MINERAÇÃO DE TEXTO

Conceitos & Aplicações

Prof. Walmes M. Zeviani
walmes@ufpr.br

Laboratório de Estatística e Geoinformação
Departamento de Estatística
Universidade Federal do Paraná

06 de Fevereiro de 2018

Humanos como sensores do mundo

Figura 1. Dispositivos e humanos como sensores do mundo. Fonte: o autor.

Dados não estruturados

Estrutura

- **Estruturados**: Numéricos, cronológicos, espaciais, categóricos.
- **Não estruturados**: Documentos, correspondências, jornais, blogs, imagens, vídeos, áudios.
- **Semi estruturados**: Arquivos de registro, notas fiscais, tweets.

Dados na forma de texto

- A maioria dos dados é não estruturado.
- Crescem em abundância e variedade.
- Contém informação de valor estratégico.
  - Para indivíduos, empresas e governos.
  - Implicações no comércio, indústria, saúde, etc.

Mineração de texto

Definição

Mineração ou análise de texto é o processo de extrair informação transformá-la de forma que possa ser aproveitada ou consumida.

Tecnologias

- Emerge do contato de disciplinas isoladas.
  - Estatística.
  - Linguistica computacional.
  - Ciência da computação.
- Diferentes níveis de maturidade.
  - Algumas metodologias já convencionais.
  - Várias em fase experimental.
  - Tanto abordagens simples quanto complexas.
  - Todas são centradas em **transformar texto em números**.
- Diferentes níveis de acesso para o praticante.

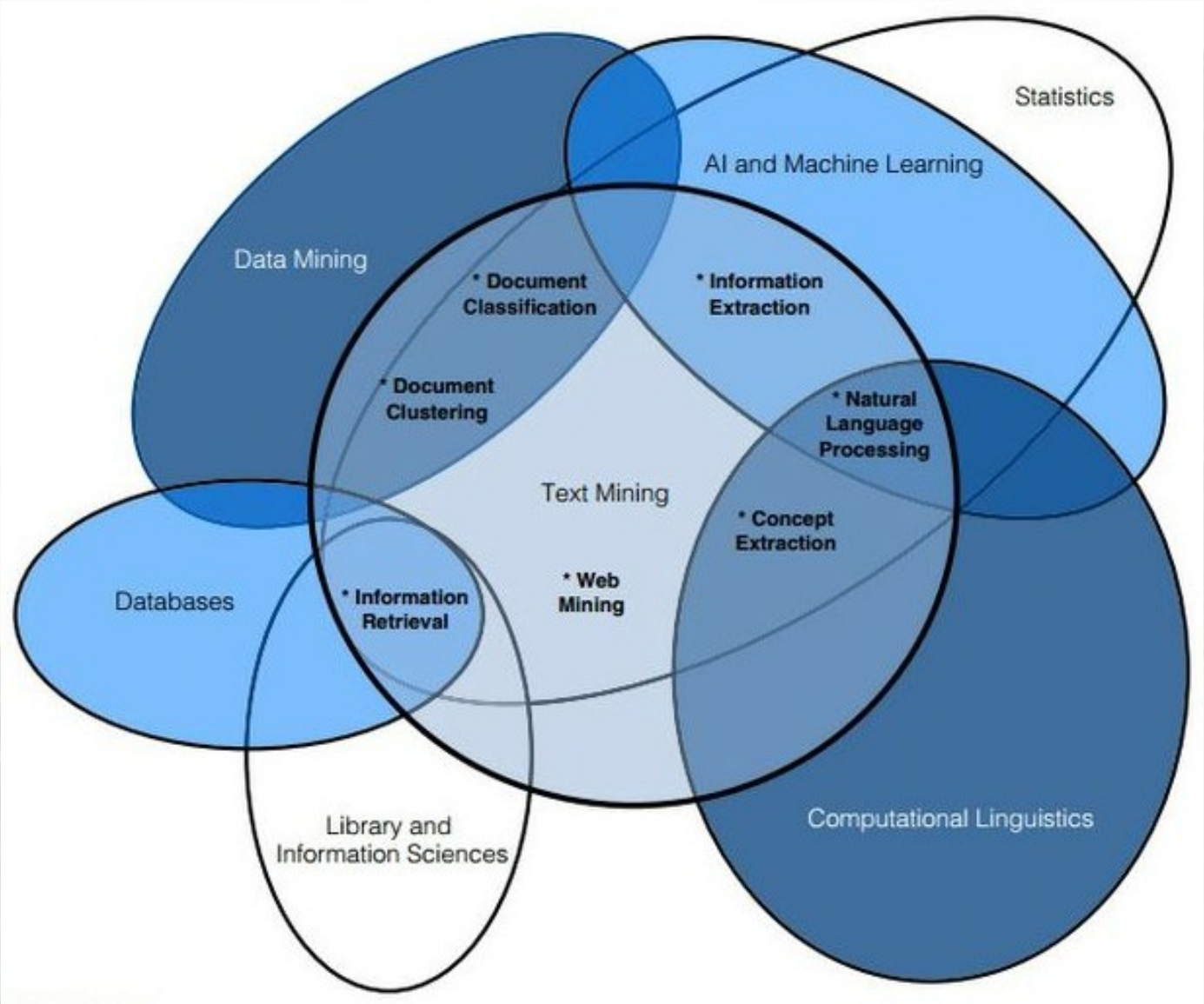


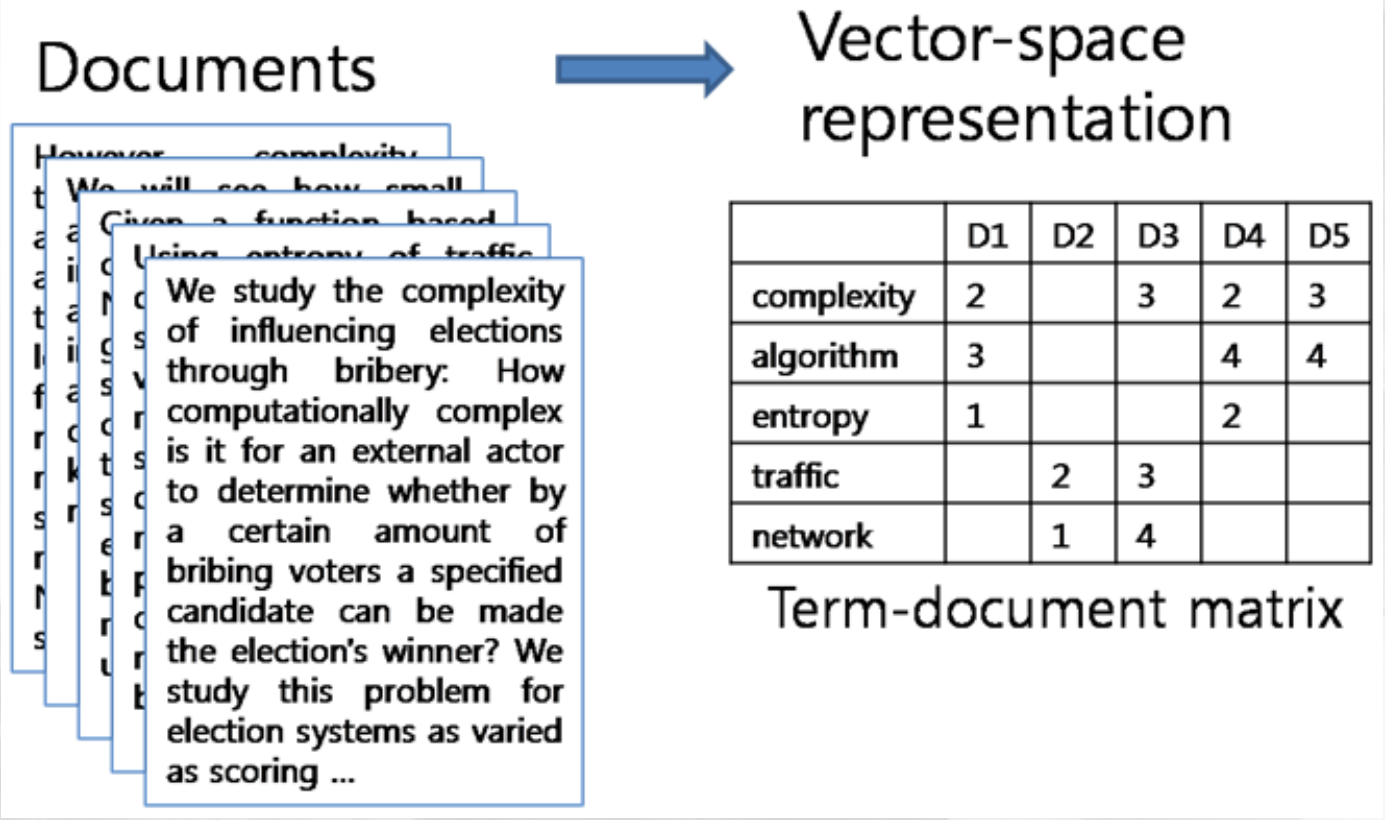
Figura 2. Disciplinas relacionadas com as 7 áreas da mineração de texto (MINER; ELDER; HILL, 2012).

As 2 abordagens principais

Bag of Words

- Baseada nos termos que o documento contém.
- Não depende da estrutura ou ordem.
- Quase sempre é idioma agnóstico.
- Etapas de pré-processamento.
- Engenharia de características é crucial.
- Abordagem bem simples em termos computacionais.

As 2 abordagens principais



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

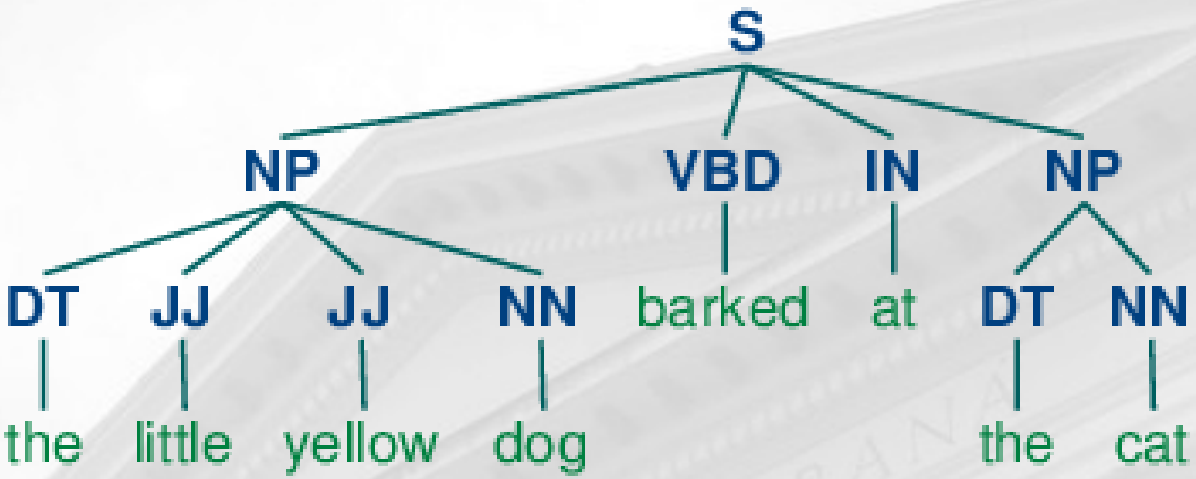
Figura 3. Matriz de documentos e termos gerada a partir da fragmentação do documento em palavras.

As 2 abordagens principais

Natural Language Processing

- Extrair significado do texto.
- Envolve conceitos linguísticos, estrutura gramatical.
- Quase sempre depende do idioma.
- Determina: quem, com quem, quando, onde, como e por quê.
- Bastante complexo.

As 2 abordagens principais



Four score and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate, we can not consecrate, we can not hallow this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us—that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion—that we here highly resolve that these dead shall not have died in vain—that this nation, under God, shall have a new birth of freedom—and that government of the people, by the people, for the people, shall not perish from the earth.

Figura 4. Análise baseada em processamento natural da linguagem.

Níveis de análise

- Lexical/morfológica
- Sintática
- Semântica
- No âmbito do discurso

Ferramentas e Softwares

Softwares comerciais

- STATISTICA Text Miner.
- SAS Text Miner
- IBM SPSS Text Analytics .
- Clarabridge.

Mais em
⟨https://en.wikipedia.org/wiki/List_of_text_mining_software⟩.

Softwares livres

- R: pacote `tm` e *task view* de Natural Language Processing.
- Python.
- Weka.

Ferramentas online

- ⟨http://news-explorer.mybluemix.net/⟩.
- ⟨https://www.paperrater.com/⟩.
- ⟨http://www.articlegeneratorpro.com/⟩.
- ⟨http://articlegenerator.org⟩.
- ⟨http://parts-of-speech.info/⟩.
- ⟨https://iwl.me⟩.
- ⟨http://textalyser.net/⟩.

Aplicação

Precificação de Coberturas em CWB

- Coberturas a venda em Curitiba.
- Dados extraídos do ⟨https://www.vivareal.com.br/⟩.
- ⟨http://leg.ufpr.br/~walmes/data/coberturas-venda-cwb-26Jan2018.txt⟩.
- Web scraping com R, $\sim$ 5 minutos.
- Informações.
  - Preço.
  - Área.
  - Quartos, banheiros, suítes, vagas.
  - Endereço $\rightarrow$ coordenadas lat/lon.
  - **Texto da descrição**.
- **Objetivo**: produzir um modelo de precificação considerando as variáveis métricas e textuais dos anúncios.

```r
dim(da)

## [1] 2634   16

names(da)

##  [1] "ID"       "title"    "address"  "price"    "condo"
##  [6] "area"     "rooms"    "suites"   "bathroom" "garages"
## [11] "url"      "descr"    "lon"      "lat"      "lprice"
## [16] "larea"
```

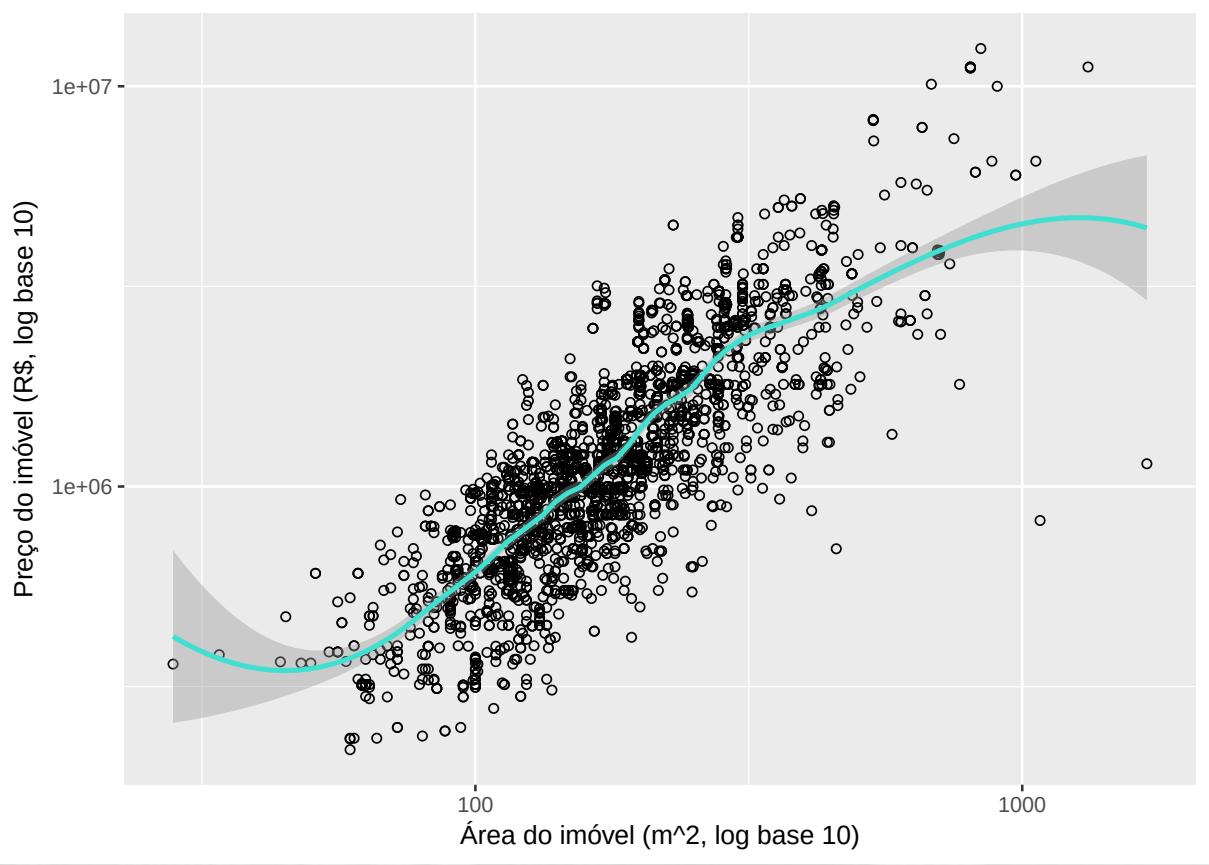


Figura 5. Diagrama de dispersão do preço em função da área das coberturas em escala log base 10.

Etapas

- Criação do corpus.
- Pré-processamento.
  - Caixa baixa.
  - Remoção da pontuação.
  - Remoção dos números.
  - Remoção dos espaços extras.
  - Remoção das *stopwords*.
  - Redução de inflexão (stemming).
  - Remoção dos acentos (ASCII).
- Criação da matriz de documentos e termos.
- Redução de esparsidade.
- Uso como matriz de variáveis *dummy* no modelo de regressão.
- Ajuste dos modelos e seleção de covariáveis.

Figura 6. Núvem de palavras gerada a partir das descrições das coberturas.

```
# Matriz de documentos e termos original.
dtm

## <<DocumentTermMatrix (documents: 2634, terms: 7843)>>
## Non-/sparse entries: 167312/20491150
## Sparsity           : 99%
## Maximal term length: 97
## Weighting          : binary (bin)

# Matriz de documentos e termos após remoção de esparsidade.
rst

## <<DocumentTermMatrix (documents: 2634, terms: 29)>>
## Non-/sparse entries: 36754/39632
## Sparsity           : 52%
## Maximal term length: 13
## Weighting          : binary (bin)
```

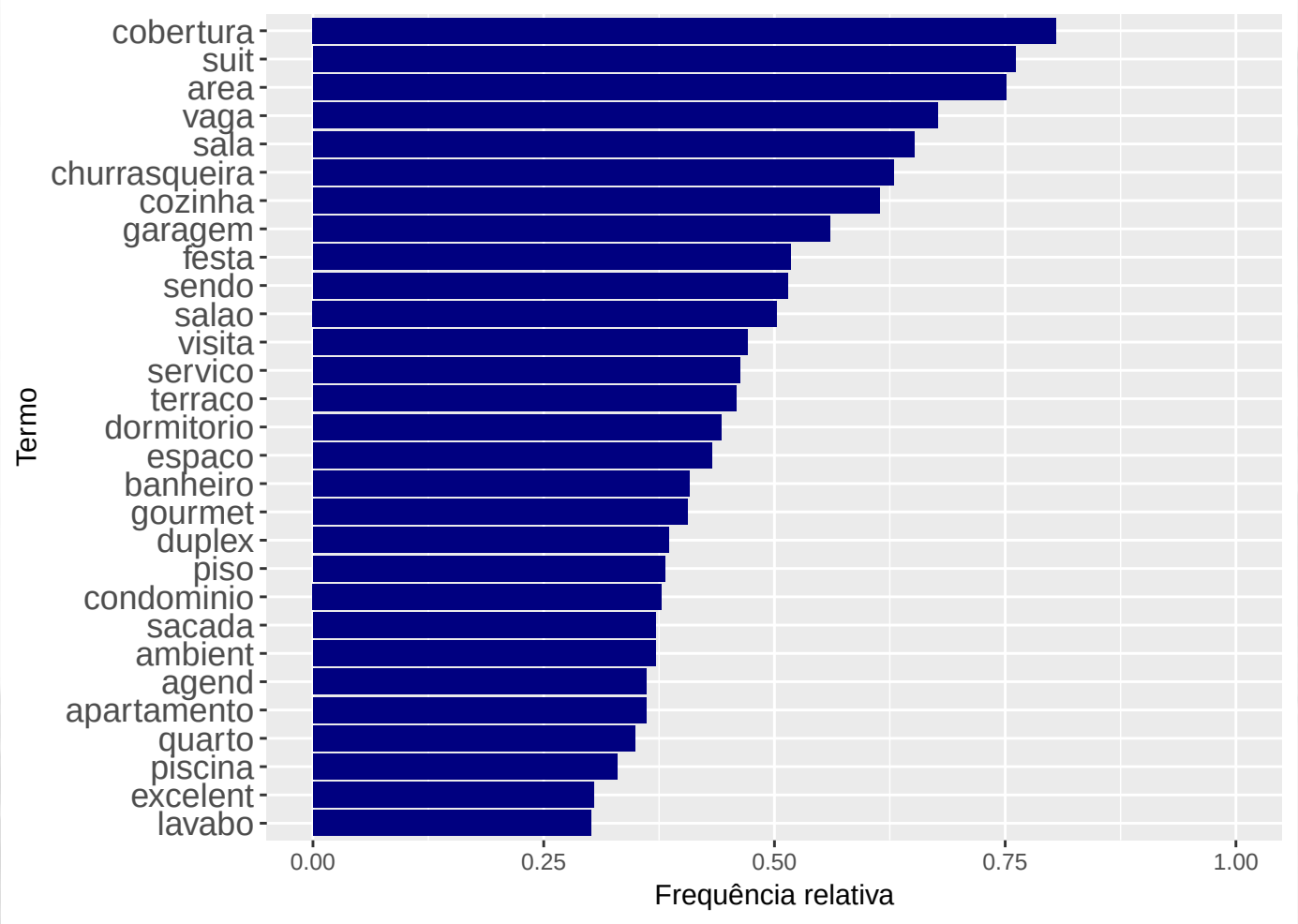


Figura 7. Termos mantidos após redução de esparcidade aplicada na matriz de documentos e termos.

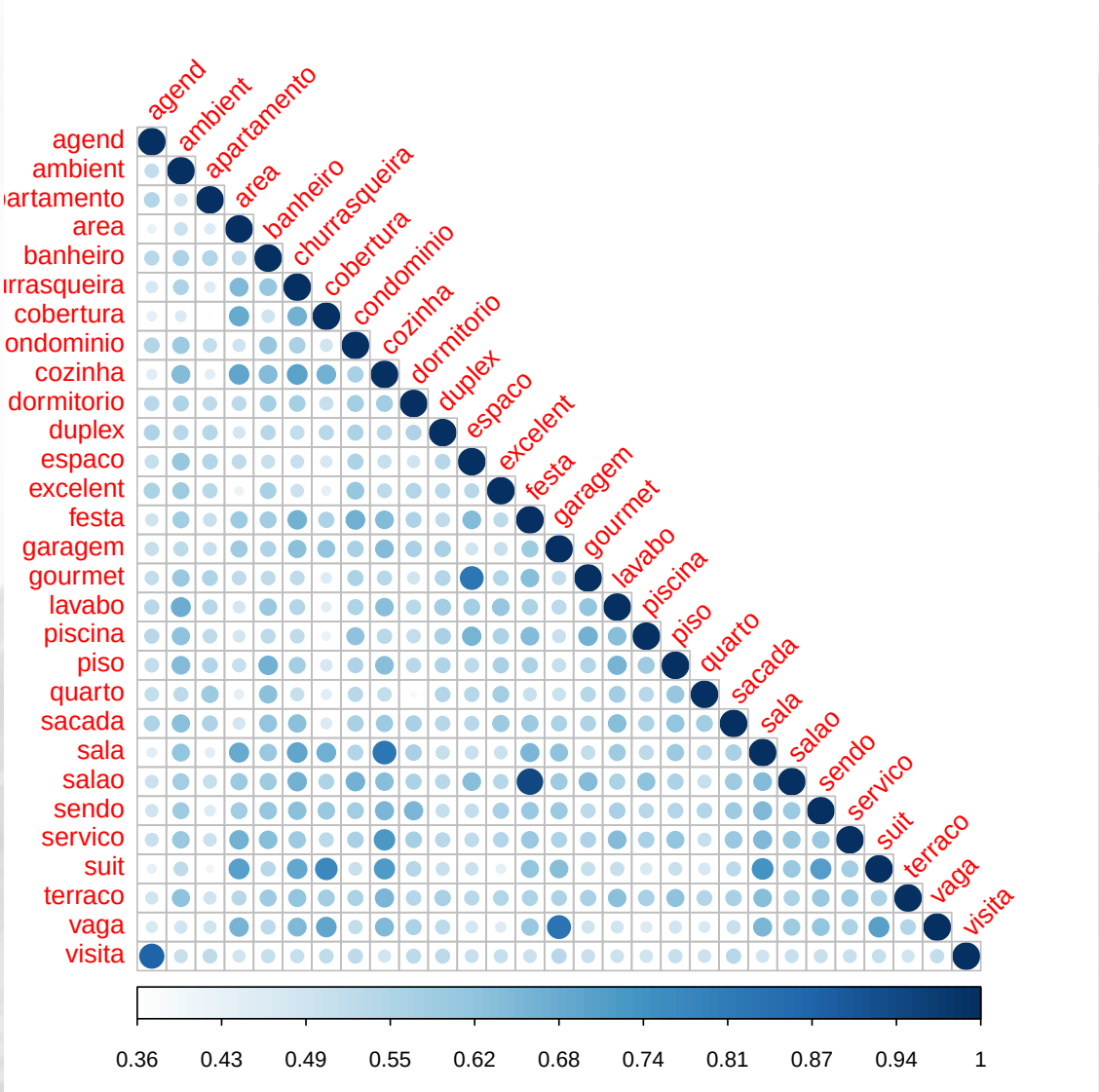


Figura 8. Diagrama de porcentagem de co-corrência entre os termos.

Modelos

- `m0`: modelo apenas com as variáveis métricas.
- `m1`: `m0` + 29 variáveis textuais.
- `m2`: aplicação de stepwise em `m1`.

```r
tb[, c("Model", "param", "R2", "PRESS", "improv")]

##   Model param       R2    PRESS   improv
## 1    m1    35 80.72794 30.27365 7.640916
## 2    m2    13 80.32628 30.58476 6.691780
## 3    m0     6 78.84192 32.77820 0.000000
```

```r
names(coef(m2))

##  [1] "(Intercept)"   "LAREA"         "SUITES"
##  [4] "BATHROOM"      "GARAGES"       "banheiro"
##  [7] "churrasqueira" "cobertura"     "condominio"
## [10] "cozinha"       "gourmet"       "lavabo"
## [13] "sendo"
```

- Todo imóvel tem texto de descrição.
- É coletado junto com as variáveis métricas.
- Por que não aproveitá-lo?
- Com uma análise legeira teve-se melhora $\geq 6\%$.
- Maior ganho com a engenharia de características.
- Pode-se usar para todo tipo de imóveis.
- Pode-se empregar outros modelos.

Complementos, dúvidas, reclamações?

Obrigado!

Referências bibliográficas

MINER, G. et al. **Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications**. Academic Press, 2012. ISBN 9780123869791. Disponível em: ⟨https://books.google.com.br/books?id=-B6amxqygTMC⟩.