1. Vector (space)model Introduction

\[ D \subseteq \mathbb{R}^{n^+} \]

\[ Q \subseteq \mathbb{R}^n \]

Retrieval functions

\[ f : D \times Q \to \mathbb{R} \]

\[ \mathbf{d} = (d_1, d_2, \ldots, d_n) \]

\[ \mathbf{q} = (q_1, q_2, \ldots, q_n) \]

Dot product function

\[ \mathbf{d}\mathbf{q}^T = \sum_{i=1}^{n} d_iq_i \]
2. **TWO VIEWS OF VECTOR CONCEPT**

-- VECTOR (PROCESSING) “MODEL”

NOTATIONAL OR DATA STRUCTURAL ASPECT

-- VECTOR **SPACE** MODEL

- DOCUMENTS, QUERIES, ETC. ARE ELEMENTS OF A VECTOR SPECE

- ANALYTICAL TOOL
3. **THE VECTOR SPACE MODEL**

- **MATHEMATICAL ASPECTS**

- **MAPPING OF DATA ELEMENTS TO MODEL CONSTRUCTS**
3.1 MATHEMATICAL ASPECTS
3.1.1 BASIC CONCEPTS

• IR OBJECTS (e.g. KEYWORDS DOCUMENTS) CONSTITUTE A VECTOR SPACE

• THAT IS, WE HAVE A SYSTEM WITH LINEAR PROPERTIES:
(i) ADDITION OF VECTORS
(ii) MULTIPLICATION BY SCALAR

CLOSURE

• BASIC ALGEBRAIC AXIOMS
  e.g. \( x + y = y + x \)
  \( x + o = x \) i.e. \( o \) exists
  For each \( x \), \( \exists -x \)
  \( \alpha (x + y) = \alpha x + \alpha y \)
  .
  .
  etc
LINEAR INDEPENDENCE

A SET OF VECTORS $y_1, y_2 \ldots y_k$ IS LINEARLY INDEPENDENT (L.I.) IF

$$\alpha_1 y_1 + \alpha_2 y_2 + \ldots + \alpha_k y_k = 0,$$

WHERE $\alpha_i$'S ARE SCALARS,

ONLY IF $\alpha_1 = \alpha_2 = \ldots \alpha_K = 0$

- BASIS: A GENERATING SET CONSISTING OF L.I. VECTORS
- DIMENSION: $n' \leq n$, where $n$ is the size of the generating set
- \{ $t_{i1}, t_{i2}, \ldots, t_{in'}$ \}
- ANY subset of L.I. VECTORS of the generating set of size $n'$ FORM A BASIS
(Inner) SCALAR PRODUCT

\[ \mathbf{x} \cdot \mathbf{y} = ||\mathbf{x}|| \ ||\mathbf{y}|| \cos \theta, \]

WHERE,

\[ \theta \] is the angle between \( \mathbf{x} \) and \( \mathbf{y} \),

\[ ||\mathbf{x}|| = \sqrt{\mathbf{x} \cdot \mathbf{x}} \]

\[ \]

- The above is an instance of a scalar product
- EUCLIDEAN SPACE: A VECTOR SPACE EQUIPPED WITH A SCALAR PRODUCT
- ORTHOGONAL: \( \mathbf{x} \cdot \mathbf{y} = 0 \)
- NORMALIZING: \( \mathbf{x} / ||\mathbf{x}|| \)
- ORTHONORMAL BASIS

If underlying basis is orthonormal,

\[ \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i \]
3.1.2 **LINEAR INDEPENDENCE VS. ORTHOGONALITY**

IF A SET OF NON-ZERO VECTORS $y_1, y_2 \ldots y_k$ are **MUTUALLY ORTHOGONAL** ($x_i \cdot y_j = 0$ for all $i \neq j$), then they are LINEARLY INDEPENDENT. But a set of linearly independent vectors is not necessarily mutually orthogonal.

UNDER THE SITUATION OF NON-ORTHOGONAL Generating set, issues of

(i) linear dependence, and

(ii) correlation *

MUST BE CONSIDERED.

* (term, term) relationship
3.1.3 REPRESENTATION IN IR

KEYWORDS:

\[ t_1, t_2, t_3 \ldots t_n \]

VECTORS:

\[ \frac{t_1, t_2, t_3 \ldots t_n}{\text{Generating set}} \]

\[ d_\alpha = (a_{1\alpha}, a_{2\alpha}, \ldots a_{n\alpha}) \]

OR

\[ d_\alpha = \sum_{i=1}^{n} a_{i\alpha} t_i \]
3.1.4 **IMPORTANT RELATIONSHIPS**

**ASSUME:**

\[ n' = n = p \]
\[ t_1, t_2, \ldots, t_n \]
\[ d_1, d_2, \ldots, d_n \]

Basis can be either

\[ ||t_i|| = 1, \ I = 1, 2, \ldots, n \]

**THUS,**

\[ d_\alpha = \sum_{i=1}^{n} a_{i\alpha} t_i \quad \ldots \quad (1) \]

**OR**

\[ t_i = \sum_{\alpha=1}^{n} b_{\alpha i} d_\alpha \quad \ldots \quad (2) \]
Projection and component are NOT the same, when the basis vectors are non-orthogonal.
3.1.5 **PROJECTION** **VS.** **COMPONENTS**

FOR VECTORS, \( \mathbf{x}, \mathbf{y} \)  
\[
\left( \frac{\mathbf{x}}{||\mathbf{x}||} \right) \cdot \mathbf{y} \text{ IS THE PROJECTION OF } \mathbf{y} \text{ ONTO } \mathbf{x}.
\]

3.1.4 (Contd.)

By MULTIPLYING equ. (1) by \( t_j \) ON BOTH SIDES,

\[
t_j \cdot \mathbf{d}_{\alpha} = \sum_{i=1}^{n} a_i \alpha t_j \cdot t_i,
\]

\( 1 \leq \alpha, j \leq n \ldots (3) \)

If \( t \)'s ARE NORMALIZED, THE LEFT HAND SIDE IS THE PROJECTION OF \( \mathbf{d}_{\alpha} \) ONTO \( t_j \)
WRITING EQN. (3) IN A MATRIX FORM, WE HAVE

\[ P = G_t A \ldots (4) \]

WHERE

\[ (P)_{j\alpha} = t_j \cdot d_\alpha \]
\[ (G_t)_{ji} = t_j \cdot t_i \]
\[ (A)_{i\alpha} = a_{i\alpha} \]

RESPECTIVELY,

PROJECTIONS,

TERM CORRELATIONS

&

COMPONENTS OF \( d \)'s

EXAMPLE 1
\[ n = 2 \]

\[ t_2 \cdot D_\alpha \]

\[ a_{2\alpha} \]

\[ d_\alpha = a_{1\alpha} t_1 + a_{2\alpha} t_2 \ldots \quad (5) \]

LET \( d_1, d_2 \) BE A BASIS (L.I.)

THEN,

\[ G_t A = \begin{bmatrix} t_1 \cdot t_1 & t_1 \cdot t_2 \\ t_2 \cdot t_1 & t_2 \cdot t_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \]
\[
= \begin{bmatrix}
t_1 \cdot (a_{11} t_1 + a_{21} t_2) & t_1 \cdot (a_{12} t_1 + a_{22} t_2) \\
t_2 \cdot (a_{11} t_1 + a_{21} t_2) & t_2 \cdot (a_{12} t_1 + a_{22} t_2)
\end{bmatrix}
\]

USING EQN. (5), WE HAVE

\[
= \begin{bmatrix}
t_1 \cdot d_1 & t_1 \cdot d_2 \\
t_2 \cdot d_1 & t_2 \cdot d_2
\end{bmatrix}
\]

= P
SIMILARLY,

STARTING FROM EQN. (2) AND MULTIPLYING BOTH SIDES BY $d_\beta$, AND WRITING IN MATRIX FORM.

$$P^T = G_d B \ldots (6)$$

WHERE

$$(G_d)_{\beta\alpha} = d_\beta \cdot d_\alpha$$

$$(B)_{\alpha i} = b_{\alpha i}$$

THAT IS,

DOCUMENT CORRELATIONS AND

COMPONENTS OF t’s ALONG DOCUMENTS

CAN further SHOW,

$$PB = G_t \ldots (7)$$

$$P^TA = G_d \ldots (8)$$
3.1.6 DOCUMENT RANKING

\[ q = \sum_{i=1}^{n} q_i t_i \]

\[ d_{\alpha} \cdot q = \left( \sum_{i=1}^{n} a_{i\alpha} t_i \right) \cdot \left( \sum_{j=1}^{n} q_j t_j \right) \]

\[ = \sum_{i,j=1}^{n} a_{i\alpha} q_j t_i t_j \quad (9) \]

EXAMPLE 2

\[ n=2 \]

\[ A^T G_t q^T \]

\[ q = q_1 t_1 + q_2 t_2 \]

\[ d_{\alpha} = a_{1\alpha} t_1 + a_{2\alpha} t_2 \]

\[ d_{\alpha} q = a_{1\alpha} q_1 t_1 \cdot t_1 \]

\[ + a_{2\alpha} q_2 t_2 \cdot t_2 \]

\[ + a_{1\alpha} q_2 t_2 \cdot t_1 \]

\[ + a_{2\alpha} q_1 t_1 \cdot t_2 \]
3.2 MAPPING OF DATA ELEMENTS TO MODEL CONSTRUCTS

Term Frequency Data

\[
W = u_m_e_n_t \quad \text{term}
\]

\[
d_o_c_w = u_{\alpha_i}
\]

May be interpreted as

\[A^T \text{ or } B \text{ or } P^T\]

But, this alone is **NOT** enough

*By interpretation we mean how data obtained from real-world documents are mapped to model constructs such as, A, B and G_t.*
Text Analysis

- Controlled vs. Free vocabulary
- Single term Indexing
  a. Extract words
  b. Stop list
  c. Stemming
  d. Term weight assignment

\[
\text{RSV} (q, d_{\alpha}) = \sum_i \left( 0.5 + 0.5 \frac{f_{ci}}{\max_j (f_{cj})} \right) \log \left( \frac{N}{n_i} \right)
\]

- More general descriptions
  a. phrases
  b. thesaurus entries
3.2.1 **TWO WAYS OF MAPPING W TO THE MODEL**

Method I. Mapping $W^T$ to $A$

$A = W^T$

$RSV_q = (d_1 \cdot q, d_2 \cdot q, \ldots, d_p \cdot q)$

$q = (q_1, q, \ldots q_n)$

$q_i$ is the component of $q$ along $t_i$

$RSV_q^T = W G_t q^T$

$= P^T q^T$, since

$P^T = A^T G_t = W G_t$, then

$P = G_t A \quad RSV_q^T = W q^T$
\[ n = 2 \]

\[ \begin{bmatrix}
    t_1 \\
    t_2
\end{bmatrix}
\]

\[ = G_t \]

\[ \begin{bmatrix}
    t_1 \\
    t_2
\end{bmatrix}
\]

\[ \begin{bmatrix}
    1 & 0 \\
    0 & 1
\end{bmatrix} \]

\[ d_\alpha = (3, 3) \]

\[ q = (3, 1) \]

\[ d_\alpha = 3t_1 + 3t_2 \]

\[ q = 3t_1 + t_2 \]

\[ d_\alpha \cdot q = |d_\alpha| |q| \cos \theta \]

\[ = \sum_{i=1}^{2} a_{\alpha i} q_i \]

\[ (3t_1 + t_2) \cdot (3t_1 + 3t_2) \]

\[ = 9t_1 \cdot t_1 + 9t_1 \cdot t_2 + 3t_2 t_1 + 3t_2 t_2 \]

\[ = 12 \]
Method II. \( B = W \)  

\[
R S V_q^T = P^T q^T \\
\downarrow \\
G_d B
\]

\[
R S V_q^T = G_d B q^T \\
= G_d w q^T
\]

- Columns of \( W \) are used as components of term vectors along document vectors
- Elements of \( q \) are components of \( q \) along term vectors
3.2.2 USING THE MODEL
COMPARISON TO EARLIER WORK

I. THE STANDARD SPECIAL CASE

• TERMS FORM AN ORTHONORMAL BASIS, \( G_t = I \)
• HERE, \( P = A \) (FROM(4) )
• \( W \) IS INTERPRETED AS
\[
A^T ( = P^T ) \sum_{i=1}^{n} a_{i\alpha} \cdot q_i \quad \text{when} \quad G_t = I
\]

In this case
\[
d_{\alpha} \cdot q = \sum_{i=1}^{n} a_{i\alpha} \cdot q_i
= \sum_{i=1}^{n} w_{\alpha i} \cdot q_i
\]
II. While the above restrictions appear compatible, one of the practices defines term vector $t_i$ as follow:

$$t_i = (w_{1i}, w_{2i}, \ldots w_{ni})$$

This suggests,

$$A^t = B$$

But, according to the vector space model,

$$P = G_tA$$

and

$$PB = G_t$$

Thus, $A^{-1} = B$

If each row of $W$ represents documents, then each column does not represent term vector, thus, what is known to be common practice is contradictory to what we show to be the relationship between $A$ and $B$ matrices.
Can Projection be negative?

Projection of \( \vec{x} \) on \( \vec{y} \) is +

Projection of \( \vec{x} \) on \( \vec{y} \) is -
### 3.2.3 Other commonly used retrieved functions

<table>
<thead>
<tr>
<th>Similarity Measure</th>
<th>Measures of vector similarity</th>
<th>Evaluation for Binary Term Vectors</th>
<th>Evaluation for Weighted Term Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>sim(X, Y)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inner product</td>
<td></td>
<td>X \cap Y</td>
<td>\sum_{i=1}^{t} x_i y_i</td>
</tr>
<tr>
<td>Dice coefficient</td>
<td>\frac{2</td>
<td>X \cap Y</td>
<td>}{</td>
</tr>
<tr>
<td>Cosine coefficient</td>
<td>\frac{</td>
<td>X \cap Y</td>
<td>}{</td>
</tr>
<tr>
<td>Jaccard coefficient</td>
<td>\frac{</td>
<td>X \cap Y</td>
<td>}{</td>
</tr>
</tbody>
</table>

X = \{t_i\}
Y = \{t_j\}

\(X = (x_1, x_2, \ldots, x_t)\)
|X| = number of terms in X
|X \cap Y| = number of terms appearing jointly in X and Y