Curs 12:

Analiza documentelor de tip text (Text Mining)

Structura

- Reminder: reprezentarea documentelor de tip text
- Analiza similarităţii între documente
- Particularităţi ale grupării documentelor

Reminder

Extragerea caracteristicilor dintr-un document: (fişier text – date nestructurate)

Abordarea bazată pe reprezentarea de tip bag-of-words:

“In document classification, a bag of words is a sparse vector of occurrence counts of words; that is, a sparse histogram over the vocabulary. In computer vision, a bag of visual words is a vector of occurrence counts of a vocabulary of local image features.”

Reminder

Extragerea caracteristicilor dintr-un document – fişier text: abordarea de tip bag-of-words

a) Eliminarea cuvintelor de legătură (stop words)

"In document classification, a bag of words is a sparse vector of occurrence counts of words; that is, a sparse histogram over the vocabulary. In computer vision, a bag of visual words is a vector of occurrence counts of a vocabulary of local image features."

"document classification bag words sparse vector occurrence counts words sparse histogram vocabulary computer vision bag visual words vector occurrence counts vocabulary local image features."

Reminder

Extragerea caracteristicilor dintr-un document – fişier text: abordarea de tip bag-of-words

b) Reducerea cuvintelor la rădăcina lor – stemming (algoritm Porter)

“document classification bag words sparse vector occurrence counts words sparse histogram vocabulariy computer vision bag visual words vector occurrence counts vocabulariy local image features”

[http://textanalysisonline.com/nltk-porter-stemmer]

“document classif bag word spars vector occurr count word spars histogram vocabulari comput vision bag visual word vector occurr count vocabulari local imag featur”

Reminder

Extragerea caracteristicilor dintr-un document – fişier text: abordarea de tip bag-of-words

c) Calculul frecvenţelor:

“document classif bag word spars vector occurr count word spars histogram vocabulari comput vision bag visual word vector occurr count vocabulari local imag featur”

Caracteristici extrase:

(bag,2), (classif,1), (comput,1), (count,2), (document,1), (featur,1), (histogram,1), (imag,1), (local,1), (occurr,2), (spars,2), (vector,2), (vision,1), (visual,1), (vocabulari,2), (word,3)

Reminder

Extragerea caracteristicilor dintr-un document – fişier text: abordarea de tip bag-of-words

c) Calculul frecvenţelor:

“document classif bag word spars vector occurr count word spars histogram vocabulari comput vision bag visual word vector occurr count vocabulari local imag featur”

Caracteristici extrase:

(bag,2), (classif,1), (comput,1), (count,2), (document,1), (featur,1), (histogram,1), (imag,1), (local,1), (occurr,2), (spars,2), (vector,2), (vision,1), (visual,1), (vocabulari,2), (word,3)

Reprezentarea datelor

Concepte:

- Corpus - colecţia de documente disponibile
- Lexicon – setul de cuvinte (termeni) folosite – în varianta pre-procesată (după stemming)

Notaţii:

- n - nr de documente
- d – nr de termeni în lexicon

Reprezentarea unui document:

- vector v cu d componente
- v(i) reprezintă
  - numărul de apariţii ale termenului i din lexicon în cadrul documentului
  - frecvenţa ajustată a termenului i (reprezentarea de tip TF-IDF)

Reprezentarea datelor

Reprezentarea TF-IDF:

TF(i) = Term Frequency = numărul de apariţii ale termenului i în document

IDF(i) = Inverse Document Frequency = 1/numărul de documente care conţine termenul i

v(i)=TF(i)*IDF(i)

Obs:

- Cu cât TF(i) este mai mare cu atât este mai reprezentativ termenul i pentru document
- Cu cât IDF(i) este mai mare cu atât termenul i are putere mai mică de discriminare (dacă un termen este comun mai multor documente atunci el nu furnizează informaţie utilă pentru discriminarea între documente)

Reprezentarea datelor

Variante ale reprezentării TF-IDF

- În calculul lui TF şi IDF se folosesc frecvenţe relative în locul frecvenţelor absolute

  rTF(i) = nr de apariţii ale termenului i în document / nr total de termeni din document

  rIDF(i) = nr total de documente/ nr de documente care conţine termenul i

  v(i)=rTF(i)*rIDF(i)

Obs: în cazul în care un termen apare foarte frecvent într-un document atunci contribuţia lui va fi dominantă în calculul măsurilor de similaritate sau disimilaritate – pentru a limita acest lucru se aplica funcţii care atenuează diferenţele între frecvenţe (ex: funcţia logaritmică, funcţia radical)

v(i)=log(TF(i))*log(IDF(i))

Reprezentarea datelor

Caracteristici ale reprezentării de tip bag-of-words bazată pe frecvenţe absolute sau pe TF-IDF:

- Reprezentare rară (sparsity): numărul de elemente nenule din vector este mic (un document conţine puţine cuvinte din lexicon)
- Pozitivitate (non-negativity): toate componentele vectorului asociat unui document sunt pozitive
- Informaţii adiţionale (side information): în anumite cazuri documentele au ataşate metadate care pot fi incluse în analiză (de exemplu link-uri asociate documentelor web)

Măsuri de similaritate

Măsurile de similaritate sunt utile în:

- Algoritmii de grupare - atât pt cei partiţionali (bazaţi pe reprezentanţi) cât şi pentru cei ierarhici

- Algoritmii de clasificare bazaţi pe criteriul celui mai apropiat vecin

Presupunem că u şi v sunt vectorii asociaţi unor documente

Măsura cosinus

- $\cos(u,v)= u^T v/(||u||\ ||v||)$

Măsura Jaccard

- $J(u,v)= u^T v/(||u||^2+ ||v||^2- u^T v)$

Adaptarea algoritmilor de clustering

Algoritmi partiţionali (kMeans, kMedoid)

- Reminder:
  - Fiecare cluster are asociat un reprezentant (centroid)
  - Asignarea unei date la un cluster se bazează pe analiza similarităţii (se asignează la clusterul cu cel mai similar reprezentant)
  - Reprezentantul se determină folosind media sau mediana elementelor din cluster (sau cel mai apropiat element de medie)

- Elemente care trebuie adaptate:
  - Măsura de similaritate (folosită la asignarea documentelor la un cluster)
  - Reprezentantul unui cluster se poate construi prin “concatenarea” documentelor ce aparţin clusterului (în cazul reprezentării bazate pe frecvenţe absolute este echivalentă cu adunarea vectorilor asociaţi)

Adaptarea algoritmilor de clustering

Algoritmi ierarhici (aglomerativi)

- Reminder:
  - Se construieşte matricea de similaritate
  - Se asignează fiecare dată la un cluster
  - La fiecare etapă se reunesc clusterele cele mai apropiate (similare)

- Elemente componente care trebuie adaptate:
  - Măsura de similaritate între documente (folosită la construirea matricii de similaritate)
  - Măsura de similaritate între clustere

    – se identifică termenii cei mai reprezentativi (frecvenţa cea mai mare) din documentele aflate în fiecare cluster (topical words); două clustere sunt considerate similare dacă există suprapunere mare între seturile de termeni reprezentativi

Adaptarea algoritmilor de clustering

Algoritmi probabilişti (algoritmul EM - Expectation Maximization)

- Reminder:
  - Fiecare dată este generată de o distribuţie de probabilitate (tipul de distribuţie depinde de natura datelor)
  - În procesul de generare a datelor fiecare dintre distribuţiile de probabilitate este selectată la rândul ei cu o anumită probabilitate

- Elemente componente care trebuie adaptate:
  - Distribuţia de probabilitate
    - Binomială – pt vectori cu elemente binare (indicatori de prezenţă)
    - Multinomială – pt vectori de frecvenţe