Bra-att-ha-algoritmer…

Identifiera grupper av objekt
Jämföra objekt
Mönsterigenkänning
Åskådliggöra fenomen

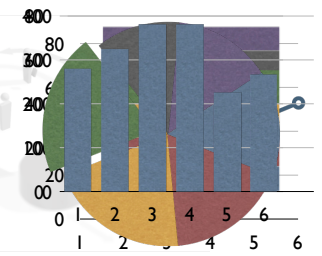
Informationssökning

Informationssökning

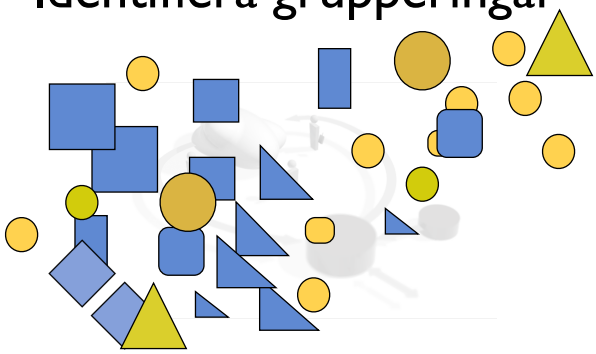
- Traditionell ansats
  - Matrisbaserad indexering (Ord x Dokument matris) (glesmatris)
  - Ordfrekvens i cellerna
    - Viktade
      - Globalt (GfIDF, IDF, 1-Entropy, Weight, Word Length, Dfi)
      - Lokalt (Term Frequency (absolute/relativ, Binarär, log(Tfi_abs+1)
  - Vektorbaserade sökning
- Mäta prestanda hos informationssökning
  - Recall - andelen av alla relevanta dokument som levereras
  - Precision - andelen av de levererade dokumenten som faktiskt är relevanta
- Problem som ställer till det
  - Synonymer - flera namn på samma begrepp (Bil, Kärra, Saab)
  - Påverkar ”recallfaktorn”
- Polysemy - flera olika begrepp med samma namn (Bilen Ford vs. President Ford)

Åskådliggöra fenomen

id	X	Y	Area
1	12	5	56
2	3	3	65
3	4	6	76
4	6	1	76
5	12	4	45
6	6	6	53



Identifiera grupperingar



K-means

Identifiera grupperingar

- Dela in datat i ett antal kluster/klasser
  - Iterativ algoritm, som jobbar med att uppfylla följande två kriterier:
    1. Varje klass har ett “tyngdpunkt” som är medel-positionen för alla punkter i den klassen.
    2. Varje punkt tillhör den klass vars tyngdpunkt som den är närmast

K-means

Identifiera grupperingar

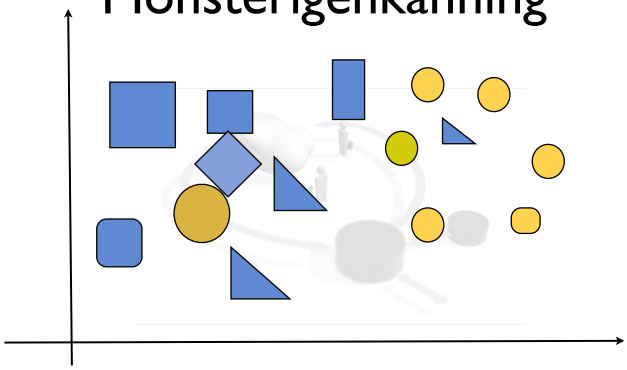
- Initialize
- Loop until termination condition is met:
  1. For each object, assign that object to a class such that the distance from this object to the center of that class is minimized.
  2. For each class, recalculate the means of the class based on the objects that belong to that class.
- End loop;

Identifiera grupperingar

K-means

- Några kritiska punkter i ansatsen
  - Hur många klasser ska man söka efter
  - Hantering av "döda klasser"
  - Initieringen
  - Likhet/avståndsmåttet
  - Termineringsvillkoret
    - Algoritmen terminerar, men kan ta tid...
      - Alternativa termineringsvillkor
        - max ett visst antal punkter som byter klass
        - Kör ett visst antal steg

Mönsterigenkänning



Mönsterigenkänning

KNN

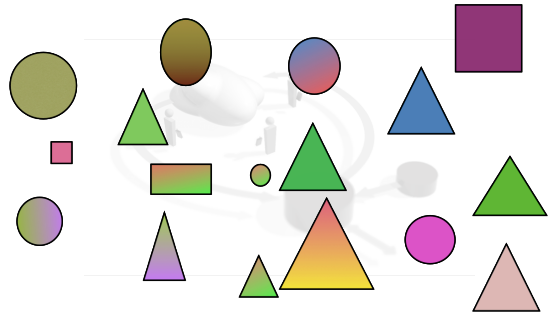
- Metod för att avgöra klasstillhörighet
- Hitta dom K närmsta grannarna i träningsmängden
- Undersöka de funna grannarna för att avgöra klasstillhörighet
  - Majoritetsbeslut
  - ???

Mönsterigenkänning

KNN

- Några kritiska punkter med KNN
  - Likhetsmått
  - Proceduren för att utvärdera grannarna
  - Antalet grannar
  - Storleken på träningsmängden

Reducering av egenskapsrymden



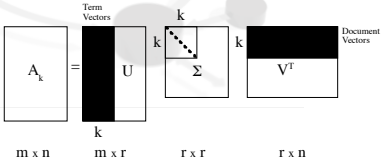
Reducering av egenskapsrymden

LSA/LSI

- LSI utvecklades slutet av 80-talet början av 90-talet av Susan Dumais & Michael Berry
  - Patenterad metod Bell Communications Research (Bellcore)
    - Computer information retrieval using latent semantic structure U.S. Patent No. 4,839,853, June 13, 1989.
- Samma grundprincip som traditionell approach
  - Matrisbaseras (ordfrekvensmatris)
  - Vektorbaseradsökning
- Stora skillnaden är indexeringsstrukturen

LSA/LSI

- K-dimensionellt semantiskt rum konstrueras där termer och dokument placeras in i samma sökrum
  - Jämföra och hitta klasser av dokument
- LSI använder en trunkerad SVD för att konstruera indexstrukturen



LSA/LSI

- För informationssökning är första man gör att representera frågan i samma indexstruktur, dvs det k-dimmensionella semantiska rummet
  - Ansatsen tillåter ord, ordlistor, dokument eller dokmentsamlingar att vara frågor.
  - Frågor är en samling av ord
  - En fråga q representeras som en term-vektor eller ett pseudodokument

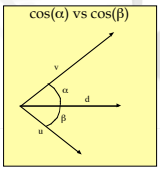
$$V^T = \Sigma^{-1} U^T A \Rightarrow V^T_{:,i} = \Sigma^{-1} U^T A_{:,i}$$

$$V_{i,:} = (A_{:,i})^T U \Sigma^{-1}$$

$$\hat{q} = q^T U_k \Sigma_k^{-1}$$

LSA/LSI

- Vidare, detta pseudo-dokument jämförs mot alla dokument som ingår i indexstrukturen, och de dokument som uppfyller ett likhets/ närhetskriterie returneras som ett söksvar



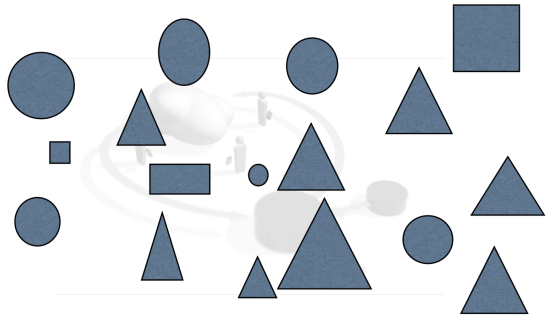
LSA/LSI

- Förändra Indexstrukturen
  - Beräkna om hela (ingen uppdatering)
    - Resurskrävande
    - Bibehåller ”korrektheten” i strukturen
  - Uppdatering är en process där man förändrar en existerande LSI-genererad indexstruktur
    - Inkorporera ”pseudo dokument” och ”pseudo termer”
    - Snabb och enkel
    - Garanterar inte ortogonaliteten i det semantiskarummet
  - SVD-updating (G.W. O’Brien (1994))
    - Addera nya dokument & nya termer
    - Korrigera för förändringar i ”termvikterna”
    - Bibehåller ”korrektheten” i strukturen
    - Snabbare än att räkna om hela

LSA/LSI

- Kritiska punkter för LSI
  - Kräver mycket beräkningsresurser
  - Minne
  - Beräkningskraft
- Välja dimensionen på indexstrukturen (k)
- Termviktning mm
- Mätning av närhet

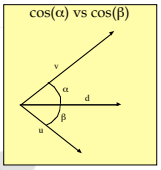
Likhetsmått



Olika sätt att jämföra

Likhetsmått

- Vinkeln mellan vektorer
- Avstånd mellan objekt
  - Ex. Euclidiskt avstånd
- Överlappning
- Area
- Form
-



Lite grann om komplexitet

Komplexitet

- Rumskomplexitet
  - Term*dokument matriser
- Tidskomplexitet
  - LSA/LSI ≈ $O(n^3)$
  - K-means ≈ O(ndcT)
  - KNN ≈ O(n*m)