Distance and Similarity

Andre Salvaro Furtado

Department of Informatics and Statistics (INE)
Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Santa Catarina, Brazil

September 15, 2015

UNIVERSIDADE FEDERAL
DE SANTA CATARINA

Topics

Topics

Introduction - Distance and Similarity

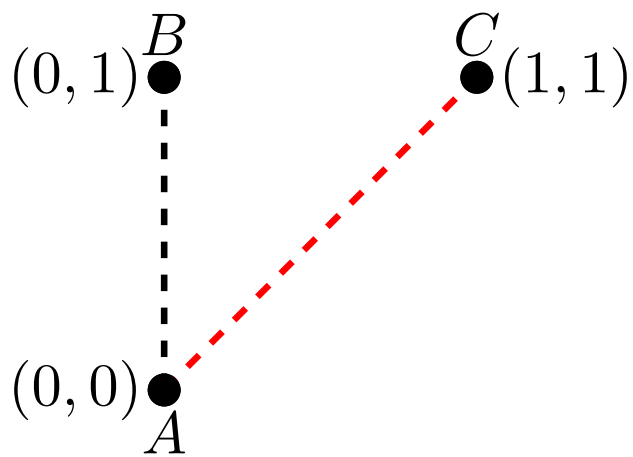
- Distance / Dissimilarity
  - Quantify the difference of two objects
  - The value is usually in the interval $[0, \infty]$
  - Lower values mean that the objects are more similar
- Similarity
  - Quantify the alikeness of two objects
  - The value is usually in the interval $[0, 1]$
  - Lower values mean that the objects are less similar

Motivation

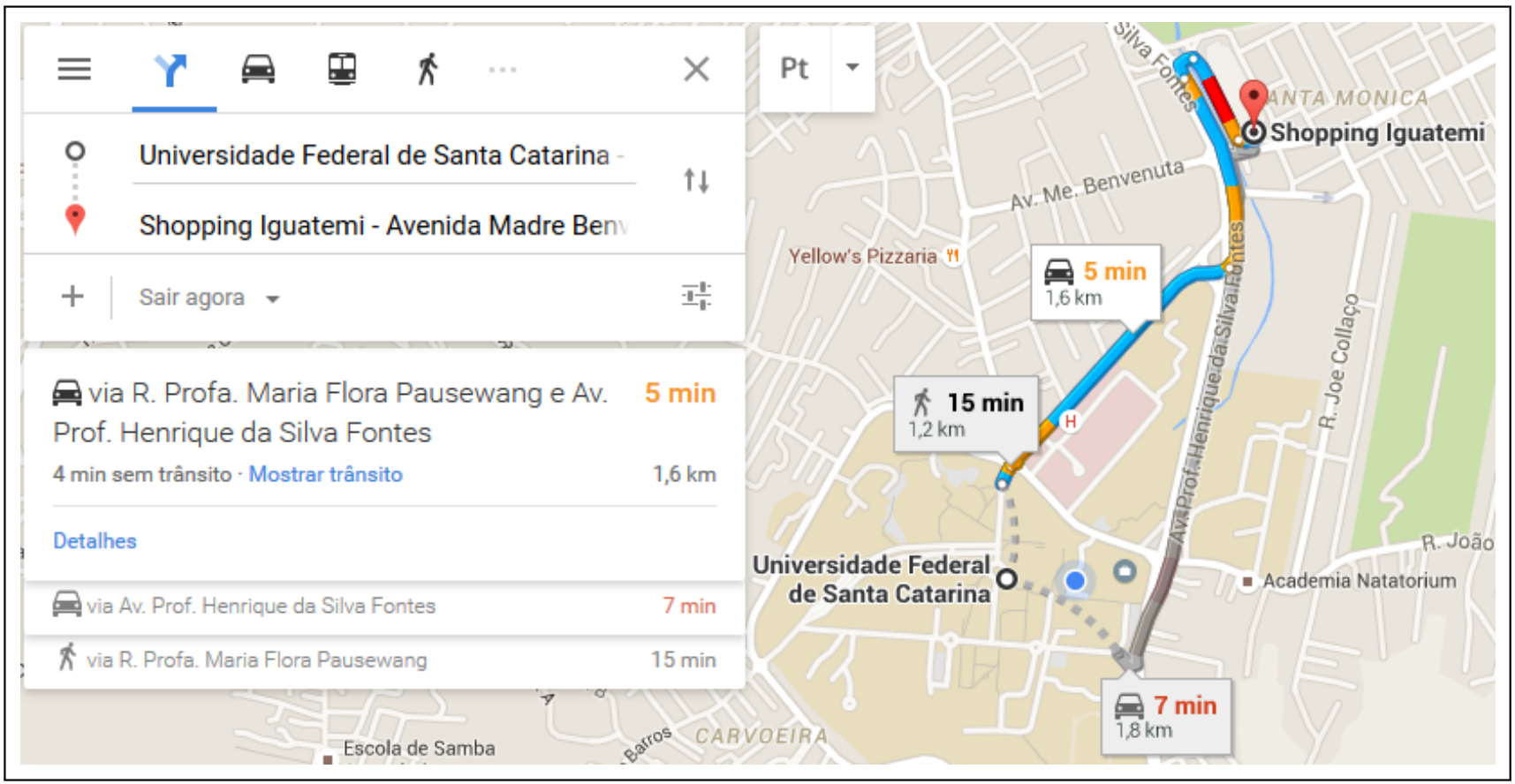
- Distance and Similarity measures are useful for several applications:
  - Calculate the distance between two points in a plane
  - Calculate the distance between two locations
  - Find the restaurants that are near a location
  - Search systems (e.g., a search in Google)
  - Given an image return the most similar images (e.g., Google Images)
  - Identify similar customers in a company database
  - ...

Example: Distance between two points in a plane

- Distance between two points in a plane
- Euclidean Distance:
  - $d(A, B) = 1$
  - $d(A, C) = \sqrt{2}$

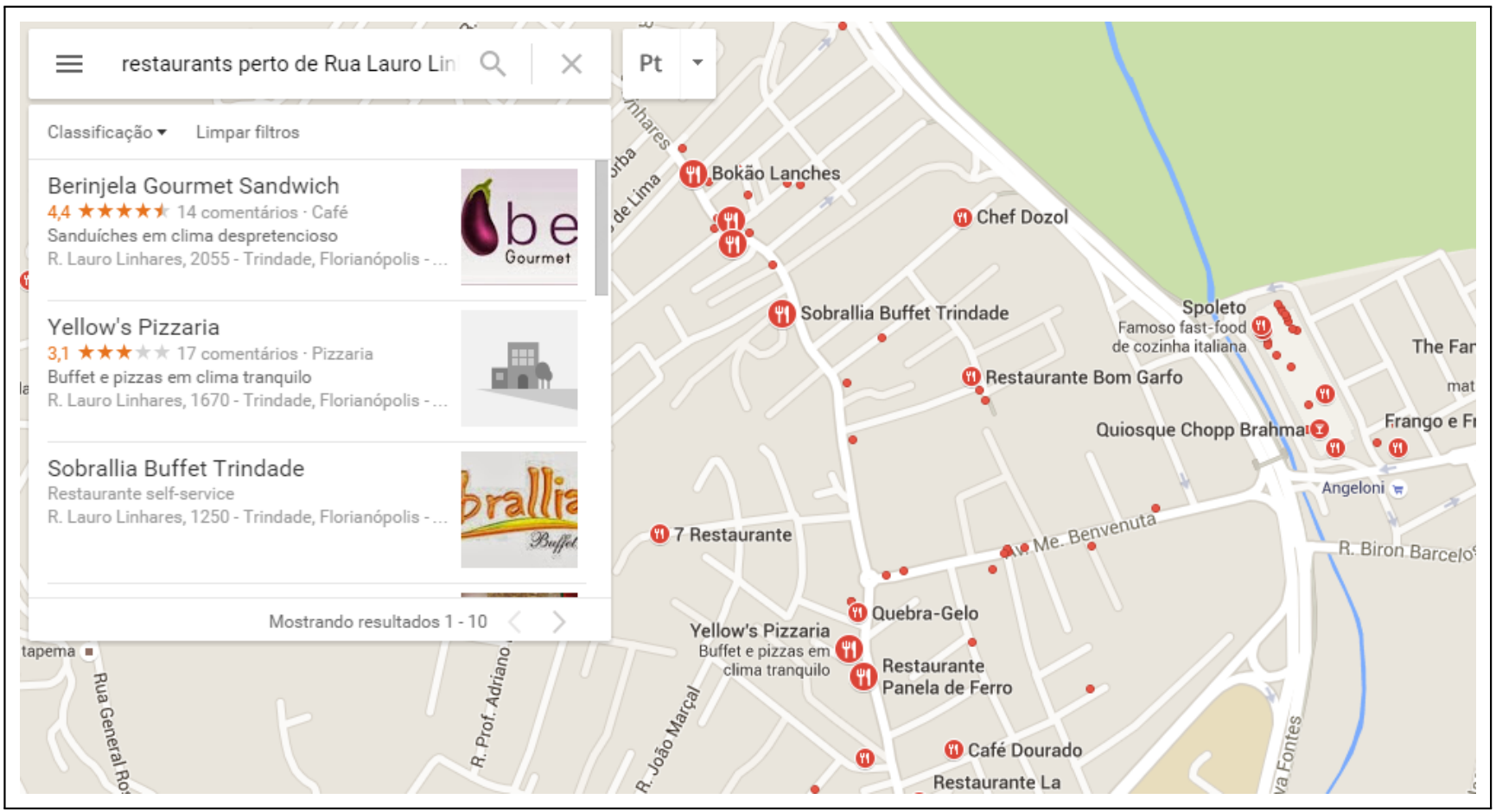


Example: Distance between two locations



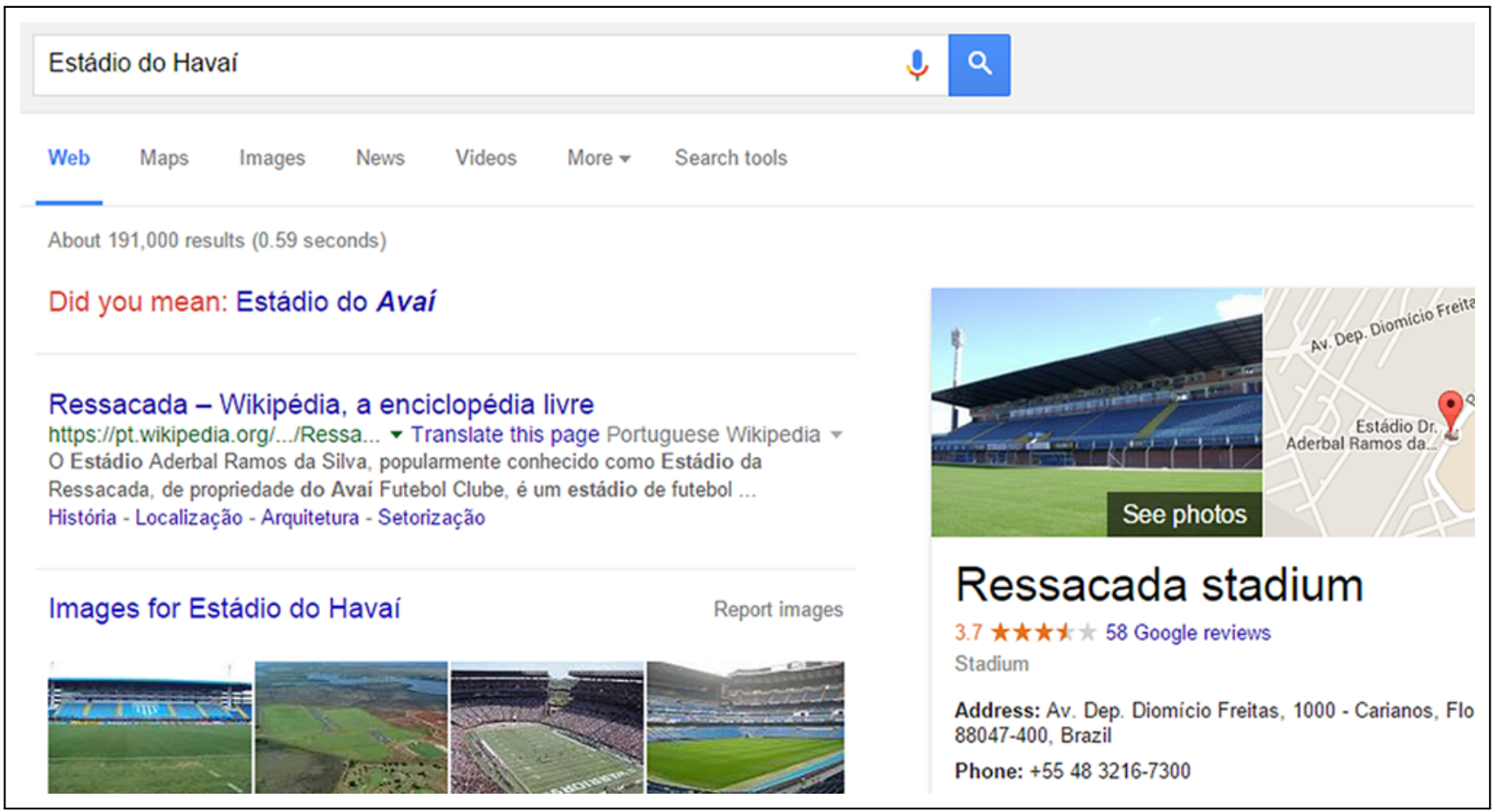
Example: Restaurants near a Location

- Similarity between sentences in keyword search

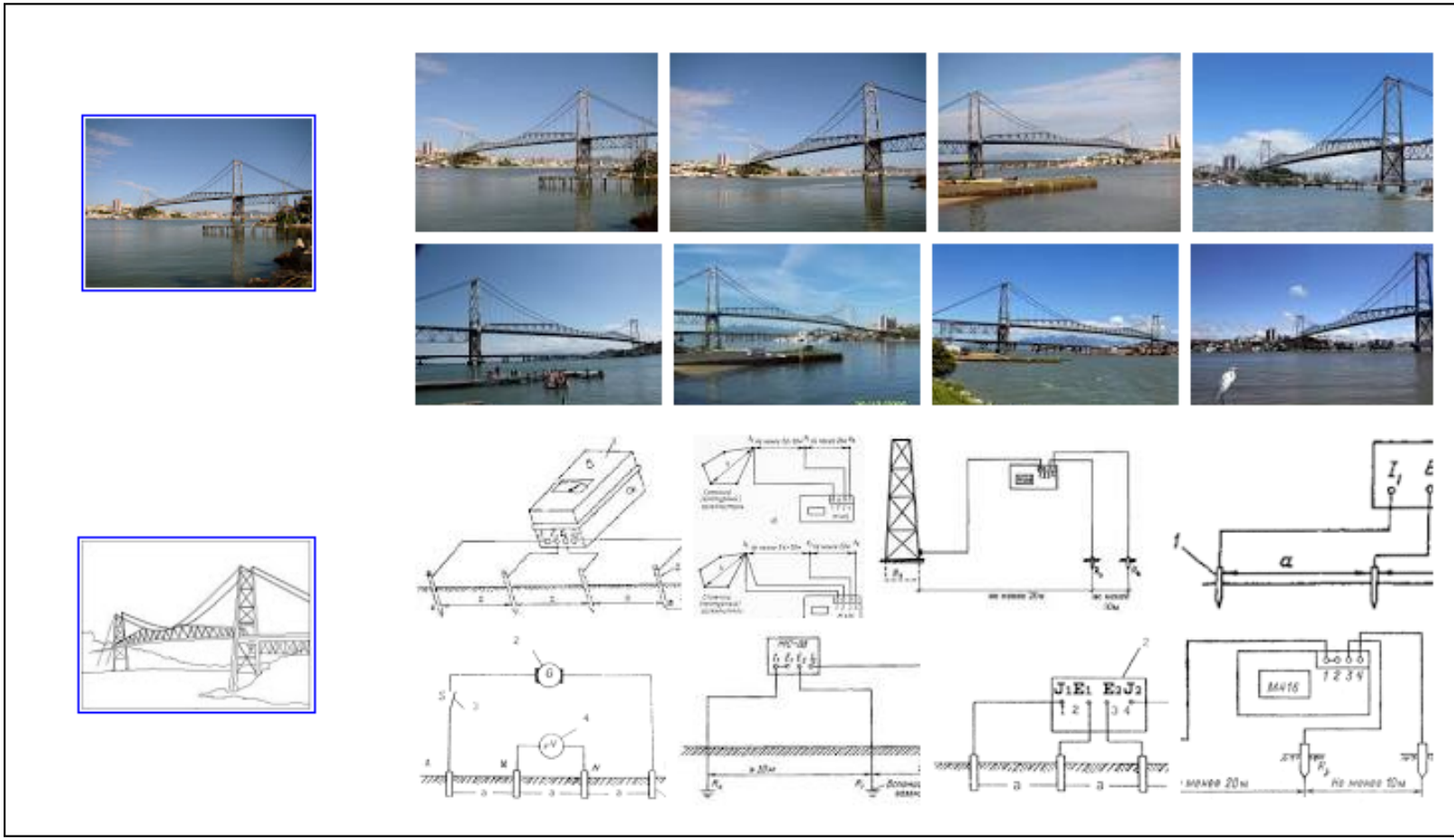


Example: Textual Similarity

- Similarity between sentences in keyword search



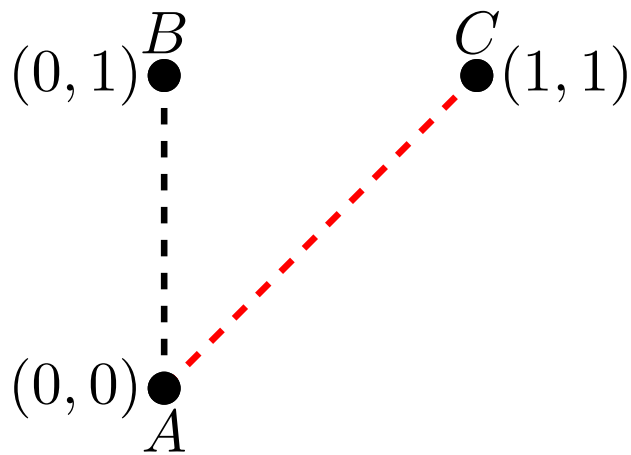
Example: Image Similarity



Topics

Euclidean Distance

- Length of the straight line between two points

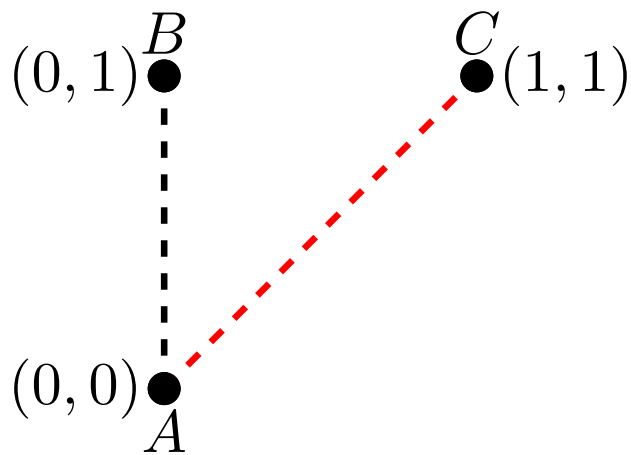


- Given two points p and q:

$$d(p,q) = \sqrt{(p.x - q.x)^2 + (p.y - q.y)^2} \quad (1)$$

Example: Euclidean Distance (A,B)

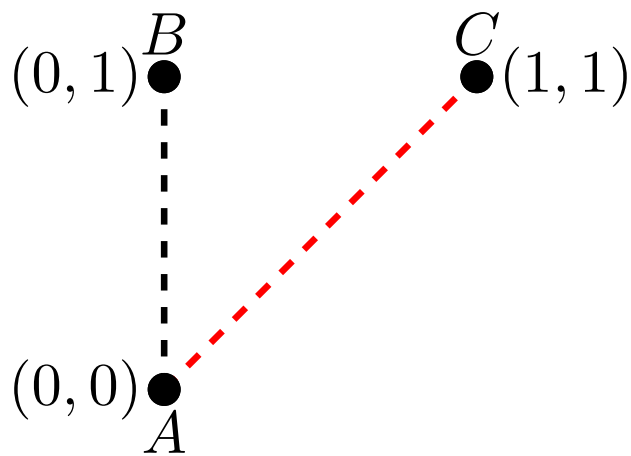
$$d(p,q) = \sqrt{(p.x - q.x)^2 + (p.y - q.y)^2}$$



$$d(A,B) = \sqrt{(A.x - B.x)^2 + (A.y - B.y)^2} \qquad \text{(ED1)}$$

Example: Euclidean Distance (A,B)

$$d(p,q) = \sqrt{(p.x - q.x)^2 + (p.y - q.y)^2}$$

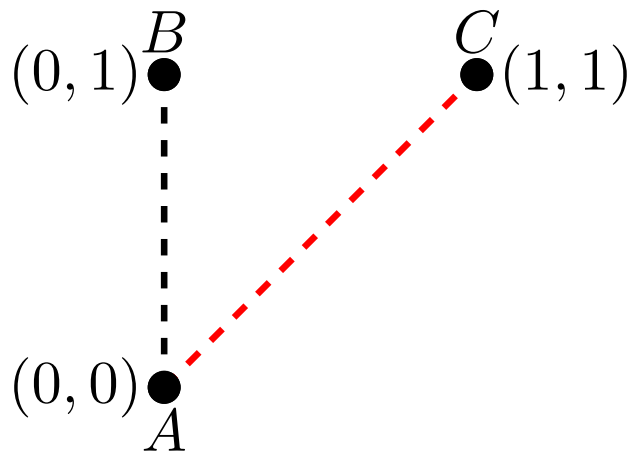


$$d(A,B) = \sqrt{(A.x - B.x)^2 + (A.y - B.y)^2} \quad \text{(ED1)}$$

$$d(A,B) = \sqrt{(0-0)^2 + (0-1)^2} \quad \text{(ED2)}$$

Example: Euclidean Distance (A,B)

$$d(p,q) = \sqrt{(p.x - q.x)^2 + (p.y - q.y)^2}$$



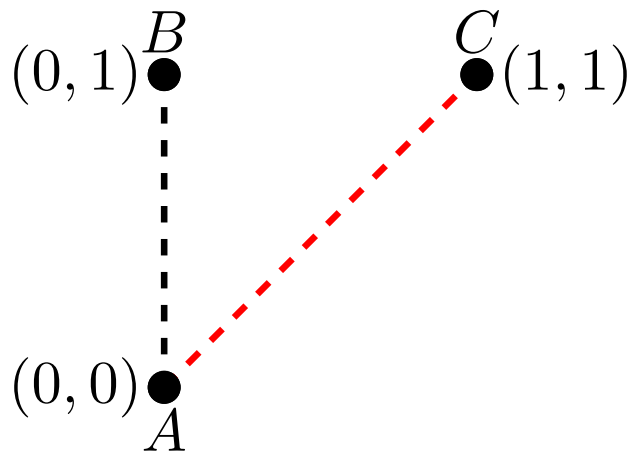
$$d(A,B) = \sqrt{(A.x - B.x)^2 + (A.y - B.y)^2} \quad \text{(ED1)}$$

$$d(A,B) = \sqrt{(0-0)^2 + (0-1)^2} \quad \text{(ED2)}$$

$$d(A,B) = \sqrt{(0)^2 + (1)^2} = 1 \quad \text{(ED3)}$$

Example: Euclidean Distance (A,C)

$$d(p,q) = \sqrt{(p.x - q.x)^2 + (p.y - q.y)^2}$$



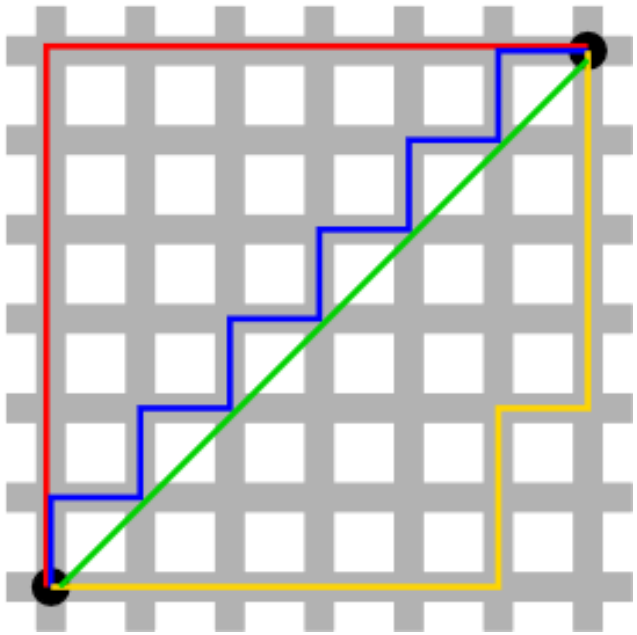
$$d(A,C) = \sqrt{(A.x - C.x)^2 + (A.y - C.y)^2} \quad \text{(ED4)}$$

$$d(A,C) = \sqrt{(0-1)^2 + (0-1)^2} \quad \text{(ED5)}$$

$$d(A,C) = \sqrt{(-1)^2 + (-1)^2} = \sqrt{2} \quad \text{(ED6)}$$

Manhattan Distance

- Absolute distance of the coordinates
- Also known as Taxicab Distance, City Block Distance...

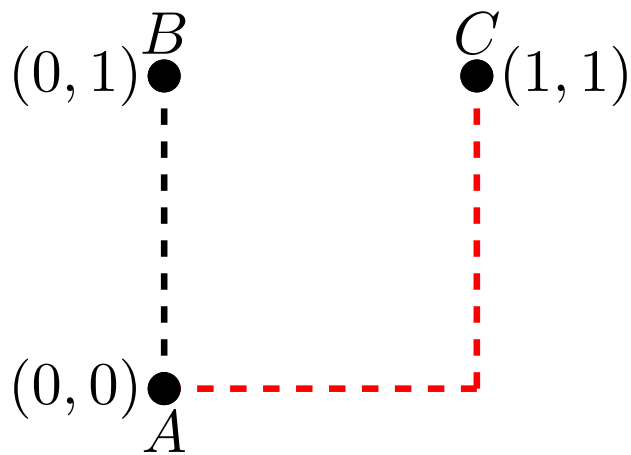


- Given two points p and q:

$$d(p, q) = |p.x - q.x| + |p.y - q.y| \tag{2}$$

Example: Manhattan Distance (A,B)

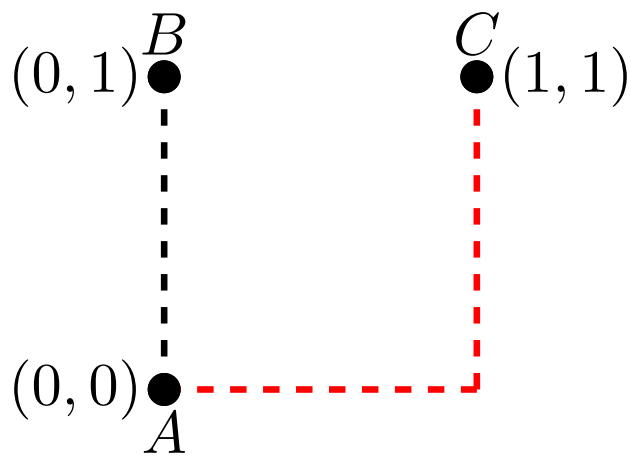
$$d(p, q) = |p.x - q.x| + |p.y - q.y|$$



$$d(A, B) = |A.x - B.x| + |A.y - B.y| \qquad \text{(MD1)}$$

Example: Manhattan Distance (A,B)

$$d(p, q) = |p.x - q.x| + |p.y - q.y|$$

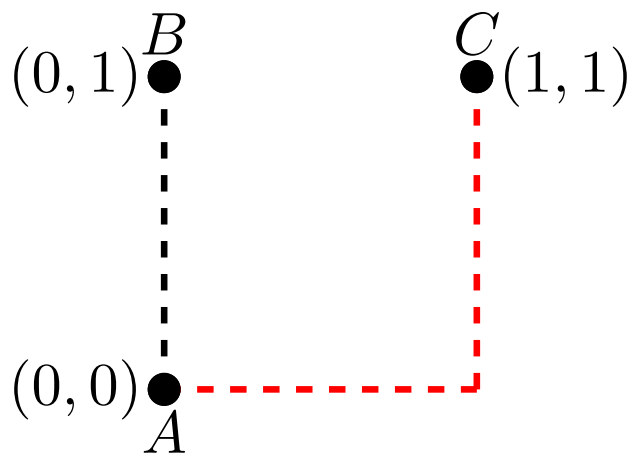


$$d(A, B) = |A.x - B.x| + |A.y - B.y| \quad \text{(MD1)}$$

$$d(A, B) = |0 - 0| + |0 - 1| \quad \text{(MD2)}$$

Example: Manhattan Distance (A,B)

$$d(p, q) = |p.x - q.x| + |p.y - q.y|$$



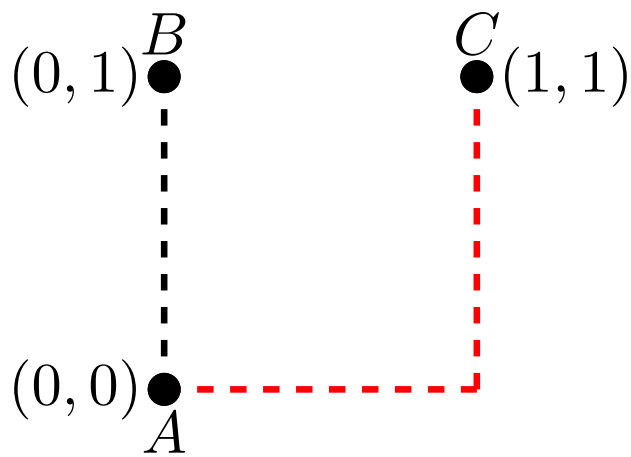
$$d(A, B) = |A.x - B.x| + |A.y - B.y| \quad \text{(MD1)}$$

$$d(A, B) = |0 - 0| + |0 - 1| \quad \text{(MD2)}$$

$$d(A, B) = 0 + 1 = 1 \quad \text{(MD3)}$$

Example: Manhattan Distance (A,C)

$$d(p, q) = |p.x - q.x| + |p.y - q.y|$$



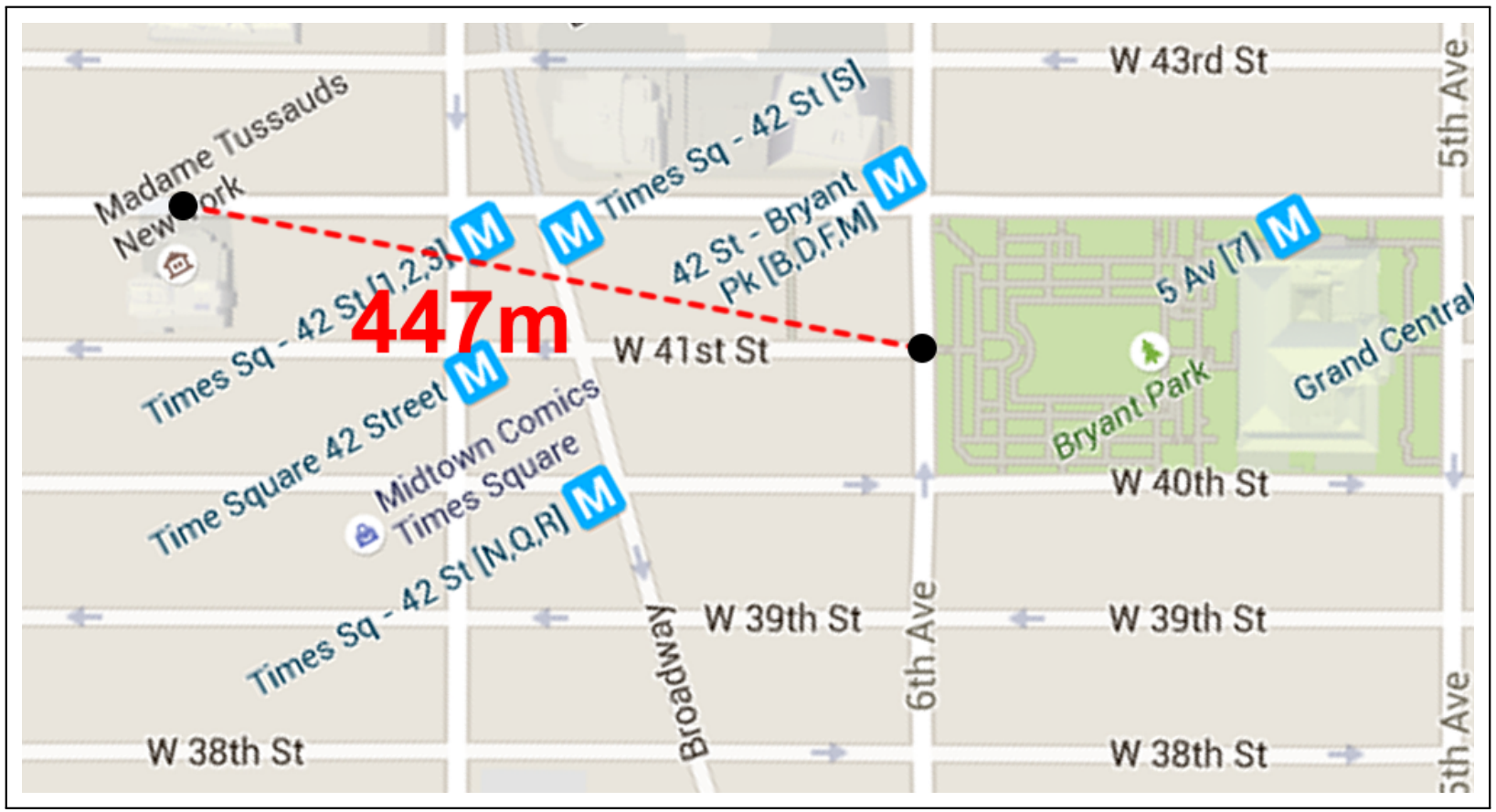
$$d(A, C) = |A.x - C.x| + |A.y - C.y| \quad \text{(MD4)}$$

$$d(A, C) = |0 - 1| + |0 - 1| \quad \text{(MD5)}$$

$$d(A, C) = 1 + 1 = 2 \quad \text{(MD6)}$$

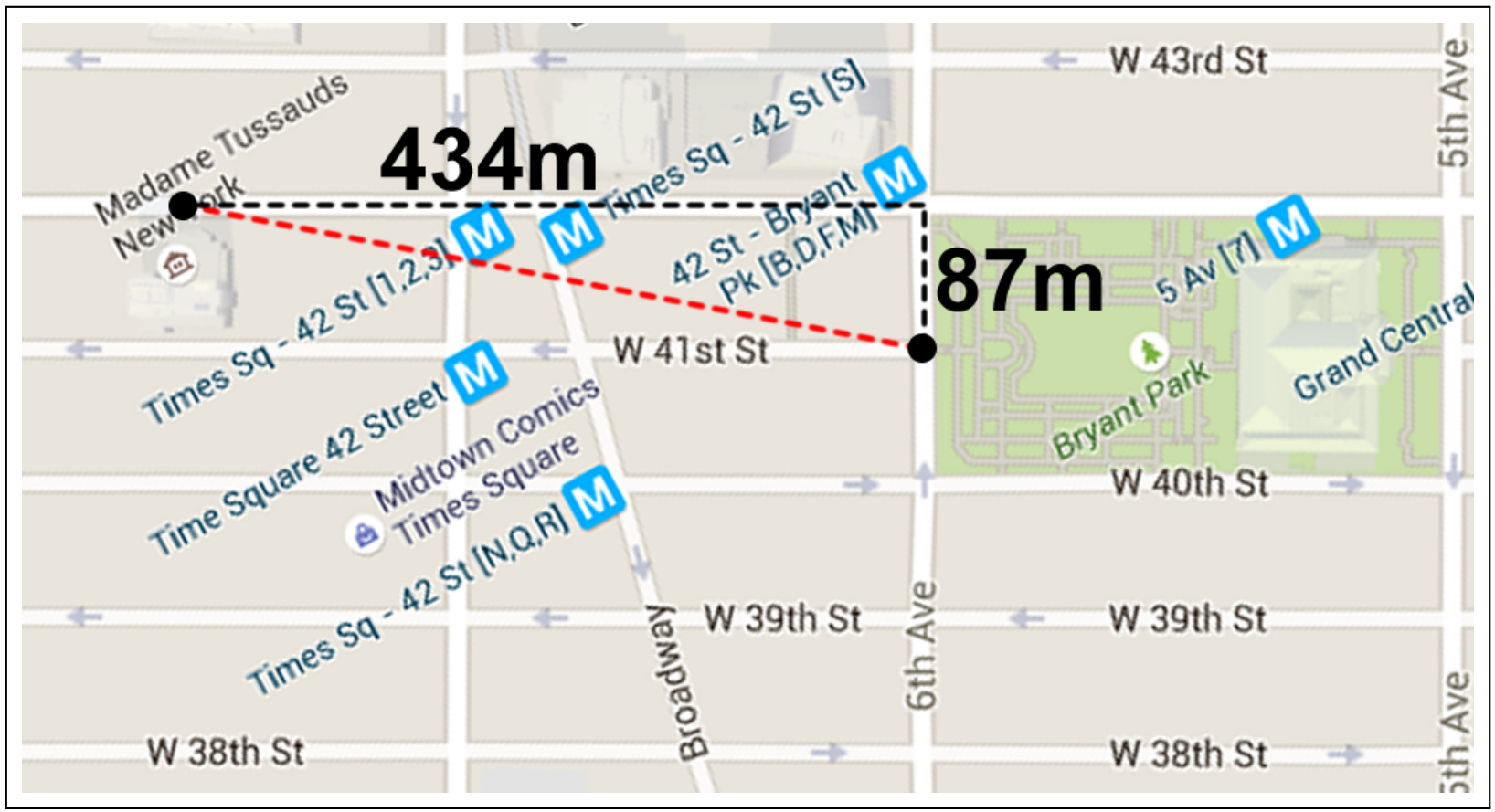
Example: Euclidean Distance in Manhattan

- Euclidean: $d(BryantPark, MadameTussaud) = 447m$



Example: Manhattan Distance in Manhattan

- Euclidean: $d(BryantPark, MadameTussaud) = 447m$
- Manhattan:
  $d(BryantPark, MadameTussaud) = 434m + 87m = 521m$



Chebyshev Distance

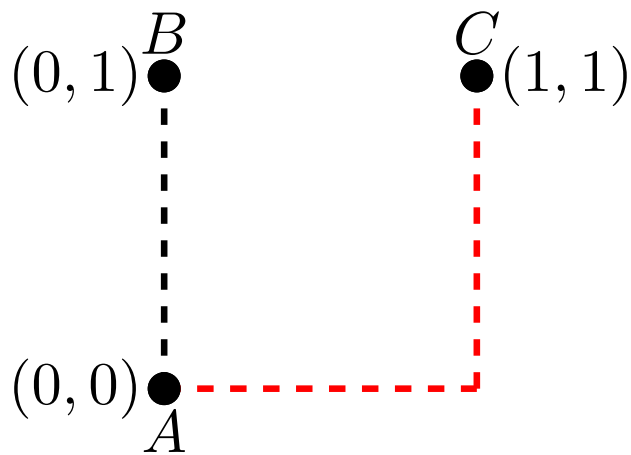
- Maximum difference in any coordinate
- Also know as Chessboard Distance

- Given two points p and q:

$$d(p,q) = max(|p.x - q.x|, |p.y - q.y|) \quad (3)$$

Example: Chebyshev Distance (A,C)

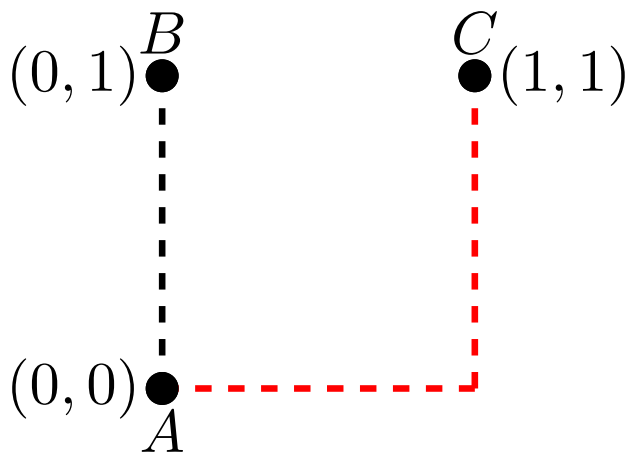
$$d(p,q) = max(|p.x - q.x|, |p.y - q.y|)$$



$$d(A,C) = max(|A.x - C.x|, |A.y - C.y|) \qquad \text{(CD1)}$$

Example: Chebyshev Distance (A,C)

$$d(p,q) = max(|p.x - q.x|, |p.y - q.y|)$$

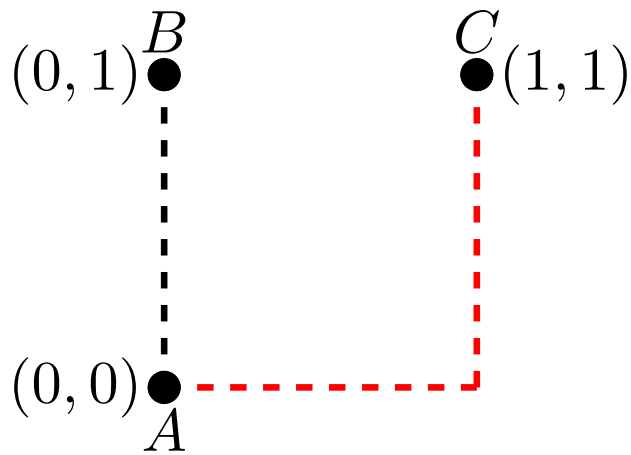


$$d(A,C) = max(|A.x - C.x|, |A.y - C.y|) \quad \text{(CD1)}$$

$$d(A,C) = max(|0 - 1|, |0 - 1|) \quad \text{(CD2)}$$

Example: Chebyshev Distance (A,C)

$$d(p,q) = max(|p.x - q.x|, |p.y - q.y|)$$



$$d(A,C) = max(|A.x - C.x|, |A.y - C.y|) \quad \text{(CD1)}$$

$$d(A,C) = max(|0 - 1|, |0 - 1|) \quad \text{(CD2)}$$

$$d(A,C) = max(1,1) = 1 \quad \text{(CD3)}$$

Minkowski Distance

- Generalization for Euclidean/Manhattan/Chebyshev Distances

$$d(p,q) = \left(\sum_{i=1}^{n} |p_i - q_i|^e \right)^{1/e} \tag{4}$$

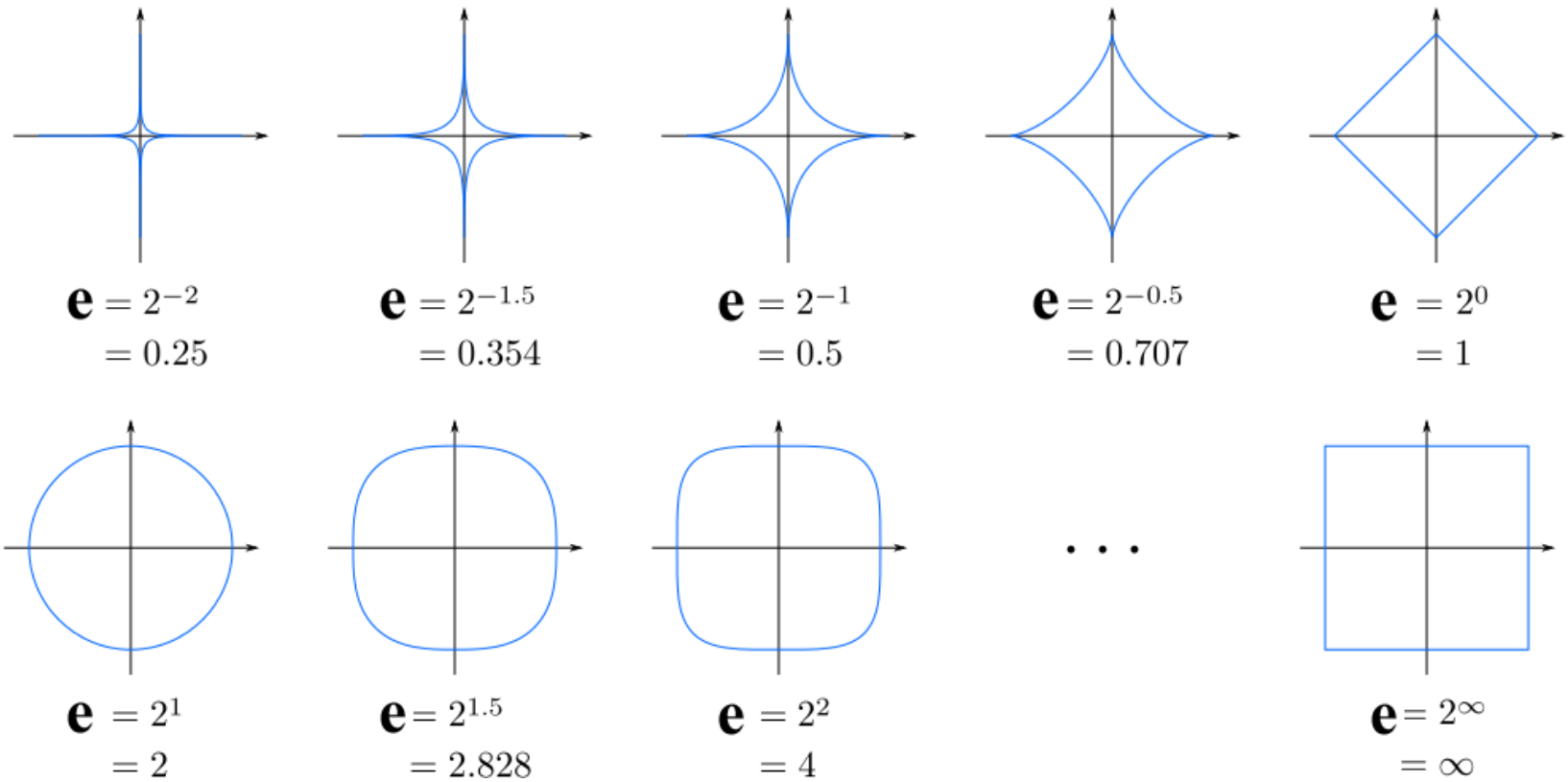
  - Manhattan (e=1)

$$d(p,q) = \left(\sum_{i=1}^{n} |p_i - q_i|^1 \right)^{1/1} = \sum_{i=1}^{n} |p_i - q_i| \tag{5}$$

  - Euclidean (e=2)

$$d(p,q) = \left(\sum_{i=1}^{n} |p_i - q_i|^2 \right)^{1/2} = \sqrt{\sum_{i=1}^{n} |p_i - q_i|^2} \tag{6}$$

Example: Minkowski Distance



Levenshtein Distance

- Distance between two sequences is given by the number of insert, delete and replace operations to transform one in another
- Also known as Edit Distance

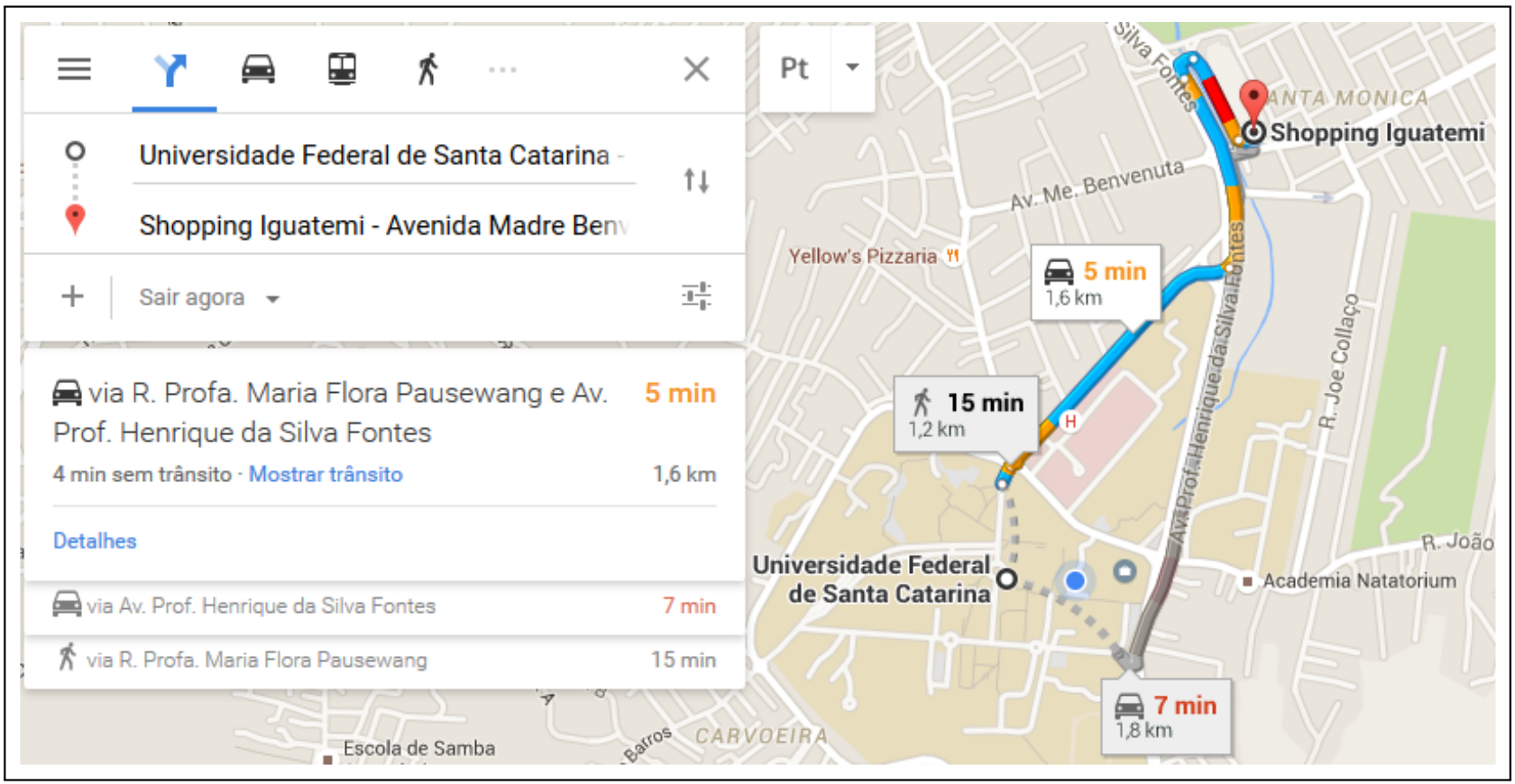
- Example $d(Avaí, ?)$:
  - Avaí
  - **Ha**vaí
  - **Hawa**ii
- Results for $d(Avaí, ?)$:
  - $d(Avaí, Avaí) = 0$
  - $d(Avaí, Havaí) = 2$
  - $d(Avaí, Hawaii) = 5$

Example: Levenshtein Distance

- Levenshtein Distance Avaí to Hawaii
- Transform Avaí into Hawaii:
  - 1 - Add H in the beggining (**H**Avaí)
  - 2 - Replace A for a (**Ha**vaí)
  - 3 - Replace v for w (**Haw**aí)
  - 4 - Replace $í$ for i (**Haw**a**i**)
  - 5 - Add i in the end (**Haw**a**ii**)
- Result for $d(Avaí, Hawaii) = 5$

More Distance Measures

- There are several other distances:
  - Road Network, Great Circle, Distance Time Warping, Mahalanobis, Jaro-Winckler, Canberra...



Topics

Initial Remarks - Distance $\rightarrow$ Similarity

- A key difference between Distance and Similarity is the score interval:
  - Distance: $[0, \infty]$
  - Similarity: $[0, 1]$
- In Similarity Measures when two objects are equivalent the score is equal to 1
- It is possible to infer similarity scores from distance measures by normalizing it when the maximum distance is known

Levenshtein - Distance → Similarity

- Maximum Levenshtein distance is the size of the longer string
  - $d(A, B)$ can be normalized by: $max(length(A), length(B))$
- Therefore, similarity score is given by:

$$sim(A, B) = 1 - \frac{d(A, B)}{max(length(A), length(B)}$$

- Example $sim(Avaí, ?)$:
  - Avaí
  - **H**avaí
  - **Hawaii**

Example: Levenshtein - Distance → Similarity

$$sim(A, B) = 1 - \frac{d(A, B)}{max(length(A), length(B)}$$

- Similarity Avaí to Hawaii:

$$\begin{aligned} sim(Avaí, Hawaii) &= 1 - \frac{d(Avaí, Hawaii)}{max(\mathbf{length(Avaí), length(Hawaii)}} \\ &= 1 - \frac{d(Avaí, Hawaii)}{\mathbf{max(4, 6)}} \\ &= 1 - \frac{\mathbf{d(Avaí, Hawaii)}}{6} \\ &= 1 - \frac{5}{6} \\ &= 1 - 0.86 = 0.17 \end{aligned} \tag{7}$$

Levenshtein - Distance → Similarity

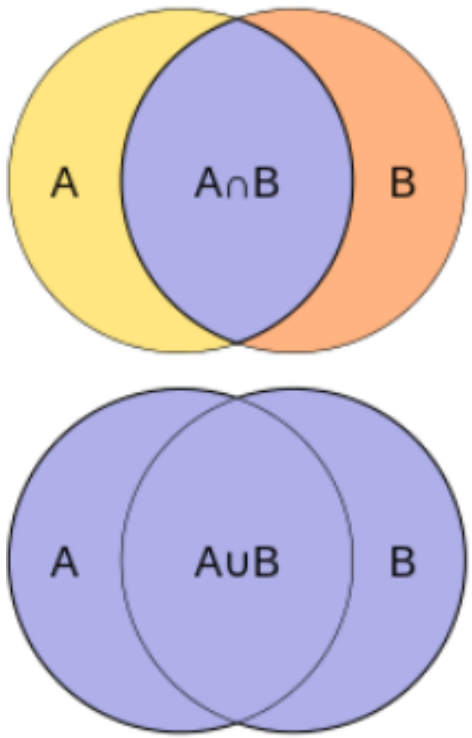
- Maximum Levenshtein distance is the size of the longer string
  - $d(A, B)$ can be normalized by: $max(length(A), length(B))$
- Therefore, similarity score is given by:

$$sim(A, B) = 1 - \frac{d(A, B)}{max(length(A), length(B)}$$

- Example $sim(Avaí, ?)$:
  - Avaí
  - **H**avaí
  - **Hawaii**
- Results for $sim(Avaí, ?)$:
  - $sim(Avaí, Avaí) = 1 - \frac{0}{4} = 1$
  - $sim(Avaí, Havaí) = 1 - \frac{2}{5} = 0.6$
  - $sim(Avaí, Hawaii) = 1 - \frac{5}{6} = 0.17$

Jaccard Similarity

- Similarity between two finite sets
- $sim(A, B) = \frac{|A \cap B|}{|A \cup B|}$



Example: Jaccard Similarity

- $sim(A, B) = \frac{|A \cap B|}{|A \cup B|}$
- Example Sets
  - $A = \{Giraffe, Monkey, Elephant, Bird\}$
  - $B = \{Monkey, Crocodile\}$
  - $C = \{Horse, Dog, Parrot\}$
  - $D = \{Monkey\}$
- Results for $sim(A, ?)$:
  - $sim(A, A) = \frac{4}{4} = 1$
  - $sim(A, B) = \frac{1}{5} = 0.2$
  - $sim(A, C) = \frac{0}{7} = 0$
  - $sim(A, D) = \frac{1}{4} = 0.25$

More Similarity Measures

- There are several other similarity measures:
  - Cosine Similarity, Longest Common Subsequence, Location In-Between Polylines, SimRank, Overlap Coefficient, Sorensen-Dice Coefficient...
- The applicability range is wide:
  - Sets, Sequences, Strings, Time-Series, Trajectories, WebPages, Documents, Images...

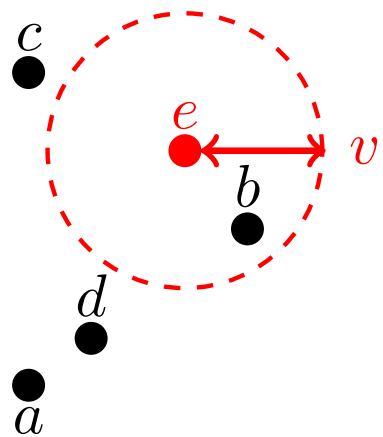
Topics

Similarity Queries

- Range
  - Given an element e and a minimum similarity score v return all elements $e\prime$ such that $sim(e, e\prime) < v$
- k-Nn (k-Nearest Neighbors)
  - Given an element e return the k most similar elements

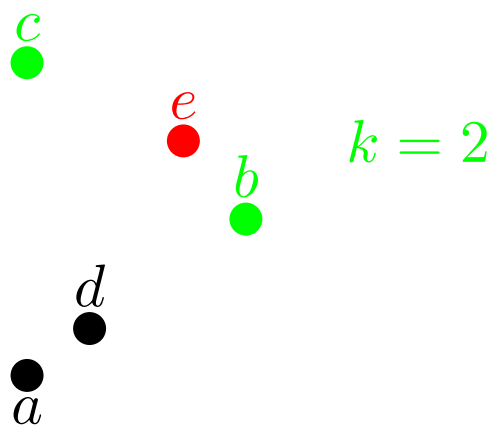
Range Query

- Range
  - Given an element e and a minimum similarity score v return all elements $e\prime$ such that $sim(e, e\prime) < v$



k-Nn Query

- k-Nn
  - Given an element e return the k most similar elements



Topics

Precision @ Recall

- Wide-used evaluation technique
- Input:
  - A query object q
  - A set of objects O
  - Let $q =$ São Francisco do Sul be the query object.
  - The other objects:

Object
São Francisco do Sul
São Franc S
San Francisco
Sao Francisco Sul
São Bento do Sul
São José
São Chico do Sul
São Caetano do Sul
São Paulo do Oeste
S Chico do Sul

Precision @ Recall - Relevant Objects

- Step 1: Given a query object q and a set of objects O, mark which objects are relevant (similar to q).
  - Let $q =$ São Francisco do Sul be the query object.
  - Which objects are relevant?

Object	Relevant
São Francisco do Sul	✓
São Franc S	✓
San Francisco	
Sao Francisco Sul	✓
São Bento do Sul	
São José	
São Chico do Sul	✓
São Caetano do Sul	
São Paulo do Oeste	
S Chico do Sul	✓

Example: Precision @ Recall - Calculate Similarity

- Step 2: Calculate the similarity of q to all objects in O
- Levenshtein (Lev.) Similarity Calculated:

Object	Relevant	Levenshtein (Lev.)
São Francisco do Sul	✓	1
São Franc S	✓	0.55
San Francisco		0.55
Sao Francisco Sul	✓	0.8
São Bento do Sul		0.65
São José		0.25
São Chico do Sul	✓	0.7
São Caetano do Sul		0.65
São Paulo do Oeste		0.4
S Chico do Sul	✓	0.6

Example: Precision @ Recall - Similarity Ranking

- Step 3: Rank the objects according to the similarity score
- Objects ranked:

Rank	Object	Relevant	Lev.
1	São Francisco do Sul	✓	1
2	Sao Francisco Sul	✓	0.8
3	São Chico do Sul	✓	0.7
4	São Bento do Sul		0.65
5	Sao Caetano do Sul		0.65
6	S Chico do Sul	✓	0.6
7	San Francisco		0.55
8	São Franc S	✓	0.55
9	São Paulo do Oeste		0.4
10	São José		0.25

Example: Precision @ Recall - Mark Recall Levels

- Step 4: Mark the recall levels
  - Example: recall level 0.2 means that $\frac{1}{5}$ of the relevant object were retrieved until that position in the ranking
  - Recall levels $\rightarrow \{0, 0.2, 0.4, 0.6, 0.8, 1\}$

Rank	Object	Relevant	Lev.	Recall
1	São Francisco do Sul	✓	1	0.2
2	Sao Francisco Sul	✓	0.8	0.4
3	São Chico do Sul	✓	0.7	0.6
4	São Bento do Sul		0.65	-
5	Sao Caetano do Sul		0.65	-
6	S Chico do Sul	✓	0.6	0.8
7	San Francisco		0.55	-
8	São Franc S	✓	0.55	1
9	São Paulo do Oeste		0.4	-
10	São José		0.25	-

Example: Precision @ Recall - Levenshtein

- Step 5: Calculate the precision for each level of recall
  - Precision at recall: $\frac{|R_l|}{|O_l|}$ where R_l is the number of relevant objects and O_l is the total number of objects up to recall l
  - Example: for $l = 0.8$ there are 6 objects and 4 are relevant. Therefore precision at recall level 0.8 is $\frac{4}{6} = 0.66....$

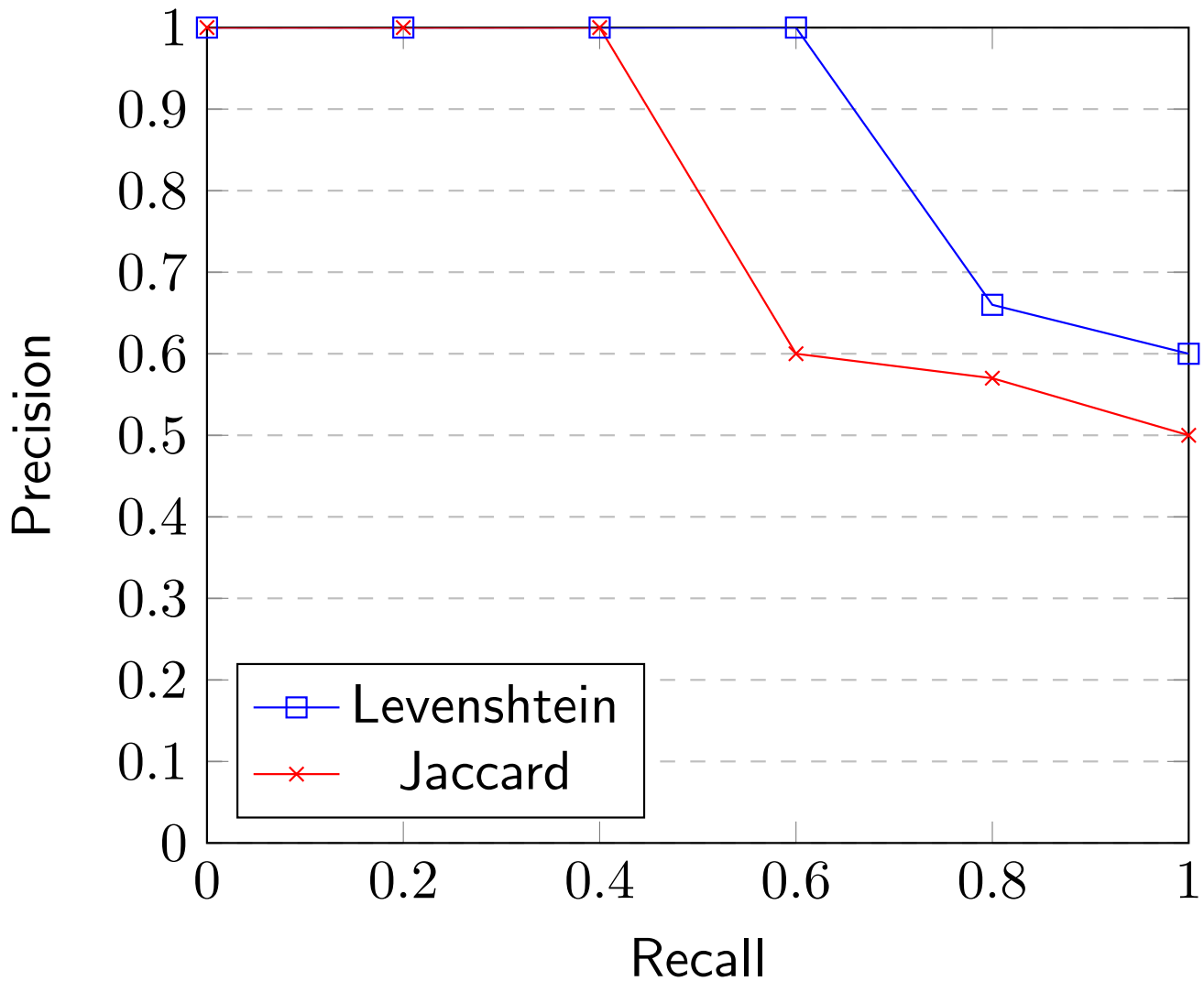
Rank	Object	Relevant	Lev.	Recall	Precision
1	São Francisco do Sul	✓	1	0.2	1
2	Sao Francisco Sul	✓	0.8	0.4	1
3	São Chico do Sul	✓	0.7	0.6	1
4	São Bento do Sul		0.65	-	-
5	Sao Caetano do Sul		0.65	-	-
6	S Chico do Sul	✓	0.6	0.8	0.66
7	San Francisco		0.55	-	-
8	São Franc S	✓	0.55	1	0.62
9	São Paulo do Oeste		0.4	-	-
10	São José		0.25	-	-

Example: Precision @ Recall - Jaccard

- Repeat Steps 2, 3, 4 and 5 for Jaccard
- Results for Jaccard:

Rank	Object	Relevant	Jac.	Recall	Precision
1	São Francisco do Sul	✓	1	0.2	1
2	S Chico do Sul	✓	0.6	0.4	1
3	São Bento do Sul		0.6	0.4	-
4	Sao Caetano do Sul		0.6	0.4	-
5	Sao Francisco Sul	✓	0.4	0.6	0.6
6	São Paulo do Oeste		0.33	0.6	-
7	São Franc S	✓	0.33	0.8	0.57
8	San Francisco		0.2	0.8	-
9	São José		0.2	0.8	-
10	São Franc do S	✓	0.16	1	0.5

Example: Precision @ Recall - Graph



Topics

Final Remarks

- There are several distance/similarity measures for different purpouses
- The choice of an adequate similarity measure is crucial to the results of classification or clustering algorithms
- Precision @ Recall is the most used evaluation technique, but there are several others to be applied in different contexts
- Most of these measures are implemented in the Java Library Simmetrics: http://sourceforge.net/projects/simmetrics/

Exercises

- 1) Dados os pontos $A = (4, 2)$, $B = (5, 4)$ e $C = (3, 5)$ calcule as distâncias Euclidiana, Manhattan e Chebyshev entre os elementos.

- 2) Dado o conjunto $A = \{Preto, Azul, Branco, Cinza\}$ calcule a similaridade Jaccard para os conjuntos $B = \{Azul\}$, $C = \{Verde\}$, $D = \{Laranja, Cinza, Preto\}$ e $E = \{Preto, Azul\}$.