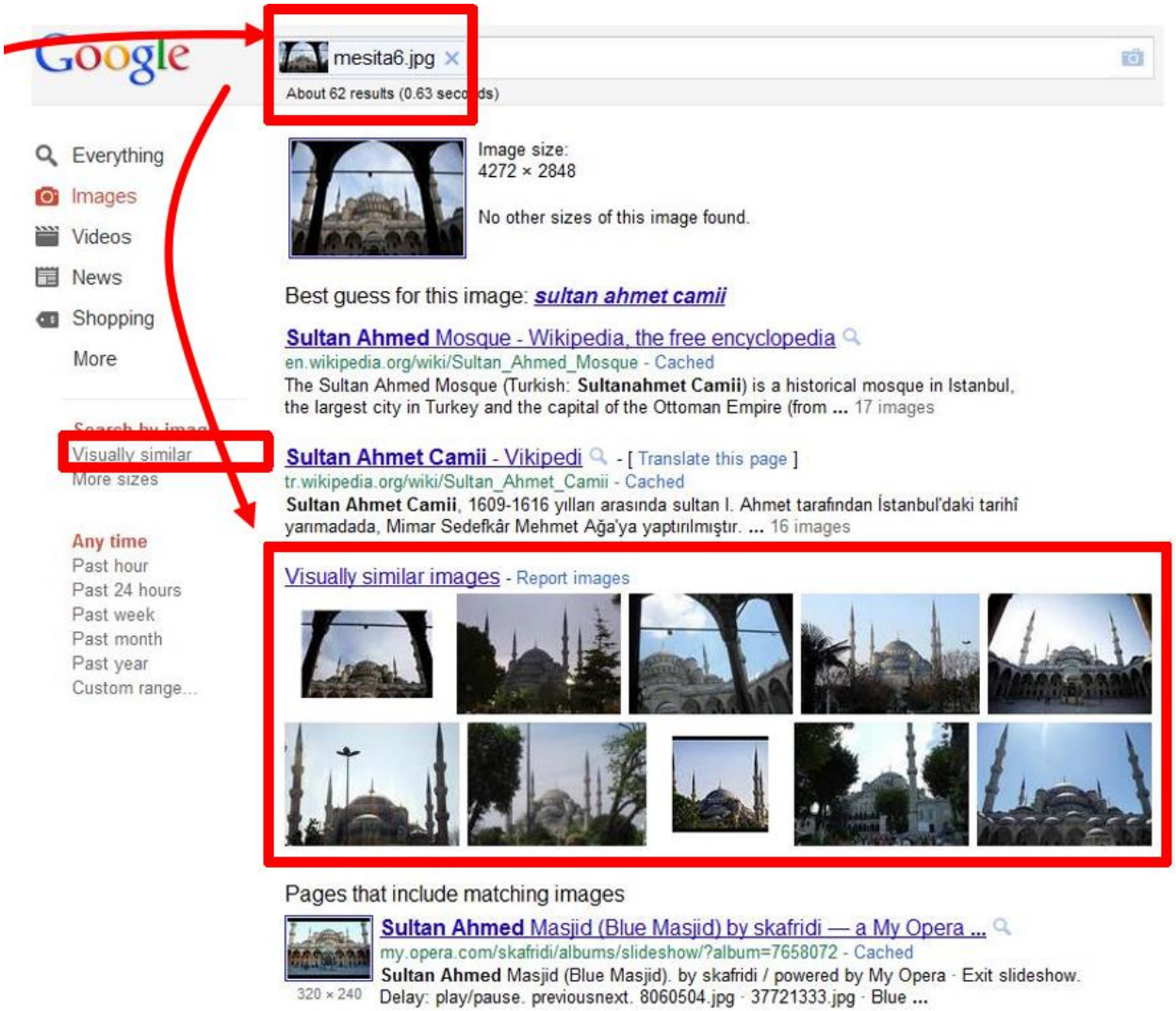
Zápočtový test
Podobnostní hledání

Programování v C++
ZS 2016/2017
Přemysl Čech

Podobnostní hledání

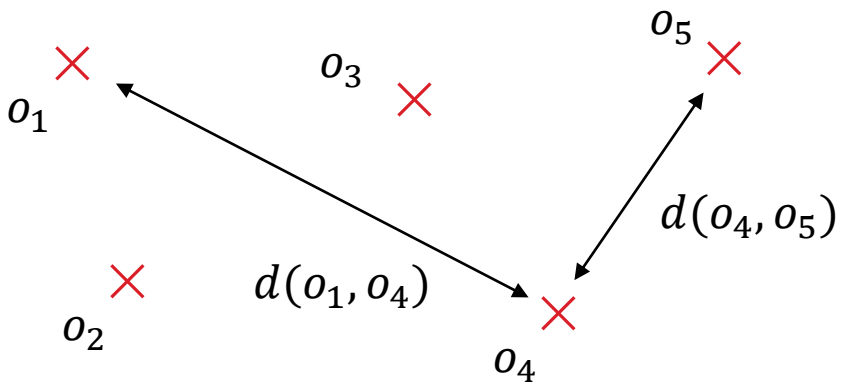
- Motivace – chceme najít nejpodobnější objekty k dotazu v databázi



Podobnostní hledání

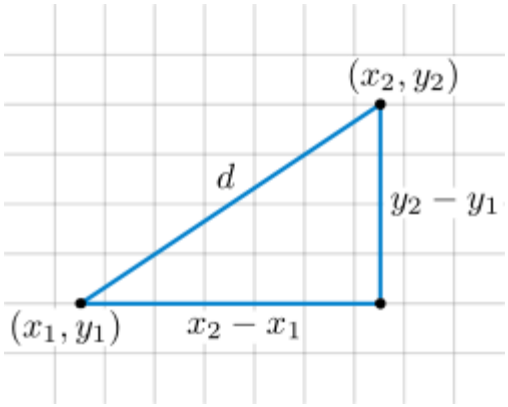
- Máme již předpřipravené klíčové vlastnosti
  - 2 soubory – databáze a dotazy

- Klíčové vlastnosti jsou mapované v metrickém prostoru: (M, d)
  - M … objekty v prostoru – v našem případě vektory čísel
  - d … podobnostní funkce (metrika), $d: M \times M \to \mathbb{R}$



Podobnostní hledání

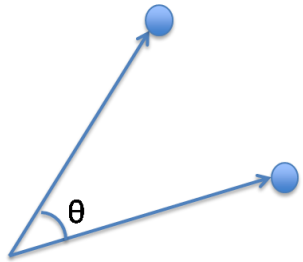
- Příklady podobnostních funkcí $d: M \times M \rightarrow \mathbb{R}$
  - n je dimenze prostoru M



- L2 (Euklidovská) metrika
  - $d_2(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

- Kosinová metrika
  - $sim(x, y) = \frac{\sum_{i=1}^{n}(x_i y_i)}{\sqrt{\sum_{i=1}^{n} {x_i}^2}\sqrt{\sum_{i=1}^{n} {y_i}^2}}$
  - $d_{cos}(x, y) = 1 - \frac{cos^{-1}(sim(x,y))}{\pi}$

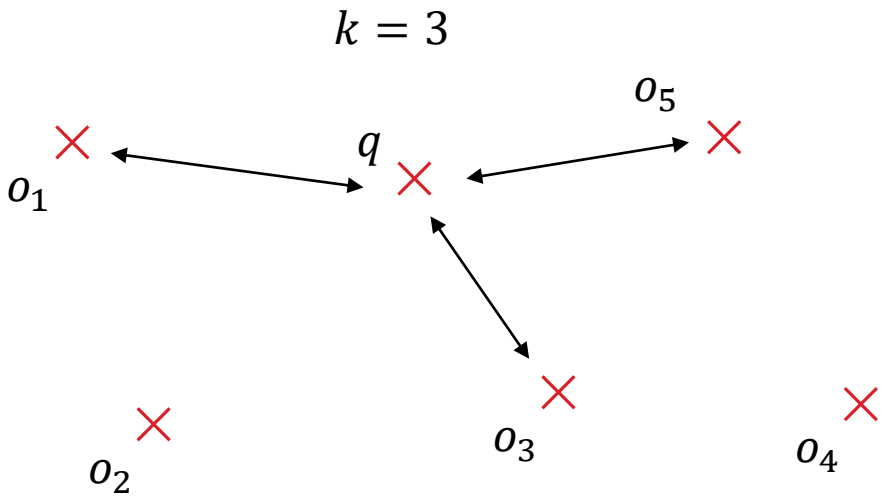
$$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



Podobnostní hledání

- k−nearest neighbors search
  - Hledáme k nejpodobnějších objektů v metrickém prostoru (M, d) k zadanému dotazu q

$$kNN(q) = \{\mathbb{Q} \subseteq M; |\mathbb{Q}| = k \ \wedge \ \forall x \in \mathbb{Q}, y \in M - \mathbb{Q}: d(q,x) \leq d(q,y)\}$$



Podobnostní hledání

- Detailněji:

```
program.exe m.csv q.csv 10 output.txt
```

- Program má čtyři parametry:
  - 1) soubor s databází
  - 2) soubor s dotazy
  - 3) hodnotu k
  - 4) výstupní soubor

M

```
1;-0.1;0
2;1;0
3;1.1;1.1
```

Q

```
4;0.5;0
5;0.5;0.5
```

- Oba vstupní soubory mají stejný formát:
  - ID objektu následované středníkem (;)
  - Vektor čísel (double) oddělený (;) – jsou stejně dlouhé

- Výstup

```
4    2|0.5    1|0.6    3|1.253
5    2|0.707107    1|0.781025    3|0.848528
```

  - Pro všechny dotazy vypsaná sekvence nejbližších k sousedů (ID) a vzdálenost k nim od nejbližších

Podobnostní hledání

- Podmínka:
  - chceme umět pracovat s různými metrikami
  - implementujte L2 a Kosinovou metriku
  - použitá metrika by měla jít snadno vyměnit (přes rozhraní) a měla by být lehce rozšiřitelná o další implementace podobnostních funkcí

- Větší testovací sobory:
  - http://siret.ms.mff.cuni.cz/cech/cpp/m.csv
  - http://siret.ms.mff.cuni.cz/cech/cpp/q.csv

Podobnostní hledání

- Program by neměl spadnout na nevalidní vstup
- Program by neměl běžet moc dlouho – efektivní řešení
- Nutné dodržet vhodné konvence kódu:
  - zapouzdření do tříd(y), hlavičkové + implementační soubory
  - rozumné rozčleněné kódu do funkcí
  - správná alokace (případně i dealokace) proměnných
  - efektivní předávání parametrů
  - …
- Můžete používat internet a vlastní poznámky pro REFERENČNÍ hledání (nesmíte kopírovat kusy kódu)
  - můžete dohledávat popisy knihovních funkcí, parametrů, atd.
- Můžete používat vlastní laptop
- Nesmíte komunikovat s kolegy
- Na řešení máte **3 hodiny**

Nějaké dotazy?