

Análise de Dados Textuais em Pesquisas de Mobilidade Urbana e Transporte com IRaMuteQ

Kevin Masinda Mahema

kevinmasinda16@gmail.com



Grupo de pesquisa comportamento em transportes e novas tecnologias - GCTNT
Programa de Pós Graduação em Transportes - PPGT/UnB
Departamento de Eng. Civil e Ambientais
Universidade de Brasília - UnB

10 de janeiro de 2022

Software IRaMuTeQ é um software livre sob licença de GNU GPL (v2). É um software que usa a interface do R (www.r-project.org) e pode ser expandido a partir da linguagem de programação Python (www.python.org). Permitindo assim, uma análise textual multidimensional e estatística. A análise textual ou análise sobre os corpus de textos, é feito a partir do método de classificação hierárquica descendente descrito por Reinert; além de outras análise lexicais tais como a análise de similitude de um segmento de texto, a estatística clássica textual e uma busca específica a partir de um segmento de texto bem definido, a fim de auxiliar na explicação/interpretação de textos.

Este manual (em formato de minicurso) tem como objetivo de apresentar a ferramenta IRaMuTeQ como uma ferramenta de análise de conteúdo textual que ajuda na engenharia comportamental para tomada de decisão, na análise de dados textuais ou lexical. O software IRaMuTeQ será apresentada desde a sua instalação, passando pela preparação de corpus, análise do corpus, manipulação de software e por fim a obtenção e exportação dos resultados finais.

Ao final deste manual, o aprendiz deverá ser capaz de:

Objetivos (cont.)

- ▶ Analisar dados textuais ou lexical;
- ▶ Compreender a dicotomia clássica entre quantitativo e qualitativo na análise de dados;
- ▶ Quantificar e empregar cálculos estatísticos sobre variáveis essencialmente qualitativas - os textos;
- ▶ Tornar possível, a partir da análise textual, a descrição material produzido por um determinado produtor;
- ▶ Utilizar a análise textual com a finalidade comparativa, relacional, comparando produções diferentes em função de variáveis específicas que descrevem quem produziu o texto;
- ▶ Interpretar os dados resultantes da análise textual.

O presente manual (minicurso) é destinado aos

- ▶ Alunos tanto da graduação como da pós graduação;
- ▶ Professores;
- ▶ Empresas que trabalham com análise qualitativa de textos tais como:
 1. Entrevista;
 2. Livro;
 3. Pesquisa de opinião;
 4. Artigo científicos;
 5. Jornais;
 6. Etc.

- ▶ Pesquisadores voltados não somente nas áreas de ciências de saúdes, humanas, sociais, mas também na área de ciências exatas e das engenharias.

Não é necessário ter conhecimento em linguagem de programação Python e/ou R

IRaMuteQ é uma interface do software R. **É indispensável a instalação do software R.**

- ▶ É recomendável baixar uma das versões do R a partir da [versão 3.6](#)
- ▶ IRaMuteQ não é compatível com as **versões 4.0 ou posterior de R.**
- ▶ IRamuTeQ pode ser baixado a partir da página do projeto no [sourceforge](#)
- ▶ Há versões disponíveis para Windows, Mac OS X et GNU/Linux.
- ▶ Código fonte do projeto no [GitHub](#)

1. Baixar e instalar R a partir www.r-project.com
2. Baixar e instalar IRaMuTeQ

1. Linux Mint: é preciso instalar o package **r-base-dev** (sudo apt-get install r-base-dev) a partir de um terminal.
2. Não é mais necessário instalar as bibliotecas R (rlg, ape, gee, igraph, proxy, rgl) como nas primeiras versões de IRaMuteQ
3. Na versão atual de IRaMuTeQ, pode acontecer que o pacote rgl não seja instalado. PS: não preocupa-se, o pacote serve somente para visualização 3D.
4. Para verificar se a instalação for bem sucedida, abre IRamuTeQ, vá em edição - preferência - verifique instalação.

Instalação - MacOS

1. Primeiro baixar e instalar o **software R**
2. Baixar e Instalar o **xquartz**
3. Baixar e Instalar **IRaMuTeQ**
4. Baixar o arquivo **Rgraph** e em seguida ir em aplicação - IRaMuTeQ (clique direto) - mostrar o conteúdo do pacote - Rscripts e por fim, substituir o arquivo Rgraph.
5. PS: sobre MacOS X pode acontecer que IRaMuTeQ não seja aberto por ser reconhecido como aplicação vindo de um fonte desconhecido. Será preciso abrir IRaMuTeQ pelo terminal. Com isso será preciso executar os seguintes comandos:

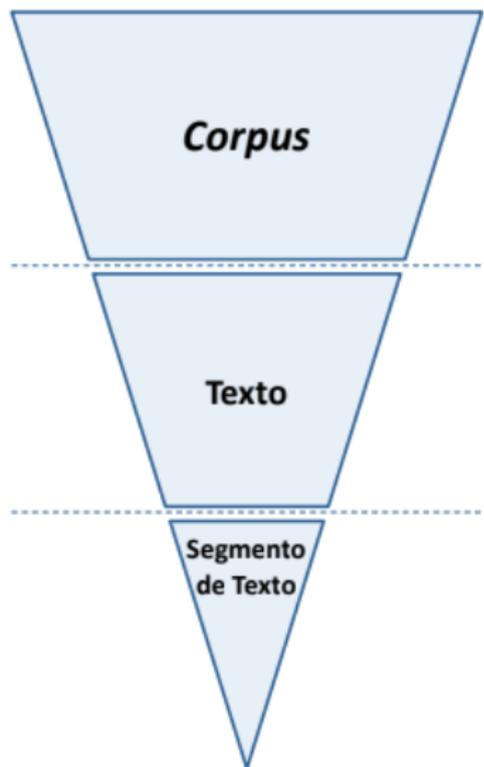
- `cd /Applications/iramuteq.app`
- `ls`
- `./iramuteq`

Por fim, quando o programa abrir, é necessário aceitar a instalação de todos os pacotes R (caso IRaMuTeQ pedir)

- ▶ **Segmentos de Texto - ST:** são geralmente os conjuntos de palavras que tem um tamanho de aproximadamente de três linhas de textos. Na maioria das vezes, é delimitado automaticamente pelo software em função do tamanho do corpus ou pode ser montado pelo pesquisador. É considerado como a unidade do corpus
- ▶ **Textos:** são os conjuntos de segmentos de texto geralmente definidos pelo pesquisador dependendo da natureza da sua pesquisa. O texto deve obrigatoriamente começar sempre por quatro asterisco: ******** e conter pelo menos uma variável. A separação entre dois textos é feito por uma **linha em branco**
- ▶ **Corpus:** É o conjunto de textos. De forma mais simples, corpus, é o nosso arquivo de análise. Só existe e tem um único corpus.

Um corpus contém vários textos, e esse último contém vários segmentos de texto

Introdução: Conceitos básicos (cont.)



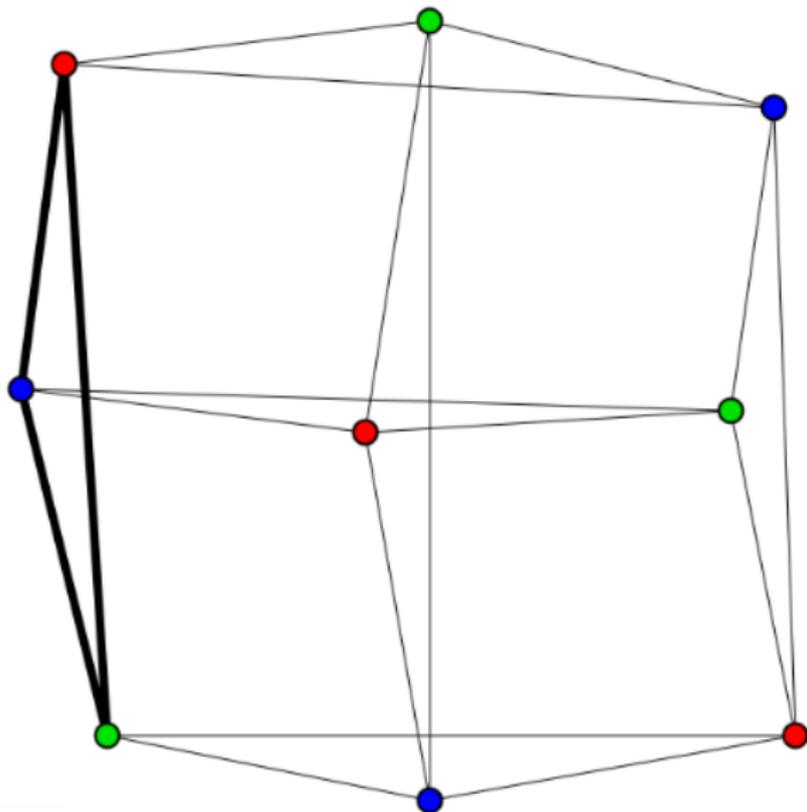
- ▶ **Classe:** pode ser definido como um conjunto de segmento de texto ou vocabulários de mesma homogeneidade ou natureza.
- ▶ **Lematização:** é o processo de trazer os substantivos nas suas formas masculino - singular, verbos nas suas formas no infinitivo e adjetivos nas suas formas masculino - singular também. O software realiza a lematização de forma automática (precisa ser ativado durante a análise pelo pesquisador)
- ▶ **Hapax:** são palavras que ocorrem/aparecem somente uma única vez em todo o corpus.

- ▶ **Grafos:** é um ramo da matemática que estuda as relações entre objetos de um mesmo conjunto. Por definição, um grafo denominado por $G(V,E)$ é um conjunto não vazio de vértices (nós) - V interligados por um conjunto de arestas - E . [1]

Dependendo da aplicação, arestas podem ou não ter direção, pode ser permitido ou não arestas ligarem um vértice a ele próprio e vértices e/ou arestas podem ter um peso (numérico) associado.

- ▶ **subgrafos:** um subgrafo de um grafo G é um pedaço de G . O conjunto de vértices e o conjunto de arestas do pedaço de G devem ser coerentes. Assim, é melhor formular o conceito como uma relação entre dois grafos: Um grafo H é subgrafo de um grafo G se todos os vértices e todas as arestas de H são vértices e arestas de G também. [1]

Noção de grafos (cont.)



- ▶ IRaMuTeq é uma interface de R que permite a inserção/anexação de um arquivo .txt ou em outras palavras, a inserção de umas cadeias de caracteres em um buffer de sistema de arquivos.
- ▶ Quando o arquivo é salvo, ele usa uma codificação de texto para decidir quais bytes cada caractere se tornará.

Um caráter na computação é o nome que se dá a cada um dos símbolos que se podem usar para produzir um programa de computador, bem como os textos e imagens apresentados na tela quando se executa um programa em modo texto. Já, uma **codificação de caracteres** é um padrão de relacionamento entre um conjunto de caracteres. [2]

- ▶ **Bit:** é a unidade básica que os computadores e sistemas digitais utilizam. Pode assumir somente dois valores, 0 ou 1. [3]
- ▶ **Bytes:** é uma unidade de informação digital equivalente a oito bits [3]

Existem vários tipos de codificação de caracteres que o IRaMuTeQ aceita. entre eles, pode-se citar:

1. **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. [4]
2. **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. [5]
3. etc.

- ▶ **Pixels:** a palavra pixel é uma junção de termos “**picture**” e “**element**”. Ou seja, “elemento de imagem”. É a menor unidade de uma imagem digital, sendo que, um conjunto de pixels com diferentes cores formam a imagem inteira. [6]
- ▶ **Default:** é um termo de origem inglês, que em português, significa padrão. A palavra default é um termo técnico em computação que quer dizer, configuração padrão, configuração pré-definida.[7]
- ▶ **permilagem ou por mil (‰):** proporção relativamente a mil; número em relação a mil; fração cujo denominador é 1000.
- ▶ **Distribuição de frequência:** uma distribuição de frequência é um agrupamento de dados em classes de modo a fornecer a quantidade e/ou a percentagem. Com isso, podemos agrupar e visualizar um conjunto de dados sem precisar levar em conta os valores individuais de dados em cada classe.

já uma **distribuição de frequência (absoluta ou relativa)** pode ser apresentada em tabelas ou gráficos. Uma distribuição de frequência agrupa os dados por classes de ocorrência. [8]

- ▶ **Distribuição hipergeométrica:** uma variável aleatória X tem uma distribuição hipergeométrica com parâmetros N , r e n quando esta representa o número de sucessos dentro de uma amostra de tamanho n extraída, sem reposição, de uma população de tamanho N formada por r sucessos e $N-r$ fracas. [9]
- ▶ **Distribuição Gama:** uma variável aleatória contínua X tem distribuição gama com parâmetros $\alpha > 0$ e $\lambda > 0$ se sua função densidade é dada por:

$$f_x(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \lambda^\alpha e^{-\alpha x}, & x > 0 \\ 0 & \text{caso contrário} \end{cases}$$

Note que, quando $\alpha = 1$, resulta a densidade exponencial com parâmetro λ , ou seja, a densidade exponencial é um caso particular da densidade gama. [9]

- ▶ **Distribuição Qui-Quadrado ou X^2 :** é um caso particular da distribuição Gama, onde, o parâmetro de forma α é igual a $\frac{n}{2}$, com n inteiro positivo, e o parâmetro λ é $\frac{1}{2}$ com n graus de liberdade. [9]

É uma das distribuições mais utilizadas em estatística inferencial, principalmente para realizar testes de X^2 . Este teste serve para avaliar quantitativamente a relação entre o resultado de um experimento e a distribuição esperada para o fenômeno. Isto é, ele nos diz com quanta certeza os valores observados podem ser aceitos como regidos pela teoria em questão. [7]

- ▶ **Lei de Zipf:** trata-se de uma lei de potências sobre a distribuição de valores de acordo com o número de ordem numa lista. Na linguística, ela conta a frequência que cada palavra aparece em um determinado texto e segue uma distribuição que pode-se aproximar pela fórmula:

$$P_n \sim 1 \cdot n^{-\alpha}$$

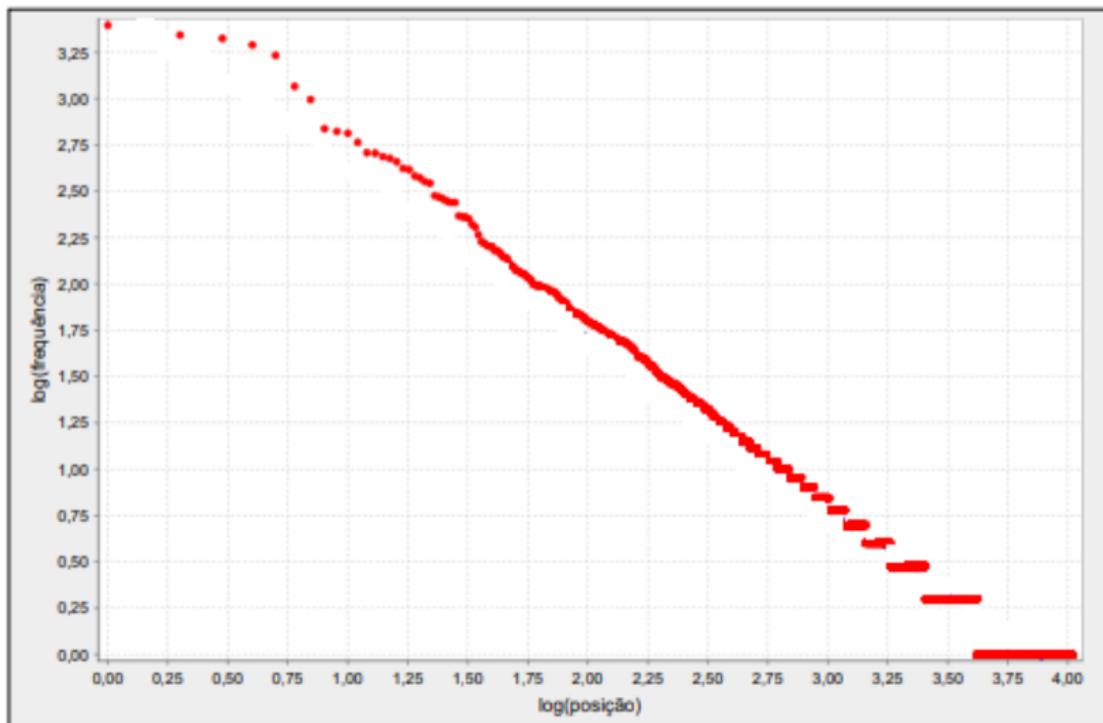
Exemplo: suponha que após a contagem de ocorrência de cada palavra dentro de um texto e classificada em ordem decrescente pela sua frequência, representada na tabela abaixo: [10]:

Noções: informática, matemática e estatística (cont.)

Posição(x)	Frequência(y)	Palavras	$\hat{x} = \log x$	$\hat{y} = \log y$	palavra
1	2483	a	0,000...	3,396...	a
2	2203	o	0,301...	3,343...	o
3	2112	que	0,477...	3,324...	que
4	1949	bom	0,602...	3,2898...	bom
⋮	⋮	⋮	⋮	⋮	⋮
178	37	mestrado	2,250...	1,568...	
⋮	⋮	⋮	⋮	⋮	⋮
10447	1	unb	4,018...	0,000...	unb
10448	1	congo	4,019...	0,000...	congo
10449	1	brasil	4,019...	0,000...	brasil

Nas leis de potencias, ao invés de analisar a posição(x) e a frequência(y), são analisados $\hat{x} = \log x$ e $\hat{y} = \log y$ e os valores são apresentados em uma figura.

Noções: informática, matemática e estatística (cont.)



Criação e preparação de corpus

A criação de um corpus é simples e segue os seguintes passos:

- ▶ Abrir um editor de texto da sua escolha. word, libreOffice, etc. (recomendo o sublime 3)
- ▶ Colar o texto a ser analisado no seu editor de texto, salvar o arquivo com extensão de texto (.txt)

Qualquer arquivo .txt pode ser considerado um corpus. Mas para que o seu corpus seja reconhecido e analisado no software IRaMuTeq, é preciso se atentar nas seguintes regras de formatação:

1. Os textos são introduzidos por quatro estrelas (****) seguido de uma variável estrelada.
2. Um texto é separado de um outro texto, pela uma quebra de linha (**uma única linha em branco**).
3. Cada texto deve conter pelo menos uma variável.
4. Uma variável é introduzido por uma estrela (*).

5. É possível colocar uma variável estrelada dentro de um texto, se é somente se, essa variável começa por um traço seguido de estrela (-*). Nesse caso, fala-se então de um **corpus com temático ou corpus temático**.
6. No corpus temático, todos os textos devem pertencer a uma temática.
7. Não usar caráter tal como traço (;). Palavras com traço, trocar por sublinha durante a montagem do corpus. ex: Jean-Pierre por Jean_Pierre ou JeanPierre.

Cuidado: IRaMuTeQ é case sensitive, ou seja, Asa_Sul e Asa_sul são formas diferentes. (a palavra sul é uma modalidade de asa)

Criação e preparação de corpus (cont.)

Para evitar incompatibilidade de codificação de caracteres entre aquela que está salva no arquivo e a escolhida no IRaMuTeQ (o que acontece muito), as figuras abaixo, mostram como deve-se escolher a codificação da sua escolha dentro das aceitas pelo IRaMuTeQ com ajuda do Sublime Text 3 e por fim, a maneira certa de salvar o corpus.

Criação e preparação de corpus (cont.)

1. Clicar em file

```

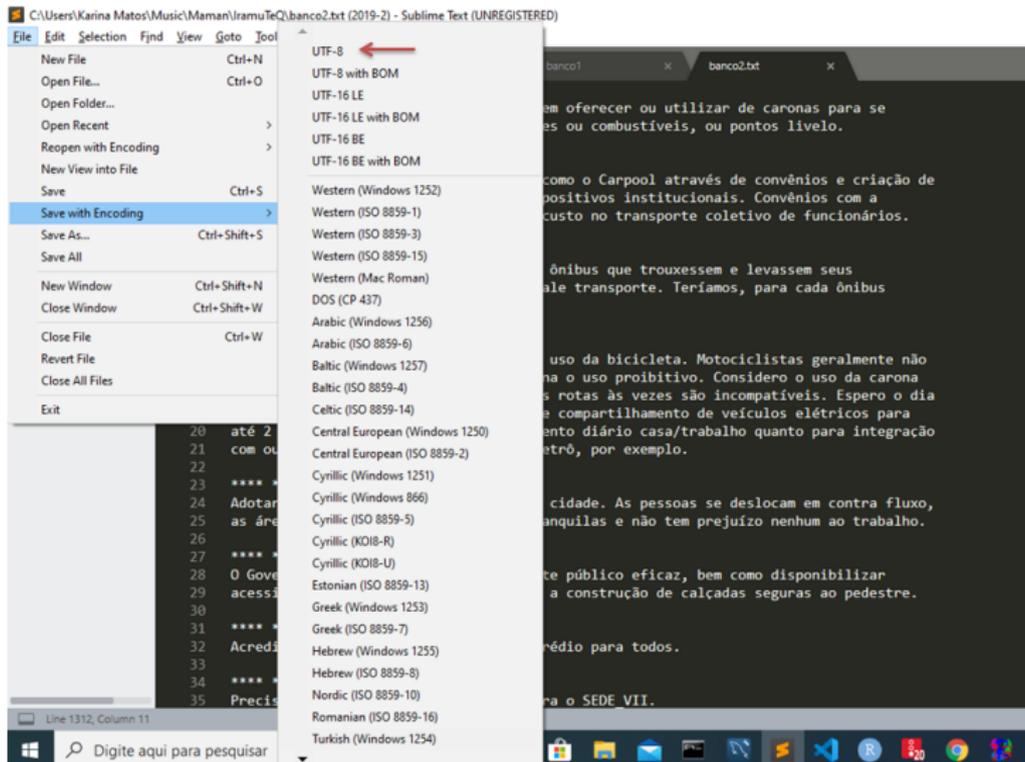
C:\Users\Kevin Mates\Music\Maman\ramu\TeC\banco.txt (2019-2) - Sublime Text (UNREGISTERED)
File Edit Selection Find View Goto Tools Project Preferences Help
New File Ctrl+N
Open File... Ctrl+O
Open Folder...
Open Recent
Reopen with Encoding
New View into File
Save Ctrl+S
Save with Encoding
Save As... Ctrl+Shift+S
Save All
New Window Ctrl+Shift+N
Close Window Ctrl+Shift+W
Close File Ctrl+W
Revert File
Close All Files
Exit

opinio_funcionarioBB
na de incentivo do Banco para quem oferecer ou utilizar de caronas para se
per até o trabalho. Voucher, vales ou combustíveis, ou pontos livrelo.
opinio_funcionarioBB
par a utilização de aplicativos como o Carpool através de convênios e criação de
de funcionários através dos dispositivos institucionais. Convênios com a
99taxi, etc para a redução do custo no transporte coletivo de funcionários.
opinio_funcionarioBB
tesas deveriam voltar a oferecer ônibus que trouxessem e levassem seus
ários, ao invés de oferecer o vale transporte. Teríamos, para cada ônibus
nado, 50carros em casa.
opinio_funcionarioBB
caso, a distância inviabiliza o uso da bicicleta. Motociclistas geralmente não
peitados no trânsito, o que torna o uso proibitivo. Considero o uso da carona
nte, mas é pouco prático pois as rotas às vezes são incompatíveis. Espero o dia
permos à disposição um sistema de compartilhamento de veículos elétricos para
20 até 2 ocupantes, tanto para o deslocamento diário casa/trabalho quanto para integração
21 com outros meios de transporte, como metrô, por exemplo.
22
23
24 **** *opinio_funcionarioBB
25 Adotar sedes fora da região central da cidade. As pessoas se deslocam em contra fluxo,
26 as áreas de estacionamento são mais tranquilas e não tem prejuízo nenhum ao trabalho.
27
28 **** *opinio_funcionarioBB
29 O Governo deveria investir no transporte público eficaz, bem como disponibilizar
30 acessibilidade nas vias públicas, como a construção de calçadas seguras ao pedestre.
31
32 **** *opinio_funcionarioBB
33 Acredito que poderíamos ter vagas no prédio para todos.
34
35 **** *opinio_funcionarioBB
36 Precisa voltar a ter a VAN do metro para o SEDE VII.

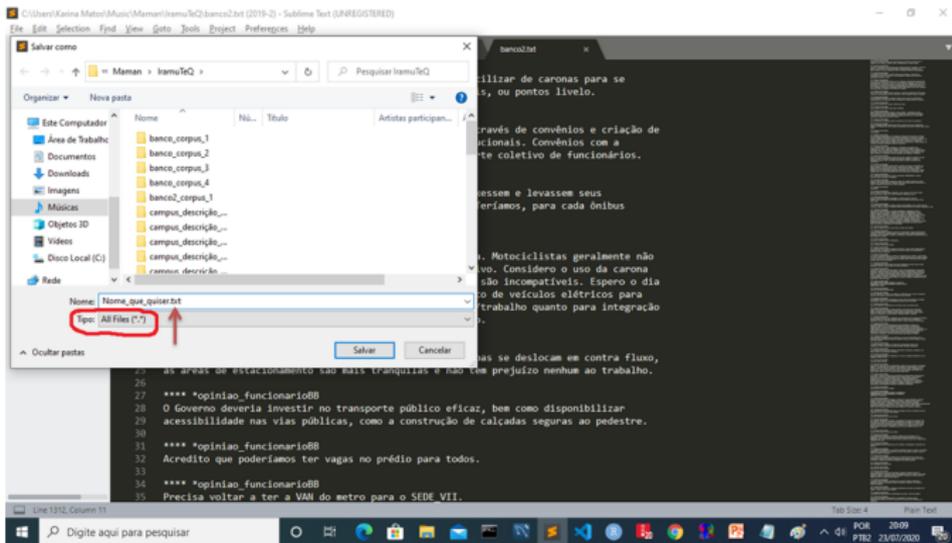
```

2. Deixar o mouse em cima

Criação e preparação de corpus (cont.)



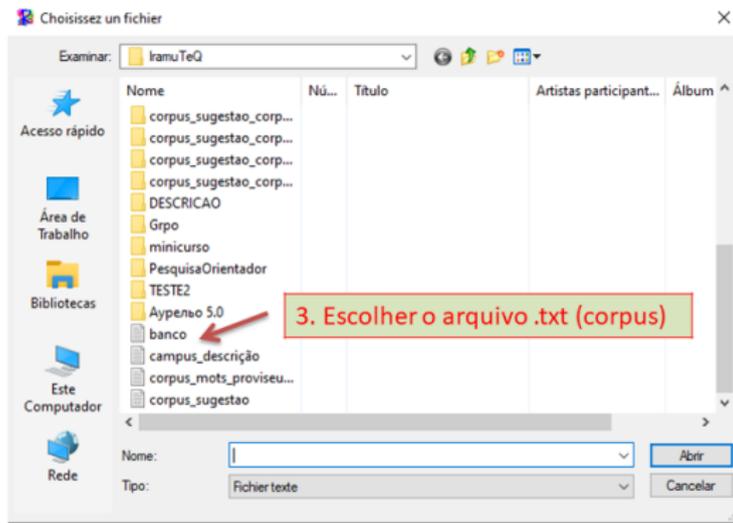
Criação e preparação de corpus (cont.)



Carregando um corpus

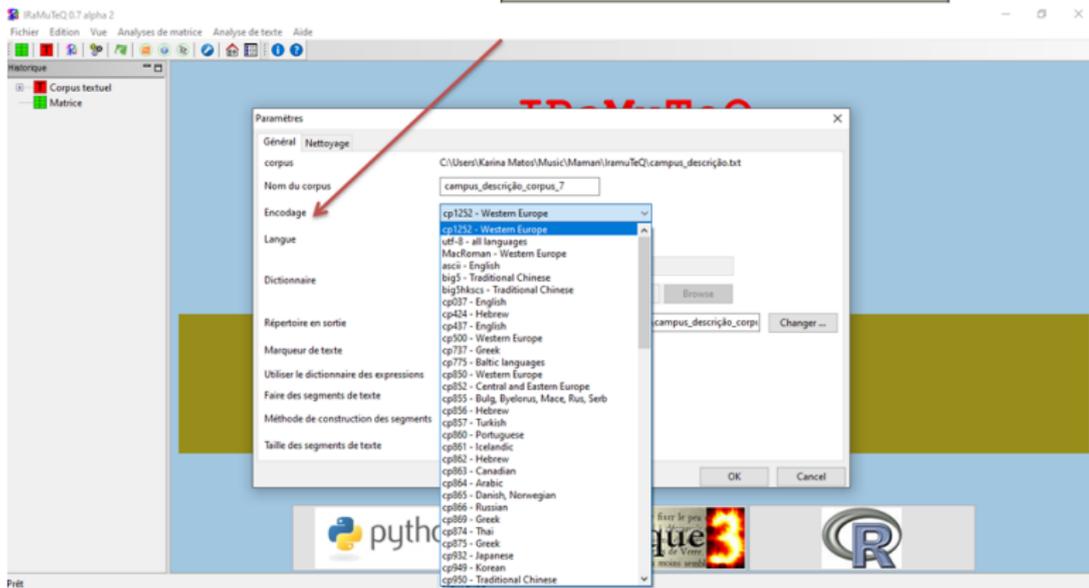
The screenshot shows the IRaMuTeQ 0.7 alpha 2 application window. The 'Fichier' menu is open, displaying options such as 'Ouvrir une matrice', 'Ouvrir un corpus texte', and 'Ouvrir une Analyse'. Two red arrows point to the 'Fichier' menu and the 'Ouvrir un corpus texte' option, with callout boxes containing the instructions: '1. Clicar no arquivo' and '2. Clicar em abri um corpus de texto'. The main window features the IRaMuTeQ logo, the URL 'http://www.iramuteq.org', and version information: 'Version 0.7 alpha 2', 'Laboratoire IRADIS', 'RIPROB', 'Licence GNU GPL', and '(c) 2000-2014 Pierre Ratinaud'. At the bottom, there are logos for Python, Lexique 3, and R.

Carregando um corpus (cont.)



Configurações: Aba Geral - IRaMuTeQ

4. Escolher a codificação adequada



Recomenda-se escolher: cp1252(windows) ou utf-8 all languages

Configurações: Aba Geral - IRaMuTeQ (cont.)

5. Escolher a língua de análise

The screenshot displays the IRaMuTeQ 0.7 alpha 2 application window. The 'Paramètres' dialog box is open, showing the 'Langue' dropdown menu with 'français' selected. A red arrow points from the text '5. Escolher a língua de análise' to the 'Langue' dropdown. The background shows the main application window with a sidebar on the left and a footer with logos for Python, Lexique 3, and R.

Configurações: Aba Geral - IRaMuTeQ (cont.)

6. Escolher o dicionário de análise

Paramètres

Général Nettoyage

corpus C:\Users\Karina Matero\Music\Maman\IramuTeQ\campus_descricao.bit

Nom du corpus campus_descricao_corpus_7

Encodage cp1252 - Western Europe

Langue français

Dictionnaire

Défaut french

Autre Browse

Répertoire en sortie C:\Users\Kevin\IramuTeQ\NOME_SUA_ESCOLHA

Marqueur de texte ****

Utiliser le dictionnaire des expressions

Faire des segments de texte

Méthode de construction des segments occurrences

Taille des segments de texte 40

OK Cancel

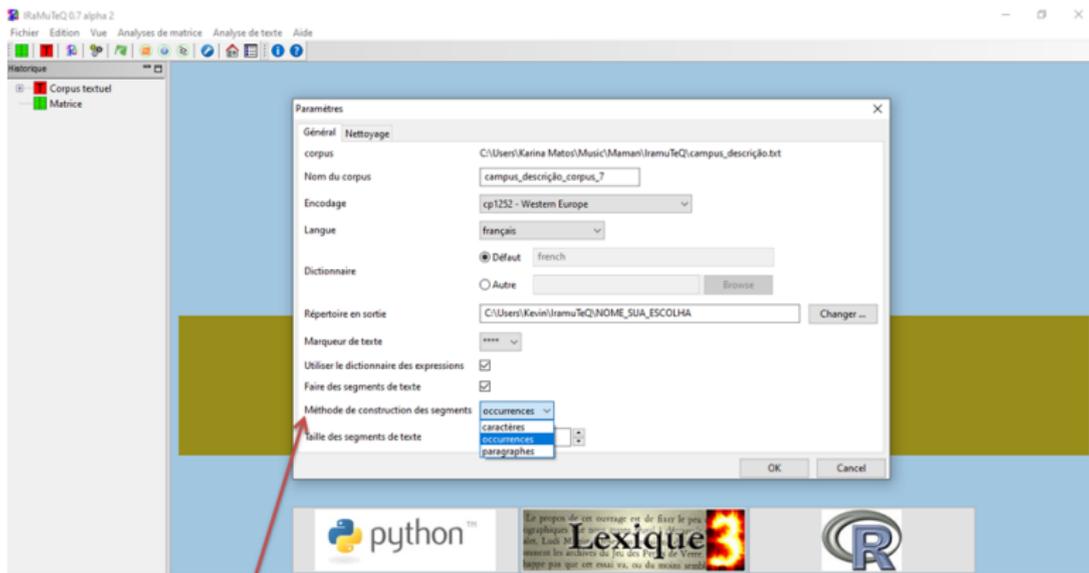
python

Lexique 3

R

7. Escolher a pasta de saída

Configurações: Aba Geral - IRaMuTeQ (cont.)



8. Escolher o método de construção de segmento de texto(ST)

Aba Limpeza: IRaMuTeQ (cont.)

- ▶ **Colocar o texto em minúsculo:** IRaMuTeQ converte de forma automática, todas as formas (palavras) em maiúsculas para minúsculas. Se desativar essa opção, IRaMuTeQ vai considerar por exemplo, **UnB** e **unb**, como duas formas diferentes.
- ▶ **Eliminar as carâteres fora dessa lista:** Por padrão, somente as palavras alfanuméricas e acentuadas são levadas em contas.
- ▶ **Trocar apóstrofos por espaço:** Substituí apóstrofos por espaço.
- ▶ **Trocar o traço por espaço:** Substituí os - pelos espaços.
- ▶ **Manter a pontuação:** Se preferir manter a pontuação, deve-se evitar de usar (;) dentro do corpus.
- ▶ **Sem espaço entre duas formas:** Se essa opção for ativado, IRaMuTeQ não irá usar o espaço como delimitador entre formas.

A análise estatística ou análise lexicográfica, apresenta uma análise simples sobre o corpus de texto. Identifica e reformata as unidades de texto, transformando textos em ST; detecta as formas (palavras) com suas respectivas frequências, calcula a frequência média das formas e o total dos hapax no corpus; faz uma redução de palavras (forma reduzida) com base nos vocabulários que têm a mesma raiz ou os lematizam; monta um dicionário de formas reduzidas e mostra todas as formas ativas e suplementares.

Análise: Estatística (cont.)

The screenshot shows the IRaMuTeQ 0.7 alpha 2 software interface. The main window displays the 'Description minicurso' for a corpus named 'minicurso'. The interface is annotated with three numbered steps:

- 1. Iniciar uma análise estatística**: Points to the 'Análise de texto' menu item in the top toolbar.
- 2. Escolher usar ou não usar a lematização**: Points to the 'Lematisation' options in the 'Paramètres' dialog box, where 'oui' is selected.
- 3. Configurações de formas**: Points to the 'Propriétés' button in the 'Paramètres' dialog box.

The 'Description minicurso' window shows the following details:

Description du corpus	
Nom	minicurso
Langue	portuguese
Encodage	cp1252
originalpath	C:\Users\Karina Mates\Music\Maman\iramuteq\corpus_sugestao.txt
pathout	C:\Users\Karina Mates\...
date	Wed Jul 22 13:10:46 2020
time	0h 0m 1s
Paramètres	
ucemethod	1
ucesize	40
keep_caract	^a-zA-Z0-9ÀÁÂÃÄÅ
expressions	1
Statistiques	
Nombre de textes	106
Nombre de segments de texte	116
occurrences	1864
Nombre de formes	525
Nombre d'hapax	313 - 59.62 % des formes - 16.75 % des occurrences

The 'Paramètres' dialog box shows the following settings:

- Lematisation: oui, non
- Propriétés: propriétés, autres
- Dictionnaire: indexation, Autre

Análise: Estatística (cont.)

The screenshot shows the IRaMuteQ 0.7 alpha 2 software interface. The main window displays the 'Description minicurso' corpus with various statistics:

Nom	minicurso
Langue	portuguese
Encodage	cp1252
originalpath	C:\Users\Karina Mateo\Music\Maman\iramuteq\corpus_sugestao.txt
pathout	C:\Users\Karina Mateo\...
date	Wed Jul 22 13:10:46 2020
time	0h 0m 1s
Paramètres	
ucemethod	1
ucesize	40
keep_caract	^a-zA-Z0-9-áâãäåäåä
expressions	1
Statistiques	
Nombre de textes	106
Nombre de segments de texte	1864
occurrences	
Nombre de formes	525
Nombre d'hapax	313 - 59.62 % des formes - 16.79 % des occurrences

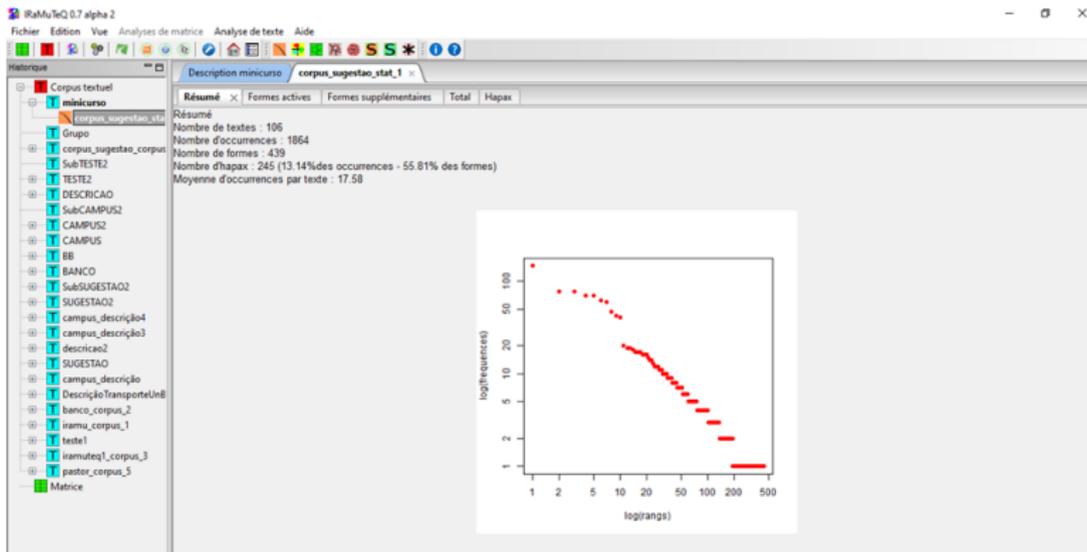
A 'Paramètres' dialog box is open, showing the following options:

- Lemmatisation: oui, non
- Paramètres des clés: propriétés
- Dictionnaire: indexation, Autre

Two red arrows point from green callout boxes to the dialog box:

- Box 4: '4. Indexar dicionário padrão ou outro' points to the 'Dictionnaire' section.
- Box 5: '5. Iniciar análise' points to the 'OK' button.

Resultado: análise estatística



Resultado: análise estatística (cont.)

1. Formas ativas

2. Categoria gramatical

3. Frequência da palavra

Forme	Freq	Types
ombus	17	nr
mais	62	adv
x	59	nom
linha	41	nom
unb	20	nr
brasil	19	nr
horarios	19	nr
rota	17	nom
ã	17	nr
passar	16	ver
quantidade	16	nom
dever	15	ver
direto	14	adj
rodovia	14	nr
ni	13	nr
cellandia	12	nr
maior	12	adj
aumentar	11	ver
jobradinho	11	nom
campus	10	nom
cidade	10	nom
noite	10	nom
satelites	10	nr
horario	9	nr
intervalo	9	nom
linha_110	9	nr
linha_1102	8	nr
pico	8	nom
vacatur	8	nr

Resultado: análise estatística (cont.)

The screenshot shows the IRaMuTeQ 0.7 alpha 2 interface. The main window displays a table with columns 'Forme', 'Freq', and 'Types'. The row 'quantidade' is selected, and a context menu is open over it, showing 'Formes associées' and 'Concordancier'. Two red arrows point from text boxes to these menu items.

Forme	Freq	Types
onibus	77	nr
mais	62	adv
e	59	nom
linha	41	nom
umb	20	nr
brasil	19	nr
horarios	19	nr
rota	17	nom
ã	17	nr
passar	16	ver
quantidade	15	
dever	15	
rodoviaria	14	nr
na	13	nr
celandia	12	nr
maior	12	adj
aumentar	11	ver
sobradinho	11	nom
campus	10	nom
cidade	10	nom
noite	10	nom
satelites	10	nr
horario	9	nr
intervalo	9	nom
linha_110	9	nr
linha_1102	8	nr
pico	8	nom
saosebartaio	8	nr

Annotations:

- Mostra as formas associadas (points to 'Formes associées')
- Mostra a forma dentro do ST (points to 'Concordancier')

Produz uma análise fatorial de correspondência em uma tabela de continência, cruzando as formas ativas com as variáveis. ou seja, possibilita a análise da produção textual em função das variáveis de caracterização com no mínimo duas modalidades.

Análise de especificidade e AFC (cont.)

The screenshot shows the IRaMuTeQ 0.7 alpha 2 interface. The main window displays the 'Description minicorso' for a corpus named 'minicorso'. The 'Paramètres' dialog box is open, showing options for 'Lemmatisation' (oui) and 'Dictionnaire' (indexation). A red arrow points from a yellow box containing the text '1. Iniciar uma análise de especificidade e AFC' to the 'Analyse de matrice' button in the top menu bar.

1. Iniciar uma análise de especificidade e AFC

Description du corpus	
Nom	minicorso
Langue	portuguese
Encodage	cp1252
originalpath	C:\Users\Karina Mates\Music\Mamao\IramuTeQ\corpus_sugestao.txt
pathout	C:\Users\Karina Mates\...
date	Wed Jul 22 13:10:46 2020
time	0h 0m 1s
Paramètres	
ucemethod	1
ucesize	40
keep_caract	^a-zA-Z0-9ââââââââââ
expressions	1
Statistiques	
Nombre de textes	106
Nombre de segments de texte	116
occurrences	1864
Nombre de formes	525
Nombre d'hapax	313 - 59.62 % des formes - 16.79 % des occurrences

Prêt

Bienvenue

Análise de especificidade e AFC (cont.)

A seleção de uma variável é obrigatória. A variável selecionada, deve ter no mínimo duas modalidades. Caso selecionar duas ou mais variáveis, somente uma será levada em conta para análise.

A Análise Fatorial de Correspondência (AFC:) é uma representação gráfica que auxilia na visualização dos dados e apresenta quão as classes ou formas se aproximam/distanciam.

Existem duas possibilidades de cálculo nessa análise, a hipergeométrica e a X^2 (qui-quadrado)

Análise de especificidade e AFC (cont.)

Escolha de formas a serem analisadas

Escolha de variáveis a serem analisadas

Escore

Choix des Variables

Formes utilisées: actives et supplémentaires

Sélection par: variables

*local
*epmiao

Choix:

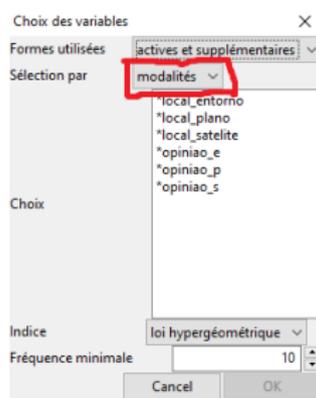
Indice: loi hypergéométrique

Fréquence minimale: 10

Cancel OK

Análise de especificidade e AFC (cont.)

Se escolhermos a seleção por modalidade, todas as modalidades devem pertencer a uma variável, caso contrário, a análise não terá sentido.



Análise de especificidade e AFC (cont.)

Formes	Formes banais	Types	Frequências des formes	Frequências des typ
formas		*opiniaio_e	*opiniaio_p	*opiniaio_s
noite	2.8988	1.3872	-4.3355	
quantidade	0.6527	-0.5261	0.2532	
horarios	0.5421	0.4159	-0.641	
mais	0.3329	-0.5488	0.4462	
aumentar	0.3257	1.2032	-1.332	
maior	0.2987	0.308	-0.3842	
direto	0.2532	-0.4046	0.346	
rodoviar	0.2532	-0.7936	0.6508	
passar	0.2163	0.9888	-0.9672	
unb	-0.1602	-1.3147	1.2258	
linha	-0.2263	-1.0445	1.0441	
onibus	-0.2481	-0.7408	0.751	
campus	-0.2522	-0.1968	0.3119	
satelites	-0.2522	-0.4787	0.7009	
cidade	-0.2522	-0.4787	0.7009	
sobradinho	-0.2777	-1.1655	1.5229	
ceilandia	-0.3033	-1.2731	1.6636	
nao	-0.3547	1.2539	-0.8357	
dever	-0.3806	-0.4641	0.7398	
rota	-0.4324	0.5402	-0.2986	
brasil	-0.4844	8.4547	-6.8024	

Formes	Formes banais	Types	Frequências des formes	Frequências des typ
formas		*opiniaio_e	*opiniaio_p	*opiniaio_s
onibus	4	13	60	
mais	4	11	47	
noite	4	5	1	
horarios	2	5	12	
quantidade	2	2	12	
linha	2	5	34	
direto	1	2	11	
unb	1	1	18	
aumentar	1	5	5	
maior	1	3	8	
passar	1	6	9	
rodoviar	1	1	12	
nao	0	6	8	
campus	0	2	8	
rota	0	5	12	
brasil	0	16	3	
ceilandia	0	0	12	
satelites	0	1	9	
dever	0	2	13	
cidade	0	1	9	
sobradinho	0	0	11	

Formes	Formes banais	Types	Frequências des formes	Frequências des typ
formas		*opiniaio_e	*opiniaio_p	*opiniaio_s
nr	1.3089	0.5499	-1.1535	
pre	0.9157	-0.748	0.3517	
pro_per	0.3378	0.9467	-0.9772	
adj	0.3312	0.3864	-0.4562	
conj	0.2469	-0.5803	0.6439	
adv_sup	-0.0238	0.7347	-0.6241	
adj_sup	-0.0238	-0.0884	0.1178	
ono	-0.0238	-0.0884	0.1178	
pro_int	-0.0477	-0.1769	0.2357	
adj_num	-0.0477	-0.1769	0.2357	
num	-0.1194	0.6372	-0.4872	
pro_pos	-0.1194	-0.4426	0.5099	
nom_sup	-0.263	-0.9756	1.3003	
nom	-0.3026	-0.9811	0.9651	
art_def	-0.3304	-0.3564	0.4489	
ver	-0.333	0.8223	-0.6631	
pro_rel	-0.4782	0.6335	-0.3842	
adv	-0.6707	0.6522	-0.3642	
pro_ind	-0.9158	-0.5654	1.0897	
ver_sup	-1.48	0.3453	0.5587	

Formes	Formes banais	Types	Frequências des formes	Frequências des types
formas		*opiniaio_e	*opiniaio_p	*opiniaio_s
onibus	166.67	141.3	191.08	
mais	166.67	119.57	148.08	
noite	166.67	54.35	3.18	
horarios	83.33	54.35	38.22	
quantidade	83.33	21.74	38.22	
linha	83.33	54.35	108.28	
direto	41.67	21.74	35.03	
unb	41.67	10.87	57.32	
aumentar	41.67	54.35	15.92	
maior	41.67	32.61	25.48	
passar	41.67	65.22	28.66	
rodoviar	41.67	10.87	38.22	
nao	0	65.22	25.48	
campus	0	21.74	25.48	
rota	0	54.35	38.22	
brasil	0	173.91	8.55	
ceilandia	0	0	38.22	
satelites	0	10.87	28.66	
dever	0	21.74	41.4	
cidade	0	10.87	28.66	
sobradinho	0	0	35.03	

Figura 2: Resultado da análise de especificidade

Análise de especificidade e AFC (cont.)

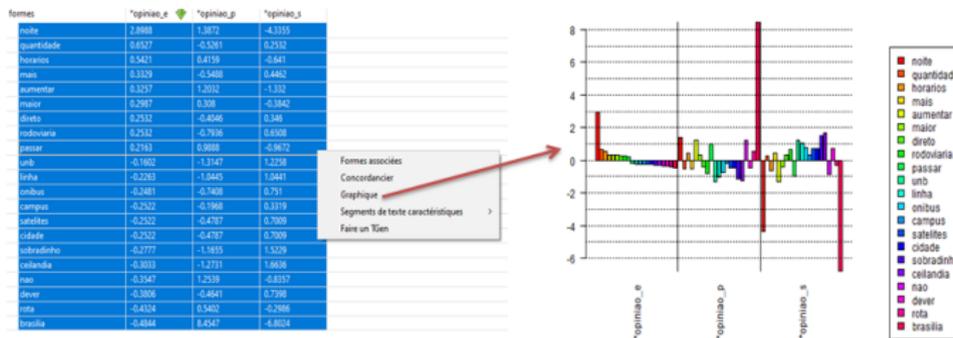


Figura 3: Gráfico de escore para cada forma

Análise de especificidade e AFC (cont.)

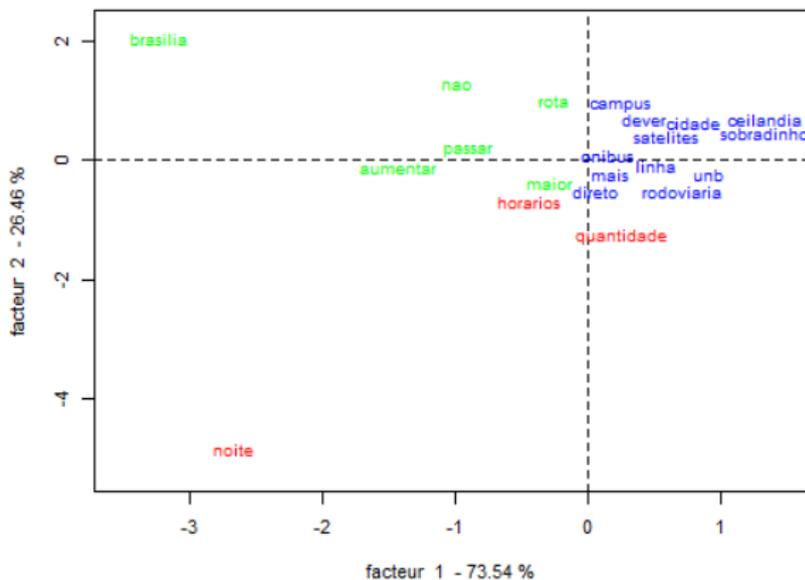


Figura 4: Gráfico AFC

Essa análise está baseado em princípios de teoria de grafos baseado na teoria descrito por Reinert, permitindo detectar as coocorrências entre formas, auxiliando na identificação da estrutura de conteúdo de um corpus. [11]

A classificação hierárquica descendente descrito por Reibert, visa obter classes de ST que apresentam vocabulários próximos ou distantes lexicalmente e com a ideia de que as palavras usadas em contexto similar, sejam associadas em um mesmo mundo léxico. [7]

Esta classificação pode ser realizada sob três formas de modalidades diferentes: **classificação simples sobre o texto**, **classificação simples sobre o segmento do texto** e por fim a **classificação dupla sobre o segmento de texto**.

Análise por classificação hierárquica descendente - CHD (cont.)

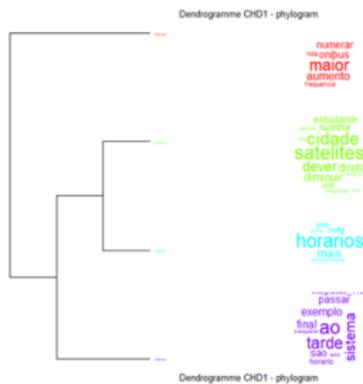
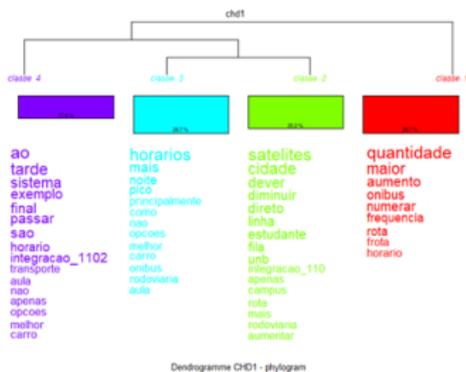
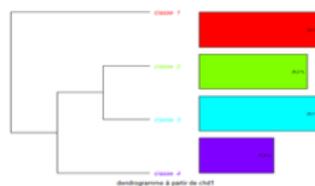
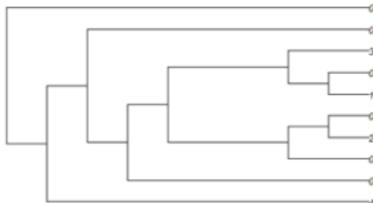
- ▶ Classificação simples sobre o texto: os textos mantêm a sua integridade, a classificação permite o reagrupamento dos textos mais próximos
- ▶ Classificação simples sobre o segmento do texto: a classificação incide sobre os segmentos de texto
- ▶ Classificação dupla sobre o segmento de texto: A classificação é feita sobre duas tabelas nas quais as linhas não são mais segmentos de texto, mas reagrupamento de segmentos de texto (RST). O mesmo tratamento é, portanto, feito duas vezes, porém mudando o número de formas ativas para RST. [11]

Análise por classificação hierárquica descendente - CHD (cont.)

```

-----
IRaMuteQ[Te]QI - Sat Aug 1 14:59:44 2020
-----
[Icons]
-----
[Icons]
-----
[Icons]
-----
Número de linhas: 197
Número de segmentos de texto: 184
Número de formas: 478
Número d'ocorrências: 1099
Número de termos: 292
Número de formas ativas: 332
Número de formas suplementares: 53
Número de formas ativas com uma frequência >= 3: 80
Moyenne de formas par segment: 9.233696
Número de classes: 4
115 segmentos classés sur 184 (62.50%)
-----
#####
tempo: 2h 5m 25s
-----

```



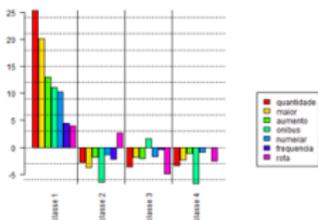
Análise por classificação hierárquica descendente - CHD (cont.)

CHD		Profilis		AFC			
1 Classe 1	2 Classe 2	3 Classe 3	4 Classe 4				
33/115	29/115	33/115	20/115				
28.7%	25.22%	28.7%	17.39%				
n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	12	14	85.71	25.33	nom	quantidade	< 0,0001
1	9	10	90.0	20.12	adj	maior	< 0,0001
2	5	5	100.0	12.99	nom	aumento	0.00031
3	25	59	42.37	11.08	nr	onibus	0.00087
4	4	4	100.0	10.3	ver	numerar	0.00133
5	4	6	66.67	4.46	nr	frequencia	0.03468
6	6	11	54.55	3.97	nom	rota	0.04625
7	2	3	66.67	2.17	nom	frota	NS (0,14067)
8	27	76	35.53	5.11	pre	de	0.02378
9	2	2	100.0	5.06		*meio_onibusoutro	0.02451
10	3	4	75.0	4.34		*deslocamento_carro	0.03717
11	9	21	42.86	2.52		*localidade_p	NS (0,11254)
12	1	1	100.0	2.51		*meio_nenhum	NS (0,11336)

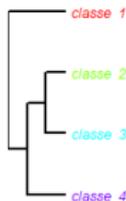
- Formes associées
- Chi2 par classe
- Chi2 par classe et dendrogramme
- Chi2 modalités de la variable
- Graphe du mot
- Concordancier
- Faire un TGen
- Outils du CNTRL (français uniquement)
- Graphe de la classe
- Segments répétés
- Segments de texte caractéristiques
- Nuage de mots de la classe
- Exporter...
- Exporter pour Tropes
- Exporter pour Owledge

Análise por classificação hierárquica descendente - CHD (cont.)

CHD / Análise - WGC		1 Classe 1		2 Classe 2		3 Classe 3		4 Classe 4	
		20/113		20/113		20/113		20/113	
		28,7%		25,2%		28,7%		27,9%	
n	nb	nb total	percentagem	nb2	nb1	nb2	nb1	nb2	nb1
1	9	10	90,0	25,12	na	meior			
2	5	5	100,0	12,08	na	numerar			
3	25	59	42,37	11,08	na	onibus			
4	4	4	100,0	10,3	na	numerar			
5	4	6	66,67	4,46	na	frequencia			
6	4	11	36,37	3,07	na	nota			
7	2	3	66,67	2,17	na	nota			
8	27	79	34,30	5,11	na	de			
9	2	2	100,0	3,96	na	*mais_judicioso			
10	3	4	75,0	4,34	na	*abdicar_nao			
11	6	21	42,86	2,52	na	*faculdade_p			
12	1	1	100,0	2,31	na	*mais_nobrem			



frequencia
onibus
aumento
quantidade
maior
numerar



```

Segmento de teste característico: Classe 1
*** "mais_judicioso,onibus" *mais,mais,onibus" *faculdade_p" *integracao_ap1
score: 49,40

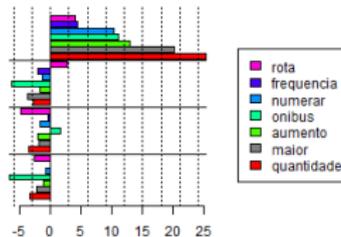
maior quantidade de onibus at 100,00
*** "mais_judicioso,onibus" *mais,mais,onibus" *integracao_ap1
score: 49,40

aumento e quantidade de onibus no decorrer do dia e no trajeto dentro do campus da ufrj
*** "mais_judicioso,onibus" *mais,mais,onibus" *integracao_ap1
score: 49,40

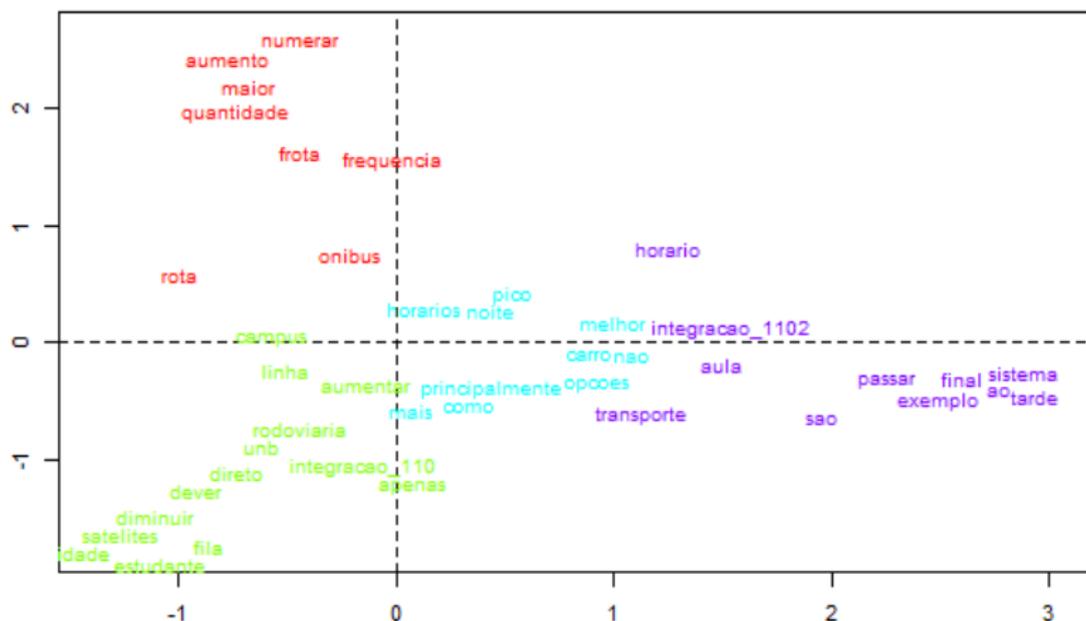
de pois e uma quantidade maior no horario da noite apos as 22h,230 principal e integracao_102
*** "mais_judicioso,onibus" *mais,mais,onibus" *integracao_ap1
score: 49,40

numerar onibus de linha e notas de onibus
*** "mais_judicioso,onibus" *mais,mais,onibus" *integracao_ap1
score: 49,40

quantidade e frequencia de onibus
*** "mais_judicioso,onibus" *mais,mais,onibus" *integracao_ap1
score: 49,38
    
```



Análise por classificação hierárquica descendente - CHD (cont.)



A Análise de similitude é baseada na teoria dos grafos cujos resultados auxiliam no estudo das relações entre objetos de um modelo matemático. [7]

A análise de similitude no software IRaMuTeQ apresenta um grafo que mostra a ligação entre formas de um corpus. 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. [11]

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.

Análise de similitude (cont.)

Paramètres

compter 80

formes	eff
onibus	80
mais	61
linha	45
horarios	20
unb	20
rota	17
passar	16
quantidade	16
dever	15
nao	14
direto	14
rodoviaria	14
maior	12
campus	11
aumentar	10
satelites	10
noite	10
cidade	10
horario	9
intervalo	9
integracao_1102	8
integracao_110	8
pico	8
ao	7
frequencia	7

Paramètres du graphe Paramètres graphiques

Indice cooccurrence

Présentation fruchterman_reingold

Type de graphique statique Format de l'image png

Arbre maximum

Seuil pour les arêtes 1

Texte sur les sommets

Indices sur les arêtes

Arêtes courbées

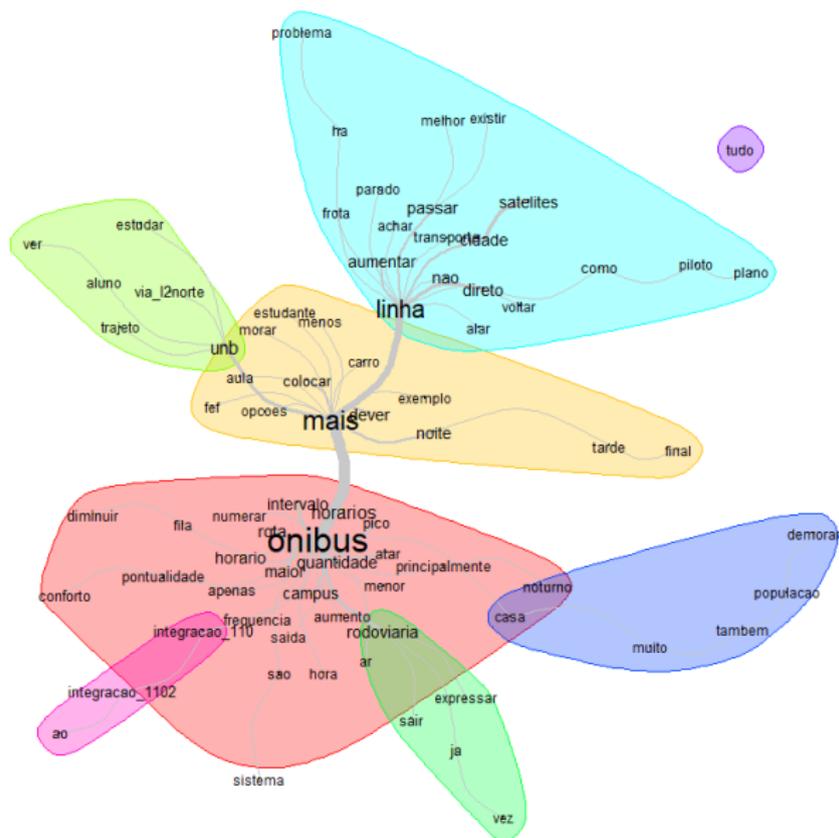
Taille du texte 10

Communautés edge.betweenness.community halo

Sélectionnez une variable

OK Cancel

Análise de similitude (cont.)



- [1] P. F. .-. IME-USP. (2014), Teoria de grafos, URL: https://www.ime.usp.br/~pf/algoritmos_para_grafos/aulas/graphs.html.
- [2] Microsoft. (2020), Codificação de caracteres, URL: <https://docs.microsoft.com/pt-br/powershell/scripting/dev-cross-plat/vscode/understanding-file-encoding?view=powershell-7>.
- [3] significados. (2020), bit - Significado de Byte, URL: <https://www.significados.com.br/>.
- [4] M. E. Salviati. (2017), Manual do Aplicativo Iramuteq, URL: <http://www.iramuteq.org/documentation/fichiers/manual-do-aplicativo-iramuteq-par-maria-elisabeth-salviati>.
- [5] P. Feofiloff. (2018), Unicode e UTF-8, URL: <https://www.ime.usp.br/~pf/algoritmos/>.
- [6] wikipedia. (2015), Pixels, URL: <https://pt.wikipedia.org/wiki/Pixel>.
- [7] M. E. SALVIATI, “Manual do aplicativo Iramuteq”, *UNB planaltina. Recuperado em novembro*, vol. 20, p. 2018, 2017.

- [8] eecis. (2016), distribuição de frequência, URL:
https://www.eecis.udel.edu/~portnoi/classroom/prob_estatistica/2006_1/lecture_slides/aula04.pdf.
- [9] E. Reis, P. Melo, R. Andrade e T. Calapez, “Estatística aplicada”, *Lisboa: Edições Sílabo*, 1999.
- [10] H. J. Bortolossi, J. J. D. B. Queiroz e M. M. da Silva, “A Lei de Zipf e Outras Leis de Potência em Dados Empíricos”, 2012.
- [11] P. Marchand e P. Ratinaud, “L’analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l’élection présidentielle française (septembre-octobre 2011)”, *Actes des 11eme Journées internationales d’Analyse statistique des Données Textuelles. JADT*, vol. 2012, pp. 687–699, 2012.