

Agrupamento e Similaridade de Documentos

Prof. Walmes Zeviani
walmes@ufpr.br

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

Justificativa e objetivos

- ▶ Agrupamento e similaridade de documentos são técnicas bastante utilizadas em mineração de texto.
- ▶ Os resultados dependem da escolha de medidas de distância e formas de agrupamento.
- ▶ Exemplificar em como realizar uma análise de agrupamento de documentos.

Medidas de distância/similaridade

Agrupamento e similaridade

- ▶ Agrupamento é uma técnica útil para organizar grandes coleções de documentos em pequenos grupos de documentos similares.
- ▶ Serve de base para os algoritmos de motores de busca.
- ▶ Também é utilizado por algoritmos de recomendação.
- ▶ Um agrupamento acurado requer precisa definição de proximidade entre pares de objetos.
- ▶ Existem várias medidas de distância/similaridade que podem ser usadas em coleções de documentos.

Distância entre documentos

- ▶ Os documentos são sacos de palavras.
- ▶ Documentos são pontos no espaço vetor.
 - ▶ Termos são as características.
 - ▶ Documentos são os objetos.
 - ▶ Scores são os valores das características nos objetos.
- ▶ Objetos similares têm valores próximos para as mesmas características.
- ▶ Distância e similaridade são conceitos ligados.
- ▶ A distância entre objetos depende:
 - ▶ de como os objetos são representados e
 - ▶ da medida de distância empregada.

Requisitos de uma medida de distância

- ▶ A distância é não negativa: $D(\mathbf{x}, \mathbf{y}) \geq 0$.
- ▶ A distância só é 0 se os objetos forem iguais: $D(\mathbf{x}, \mathbf{y}) = 0$ somente se $\mathbf{x} = \mathbf{y}$.
- ▶ A distância deve ser simétrica: $D(\mathbf{x}, \mathbf{y}) = D(\mathbf{y}, \mathbf{x})$.
- ▶ A medida deve satisfazer a “desigualdade triangular”:
 $D(\mathbf{x}, \mathbf{z}) \leq D(\mathbf{x}, \mathbf{y}) + D(\mathbf{y}, \mathbf{z})$.

Medidas de distância

- ▶ Distância euclidiana.
 - ▶ Distância entre pontos no espaço.
 - ▶ Calculada por

$$D_{euc}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^v (x_i - y_i)^2} = \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})}.$$

- ▶ Distância do cosseno.
 - ▶ Baseada no ângulo entre vetores no espaço.
 - ▶ $\cos(\theta) \in [0, 1]$
 - ▶ Ângulo fechado = cosseno grande = distância pequena.
 - ▶ $\cos(\theta)$: similaridade.
 - ▶ $1 - \cos(\theta)$: distância.
 - ▶ Calculada por

$$D_{cos}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}} = 1 - \frac{\mathbf{x}'\mathbf{y}}{\sqrt{(\mathbf{x}'\mathbf{x})(\mathbf{y}'\mathbf{y})}}.$$

- ▶ Não muda com a multiplicação dos vetores por constantes.

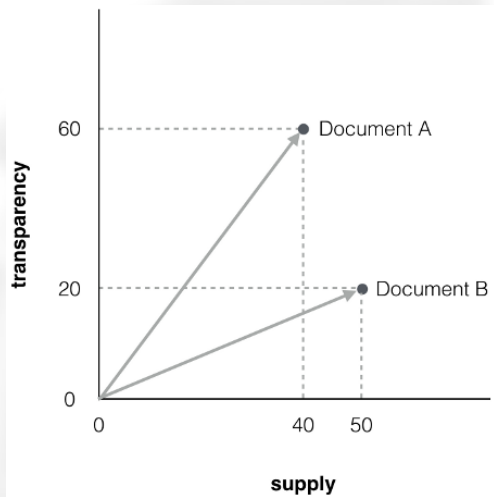


Figura 1. Figura 1: Dois documentos no espaço 2-dimensional.

Cálculo das distâncias

- ▶ A função `stats::dist()` é usada para cálculo de distância e possui diversas métricas.
- ▶ A função `proxy::dist()` adiciona mais algumas métricas, como a `cosine()`.
- ▶ Elas recebem uma matriz com objetos nas linhas.
- ▶ Retornam a matriz $n \times n$ de distância entre os pares de objetos objetos.

```
# Matriz de documentos e termos (Doc = linha, Term = coluna).
```

```
dtm <- rbind(doc1 = c(1, 1, 1.5),  
            doc2 = c(1, 1, 1),  
            doc3 = 2 * c(1, 1, 1.5))
```

```
dtm
```

```
##      [,1] [,2] [,3]  
## doc1   1   1  1.5  
## doc2   1   1  1.0  
## doc3   2   2  3.0
```

```
# Distâncias euclidianas.
```

```
stats::dist(dtm, method = "euclidian")
```

```
##          doc1      doc2  
## doc2 0.500000  
## doc3 2.061553 2.449490
```

```
# Fazendo para o primeiro par de documentos.
```

```
d <- dtm[1, ] - dtm[2, ]
```

```
sqrt(t(d) %*% d)
```

```
##      [,1]  
## [1,] 0.5
```

```
library(proxy)
```

```
# Distância coseno.
```

```
proxy::dist(dtm, method = "cosine")
```

```
##           doc1           doc2
## doc2 0.01980394
## doc3 0.00000000 0.01980394
```

```
# Fazendo para o primeiro par de documentos.
```

```
1 - sum(dtm[1, ] * dtm[2, ])/
  sqrt(sum(dtm[1, ]^2) * sum(dtm[2, ]^2))
```

```
## [1] 0.01980394
```

```
1 - (t(dtm[1, ]) %*% dtm[2, ]) /
  sqrt((t(dtm[1, ]) %*% dtm[1, ]) * t(dtm[2, ]) %*% dtm[2, ])
```

```
##           [,1]
## [1,] 0.01980394
```

Outras medidas

- ▶ Coeficiente de Jarccard (similaridade).
- ▶ Correlação de Pearson (similaridade).
- ▶ Divergência média de Kullback-Leiber (distância).