نمایه گذاری معنایی پنهان

latent semantic indexing(LSI)

نادیه آرمین

پاییز ۹۶

فهرست مطالب

- مقدمه
  - نمایه گذاری
  - فضای برداری
- تعریف نمایه گذاری معنایی پنهان
- آشنایی با SVD
- کاربردهای SVD
- مثال
- پیاده سازی
- محاسبه شباهت ها درLSI

مقدمه_ نمایه گذاری

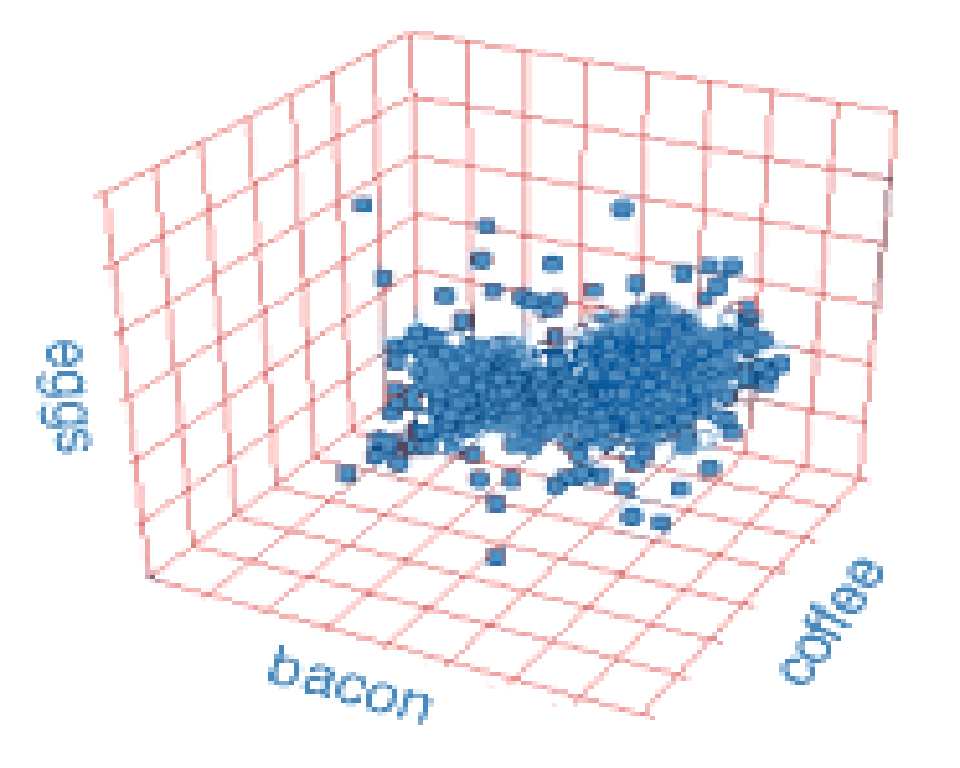
- ماتریس واژه – سند

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

– تعداد واژه

– TF_IDF

مقدمه_ فضای برداری

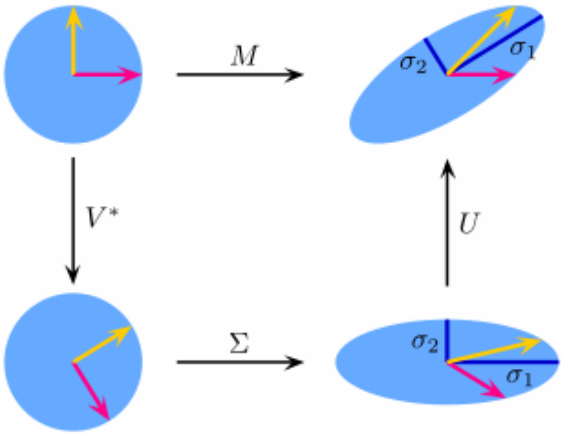


تعریف نمایه گذاری معنایی پنهان

- نگاشت ماتریس کلمه _ سند به ماتریس معنا_ سند
- مثال :
  - جستجو عبارت «درگیری‌های خاورمیانه»
  - احتمالاً سندی با عنوان «جنگ عراق» مفید است
  - جنگ هم‌معنی درگیری و عراق در خاورمیانه قرار دارد.
- مزایا:
  - عدم نیاز به استفاده از تعاریف منابع خارجی
  - پیاده‌سازی به وسیله روش‌های ریاضی

آشنایی با SVD

- تجزیه مقادیر منفرد(Singular Value Decomposition)



- SVD: $M = U\Sigma V^T$
- Mماتریس سند_ واژه
- محاسبه.M′
- M′ در مقایسه با Mمقایسه بهتری بین اسناد انجام میدهد

کاربردهای SVD

- کاهش ویژگی
- پردازش تصویر
- فشرده سازی
- بازیابی اطلاعات
- و ...

مثال

- SVD: $C = U\Sigma V^T$

  (where C = term-document matrix)

- use the SVD to compute:

  a new, improved term-document matrix C'.

- better similarity values out of C' (compared to C).

مثال

Example of $C = U\Sigma V^T$: The matrix C

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

This is a standard term-document matrix. Actually, we use a non-weighted matrix here to simplify the example.

مثال

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

$=$

U	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0.00	0.73
ocean	−0.48	−0.51	−0.37	0.00	−0.61
wood	−0.70	0.35	0.15	−0.58	0.16
tree	−0.26	0.65	−0.41	0.58	−0.09

$\times$

Σ	1	2	3	4	5
1	2.16	0.00	0.00	0.00	0.00
2	0.00	1.59	0.00	0.00	0.00
3	0.00	0.00	1.28	0.00	0.00
4	0.00	0.00	0.00	1.00	0.00
5	0.00	0.00	0.00	0.00	0.39

$\times$

V^T	d_1	d_2	d_3	d_4	d_5	d_6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.28	−0.75	0.45	−0.20	0.12	−0.33
4	0.00	0.00	0.58	0.00	−0.58	0.58
5	−0.53	0.29	0.63	0.19	0.41	−0.22

$$C = U\Sigma V^T$$

پیاده سازی

- Compute SVD of term-document matrix
- Reduce the space and compute reduced document representations
- Map the query into the reduced space
- This follows from:
- Compute similarity of q_2 with all reduced documents in V_2
- Output ranked list of docun

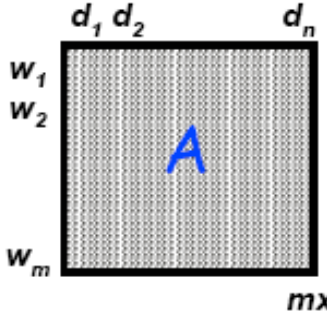
$$\vec{q}_2^T = \Sigma_2^{-1} U_2^T \vec{q}^T.$$

$$C_2 = U\Sigma_2 V^T \Rightarrow \Sigma_2^{-1} U^T C = V_2^T$$

- Exercise: What is the fundamental problem with this approach?

محاسبه شباهت ها درLSI

- Fundamental comparisons based on SVD
  - The original word-document matrix (A)



- compare two terms → dot product of two rows of A
  - or an entry in AA^T
- compare two docs → dot product of two columns of A
  - or an entry in A^TA
- compare a term and a doc → each individual entry of A

  - The new word-document matrix (A')

$U'=U_{mxk}$
$\Sigma'=\Sigma_k$
$V'=V_{nxk}$

- compare two terms → dot product of two rows of $U'\Sigma'$
- compare two docs → dot product of two rows of $V'\Sigma'$
- compare a query and a doc → each individual entry of A'

$A'A'^T=(U'\Sigma'V'^T)(U'\Sigma'V'^T)^T=U'\Sigma'V'^TV'\Sigma'^TU'^T=(U'\Sigma')(U'\Sigma')^T$

$A'^TA'=(U'\Sigma'V'^T)^T(U'\Sigma'V'^T)=V'\Sigma'^TU^TU'\Sigma'V'^T=(V'\Sigma')(V'\Sigma')^T$

I

For stretching or shrinking