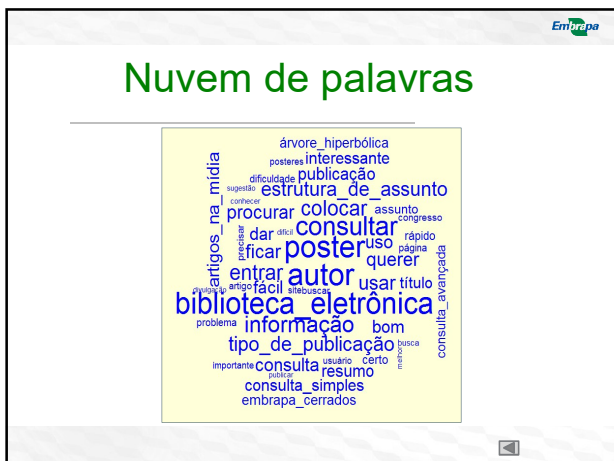
Análise de Especificidades

Formas	Formas comuns	Tipos	Formas específicas	Tipos de frequências	Frequência relativa das formas	Tipos
há	146	313				
achar	100	69				
estar	93	66				
porque	61	43				
mas	55	42				
saber	50	32				
aqui	42	30				
coisa	41	22				
já	40	27				
poster	37	20				
então	36	32				
gente	36	25				
masco	35	31				
autor	32	26				
tempo	30	30				
vez	27	16				
estar	26	13				
consultar	24	25				
colocar	24	15				
quando	23	17				
tipo_de_publicação	22	15				
estrutura_de_assunto	22	13				
biblioteca_eletrônica	21	35				
usar	19	16				
consulta_simples	19	9				
procurar	19	14				
se	19	26				



Análise de Similitudes





2ª Aula

Dia 14/03
13:00-15:00 hs
Construção do corpus
Formatação; Variáveis e temáticas; Gravação do corpus ;
Realização de exemplo real – Parte 1
✓ **Importação do corpus:** realização de exemplo real.
15:00-15:20 hs
Intervalo
15:20- 16:30 hs
Realização de exemplo real – Parte 1
✓ **Estatísticas textuais e correção do corpus:** realização de exemplo real.
✓ **Dicionário de termos**

Emprepa

Construção do *Corpus*

Emprepa

Noções de *corpus*, texto e segmento de texto

- Documentos;
- Artigos (jornal, revista);
- Entrevistas;
- Questionários abertos; e
- outros materiais discursivos

Figura adaptada de: CAMARGO & JUSTO, 2013.

Emprepa

Formatação do *corpus*

Para edição do *corpus* utilize o aplicativo **OpenOffice Writer** (para texto) e **Calc** (para planilhas).

Os aplicativos editores de texto que fazem parte do Windows: Wordpad, Bloco de Notas e Office (Word, Excel) e mesmo o Open Office, trabalham com o padrão de codificação **CP1252**.

De forma que quando se grava um texto com formato txt por esses aplicativos eles estarão obedecendo a esse padrão, a menos que o usuário, no momento de gravação, altere a codificação para outro formato, por exemplo, **UTF-8** (padrão que permite aos computadores representar e manipular, de forma consistente, texto de qualquer sistema de escrita existente).

O formato UTF-8 é melhor se o usuário trabalhar com aplicativos que não aceitam a codificação CP1252, principalmente quando se trata de software livre que não obedece ao padrão Windows.

Emprepa

Formatação do *corpus*

As entrevistas ou questões abertas devem ser formatadas sem as questões. Deve-se obedecer às seguintes regras:

Sinais proibidos: aspas; apóstrofo; cifrão; porcentagem; asterisco; reticências; travessão; negrito, itálico, grifo e outros sinais similares; recuo de parágrafo, margens ou tabulações do texto; justificação do texto.

- ✓ Pontuação permitida: ponto; dois pontos; vírgula; interrogação e exclamação.
- ✓ Formatação de texto todo corrido, sem mudança de linha.
- ✓ Uso de maiúsculas só para nomes próprios.
- ✓ Palavras compostas devem ser unidas por underline, mesmo aquelas unidas ortograficamente pelo hífen. Ex.: recém_casado; anti_inflamatório; Distrito_Federal.

Emprepa

Formatação do *corpus*

- ✓ Padronização das siglas e nomes próprios para obedecer sempre mesma grafia.
- ✓ Revisão gramatical do português, corrigindo-se grafia e concordância.
- ✓ Complementação de todas as frases incompletas: cada frase deve encerrar um sentido completo e não deve possuir palavras subtendidas.

Complementar com as palavras necessárias, sem modificar o sentido. Se necessário, reexaminar o texto original para escolher as palavras adequadas. Caso haja impossibilidade de completar determinadas frases, elas deverão ser eliminadas.

- ✓ Eliminação de expressões sem necessidade, tais como: Ahh, Uhhh, né, tá.
- ✓ Eliminação de frases não condizentes com o assunto tratado.
- ✓ Não utilizar as flexões verbo-pronominais. Ex.: No lugar de "tornei-me", a escrita deve ser: "me tornei".
- ✓ Números devem ser mantidos em forma de algarismos.

Emprepa

Formatação do *corpus*

Com que frequência você utiliza, o/ha temos as opções: diária, semanal, mensal, esporádica.

Isso varia muito. Se eu to fazendo um projeto, às vezes, diária... mas às vezes também...

Na média, você poderia dizer o que?

Eu poderia colocar semanal, mas, obviamente, não é toda semana que eu consulto.

Entendi. É na média né?!

Mas outras vezes é diária.

Eu tenho aqui diversos tipos de consulta: por assunto, por autor, tipo, título e consulta avançada. Você já utilizou todas essas formas, ou não?

Não. Eu uso a por assunto e por autor.

É?

Normalmente.

Trecho original transcrito de uma entrevista realizada

A frequência de uso da biblioteca eletrônica varia muito. Se eu estou fazendo um projeto, às vezes, a frequência de uso é diária. Eu poderia colocar frequência de uso semanal, mas, obviamente, não é toda semana que eu consulto. Mas outras vezes a frequência de uso é diária. Não utilizo todas as formas de consulta. Eu uso a consulta por assunto e por autor, normalmente.

Mesmo trecho anterior preparado para o corpus de análise automática

Entropa

Formatação do *corpus*

Algumas observações sobre construção do *corpus*: Celina Tomaz de Carvalho

- Questionário
- Variáveis
- Corpus

Entropa

Variáveis e Temáticas

Cada texto deve ser separado por linhas de comando. No caso de entrevistas, por exemplo, cada uma delas deve iniciar com uma linha de comando. Esta linha informa algumas variáveis como: o número de identificação do entrevistado; sexo; faixa etária; afiliação a determinados grupos; nível social e cultural, etc. Isto depende de cada pesquisa.

Já no caso de tabelas efetuadas no Office Calc, cada linha representa um indivíduo, isto é, formato de planilhas e bancos de dados (arquivos ods).

A linha de comando inicia por quatro estrelas (****) seguidas pelas variáveis que são introduzidas com uma * (estrela) separada por um espaço.

Um texto terá que ter obrigatoriamente, pelo menos uma variável. Ex.:

```
**** *suj_001 *sex_1 *ida_21 *escol_2
```

Uma variável temática é indicada por um hífen e uma estrela (-*) e indica pedaços de textos que se referem a temas ou aspectos diferentes. Ex.:

```
**** *var1_1 *var2_2
    -*tematica1
    texto
```

Atenção é hífen

Entropa

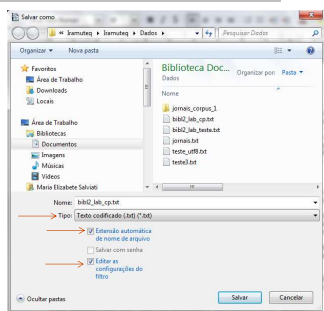
Gravação do *corpus*

O formato CP1252 - no Open Office está designado como Europa Ocidental (Windows-1252/WinLatin 1), como já explicado anteriormente, não precisa ser informado no momento da gravação, basta escolher o formato texto (.txt).

Porém, a gravação de texto, utilizando o formato UTF-8, deve seguir codificação especial, portanto, antes de gravá-lo, siga as seguintes orientações:

- ✓ Salvar o *corpus* com a opção "Salvar como"
- ✓ Escolher o tipo "Texto codificado (.txt)"
- ✓ Marcar a opção de "Extensão automática de nome de arquivo"
- ✓ Marcar também a opção "Editar as configurações do filtro".

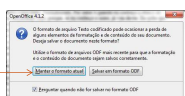
Gravação do corpus



Janela de "Salvar como" do Open Office Writer

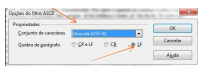
Gravação do corpus

- ✓ Clicar em Manter o formato atual na nova janela.



O sistema mostra o menu para escolha da codificação desejada.

- ✓ Escolher Unicode UTF-8 e a Quebra de parágrafo por LF e clicar em Ok.



Toda vez que o corpus for aberto com o OpenOffice Writer, vai aparecer essa mesma janela e sempre deverão ser escolhidas essas opções.

Importação do corpus

➔ Visual, Menus e Janelas do Iramuteq

Emprespa

Importação do *corpus*

➔ **Mostrar importação e abertura de análise o Iramuteq**

- ✓ Abrir um corpus textual (Iramuteq\Dados\teste\artigos_novo2.txt)
- ✓ Abrir uma análise (ex: artigos_novo2)

Emprespa

Importação do *corpus*

Exercício: Importação do *corpus*

- Arquivo: celina_teste.txt
- Idioma: português
- Caracteres: CP1251 Windows
- Dicionário: padrão português

Usar as outras configurações padrão, tanto nessa aba como na aba limpando

Emprespa

Análises Textuais

- ✓ Estatísticas: estatísticas simples sobre o corpus
- ✓ Especificidades e AFC: associa palavras (formas) com as variáveis ou modalidades da variável
- ✓ Classificação: CHD do Alceste
- ✓ Análises de similitudes: grafo de ligação entre formas do corpus
- ✓ Nuvem de palavras: formas organizadas em formato de nuvem
- ✓ Criação de subcorpus: criação de um corpus parcial formado por variáveis, modalidade das variáveis ou temáticas.
- ✓ Exportação da tabela de metadados (exporta para formato de planilha as variáveis existentes no corpus).

➔ **Mostrar menu de Análises no Iramuteq**

Estatísticas textuais

Esta análise executa estatísticas simples sobre o "corpus" textual. Ele executa os seguintes procedimentos:

- ✓ identificação das palavras
- ✓ pesquisa no vocabulário e redução das palavras com base em suas raízes (formas reduzidas)
- ✓ identificação da quantidade de palavras, quantidade de ocorrências, média de ocorrências por texto e quantidade de hápax
- ✓ identificação das formas ativas e suplementares
- ✓ criação do dicionário de formas reduzidas do corpus

O Iramuteq fornece ainda as listas de formas ativas (principais), suplementares e total com suas respectivas frequências, bem como sua classificação gramatical, de acordo com o dicionário e o dicionário de formas reduzidas do corpus.

Estatísticas textuais

Parâmetros comuns a todas as análises

Estatísticas textuais

- ✓ **Resultados**

Traz informações nas abas:

- ✓ Resumo;
- ✓ Actives Formes fomas principais);
- ✓ Supplementary Formes (formas secundárias);
- ✓ Total (todas as formas); e
- ✓ Hápax (formas com frequência 1)

➡ Executar e Mostrar os resultados no Iramuteq (artigos_novo2_corpus_1)

➡ Mostrar a pasta onde foram gravadas as Estatísticas

Emprespa

Estatísticas textuais

✓ **Resultados**

Procedimentos complementares

➔ Mostrar menu de procedimentos complementares no Iramuteq

Emprespa

Estatísticas textuais

Exercício: realizar as Estatísticas textuais do corpus importado (celina_teste)

- Propriedades: usar a mesma configuração padrão, exceto para:

Advérbio suplementar: 0

Artigo definido: 0

Artigo indefinido: 0

Auxiliar: 0

Número (chiffre): 0

Emprespa

Estatísticas textuais

Correção do corpus

✓ Examine a lista de palavras ativas e suplementares geradas. Observe os sinônimos, palavras com erro de grafia.

dar	32	ver	abrir	6	ver
fácil	32	adj	acessar	6	nr
artigos_ea_midia	31	nr	artigos_ea_midia	6	nr
preços	31	nom	boletim	6	nom
publicação	31	nom			
resumo	31	nom			

artigos_ea	3	nr
atualizar	3	ver
avaliar	3	ver

✓ Corrija o corpus (arquivo txt)

✓ Importe o novo corpus.

✓ Gere novas Estatísticas textuais e veja o resultado.

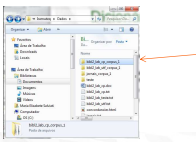
Dicionário de termos

O LACCOS - Laboratório de Psicologia da Comunicação e Cognição da Universidade Federal de Santa Catarina em parceria com a Fundação Carlos Chagas e a UNESP estão aprimorando o dicionário experimental em língua portuguesa, garantindo análises mais estáveis (CAMARGO & JUSTO, 2013).

Mesmo estando o dicionário atualizado, fatalmente ele não encontrará todas as palavras existentes no seu texto.

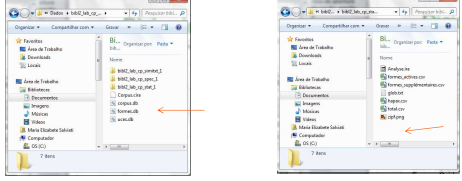
Para atualizar e corrigir o dicionário, após submissão de uma análise:

- ✓ Localizar a pasta com a análise do corpus (nome do *corpus* seguido por “_corpus_1”)



Dicionário de termos

- ✓ Abrir a subpasta cujo nome é o mesmo do *corpus* seguido por “_stat_1”



- ✓ Abrir a planilha **Total** com o Calc (clicar com o botão direito do mouse sobre a planilha e escolher Abrir com Office Calc com a codificação Europa Ocidental (Windows 1252/WinLatin 1))
Esta planilha contém todas as palavras encontradas no texto, seguidas pela frequência de ocorrência e classe gramatical. As palavras não existentes no dicionário estarão com a identificação de “nr” (não reconhecidas).

➡ Mostrar no Iramuteq

Dicionário de termos

Para atualizar o dicionário acrescentando essas palavras, proceda da seguinte maneira:

- ✓ Selecione as três colunas da planilha e peça a ordenação de A/Z pela última coluna e clique Ok.
- ✓ Selecione todas as palavras classificadas como nr (não reconhecidas), copie e cole em outra planilha do Calc.
- ✓ Verifique o resto da lista se existe outra palavra mal classificada. Nesse caso, copie-a e cole-a também na nova planilha.
- ✓ Substitua o nr ou a classificação errada pela correta conforme tabela de tipos gramaticais. ➡ (acesso à tabela)
- ✓ Classifique as palavras restantes conforme tabela, pela primeira coluna: Dados/Classificar/Crescente (Coluna A)
- ✓ Excluir a planilha 1 que está completa. Acessar a planilha 1 e escolher: Editar/Planilha/Excluir.
- ✓ Salvar esse arquivo com novo nome: SeuNome"_data (formato ddmmaaaa)


Empresa

Dicionário de termos

✓ Envie cópia da nova planilha gravada brigido.camargo@gmail.com

Codificação das formas gramaticais

- adj = adjetivo
- adj_num = adjetivo numeral
- adj_sup = adjetivo colocado em forma suplementar
- adv = advérbio
- adv_sup = advérbio colocado em forma suplementar
- art_def = artigo definido
- conj = conjunção
- nom = nome
- nom_sup = nome colocado em forma suplementar
- nr = não reconhecida
- ono = onomatopéia
- pro_ind = pronome indefinido
- pre = preposição
- ver = verbo
- verbe_sup = verbo colocado em forma suplementar



Empresa

3ª Aula

Empresa

Dia 15/03

13:00-15:00 hs

Realização de exemplo real – Parte 2

- ✓ dendrograma
- ✓ Listas de segmentos de texto (UCE) de cada classe

15:00-15:20 hs

Intervalo

15:20- 16:30 hs

Realização de exemplo real – Parte 3

- ✓ Análise de especificidades e AFC

Dendrograma: Método Reinert

Esta é uma das análises mais importantes do Iramuteq, nela o software processa o texto de modo que possam ser identificadas classes de vocabulário que permitem inferir quais são as ideias principais do *corpus* textual.

Ela visa obter classes de segmentos de texto (ST) que, apresentam vocabulário semelhante entre si e vocabulário diferente das ST das outras classes. Esta análise é baseada na proximidade léxica e na ideia que palavras usadas em contexto similar estão associadas ao mesmo mundo léxico e são parte de mundos mentais específicos ou sistemas de representação.

A classificação pode ser:

- ✓ Dupla sobre RST
- ✓ Simples sobre ST
- ✓ Simples sobre textos

Dendrograma: Método Reinert

Dendrograma: Método Reinert

✓ **Resultados**

Os resultados diretamente disponíveis apresentam:

- ✓ resumo da classificação (aba CHD);
- ✓ os perfis das classes (aba perfis) permitem consultar os ST (UCes) de cada classe;
- ✓ análise fatorial das correspondências (AFC) realizadas sobre a tabela de contingência cruzando formas e classes no (aba AFC).

➔ Executar e Mostrar resultados no Iramuteq (artigos_novo2_corpus_1)

Dendrograma: Método Reinert

Como editar os Dendrogramas

Indicar o tamanho da imagem em pixels

Os formatos possíveis são png e svg

1) Mesmo formato apresentado
2) Formato horizontal
3) Formato vertical

Phylogram
Cladogram
Fan
Unrooted
Radial

Dendrograma: AFC

- 3D não funciona
- Png x svg
- Representação por coordenadas ou por correlação
- Variáveis: ativas, suplementares, variáveis ou classes
- Altura e Largura em pixels
- Tamanho do texto: melhor escolher opção mais abaixo

Essas três opções permitem limitar o tamanho do gráfico. É interessante usar uma delas para não gerar um gráfico de difícil visualização.

Evitar sobreposição: sempre checar essa opção
- Tamanho do texto proporcional à frequência ou ao qui-quadrado

Não utilizar as demais opções

Dendrograma: Método Reinert

✓ **Resultados**

- Como gerar o relatório de STs (UCEs) de cada classe
- Como executar análises e procedimentos complementares

➔ **Mostrar menu no Iramuteq**

Empirica

Dendrograma: Método Reinert

Exercício: realizar a classificação de Reinert do *corpus* (celina_teste).

- Propriedades: usar a mesma configuração anterior:

Advérbio suplementar: 0
 Artigo definido: 0
 Artigo indefinido: 0
 Auxiliar: 0
 Número (chiffre): 0

- Configurações de indexação: Usar a padrão do aplicativo: indexação simples sobre ST, tamanhos já estabelecidos e método.
- Elaborar um filograma (utilizando o segundo ícone do dendrograma)
- Elaborar um filograma (utilizando o terceiro ícone do dendrograma)

Empirica

Especificidades e AFC

Esta análise possibilita analisar o *corpus* textual em função das variáveis de caracterização. Quando o *corpus* é preparado, associam-se, variáveis que o pesquisador deseja analisar.

Nessa análise, a base de dados é dividida de acordo com a variável selecionada. Por exemplo, a comparação entre homens e mulheres em um questionário aplicado.

A Análise Fatorial de Correspondência é uma representação gráfica dos dados para ajudar a visualização da proximidade entre classes ou palavras.

Procedimentos executados pelo Iramuteq:

- ✓ cálculo das frequências e dos valores de correlação qui-quadrado de cada palavra do *corpus*, a partir de frequência mínima escolhida;
- ✓ execução da análise fatorial de correspondências (AFC) numa tabela de contingência que cruza as formas ativas e as variáveis.

Empirica

Especificidades e AFC

Emprepa

Especificidades e AFC

✓ **Resultados**

Os resultados diretamente disponíveis apresentam:

- ✓ Apresenta a correlação e frequência das formas e dos tipos gramaticais com as modalidades da variável escolhida
- ✓ análise fatorial das correspondências (AFC) realizadas sobre a tabela de contingência cruzando formas/lemas e as modalidades da variável escolhida (apresentado em 2 gráficos: só formas e só modalidade das variáveis).

➡ Executar e Mostrar os resultados no Iramuteq (artigos_novo2_corpus_1, variável jornal)

Emprepa

Especificidades e AFC

Exercício: realizar a análise de especificidade e AFC do corpus (celina_teste)

- Propriedades: usar a mesma configuração anterior:

Advérbio suplementar: 0
Artigo definido: 0
Artigo indefinido: 0
Auxiliar: 0
Número (chiffre): 0

- Escolher: a) formas ativas; b) modalidades seguintes da variável que: van, dif e fac; c) escore: qui-quadrado.
- Editar o gráfico da AFC com os seguintes parâmetros:
 - a) evitar sobreposição
 - b) tamanho proporcional à frequência: mínimo 8 e máximo 20

Emprepa

4ª Aula

Emprepa

Dia 16/03

13:00-15:00 hs
Realização de exemplo real – Parte 4
 ✓ Análise de similitudes

15:00-15:20 hs
Intervalo

15:20- 16:30 hs
Realização de exemplo real – Parte 4
 ✓ Nuvem de palavras

Emprepa

Similitudes

Esta análise é baseada na teoria dos grafos cujos resultados auxiliam no estudo das relações entre objetos de um modelo matemático.

No Iramuteq, a análise de similitude mostra um grafo que representa a ligação entre palavras do *corpus* textual.

O grafo é formado por vértices ou nós (contém as formas) e as arestas (mostram as ligações com outros nós).

A partir desta análise é possível inferir a estrutura de construção do texto e os temas de relativa importância, a partir da coocorrência entre as palavras.

Emprepa

Similitudes

The screenshot shows the 'Similitudes' configuration window in Iramuteq. It includes a list of variables on the left, a 'Configurações gráficas' section with options for 'Escora', 'Apresentação', 'Tipos de gráficos', 'Árvore máxima', 'Bordas limitadas', 'Texto sobre os vértices', 'Escora nas bordas', 'Edge curved', 'Tamanho do texto', and 'Comunidades'. Several callouts provide additional information:

- Escolha as formas mais importantes para ter um grafo mais legível** (points to the 'Formas' list)
- Edição da análise** (points to the 'Análise' dropdown)
- Edição do gráfico** (points to the 'Gráfico' dropdown)
- Essa é o mais utilizado, porém existem outros escores** (points to the 'Escora' dropdown)
- Existem vários formatos. É interessante testar** (points to the 'Tipos de gráficos' dropdown)
- Escolher dinamicamente para editar o gráfico** (points to the 'Árvore máxima' checkbox)
- Árvore máxima pode gerar um grafo difícil de enxergar** (points to the 'Árvore máxima' checkbox)
- É interessante marcar essa opção para diminuir o número de arestas e melhorar a visibilidade do gráfico** (points to the 'Bordas limitadas' checkbox)
- Delimita agrupamentos, conforme o formato escolhido** (points to the 'Comunidades' dropdown)
- Mostra os agrupamentos com cores** (points to the 'Comunidades' dropdown)
- Se desajustado, pode-se examinar só uma variável ou uma modalidade** (points to the 'Selecione uma variável' field)

Similitudes

Similitudes

Resultados:

- ✓ Quando optou-se por um **grafo estático**, o Iramuteq traz o grafo e os ícones abaixo que permite refazer o grafo ou exportá-lo.

Clique nessa opção para editar novamente o grafo, mudando as opções, tamanho e cor dos vértices e das arestas.

Essa opção permite exportá-lo em formato graphml (é armazenado na pasta de resultados). Esse grafo poderá ser editado pelos softwares Gephi ou Visone, ambos livres.

➡ Executar e Mostrar os resultados no Iramuteq (artigos_novo2_corpus_1)

- ✓ Quando optou-se por um **grafo dinâmico**, o Iramuteq abre uma janela contendo o grafo e um menu que permite editar alguns elementos do grafo. Depois de editá-lo, exporte o grafo antes de fechar a janela, porque não será mais possível reeditar o grafo depois de fechá-la.

Similitudes

Resultados:

Edição de grafos dinâmicos:

➡ Reeditar artigos_novo2 gerado em dinâmico

- ✓ Close: só use essa opção depois de exportar o grafo editado.
- ✓ Select: seleciona vértices e arestas para edição: todos, ou clicar no desejado e os demais com CTRL. Com botão direito escolhe-se a cor do vértice e da aresta.
- ✓ Layout: Permite mudar o formato do grafo
- ✓ View: Permite centralizar ou rodar o grafo
- ✓ Export: Permite salvar em formato eps

Similitudes Emprepa

Exercícios: 1) Realizar a análise de similitudes do corpus celina_teste

- Propriedades:** usar a mesma configuração anterior. Advérbio suplementar: 0; Artigo definido: 0; Artigo indefinido: 0; Auxiliar: 0; e Número (chiffre): 0.
- Seleção de formas:** frequência acima de 10, porém retirando as seguintes palavras: não, gente, mais, porque, achar, então, assim, aqui, também, aí, como, mesmo, ver, já, tudo, coisa, só, lá, pessoa, vez, parte, quando, ainda, até, próprio, ali, ao.
- Configurações e ajustes gráficos:** padrão (isto é, sem alterações)

2) Editar o gráfico resultante, fazendo as seguintes modificações:

- Apresentação:** Kamada Kawai
- escore nas arestas**
- tamanho: 1500 x 1200**
- texto do vértice proporcional à frequência (eff)**
- arestas com largura proporcional ao escore**
- escolha livre da cor do vértice e cor das bordas**

3) Reedite novamente e marque Comunidades e Halo.

Nuvem de palavras Emprepa

Esta análise traz um conjunto de palavras agrupadas, organizadas e estruturadas em forma de nuvem.

As palavras são apresentadas com tamanhos diferentes, ou seja, as palavras maiores são aquelas que detêm maior importância no *corpus* textual, a partir do indicador de frequência ou outro escore estatístico escolhido.

É uma análise lexical mais simples, porém, bastante interessante, na medida em que possibilita rápida identificação das palavras-chaves de um *corpus*, isto é, a rápida visualização de seu conteúdo, pois as palavras mais importantes estão mais perto do centro e graficamente são escritas com fonte maiores.

Nuvem de palavras Emprepa

Parâmetros

Configuração da nuvem de palavras 32

altura	800	largura	800
Formato de imagem	png		
Número máximo de	600		
Formas utilizadas	ativas		
Tamanho do texto	Min 5	Max 50	
Cor do texto	[Preto]		
Cor do fundo	[Branco]		
OK		Cancel	

Para maior visibilidade é melhor escolher um tamanho compatível com o tamanho da imagem.

Seguir as mesmas orientações dos gráficos: tamanho conforme resolução de tela.

Escolher o formato de gravação para edição em aplicativos gráficos.

Escolher o mínimo de 5 e o máximo menor de 50, na maioria das vezes.

Escolher quais formas se deseja fazer a nuvem.

O default é preto e branco, mas pode-se mudar as cores conforme o desejo.

Nuvem de palavras

Parâmetros

O sistema traz a lista de formas escolhida (ativas, complementares ou ambas) com a respectiva frequência para se marcar quais se deseja mostrar na nuvem.

Forma	Freqüência
ingress	10
abstrage	10
relat	10
circu	8
abstrand	7
clifre	7
con	7
avtar	6
part	6
lks	6
concre	6
hauar	6
constru	6
lks	5
pa	5
con	5
constru	4
part	4
lks	4
lks	4
medal	4


Nuvem de palavras

Resultados:

O sistema traz a nuvem de acordo com as configurações escolhidas.

O arquivo em formato png é gravado na pasta de resultados dentro da pasta criada com o nome do corpus: corpus_wordcloud_1.

Ele pode ser editado por um aplicativo gráfico que aceita o formato png.

 Executar e Mostrar os resultados no Iramuteq (artigos_novo2_corpus_1)

Nuvem de palavras

Exercício:

1) Realizar a nuvem de palavras do corpus importado, conforme configurações abaixo:

a) Propriedades: usar a mesma configuração anterior: Advérbio suplementar: 0; Artigo definido: 0; Artigo indefinido: 0; Auxiliar: 0; e Número (chifre): 0.

b) Configurações gráficas: manter as configurações padrão

c) Selecionar formas com frequência acima de 10, porém retirando as seguintes palavras: não, gente, mais, porque, achar, então, assim, aqui, também, aí, como, mesmo, ver, já, tudo, coisa, só, lá, pessoa, vez, parte, quando, ainda, até, próprio, ali, ao.