Latent Semantic Analysis

Kuan-Yu Chen (陳冠宇)

2020/11/06 @ TR-313, NTUST

Review

- Recall
- Precision
- Mean Average Precision (MAP)
- Normalized Discounting Cumulated Gain (NDCG)

Homework 2

1	—	我們不要BM25了	0.77480
2	—	這是VSM	0.75719
3	—	Dhye6373ryd	0.75714
4	—	M10915010_盧克函	0.74720
5	—	M10915080_羅笠程	0.74603
6	—	M10907505_游照臨	0.74532
7	—	B10632026_吳苡瑄	0.74395
8	—	B10615017_林辰叡	0.74150
9	—	M10909109_陳兆炫	0.74041
10	—	60947007S_翁詩諺	0.73798

Introduction

- Classic IR might lead to poor retrieval due to:
  - Relevant documents that do not contain at least one index term are not retrieved
  - Synonymy (同義詞) and polysemy (一詞多義) are crucial for IR
    - Car vs. Automobile

      The prevalence of synonyms tends to decrease the **recall** performance of retrieval systems
    - Bank

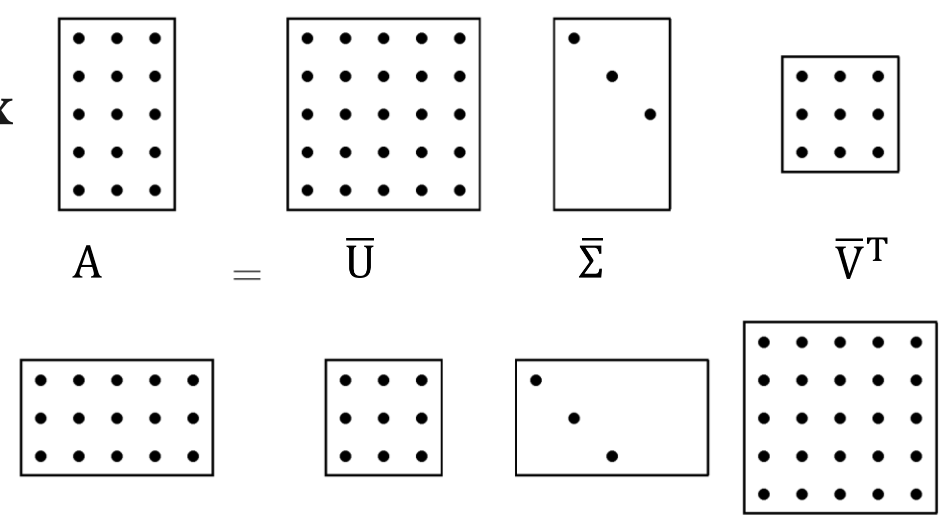
      Polysemy is one factor underlying poor **precision**
  - Retrieval based on index terms is vague and noisy
    - The user information need is more related to **concepts** and ideas than to **index terms**

Latent Semantic Analysis

Singular Value Decomposition

- In linear algebra, the singular-value decomposition (SVD) is a factorization of a real or complex matrix

- Formally, the SVD of an $m \times n$ matrix A is a factorization of the form $\overline{\mathrm{U}}\overline{\Sigma}\overline{\mathrm{V}}^{\mathrm{T}}$
  - $\overline{\mathrm{U}}$ is an $m \times m$ **unitary matrix** (i.e.,$\overline{\mathrm{U}}\overline{\mathrm{U}}^{\mathrm{T}} = \mathrm{I} = \overline{\mathrm{U}}^{\mathrm{T}}\overline{\mathrm{U}}$)
  - $\overline{\Sigma}$ is an $m \times n$ rectangular **diagonal matrix** with non-negative real numbers on the diagonal
    - Singular value
  - $\overline{\mathrm{V}}$ is an $n \times n$ **unitary matrix**



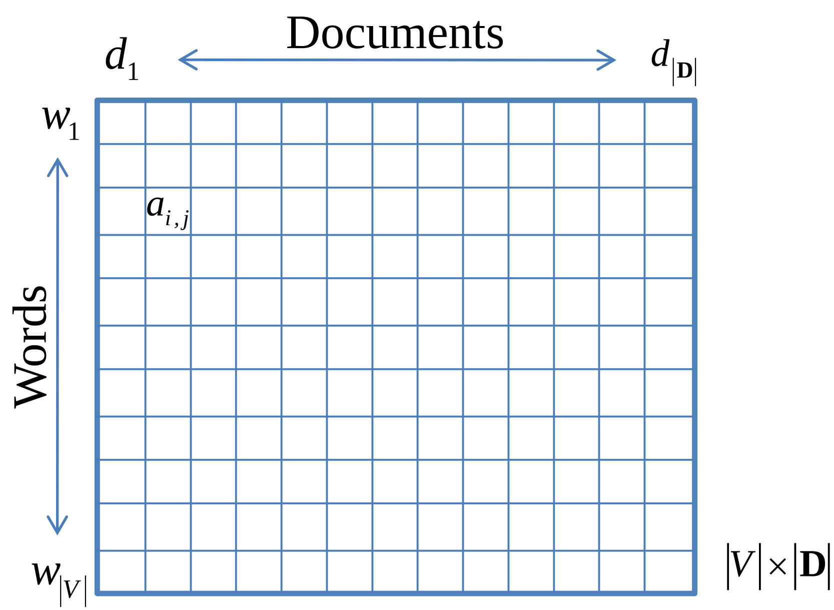
Introduction - LSA

- **Latent Semantic Analysis** also called
  - Latent Semantic Indexing (LSI)
  - Latent Semantic Mapping (LSM)
  - Two-Mode Factor Analysis

- The LSA paradigm operates under the assumption that there is some underlying **latent semantic structure** in the data
  - Algebraic and/or statistical techniques are brought to bear to estimate this latent structure and get rid of the obscuring "noise"

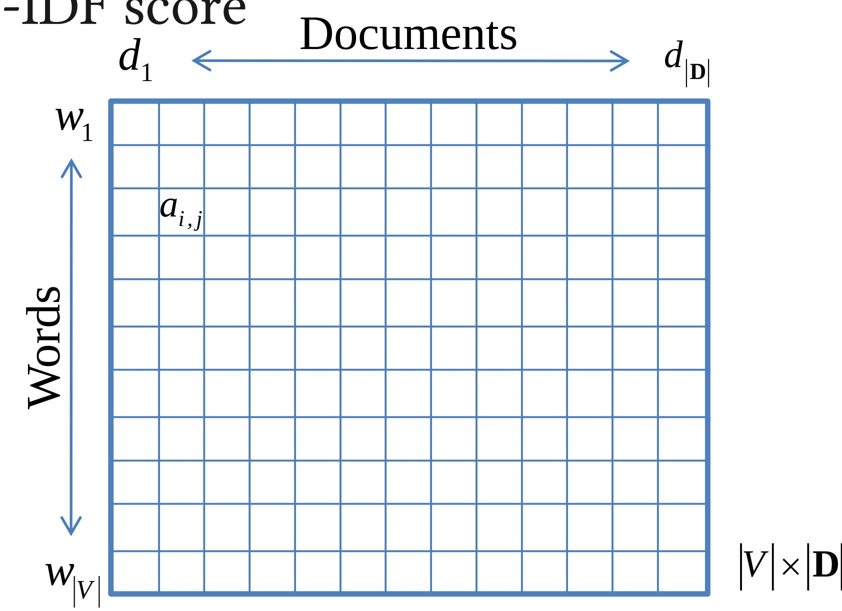
Latent Semantic Analysis.

- A given document collection can be represented as a word-document matrix
  - Row: composed of **words** (index terms), which are the individual components making up a document
  - Column: composed of **documents** which are of a predetermined size of text such as paragraphs, collections of paragraphs, sentences, etc.



Latent Semantic Analysis..

- In the word-by-document, each element $a_{i,j}$ is represented the importance of word w_i in document d_j
  - $a_{i,j}$ can be determined by the TF-IDF score

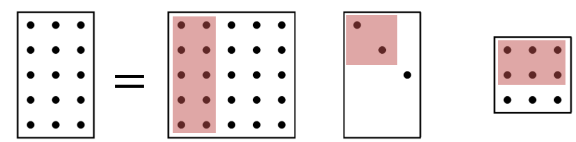


- The properties of the matrix
  - the dimensions and can be extremely large
  - the column vectors are typically very sparse
  - the two spaces (for words and documents) are distinct from one other

Latent Semantic Analysis...

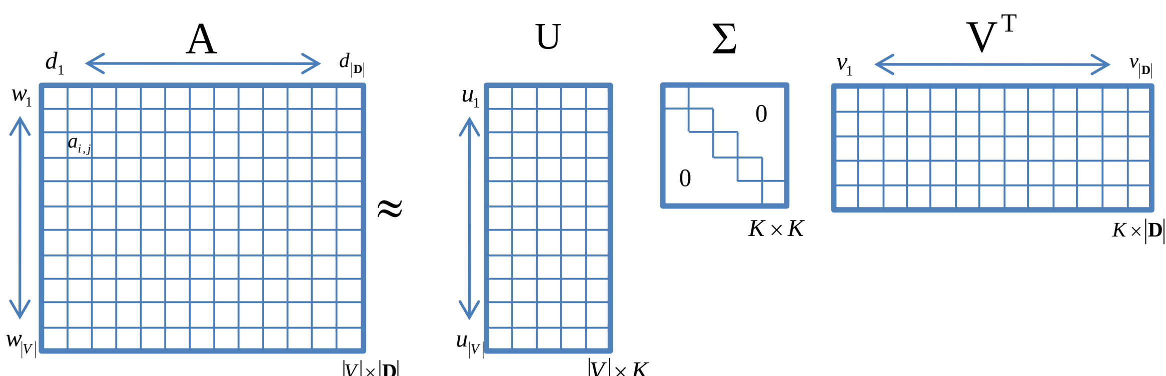
- In order to explore the latent semantic space, to project word and document vectors in the space, and to reduce the size of the vectors, the **Singular Value Decomposition** (SVD) can be employed
  - $K \leq \min(|V|, |\mathbf{D}|)$: **low-rank approximation**

$$\mathrm{A}_{m\times n} = \bar{\mathrm{U}}_{m\times m}\bar{\Sigma}_{m\times n}\bar{\mathrm{V}}^{\mathrm{T}}_{n\times n}$$



$$\mathrm{A}_{|V|\times|\mathbf{D}|} = \bar{\mathrm{U}}_{|V|\times|V|}\bar{\Sigma}_{|V|\times|\mathbf{D}|}\bar{\mathrm{V}}^{\mathrm{T}}_{|\mathbf{D}|\times|\mathbf{D}|} \approx \mathrm{U}_{|V|\times K}\Sigma_{K\times K}\mathrm{V}^{\mathrm{T}}_{K\times|\mathbf{D}|} = \mathrm{A}'_{|V|\times|\mathbf{D}|}$$

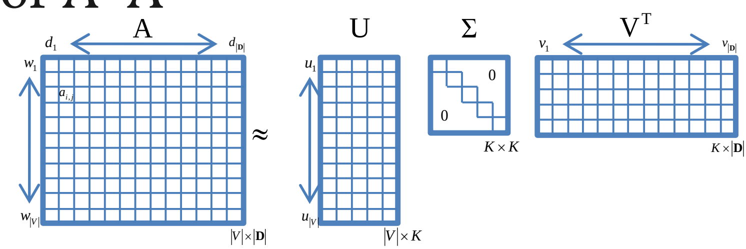
- The result is equivalent to the minimize the objective function $min\|\mathrm{A}-\mathrm{A}'\|_F^2$, for a given K



$$\|\mathrm{B}\|_F^2 = \sum_i\sum_j b_{i,j}^2$$

Latent Semantic Analysis….

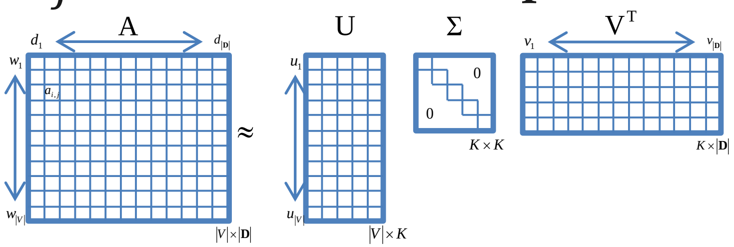
- Properties of SVD decomposition
  - Both left and right singular matrices (i.e., U and V) are column-orthonormal
    - $\mathrm{U^TU = V^TV = I}$
  - Values (nonnegative real numbers) in diagonal matrix are square roots of the eigenvalues of $\mathrm{A^TA}$
    - $\Sigma^2 = \mathrm{diag}\{\lambda_1, \lambda_2, \dots, \lambda_K\}$
    - $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K \geq 0$



  - The column vectors of U define an orthonormal basis for d_j
    - $\mathrm{A \approx U\Sigma V^T \ \Rightarrow\ A^TU \approx (U\Sigma V^T)^T U = V\Sigma U^TU = V\Sigma \Rightarrow U^TA = \Sigma V^T}$
  - The column vectors of V define an orthonormal basis for w_i
    - $\mathrm{A \approx U\Sigma V^T \ \Rightarrow\ AV \approx (U\Sigma V^T)V = U\Sigma V^TV = U\Sigma \Rightarrow V^TA^T = \Sigma U^T}$

Latent Semantic Analysis.....

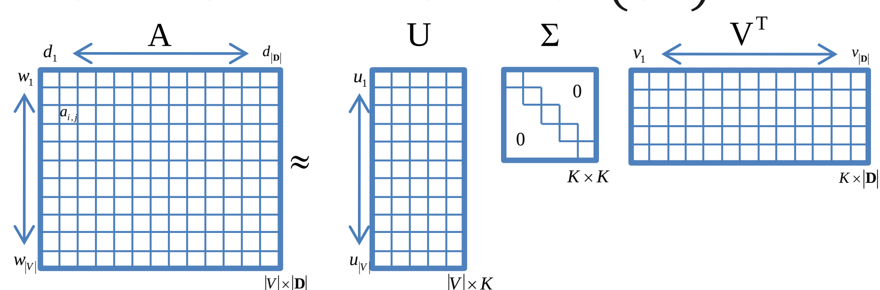
- New representations
  - For word w_i, the new vector representation is Σu_i^{T}
  - For document d_j, the new vector representation is Σv_j^{T}



- While the original high-dimensional vectors are sparse, the corresponding low-dimensional latent vectors will typically not be sparse
  - It is possible to compute meaningful association values between pairs of documents, even if the documents do not have any terms in common
    - The hope is that terms having a common meaning (synonyms), are roughly mapped to the same direction in the latent space

Latent Semantic Analysis……

- Based on LSA
  - Compare two words w_i and w_j
    - $(\Sigma u_i^{\mathrm{T}})^{\mathrm{T}} \Sigma u_j^{\mathrm{T}} = u_i \Sigma \Sigma u_j^{\mathrm{T}} = u_i \Sigma (u_j \Sigma)^{\mathrm{T}}$
    - $\mathrm{A} \approx \mathrm{U} \Sigma \mathrm{V}^{\mathrm{T}} \Rightarrow \mathrm{A}\mathrm{A}^{\mathrm{T}} \approx \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}(\mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}})^{\mathrm{T}} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}\mathrm{V}\Sigma^{\mathrm{T}}\mathrm{U}^{\mathrm{T}} = \mathrm{U}\Sigma(\mathrm{U}\Sigma)^{\mathrm{T}}$
  - Compare two documents
    - $(\Sigma v_i^{\mathrm{T}})^{\mathrm{T}} \Sigma v_j^{\mathrm{T}} = v_i \Sigma \Sigma v_j^{\mathrm{T}} = v_i \Sigma (v_j \Sigma)^{\mathrm{T}}$
    - $\mathrm{A} \approx \mathrm{U} \Sigma \mathrm{V}^{\mathrm{T}} \Rightarrow \mathrm{A}^{\mathrm{T}}\mathrm{A} \approx (\mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}})^{\mathrm{T}}\mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}} = \mathrm{V}\Sigma^{\mathrm{T}}\mathrm{U}^{\mathrm{T}}\mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}} = \mathrm{V}\Sigma(\mathrm{V}\Sigma)^{\mathrm{T}}$



  - Compare words and documents
    - $(\Sigma u_i^{\mathrm{T}})^{\mathrm{T}} \Sigma v_j^{\mathrm{T}} = u_i \Sigma \Sigma v_j^{\mathrm{T}} = u_i \Sigma (v_j \Sigma)^{\mathrm{T}}$?
    - **Usually the reconstructed matrix is used**

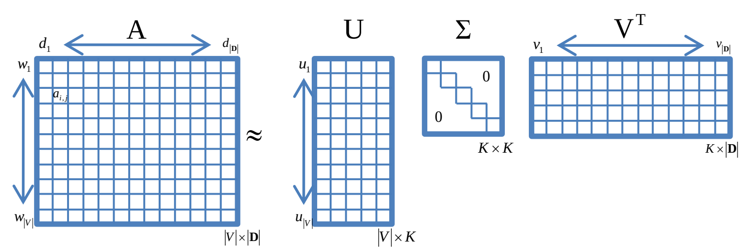
$$\mathrm{A}_{|V|\times|\mathbf{D}|} = \bar{\mathrm{U}}_{|V|\times|V|}\bar{\Sigma}_{|V|\times|\mathbf{D}|}\bar{\mathrm{V}}^{\mathrm{T}}_{|\mathbf{D}|\times|\mathbf{D}|} \approx \mathrm{U}_{|V|\times K}\Sigma_{K\times K}\mathrm{V}^{\mathrm{T}}_{K\times|\mathbf{D}|} = \mathrm{A}'_{|V|\times|\mathbf{D}|}$$

LSA for IR.

- For a given query (as a document), a low-dimensional representation should be inferred
  - The low-dimensional representation can be obtained by using the **fold-in** strategy
    - The column vectors of U define an orthonormal basis for d_j
      - $\mathrm{A} \approx \mathrm{U\Sigma V^T}$
        $\Rightarrow \mathrm{A^T U} \approx \left(\mathrm{U\Sigma V^T}\right)^{\mathrm{T}} \mathrm{U} = \mathrm{V\Sigma U^T U} = \mathrm{V\Sigma}$
        $\Rightarrow \mathrm{U^T A} = \mathrm{\Sigma V^T}$
      - For each document, the new representation is $\Sigma {v_i}^{\mathrm{T}}$
    - For the new query, the low-dimensional representation can be derived by the same mechanism

$$\left(\mathrm{U}_{|V|\times K}\right)^{\mathrm{T}} (\vec{q})_{|V|\times 1} = \Sigma_{K\times K} v_q^{\mathrm{T}}$$

$$\Sigma_{K\times K}^{-1} \left(\mathrm{U}_{|V|\times K}\right)^{\mathrm{T}} (\vec{q})_{|V|\times 1} = (v_q^{\mathrm{T}})_{1\times K}$$



Each dimension has its own weight

Weighted sum of the word vectors

LSA for IR..

- For a given query (as a document), a low-dimensional representation should be inferred
  - The low-dimensional representation can be obtained by using the **fold-in** strategy
    - For the new query, the low-dimensional representation can be derived by the same mechanism

$$\left(\mathrm{U}_{|V|\times K}\right)^{\mathrm{T}}(\vec{q})_{|V|\times 1} = \Sigma_{K\times K} v_q^{\mathrm{T}}$$

$$\Sigma_{K\times K}^{-1}\left(\mathrm{U}_{|V|\times K}\right)^{\mathrm{T}}(\vec{q})_{|V|\times 1} = (v_q^{\mathrm{T}})_{1\times K}$$

    - Notably, for a new document, the representation can also be derived by the fold-in strategy
  - Consequently, the relevance degree can be computed:

$$sim(q,d) = cos\left(\Sigma v_q^{\mathrm{T}}, \Sigma v_d^{\mathrm{T}}\right) = \frac{\Sigma v_q^{\mathrm{T}} \cdot \Sigma v_d^{\mathrm{T}}}{|\Sigma v_q^{\mathrm{T}}||\Sigma v_d^{\mathrm{T}}|}$$

Example – 1.

A =

	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
voyage	1	0	0	1	1	0
trip	0	0	0	1	0	1

U =

	1	2
ship	−0.44	−0.30
boat	−0.13	−0.33
ocean	−0.48	−0.51
voyage	−0.70	0.35
trip	−0.26	0.65

Σ =

2.16	0.00
0.00	1.59

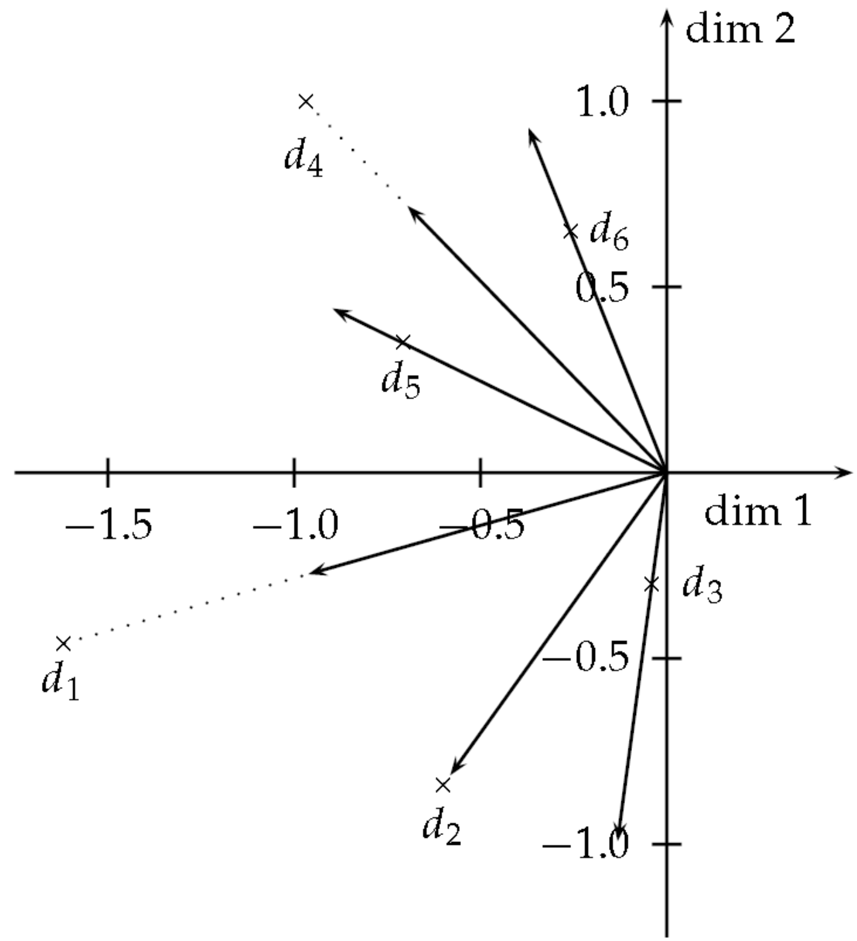
V^T =

	d_1	d_2	d_3	d_4	d_5	d_6
1	−1.62	−0.60	−0.44	−0.97	−0.70	−0.26
2	−0.46	−0.84	−0.30	1.00	0.35	0.65

Example – 1..

- The relationship between d_2 and d_5 can be reasonable determined

	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
voyage	1	0	0	1	1	0
trip	0	0	0	1	0	1



Example – 2.

c1: *Human* machine *interface* for Lab ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

	Terms	Documents								
		c1	c2	c3	c4	c5	m1	m2	m3	m4
1	*human*	1	0	0	1	0	0	0	0	0
2	*interface*	1	0	1	0	0	0	0	0	0
3	*computer*	1	1	0	0	0	0	0	0	0
4	*user*	0	1	1	0	1	0	0	0	0
5	*system*	0	1	1	2	0	0	0	0	0
6	*response*	0	1	0	0	1	0	0	0	0
7	*time*	0	1	0	0	1	0	0	0	0
8	*EPS*	0	0	1	1	0	0	0	0	0
9	*survey*	0	1	0	0	0	0	0	0	1
10	*trees*	0	0	0	0	0	1	1	1	0
11	*graph*	0	0	0	0	0	0	1	1	1
12	*minors*	0	0	0	0	0	0	0	1	1

Example – 2..

Query="human computer interaction"

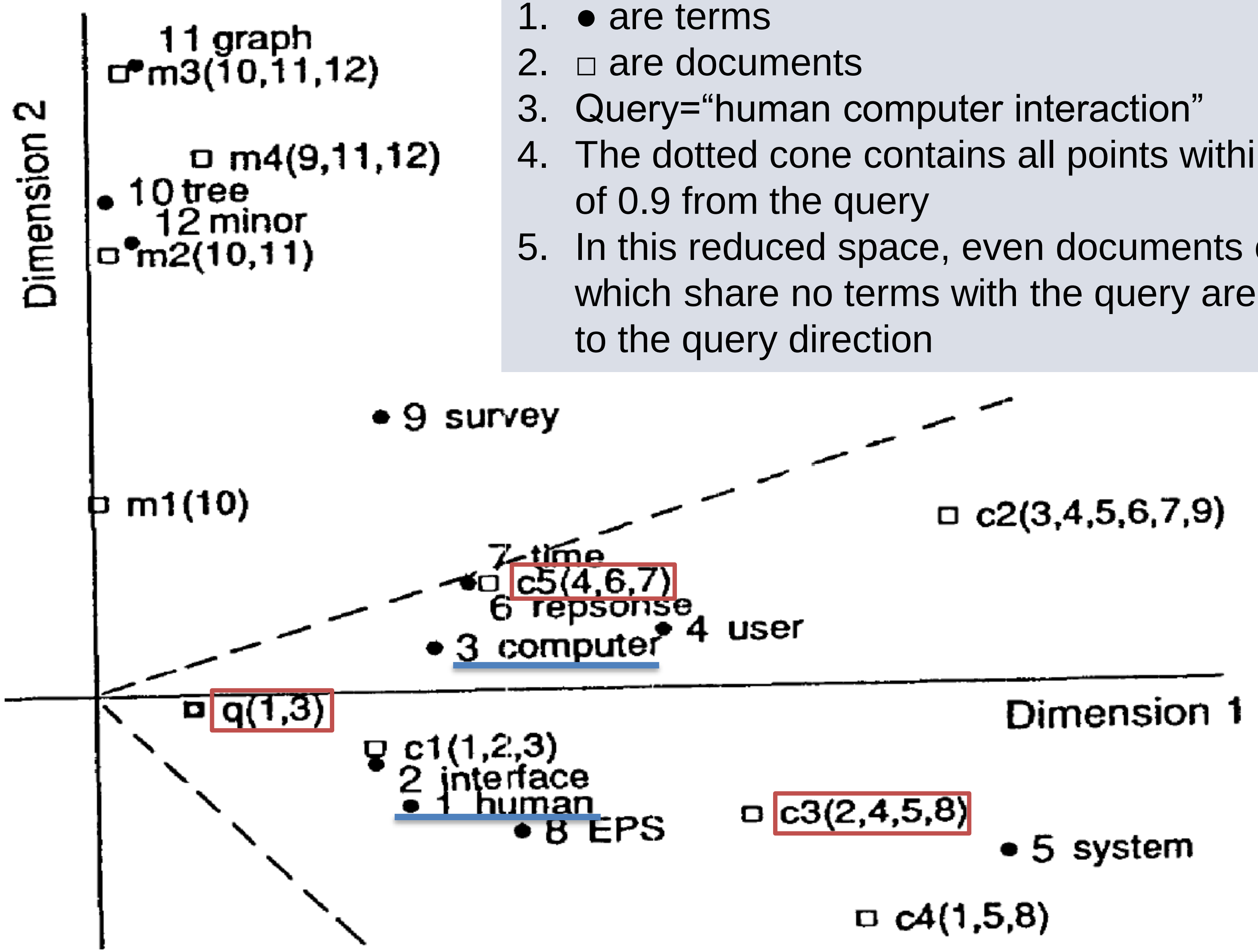
c1: *Human* machine *interface* for Lab ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

	Terms	Documents								
		c1	c2	c3	c4	c5	m1	m2	m3	m4
1	*human*	1	0	0	1	0	0	0	0	0
2	*interface*	1	0	1	0	0	0	0	0	0
3	*computer*	1	1	0	0	0	0	0	0	0
4	*user*	0	1	1	0	1	0	0	0	0
5	*system*	0	1	1	2	0	0	0	0	0
6	*response*	0	1	0	0	1	0	0	0	0
7	*time*	0	1	0	0	1	0	0	0	0
8	*EPS*	0	0	1	1	0	0	0	0	0
9	*survey*	0	1	0	0	0	0	0	0	1
10	*trees*	0	0	0	0	0	1	1	1	0
11	*graph*	0	0	0	0	0	0	1	1	1
12	*minors*	0	0	0	0	0	0	0	1	1

Example – 2..

1. ● are terms
2. □ are documents
3. Query="human computer interaction"
4. The dotted cone contains all points within a cosine of 0.9 from the query
5. In this reduced space, even documents c3 and c5, which share no terms with the query are very close to the query direction



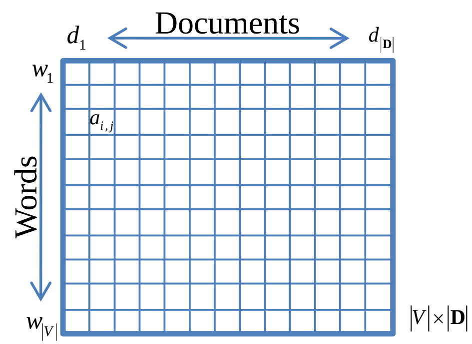
Pros and Cons

- Advantages
  - As we reduce K, **recall tends to increase**, as expected
  - Most surprisingly, a value of K in the low hundreds can actually **increase precision** on some query benchmarks
  - Finding new spaces for words and documents

- Disadvantages
  - The computational cost of the SVD is significant
  - Irrelevant or Antonymous
  - The reconstruction has negative entities

Entropy-based Weighting Method.

- In the word-by-document, each element $a_{i,j}$ is represented the importance of word w_i in document d_j
  - The TF-IDF score
  - The Entropy-based method



$$a_{i,j} = (1 - \varepsilon_i) \frac{c(w_i, d_j)}{|d_j|}$$

$$\varepsilon_i = -\frac{1}{log\,|\mathbf{D}|} \sum_{j=1}^{|\mathbf{D}|} \left(\frac{c(w_i, d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i, d_{j'})} log \frac{c(w_i, d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i, d_{j'})} \right)$$

- $0 \le \varepsilon_i \le 1$
  - $\varepsilon_i = 1 \Rightarrow \forall d_j,\ \ c(w_i, d_j) = \frac{\sum_{j'=1}^{|\mathbf{D}|} c(w_i, d_{j'})}{|\mathbf{D}|}$: the word distributed across many documents throughout the corpus
  - $\varepsilon_i = 0 \Rightarrow \exists d_j,\ \ c(w_i, d_j) \approx \sum_{j'=1}^{|\mathbf{D}|} c(w_i, d_{j'})$: the word is present only in a few specific documents

Entropy-based Weighting Method..

$$\varepsilon_i = -\frac{1}{log\,|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\,log\,\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\right)$$

- $\varepsilon_i = 1 \Rightarrow \forall d_j,\ \ c(w_i,d_j) = \frac{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}{|\mathbf{D}|}$: the word distributed across many documents throughout the corpus

$$\varepsilon_i = -\frac{1}{log|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\,log\,\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\right)$$

$$= -\frac{1}{log|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{\frac{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}{|\mathbf{D}|}}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\,log\,\frac{\frac{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}{|\mathbf{D}|}}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\right)$$

$$= -\frac{1}{log|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{1}{|\mathbf{D}|}\,log\,\frac{1}{|\mathbf{D}|}\right) = -\frac{1}{log|\mathbf{D}|}\left(log\,\frac{1}{|\mathbf{D}|}\right) = -\frac{1}{log|\mathbf{D}|}(-\,log|\mathbf{D}|) = 1$$

$$a_{i,j} = (1-\varepsilon_i)\frac{c(w_i,d_j)}{|d_j|}$$

Entropy-based Weighting Method…

$$\varepsilon_i = -\frac{1}{log\,|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\,log\,\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\right)$$

- $\varepsilon_i = 0 \Rightarrow \exists d_j,\ \ c(w_i,d_j) \approx \sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})$: the word is present only in a few specific documents

$$\begin{aligned}\varepsilon_i &= -\frac{1}{\log|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\,log\,\frac{c(w_i,d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\right)\\ &= -\frac{1}{\log|\mathbf{D}|}\times(|\mathbf{D}|-1)\times\left(\frac{0}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\,log\,\frac{0}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\right)\\ &\quad -\frac{1}{\log|\mathbf{D}|}\times\left(\frac{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\,log\,\frac{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i,d_{j'})}\right)\\ &= 0\end{aligned}$$

$$a_{i,j} = (1-\varepsilon_i)\frac{c(w_i,d_j)}{|d_j|}$$

LSA-based Language Modeling – 1

- A goal of statistical language modeling is to learn the joint probability function of sequences of words in a language
  - By using n-gram model

$$P(w_1, w_2, \ldots, w_T) \approx \prod_{t=1}^{T} P(w_t | w_{t-n+1}, \ldots, w_{t-1})$$

  - By incorporating n-gram model and LSA-based model

$$P(w_1, w_2, \ldots, w_T) \approx \prod_{t=1}^{T} P\left(w_t | H_{t-1}^{n,l}\right) = \prod_{t=1}^{T} P\left(w_t | \underline{H_{t-1}^{n}}, \underline{H_{t-1}^{l}}\right)$$

Lexical Information

Semantic Information

LSA-based Language Modeling – 2

- The probability can further be decomposed:

$$P\left(w_t | H_{t-1}^n, H_{t-1}^l\right) = \frac{P\left(w_t, H_{t-1}^l | H_{t-1}^n\right)}{\sum_{w_i \in V} P\left(w_i, H_{t-1}^l | H_{t-1}^n\right)}$$

- Expanding and rearranging, the numerator is seen to be:

$$\begin{aligned}
P\left(w_t, H_{t-1}^l | H_{t-1}^n\right) &= \frac{P\left(w_t, H_{t-1}^l, H_{t-1}^n\right)}{P\left(H_{t-1}^n\right)} \\
&= \frac{P\left(w_t, H_{t-1}^l, H_{t-1}^n\right) P\left(w_t, H_{t-1}^n\right)}{P\left(H_{t-1}^n\right) P\left(w_t, H_{t-1}^n\right)} \\
&= P\left(w_t | H_{t-1}^n\right) P\left(H_{t-1}^l | w_t, H_{t-1}^n\right) \\
&= P\left(w_t | w_{t-n+1}, \dots, w_{t-1}\right) P\left(H_{t-1}^l \middle| w_{t-n+1}, \dots, w_{t-1}, w_t\right) \\
&= P\left(w_t | w_{t-n+1}, \dots, w_{t-1}\right) P\left(H_{t-1}^l \middle| w_t\right)
\end{aligned}$$

We assume the probability of the document history given the current word is not affected by other context words

LSA-based Language Modeling – 3

$$P\left(w_t, H_{t-1}^l \middle| H_{t-1}^n\right) = P(w_t | w_{t-n+1}, \dots, w_{t-1}) P\left(H_{t-1}^l \middle| w_t\right)$$

- Consequently, we can obtain:

$$P\left(w_t \middle| H_{t-1}^{n,l}\right) = P\left(w_t \middle| H_{t-1}^n, H_{t-1}^l\right) = \frac{P\left(w_t, H_{t-1}^l \middle| H_{t-1}^n\right)}{\sum_{w_i \in V} P\left(w_i, H_{t-1}^l \middle| H_{t-1}^n\right)}$$

$$= \frac{P(w_t | w_{t-n+1}, \dots, w_{t-1}) P\left(H_{t-1}^l \middle| w_t\right)}{\sum_{w_i \in V} P(w_i | w_{t-n+1}, \dots, w_{t-1}) P\left(H_{t-1}^l \middle| w_i\right)} = \frac{P(w_t | w_{t-n+1}, \dots, w_{t-1}) \frac{P(w_t | H_{t-1}^l)}{P(w_t)}}{\sum_{w_i \in V} P(w_i | w_{t-n+1}, \dots, w_{t-1}) \frac{P(w_i | H_{t-1}^l)}{P(w_i)}}$$

- H_{t-1}^l can be represented by a vector in the semantic space

$$\left(\overrightarrow{H_{t-1}^l}'^{\mathrm{T}}\right)_{1 \times K} = \left(\overrightarrow{H_{t-1}^l}^{\mathrm{T}}\right)_{1 \times |V|} \mathrm{U}_{|V| \times K} \Sigma_{K \times K}^{-1}$$

- Thus, the semantic smoothing factor can be estimated by:

$$P(w_t | H_{t-1}^l) \propto \cos(\Sigma^{\frac{1}{2}} \overrightarrow{H_{t-1}^l}', \Sigma^{\frac{1}{2}} u_{w_t}^{\mathrm{T}})$$

LSA-based Language Modeling – Appendix

- By using the entropy-based method to score each element in the vector, a fast strategy can be derived for sequential data

$$\overrightarrow{H_t^l} = \frac{|H_t^l| - 1}{|H_t^l|}\overrightarrow{H_{t-1}^l} + \frac{1 - \varepsilon_{w_t}}{|H_t^l|}\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$a_{i,j} = (1 - \varepsilon_i)\frac{c(w_i, d_j)}{|d_j|}$$

$$\varepsilon_i = -\frac{1}{log\,|\mathbf{D}|}\sum_{j=1}^{|\mathbf{D}|}\left(\frac{c(w_i, d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i, d_{j'})}\,log\,\frac{c(w_i, d_j)}{\sum_{j'=1}^{|\mathbf{D}|} c(w_i, d_{j'})}\right)$$

The Evolution.

Scott Deerwester

- 1998 Language Modeling Approaches
- 1994 Best Match Models (Okapi Systems)
- **1988 Latent Semantic Analysis**
- 1976 Probabilistic Model
- 1975 Vector Space Model
- 1973 Boolean Model
- 1972 Inverse Document Frequency
- 1957 Term Frequency

September 1991 - June 1995
The Hong Kong University of Science and Technology
Department of Computer Science and Engineering · null, Hong Kong
Position
Lecturer

January 1985 - August 1991
University of Chicago
Chicago, United States
Position
Professor (Assistant)

August 1983 - December 1984
Colgate University
null, United States
Position
Professor (Assistant)

The Evolution..

Scott Deerwester

1998 Language Modeling Approaches

1994 Best Match Models (Okapi Systems)

1988 Latent Semantic Analysis

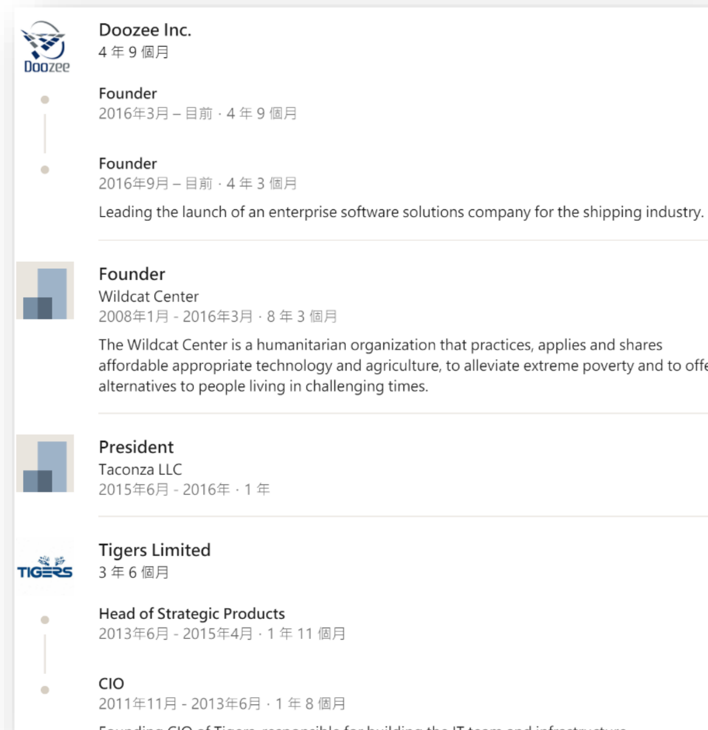
1976 Probabilistic Model

1975 Vector Space Model

1973 Boolean Model

1972 Inverse Document Frequency

1957 Term Frequency



Doozee Inc.
4 年 9 個月

Founder
2016年3月 – 目前 · 4 年 9 個月

Founder
2016年9月 – 目前 · 4 年 3 個月
Leading the launch of an enterprise software solutions company for the shipping industry.

Founder
Wildcat Center
2008年1月 - 2016年3月 · 8 年 3 個月
The Wildcat Center is a humanitarian organization that practices, applies and shares affordable appropriate technology and agriculture, to alleviate extreme poverty and to offer alternatives to people living in challenging times.

President
Taconza LLC
2015年6月 - 2016年 · 1 年

Tigers Limited
3 年 6 個月

Head of Strategic Products
2013年6月 - 2015年4月 · 1 年 11 個月

CIO
2011年11月 - 2013年6月 · 1 年 8 個月
Founding CIO of Tigers, responsible for building the IT team and infrastructure.

Questions?

kychen@mail.ntust.edu.tw