นายภัทรชัย วรรณสินทัต 65010815

# Boolean Model → No Ranking

Ex. D1 = (0.8A , 0.5B , 0.6C) * and ใช้ min

D2 = (0.4A , 0.5B , 0.1C , 0.8D) or ใช้ max

Query = (0.5A and 0.2B) or (not D or 0.3C)

| | A and B | Not D | Relevance |
|---|---|---|---|
| D1 | min (0.5×0.8 , 0.2×0.5) = 0.1 | max (1 , 0.3×0.6) = 1 | max (0.1 , 1) = 1 |
| D2 | min (0.5×0.4 , 0.2×0.4) = 0.08 | max (0.2 , 0.3×0.1) = 0.2 | max (0.08 , 0.2) = 0.2 |

# Vector Model → ไม่มี ดสพ. ระหว่างข้อมูล

query weight : $w_{i,q} = (0.5 + 0.5\, f_{iq} / \max f_{iq}) \times idf_i$

doc weight : $w_{ij} = tf_{ij} \times idf_i$

similarity score : $sim(q, d_i) = \dfrac{\sum w_{q_j} w_{d_{ij}}}{\sqrt{\sum (w_{q_j})^2 \sum (w_{d_{ij}})^2}}$

kw/doc : $tf_{ij} = f_{ij} / \max f_{ij}$

All Doc/Kw Doc : $idf_i = \log(N/n_i)$

Ex. Query = "cat dog tiger cat"

| | Bird | Cat | Dog | Tiger | Max |
|---|---|---|---|---|---|
| Doc1 | 3 | 2 | 2 | 0 | 3 |
| Doc2 | 0 | 2 | 1 | 1 | 2 |
| Doc3 | 2 | 0 | 1 | 0 | 2 |
| Doc4 | 0 | 1 | 0 | 1 | 1 |
| Doc5 | 0 | 1 | 1 | 3 | 3 |
| Doc6 | 3 | 2 | 0 | 2 | 3 |
| Doc7 | 1 | 1 | 1 | 1 | 1 |
| Doc8 | 1 | 1 | 1 | 0 | 1 |
| Doc9 | 0 | 1 | 1 | 1 | 1 |
| Doc10 | 0 | 0 | 0 | 3 | 3 |
| n | 5 | 8 | 7 | 7 | |

→ เติมที่เขียนตารางก็ได้ เพื่อความสบาย

① หา $w_{ij}$ จาก $tf_i$ , $idf_{ij}$

หารวม
$idf_{bird} = \log(10/5) = 0.301$
$idf_{cat} = \log(10/8) = 0.097$
$idf_{dog} = \log(10/7) = 0.155$
$idf_{tiger} = \log(10/7) = 0.155$

หาทีละ doc คนดูทุก doc

↳ Doc1 : $tf_{bird} = 3/3 = 1$
$tf_{cat} = 2/3 = 0.67$
$tf_{dog} = 2/3 = 0.67$
$tf_{tiger} = 0/3 = 0$

② คำนวณ w ทุก doc

Doc1 : $w_{bird,1} = 1 \times 0.301 = 0.301$
$w_{cat,1} = 0.67 \times 0.097 = 0.065$
$w_{dog,1} = 0.67 \times 0.155 = 0.103$
$w_{tiger,1} = 0 \times 0.155 = 0.000$

สามารถตั้งให้เป็น 0 ได้ เพราะไม่มีการสืบค้น ทางฝั่ง ให้ตรวจประเด็นกา

③ คำนวณ w ของ query

$w_{bird,q} = (0.5 + 0.5(0)/2) \times 0.301 = 0.151$ , 0
$w_{cat,q} = (0.5 + 0.5(2)/2) \times 0.097 = 0.097$
$w_{dog,q} = (0.5 + 0.5(1)/2) \times 0.155 = 0.117$
$w_{tiger,q} = (0.5 + 0.5(1)/2) \times 0.155 = 0.117$

| | Bird | Cat | Dog | Tiger |
|---|---|---|---|---|
| Doc1 | 0.301 | 0.065 | 0.103 | 0.000 |
| Doc2 | 0.000 | 0.097 | 0.077 | 0.077 |
| Doc3 | 0.301 | 0.000 | 0.077 | 0.000 |
| Doc4 | 0.000 | 0.097 | 0.000 | 0.155 |
| Doc5 | 0.000 | 0.032 | 0.052 | 0.155 |
| Doc6 | 0.301 | 0.065 | 0.000 | 0.103 |
| Doc7 | 0.301 | 0.097 | 0.155 | 0.155 |
| Doc8 | 0.301 | 0.097 | 0.155 | 0.000 |
| Doc9 | 0.000 | 0.097 | 0.155 | 0.155 |
| Doc10 | 0.000 | 0.000 | 0.000 | 0.155 |
| q | 0.151 | 0.097 | 0.117 | 0.117 |

④ คำนวณ sim แต่ละ doc

$$sim(d_1, q) = \frac{(0.151)(0.301) + (0.097)(0.065) + (0.117)(0.103)}{\sqrt{(0.151^2 + 0.097^2 + 0.117^2 + 0.117^2)(0.301^2 + \dots 0^2)}}$$

$= 0.806$

⑤ Ranking ตัวไหน sim มากสุด

# Fuzzy Logic → kw มีดสพ. , ค่าต่อเนื่อง 0-1

complement : $\mu_{\bar{A}}(x) = 1 - \mu_A(x)$

Intersection : $\mu_{A \cap B} = \mu_A(x) \cdot \mu_B(x)$

Union : $\mu_{A \cup B} = 1 - (1 - \mu_A(x)) \cdot (1 - \mu_B(x))$

ดสพ. kw i,j : $c_{ij} = n_{ij} / n_i + n_j - n_{ij}$ n = จน.เอกสาร

→ มีทั้ง i,j

Ex. Query = "ต้องการเอกสารที่มี cat หรือ kitty แต่ไม่ต้องการเอกสารที่มี dog"

D1 : { bird , cat , bird , cat , tiger , kitty , fish , fish }
D2 : { cat , dog , tiger , zoo }
D3 : { house , kitchen , bird , bird , cat , cat , kitty }
D4 : { dog , rubber , house , dog }
D5 : { tiger , forest , fish }

① เขียนดสพ.kw ex. bird → cat , fish , house , kitchen , kitty , tiger
ex. fish → bird , cat , forest , kitty , tiger

② หาดสพ. kw โดยนทุกดสพ. ex. bird , cat = 2/(2+3-2) = 0.67
bird , fish = 1/(2+2-1) = 0.33

③ ex doc 1

| | bird | cat | dog | fish | forest | house | kitchen | kitty | rubber | tiger | zoo |
|---|---|---|---|---|---|---|---|---|---|---|---|
| bird | 1.00 | 0.67 | 0.00 | 0.33 | 0.00 | 0.33 | 0.50 | 1.00 | 0.00 | 0.25 | 0.00 |
| cat | 0.67 | 1.00 | 0.25 | 0.25 | 0.00 | 0.25 | 0.33 | 0.67 | 0.00 | 0.50 | 0.33 |
| dog | 0.00 | 0.25 | 1.00 | 0.00 | 0.00 | 0.33 | 0.00 | 0.00 | 0.50 | 0.25 | 0.50 |
| fish | 0.33 | 0.25 | 0.00 | 1.00 | 0.50 | 0.00 | 0.00 | 0.33 | 0.00 | 0.67 | 0.00 |
| forest | 0.00 | 0.00 | 0.00 | 0.50 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.00 |
| house | 0.33 | 0.25 | 0.33 | 0.00 | 0.00 | 1.00 | 0.50 | 0.33 | 0.50 | 0.00 | 0.00 |
| kitchen | 0.50 | 0.33 | 0.00 | 0.00 | 0.00 | 0.50 | 1.00 | 0.50 | 0.00 | 0.00 | 0.00 |
| kitty | 1.00 | 0.67 | 0.00 | 0.33 | 0.00 | 0.33 | 0.50 | 1.00 | 0.00 | 0.25 | 0.00 |
| rubber | 0.00 | 0.00 | 0.50 | 0.00 | 0.00 | 0.50 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| tiger | 0.25 | 0.50 | 0.25 | 0.67 | 0.33 | 0.00 | 0.00 | 0.50 | 0.00 | 1.00 | 0.33 |
| zoo | 0.00 | 0.33 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 1.00 |

③ q Query ต้องมีตัวไหนบ้าง
cat , kitty , dog

| | Cat | Kitty | Dog |
|---|---|---|---|
| D1 | 1 | 1 | 0.25 |
| D2 | 1 | 0.67 | 1 |
| D3 | 1 | 1 | 0.33 |
| D4 | 0.25 | 0.33 | 1 |
| D5 | 0.50 | 0.33 | 0.25 |

เอกสารมี kw นั้น = 1 , ไม่มี หาตัวที่มีค่า $c_{ij}$ มากสุด เทียบกับ kw ใน doc นั้น

④ Query = (cat OR kitty) AND not dog แล้วหา Ranking โดยใช้ตรง

∴ Ranking = $(1 - (1 - \mu_{cat}) \cdot (1 - \mu_{kitty})) \cdot (1 - \mu_{dog})$

ex. Doc1 = $(1 - (1-1)(1-1)) \cdot (1 - 0.25) = 0.75$

⑤ Ranking ตัวที่ค่ามากสุด

# Extend Boolean Model → kw ไม่มีพันธ์ , เป็น Boolean

$tf\,norm_{ij} = \dfrac{tf_{ij}}{tf\,max_{ij}}$ | $idf\,norm_i = \dfrac{idf_i}{idf\,max_i}$

$sim(q_{or}, d_j) = \sqrt{\dfrac{w_{1j}^2 + w_{2j}^2}{2}}$ | $sim(q_{and}, d_j) = 1 - \sqrt{\dfrac{(1-w_{1j})^2 + (1-w_{2j})^2}{2}}$

ex. Query = รักแมวและสุนัข แต่ไม่ชอบเสือ = (cat AND dog) AND NOT tiger

| | Bird | Cat | Dog | Tiger | Max |
|---|---|---|---|---|---|
| Doc1 | 3 | 2 | 2 | 0 | 3 |
| Doc2 | 0 | 2 | 1 | 1 | 2 |
| Doc3 | 2 | 0 | 1 | 0 | 2 |
| Doc4 | 0 | 1 | 0 | 1 | 1 |
| Doc5 | 0 | 1 | 1 | 3 | 3 |
| Doc6 | 3 | 2 | 0 | 2 | 3 |
| Doc7 | 1 | 1 | 1 | 1 | 1 |
| Doc8 | 1 | 1 | 1 | 0 | 1 |
| Doc9 | 0 | 1 | 1 | 1 | 1 |
| Doc10 | 0 | 0 | 0 | 3 | 3 |
| n | 5 | 8 | 7 | 7 | |

① สร้างตารางจาก Doc เพื่อความง่าย

② คำนวณ tf , $idf_{norm}$ ทุก DOC

ex. $tf_{bird} = 3/3 = 1$
$tf_{cat} = 2/3 = 0.667$

ex. $idf_{bird} = \log(10/5) = 0.301$
$idf_{norm\ bird} = 0.301 / 0.301 = 1$

หารวมครั้งเดียว

③ หา $w_{ij}$ แต่ละ DOC ex. $w_{bird1} = 1 \times 1 = 1$
$w_{cat1} = 0.667 \times 0.322 = 0.215$

| | Bird | Cat | Dog | Tiger |
|---|---|---|---|---|
| Doc1 | 1.000 | 0.215 | 0.343 | 0.000 |
| Doc2 | 0.000 | 0.322 | 0.257 | 0.257 |
| Doc3 | 1.000 | 0.000 | 0.257 | 0.000 |
| Doc4 | 0.000 | 0.322 | 0.000 | 0.515 |
| Doc5 | 0.000 | 0.107 | 0.172 | 0.515 |
| Doc6 | 1.000 | 0.215 | 0.000 | 0.343 |
| Doc7 | 1.000 | 0.322 | 0.515 | 0.515 |
| Doc8 | 1.000 | 0.322 | 0.515 | 0.000 |
| Doc9 | 0.000 | 0.322 | 0.515 | 0.515 |
| Doc10 | 0.000 | 0.000 | 0.000 | 0.515 |

④ หาค่า sim โดยดูจาก Query
↳ (cat AND dog) AND NOT tiger
หาให้ครบทุก doc

$$sim(q_{and}, d_j) = 1 - \sqrt{\frac{\left(1 - \left(1 - \sqrt{\frac{(1-w_{cat_j})^2 + (1-w_{dog_j})^2}{2}}\right)\right)^2 + (1 - (1 - w_{tiger_j}))^2}{2}}$$

$$sim(q_{and}, d_1) = 1 - \sqrt{\frac{\left(1 - \left(1 - \sqrt{\frac{(1-0.215)^2 + (1-0.343)^2}{2}}\right)\right)^2 + (1 - (1 - 0))^2}{2}}$$

$= 0.488$

⑤ Ranking ตัวไหน sim มากสุด

นายภัทรชัย วรรณสันทัด 65010815

# Generalizes Vector Space Model

↳ kw ไม่อิสระต่อกัน, จัดหมวดหมู่, เอกสารที่มี kw เดียวกัน ความตรงประเด็นต้องใกล้กัน

≈ weight

$$k_i = \frac{\sum_{\forall r g_i(m_r)} c_{ir} m_r}{\sqrt{\sum_{\forall r g_i(m_r)} c_{ir}^2}}$$

$c_{ir} = \sum w$

↳ i คือ kw ที่ r คือ กลุ่ม m

$$sim(q, d_j) = \frac{\sum S_{dr} \cdot S_{qr}}{\sqrt{\sum S_{dr}^2 \cdot \sum S_{qr}^2}}$$

ex. Query = "cat cat cat dog dog tiger"

① ทำเหมือน VSM อาจจะเริ่มที่การตรวจนค.ข้อมูล
② เริ่มที่หา tf, idf
③ หา weight ทุก doc
④ หา weight Query
⑤ แบ่งกลุ่ม doc, kw ปรากฏเหมือนกัน = กลุ่มเดียวกัน

| | Bird | Cat | Dog | Tiger | |
|---|---|---|---|---|---|
| Doc1 | 0.398 | 0.031 | 0.103 | 0.000 | m1 |
| Doc2 | 0.000 | 0.046 | 0.077 | 0.111 | m2 |
| Doc3 | 0.398 | 0.000 | 0.077 | 0.000 | m3 |
| Doc4 | 0.000 | 0.046 | 0.000 | 0.222 | m4 |
| Doc5 | 0.000 | 0.015 | 0.052 | 0.222 | m2 |
| Doc6 | 0.000 | 0.046 | 0.000 | 0.222 | m4 |
| Doc7 | 0.398 | 0.046 | 0.155 | 0.000 | m1 |
| Doc8 | 0.398 | 0.046 | 0.155 | 0.000 | m1 |
| Doc9 | 0.000 | 0.046 | 0.155 | 0.222 | m2 |
| Doc10 | 0.000 | 0.023 | 0.000 | 0.222 | m4 |
| Q | 0 | 0.523 | 0.552 | 0.537 | |

ex. $m1 = (1,1,1,0)$
$m2 = (0,1,1,1)$
$m3 = (1,0,1,0)$
$m4 = (0,1,0,1)$

⑥ คำนวณหา $k_i$ โดยดูลำดับ kw ทำให้ครบทุก kw (4 ตัว)

ex. $k_1 = \frac{C_{11} m_1 + C_{13} m_3}{\sqrt{C_{11}^2 + C_{13}^2}}$

$= \frac{1.194 m_1 + 0.398 m_3}{\sqrt{1.194^2 + 0.398^2}}$

$= 0.949 m_1 + 0.316 m_3$

$C_{11} = w_{11} + w_{17} + w_{18}$
$= 0.398 + 0.398 + 0.398$
$= 1.194$

$C_{13} = w_{13} = 0.398$

⑦ แทนสมการ k แต่ละตัวลงใน Query

แทนสมการจากข้อ 6

$q = 0.523 k_2 + 0.552 k_3 + 0.537 k_4$

$q = 0.523(0.615 m_1 + 0.538 m_2 + 0.576 m_4) + 0.552 k_3 + 0.537 k_4$

$q = 0.771 m_1 + 0.934 m_2 + 0.084 m_3 + 0.714 m_4$

⑧ แทนสมการ k แต่ละตัวลงใน Doc แต่ละตัว

ex. $d_1 = 0.398 k_1 + 0.031 k_2 + 0.103 k_3$

$d_1 = 0.480 m_1 + 0.074 m_2 + 0.142 m_3 + 0.018 m_4$

| | m1 | m2 | m3 | m4 |
|---|---|---|---|---|
| Doc1 | 0.480 | 0.074 | 0.142 | 0.018 |
| Doc2 | 0.091 | 0.139 | 0.012 | 0.112 |
| Doc3 | 0.441 | 0.043 | 0.138 | 0.000 |
| Doc4 | 0.028 | 0.167 | 0.000 | 0.197 |
| Doc5 | 0.051 | 0.179 | 0.008 | 0.179 |
| Doc6 | 0.028 | 0.167 | 0.000 | 0.197 |
| Doc7 | 0.532 | 0.111 | 0.149 | 0.026 |
| Doc8 | 0.532 | 0.111 | 0.149 | 0.026 |
| Doc9 | 0.154 | 0.253 | 0.024 | 0.197 |
| Doc10 | 0.014 | 0.154 | 0.000 | 0.184 |
| Q | 0.771 | 0.934 | 0.084 | 0.714 |

⑨ หาค่า sim ของแต่ละ doc

ex. $sim(q, d_1) = \frac{(0.480 \times 0.771) + (0.774 \times 0.934) + \ldots + (0.018 \times 0.714)}{\sqrt{(0.480^2 + 0.074^2 + \ldots) \times (0.771^2 + 0.934^2 + \ldots)}}$

$= 0.974$

⑩ Ranking ตัวไหน sim มากสุด

# Probabilistic Model

→ user marked > kw, ไม่สน tf, idf

☆ smooth tuning กันไปหารเป็น 0 หรือไม่ก็ได้

$P(K_i|R) = \frac{r_i + 0.5}{R + 1}$

$P(\bar{K}_i|R) = R - r_i / R$

$P(K_i|\bar{R}) = \frac{n_i - r_i + 0.5}{N - R + 1}$

$P(\bar{K}_i|\bar{R}) = \frac{(N-R) - (n_i - r_i)}{N - R}$

$P(K_i|R) + P(\bar{K}_i|R) = 1$

$P(K_i|\bar{R}) + P(\bar{K}_i|\bar{R}) = 1$

$r_i$ = จน.เอกสารที่มี kw i แล้วตรงประเด็น
$n_i$ = จน. เอกสารที่มี kw i ทั้งหมด
N = All เอกสาร
R = All เอกสาร R

similarity : $sim(d_j, q) = \frac{P(R|\vec{d_j})}{P(\bar{R}|\vec{d_j})}$

$= \frac{P(\vec{d_j}|R) \times P(R)}{P(\vec{d_j}|\bar{R}) \times P(\bar{R})}$

① User ทำใส่ Query เข้ามาแล้ว mark ว่าเอกสารไหนตรงประเด็น
② แบ่งเอกสารที่ Relevance & Non R

ex. Relevance: $d1 = \{1,1\}$, $d2 = \{1,1\}$, $d3 = \{1,0\}$, $d5 = \{1,0\}$, $d8 = \{0,1\}$, $K_1 = 4$, $K_2 = 3$

Non Relevance: $d4 = \{1,0\}$, $d6 = \{0,1\}$, $d7 = \{0,1\}$, $d9 = \{0,1\}$, $K_1 = 1$, $K_2 = 3$

$K_1$ = cat, $K_2$ = dog
$N = 9$
$R = 5$
$\bar{R} = 4$

$K_1, N = 5$
$K_1, R = 4$
$K_1, \bar{R} = 1$
$\bar{K}_1, R = 1$
$\bar{K}_1, \bar{R} = 3$

$K_2, N = 6$
$K_2, R = 3$
$K_2, \bar{R} = 3$
$\bar{K}_2, R = 2$
$\bar{K}_2, \bar{R} = 1$

③ หา Prob ของแต่ละ kw

$P(K_1|R) = 4/5$
$P(\bar{K}_1|R) = 1/5$
$P(K_1|\bar{R}) = 1/4$
$P(\bar{K}_1|\bar{R}) = 3/4$

$P(K_2|R) = 3/5$
$P(\bar{K}_2|R) = 2/5$
$P(K_2|\bar{R}) = 3/4$
$P(\bar{K}_2|\bar{R}) = 1/4$

Binary Independence Retrieval Model

④ หา sim ของแต่ละ pattern kw มี 2 แบบ Simple, BIR

simple : $P(R|(1,1)) = 2/2 = 1$
$P(R|(1,0)) = 2/3 = 0.66$
$P(R|(0,1)) = 1/4 = 0.25$

BIR : $sim(d_1, q) = \frac{P(K_1|R) \times P(K_2|R)}{P(K_1|\bar{R}) \times P(K_2|\bar{R})} \times \frac{P(R)}{P(\bar{R})}$

{1,1} $\quad 5/9 \quad 4/9$

$= \frac{4/5 \times 3/5}{1/4 \times 3/4} \times \frac{5}{4} = \frac{16}{5}$

$sim(d_3, q) = \frac{P(K_1|R) \times P(\bar{K}_2|R)}{P(K_1|\bar{R}) \times P(\bar{K}_2|\bar{R})} \times \frac{P(R)}{P(\bar{R})}$

{1,0}

$= \frac{4/5 \times 2/5}{1/4 \times 1/4} \times \frac{5}{4} = \frac{32}{5}$

หลังจาก ทำจนครบทุก pattern นำไปเข้าสูตร $sim(d_j, q) = \frac{sim(d_j, q)}{1 + sim(d_j, q)}$

ex. $sim(d_1, q) = 16/5 \rightarrow \frac{16/5}{1 + 16/5} = \frac{16}{21} = 0.762$

ทำจนครบทุก pattern

ex. จะได้ประมาณนี้

| | {1,1} | {1,0} | {0,1} | {0,0} |
|---|---|---|---|---|
| Simple | 0.800 | 0.667 | 0.500 | 0.333 |
| BIR | 0.757 | 0.690 | 0.483 | 0.400 |

นายภัทรชัย วรรณสินทัต 65010815

# Extend Prob. Model (BM25)

↳ ใช้ idf, tf doc, tf query

$$sim_{bm25}(d_i, q) = \sum idf \times tf_{doc} \times tf_{query}$$

Inverse doc. freq. : $idf_i = \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)(N - n_i - R + r_i + 0.5)}$

doc. term freq. : $tf_{doc} = \frac{(k_1 + 1) f_i}{k_1((1-b) + b \times dl/avdl) + f_i}$

query term freq. : $tf_q = \frac{(k_2 + 1) qf_i}{k_2 + qf_i}$

ค่าคงที่ต่างๆ : $k_1 \approx 1.25$, $k_2 \approx 200$, $b \approx 0.75$

dl = จน. คำเอกสาร j

avdl = จน. คำเฉลี่ยของทุก เอกสาร

f → "ปรากฏในเอกสารที่สนใจ"

ex. Query = Honda, Toyota, Isuzu

① เขียนตาราง term freq.

② คำนวณ dl, avdl จากโจทย์ (อาจนิ่งนับคำเอง)

| f | Honda | Toyota | Isuzu |
|---|---|---|---|
| d1 | 0 | 3 | 6 |
| d2 | 0 | 4 | 0 |
| d3 | R 6 | 2 | 0 |
| d4 | 0 | 2 | 3 |
| d5 | R 1 | 3 | 0 |
| d6 | R 3 | 3 | 3 |
| d7 | R 2 | 2 | 3 |
| d8 | R 3 | 4 | 1 |

| | | |
|---|---|---|
| d1 | | My father is thinking about the car that she want to buy between Isuzu D-MAX X-series, Isuzu D-MAX V-Cross 4 door, Isuzu D-MAX V-Cross 2 door, Isuzu Mu-7, Isuzu Mu-X, Isuzu D-MAX Hi-Lander, Toyota Hilux vigo, Toyota Hilux revo or Toyota Innova. |
| d2 | | My uncle suggest My father that he should buy Toyota camry, Toyota vios, Toyota yaris or Toyota corolla altis. |
| d3 | R | My mother will buy a car for me and my brother, we think we should Honda city, Honda brio, Honda CR-V, Honda BR-V, Honda civic, Honda accord, Toyota Yaris, Toyota vios. |

ex. $dl = 42$, $d2 = 19$ ....

$AVR = \frac{42 + 19 + \ldots}{n} = 32.5$

โจทย์อาจไม่ให้ R (ไม่จำเป็นต้องมี) ก็ให้แทน $R, r_i$ เป็น 0

③ หา idf ของ แต่ละ จากข้อมูลตาราง

$R = 5$ $N = 8$

$r_{Honda} = 5$ $n_H = 5$

$r_{Toyota} = 5$ $r_T = 8$

$r_{Isuzu} = 3$ $r_I = 5$

ex. $idf_H = \log \frac{(5+0.5)/(5-5+0.5)}{(5-5+0.5)/(8-5-5+5+0.5} = 1.89$

④ หา sim ของ doc. ที่ต้องการ (อาจเป็นเอกสารที่นอนนิ่งอยู่ในระบบ)

| | ความถี่ Honda | ความถี่ Toyota | ความถี่ Isuzu | จำนวนคำในเอกสาร |
|---|---|---|---|---|
| d10 | 0 | 4 | 2 | 21 |
| d20 | 9 | 15 | 2 | 55 |
| d30 | 11 | 7 | 5 | 35 |
| d40 | 6 | 6 | 6 | 25 |

เป็น 1 เพราะ query มีอย่างละ 1

ex. $sim_{bim25}(d_{30}, q) = 1.89 \times \frac{(2.25)11}{1.25((1-0.75) + 0.75(\frac{35}{32.5})) + 11} \times \frac{(200+1)1}{(200+1)}$

$+ 0.2 \times \frac{(2.25)7}{1.25((1-0.75) + 0.75(\frac{35}{32.5})) + 7} \times 1$

$- 0.08 \times \frac{(2.25)5}{1.25((1-0.75) + 0.75(\frac{35}{32.5})) + 5} \times 1$

$= 3.789 + 0.371 - 0.135 = 4.026$

| | ความถี่ Honda | ความถี่ Toyota | ความถี่ Isuzu | จำนวนคำในเอกสาร | sim |
|---|---|---|---|---|---|
| d30 | 11 | 7 | 5 | 35 | 4.026 |
| d40 | 6 | 6 | 6 | 25 | 3.852 |
| d20 | 9 | 15 | 2 | 55 | 3.810 |
| d10 | 0 | 4 | 2 | 21 | 0.242 |

⑤ Ranking ค่า sim มากสุด

# Evaluate → ประเมินค.ตรงประเด็น model

| | Relevant | Not relevant |
|---|---|---|
| Retrieved | A | B |
| Not retrieved | C | D |

recall = $\frac{\text{จน. doc ที่ออก + Relevant}}{\text{จน. doc All Relevant}} = \frac{A}{A+C}$

precision = $\frac{\text{จน. doc ที่ออก + Relevant}}{\text{จน. doc All ที่ออก}} = \frac{A}{A+B}$

พลาดสิ่งดีๆ

miss = $C/A+C$

ex. ① หาค่า Precision และ Recall จากลำดับเอกสาร

| n | doc # | relevant |
|---|---|---|
| 1 | 588 | x |
| 2 | 589 | x |
| 3 | 576 | |
| 4 | 590 | x |
| 5 | 986 | |
| 6 | 592 | x |
| 7 | 984 | |
| 8 | 988 | |
| 9 | 578 | |
| 10 | 985 | |
| 11 | 103 | |
| 12 | 591 | |
| 13 | 772 | x |
| 14 | 990 | |

$R = 1/5$ $P = 1/1 = 1$

$R = 2/5$ $P = 2/2 = 1$

$R = 3/5$ $P = 3/4 = 0.75$

$R = 4/5$ $P = 4/6 = 0.67$

$R = 5/5$ $P = 5/13 = 0.38$

② หา Avg Precision → $\frac{1 + 1 + \ldots + 0.38}{5} = 0.76 = 76\%$

ของแต่ละ model หรือ ระบบ

③ เทียบผล Avg Precision

| n | doc # | relevant |
|---|---|---|
| 1 | 588 | x |
| 2 | 589 | x |
| 3 | 576 | |
| 4 | 590 | x |
| 5 | 986 | |
| 6 | 592 | x |
| 7 | 984 | |
| 8 | 988 | |
| 9 | 578 | |
| 10 | 985 | |
| 11 | 103 | |
| 12 | 591 | |
| 13 | 772 | x |
| 14 | 990 | |

R-Precision → ดู Top Rank, ตามจน. เอกสารที่กำหนด

ex. R = # of relevant docs = 5

R-Precision = 3/5 = 0.6

Precision Histograms : $RP_{A/B}(i) = RP_A(i) - RP_B(i)$

$= 60\% - 80\% = -20\%$

F-Measure : $F = \frac{2PR}{P+R} = \frac{2}{1/R + 1/P}$

E-Measure → ให้น้ำหนักได้ว่าอยากใช้ P หรือ R มากกว่า

$E = \frac{(1+\beta)^2 PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R} + \frac{1}{P}}$

User-Oriented Measure

Coverage = $|R_k| / |U|$

Novelty = $|R_u| / (|R_u| + |R_k|)$

$|U|$ = จน. R doc ที่รู้

$|R_k|$ = จน. R doc ที่รู้และได้มา

$|R_u|$ = จน. R doc ที่ไม่รู้และได้มา

Fallout → ไม่รู้ Recall และ Precision : $\frac{\text{จน. non rel ที่ออก}}{\text{All จน. non rel}}$

# Query Operations → ทำให้ Query ตรงประเด็นมากขึ้น

Relevance Feedback → 1. user feedback (explicit feedback)

2. automatically (implicit feedback)

Query Reformulation ปรับ Query

- Query Expansion เพิ่ม terms ของ R
- Term Reweighting เพิ่ม weight ของ term

Vector Space Re-weighting

↳ Rocchio $= \alpha Q + \frac{\beta}{n_1} \sum_{\forall d_j \in D_r} \vec{d}_j - \frac{\gamma}{n_2} \sum_{\forall d_j \in D_n} \vec{d}_j$

$Q_0$ = vector initial Q

$D_r$ = set R docs

$D_n$ = set $\bar{R}$ docs

$\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$

$n_1$ = จน. R ที่ถูกเลือก

$n_2$ = จน. $\bar{R}$ ที่ถูกเลือก

ex. $R_1 = (0.03, 0, 0.025)$, $R_2 = (0.02, 0.09, 0.02)$

$S_1 = (0.03, 0.01, 0.02)$, $Q = (0, 0, 0.5)$

$Q_{new} = \alpha \times Q + (\frac{\beta}{2}(R_1 + R_2)) - (\frac{\gamma}{1}(S_1)) = (0.011, 0.000875, 0.002)$ เลขสมมติ

↳ Ide Regular Method : $\vec{q}_1 = \alpha \vec{q}_0 + \beta \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$

ex. Original Query = (5, 0, 3, 0, 1), $\alpha = 0.5$, $\beta = 0.25$

Doc. $D_1$ Relevant = (2, 1, 2, 0, 0), Doc. $D_2$ NonRel. = (1, 0, 0, 0, 2)

$q' = q + 0.5 D_1 - 0.25 D_2$

$= (5, 0, \ldots) + 0.5(2, 1, \ldots) - 0.25(1, 0, \ldots) = (5.75, 0.5, \ldots)$

นายภัทรชัย วงรณสันทัด b5010815

## Automatic Global Analysis → หา itf แทนที่จะเป็น idf

กราฟเปลี่ยนเป็น key\Doc →

$$itf_j = \log t/t_j$$

☆ ค่า t = Term = key ในdoc

weight ของ kw ของแต่ละ doc. : $W_{ij} = \dfrac{(0.5+0.5\, f_{ij}/\max_i(f_{ij}))\, itf_j}{\sqrt{\sum (0.5+0.5\, f_{ij}/\max_i(f_{ij}))^2\, itf_l^2}}$

ค่าคล้าย.ของ kw. : $C_{uv} = \vec{k}_u \cdot \vec{k}_v = \sum_{j=1} W_{uj} \times W_{vj}$

sim ของ kw. : $sim(q,k_v) = \vec{q}\cdot\vec{k}_v = \sum_{k_u \in q} W_{uq} \times C_{uv}$

weight kw ของ Query : $W_{vq}' = \dfrac{sim(q,k_v)}{\sum_{k_u \in q} W_{uq}}$

ใช้ทั้ง w kw., w query

sim ของ doc. : $sim(q,d_j) \propto \sum_{k_v \in d_j} \sum_{k_u \in q} W_{ij} \times W_{uq} \times C_{uv}$

ex. Query = 2.3A + C

| Key/Doc | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| A | 3 | 0 | 0 | 1 |
| B | 2 | 0 | 1 | 0 |
| C | 1 | 1 | 0 | 0 |
| D | 0 | 2 | 0 | 1 |
| E | 0 | 0 | 2 | 1 |
| | | | | |
| Max | 3 | 2 | 2 | 1 |
| $t_j$ | 3 | 2 | 2 | 3 |
| itf(Doc) | 0.222 | 0.398 | 0.398 | 0.222 |

① สร้างตารางและหา itf ของทุกเอกสาร

ex. Term = 5

$itf_4 = 5/3 = 0.222$

② หา $W_{ij}$ ของ kw ในแต่ละ doc ให้ครบทุกช่อง

ex. $W_{1,3} = \dfrac{(0.5+0.5\frac{f_{1,3}}{\max(f_{d_3})})\, itf_3}{\sqrt{(0.5+0.5\frac{f_{1,1}}{\max(f_{d_1})})^2 itf_1^2 + (0.5+0.5\frac{f_{1,2}}{\max(f_{d_2})})^2 itf_2^2 + \dots}}$

$= \dfrac{(0.5+0.5\,{}^{0}/_{2})\,0.398}{(0.5+0.5\,{}^{3}/_{3})^2(0.222)^2 + (0.5+0.5\,{}^{0}/_{2})^2(0.398)^2 + \dots}$

$= 1.509$

| W | $D_1$ | $D_2$ | $D_3$ | $D_4$ |
|---|---|---|---|---|
| A | 1.683 | 1.509 | 1.509 | 1.683 |
| B | 1.228 | 1.322 | 1.983 | 0.737 |
| C | 0.996 | 2.010 | 1.340 | 0.747 |
| D | 0.598 | 2.146 | 1.073 | 1.197 |
| E | 0.598 | 1.073 | 2.146 | 1.197 |

③ หาค่าคล้าย.ระหว่าง kw. ให้ครบทุกคู่ สมมติคู่ 1,3

ex. $C_{1,3} = W_{11}W_{31} + W_{12}W_{32} + W_{13}W_{33} + W_{14}W_{34}$

$= 1.683(0.996) + 1.509(2.01) + 1.509(1.34) + 1.683(0.747)$

$= 7.987$

| C | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 10.218 | 8.293 | 7.987 | 7.879 | 7.879 |
| B | 8.293 | 7.728 | 7.085 | 6.581 | 7.290 |
| C | 7.987 | 7.085 | 7.383 | 7.241 | 6.522 |
| D | 7.879 | 6.581 | 7.241 | 7.548 | 6.397 |
| E | 7.879 | 7.290 | 6.522 | 6.397 | 7.548 |
| q | 2.3 | 0 | 1 | 0 | 0 |

④ ดูว่าจะใช้ kw ไหนบ้างจากการหาค่า $sim(q,k_v)$

ex. $sim(q,k_3) = W_{1q}C_{13} + W_{2q}C_{23} + W_{3q}C_{33} + \dots + W_{5q}C_{53}$

$= 2.3(7.987) + 1(7.383) = 25.753$

| Sim(q,$K_i$) |
|---|
| 31.487 |
| 26.159 |
| 25.753 |
| 25.362 |
| 24.643 |

หลังจากคำนวณครบ kw จะได้ตาราง sim ถ้ามี kw ไหนค่า sim มากกว่า kw ที่น้อยสุดต้องเพิ่ม kw นั้นเข้ามาด้วย ในที่นี้คือ $k_2$ หลังจากเพิ่ม kw ให้คำนวณ sim ของ kw ใหม่

ex. $sim(q,k_3) = 2.3(7.987) + 1(7.085) + 1(7.383) = 32.838$

| C | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 10.218 | 8.293 | 7.987 | 7.879 | 7.879 |
| B | 8.293 | 7.728 | 7.085 | 6.581 | 7.290 |
| C | 7.987 | 7.085 | 7.383 | 7.241 | 6.522 |
| D | 7.879 | 6.581 | 7.241 | 7.548 | 6.397 |
| E | 7.879 | 7.290 | 6.522 | 6.397 | 7.548 |
| q | 2.3 | 1 | 1 | 0 | 0 |

| Sim(q,$K_i$) |
|---|
| 39.780 |
| 33.887 |
| 32.838 |
| 31.942 |
| 31.933 |

⑤ คำนวณ weight ของ kw ที่ใช้ query

ex. $q = 2.3K_1 + K_2 + K_3$

sum query weight = 2.3 + 1 + 1 = 4.3

$W_{1,q}' = 39.780/4.3 = 9.251$

$W_{2,q}' = 33.887/4.3 = 7.881$

$W_{3,q}' = 32.838/4.3 = 7.637$

| | A | B | C | D | E |
|---|---|---|---|---|---|
| q' | 9.251 | 7.881 | 7.637 | - | - |

⑥ หา sim ของ doc โดยต้องใช้ W. query, W. doc, $C_{uv}$

ex. $sim(q,d_2) = W_{1,2}W_{1q}(C_{1,2}+C_{1,2}+C_{1,3}+C_{1,4}+C_{1,5}) + W_{2,2}W_{2q}(C_{2,1}+C_{2,2}+\dots+C_{2,5}) + \dots$

$\vdots$

$W_{5,2}W_{5q}(C_{5,1}+C_{5,2}+\dots+C_{5,5})$

จากตาราง

$W_{1,2} = 1.509$ $W_{1q} = 9.251$

$W_{2,2} = 1.322$ $W_{2q} = 7.881$

$W_{3,2} = 2.010$ $W_{3q} = 7.637$

$W_{4,2} = 2.146$ $W_{4q} = 0$

$W_{5,2} = 1.073$ $W_{5q} = 0$

$= 1531.123$

⑤ Ranking ตัวที่ ค่า sim มากสุด

## Automatic Local Analysis

มี 3 วิธี → association clusters, metric clusters, scalar clusters

### Association Clusters → หาจากคล้าย.ของ kw

คล้าย. kw : $C_{u,v} = \sum_{d_j \in D_k} f_{s_u,j} \times f_{s_v,j}$

normalized Corr. Matrix : $S_{u,v} = \dfrac{C_{u,v}}{C_{u,u} + C_{v,v} - C_{u,v}}$

ex. Query = A+B

① เขียนตารางจากข้อมูล

ex. $d_1$ = A, A, B, D

$d_2$ = B, A, C, C, D

⋮

{B,D,C}→A

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | $d_7$ |
|---|---|---|---|---|---|---|---|
| A | 2 | 1 | 1 | 0 | 0 | 1 | 1 |
| B | 1 | 1 | 1 | 1 | 0 | 1 | 2 |
| C | 0 | 2 | 0 | 1 | 0 | 0 | 0 |
| D | 1 | 1 | 0 | 1 | 1 | 1 | 0 |

② หาค่า $C_{u,v}$

| | A | B | C | D |
|---|---|---|---|---|
| A | 8 | 7 | 2 | 4 |
| B | 7 | 9 | 3 | 4 |
| C | 2 | 3 | 5 | 3 |
| D | 4 | 4 | 3 | 5 |

ex. $C_{1,4} = (f_{1,1}f_{4,1}) + (f_{1,2}f_{4,2}) + (f_{1,3}f_{4,3}) + \dots + (f_{1,7}f_{4,7})$

$= (2\times1) + (1\times1) + (1\times0) + (0\times1) + (0\times1) + (1\times1) + (1\times0)$

$= 4$

③ Normalized Corr. Matrix

ex. $S_{1,2} = \dfrac{7}{8+9-7} = 0.7$

| | A | B | C | D |
|---|---|---|---|---|
| A | 1 | 0.70 | 0.18 | 0.44 |
| B | 0.70 | 1 | 0.27 | 0.40 |
| C | 0.18 | 0.27 | 1 | 0.43 |
| D | 0.44 | 0.40 | 0.43 | 1 |

④ กำหนด Query ใหม่จากตาราง

ex. Term Relation

A → {A,B}
B → {B,A}
C → {C,D}
D → {D,A}

$q = A+B \rightarrow q' = (A+0.7B) + (0.7A+B)$

$= 1.7A + 1.7B$

$= A+B$ → 1.7 เท่ากันแสดงเอาออก

ex. D → {D,B,C}

ex. $q = C+2D \rightarrow q' = (C+0.43D) + (0.63B + 0.63C + D)$

$= 0.63B + 1.63C + 1.43D$

☆ ถ้ามีคู่

| | | | | |
|---|---|---|---|---|
| D | 0.44 | 0.63 | 0.63 | 1 |

เอามาคู่

### Metric Clusters → หาจาก distance ของ kw., ต้องแบ่งจากสัพท์ (stem)

คล้าย.ของ stem : $C_{u,v} = \sum_{k_i \in V(s_u)} \sum_{k_j \in V(s_v)} \dfrac{1}{r(k_i,k_j)}$

Normalized Corr. Matrix : $S_{u,v} = \dfrac{C_{u,v}}{|V(s_u)| \times |V(s_v)|}$

ex. Query = A+2D

$k_n$ = A, B, C, D, E, F

① แบ่งจากสัพท์ (stem)

| ระยะห่าง | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | 5 | ∞ | ∞ | 1 | 2 |
| B | 5 | 0 | 3 | 2 | 1 | 1 |
| C | ∞ | 3 | 0 | 3 | 4 | ∞ |
| D | ∞ | 2 | 3 | 0 | ∞ | 5 |
| E | 1 | 1 | 4 | ∞ | 0 | 1 |
| F | 2 | 1 | ∞ | 5 | 1 | 0 |

$V(s_1) = \{A,B,C\}$

$V(s_2) = \{D,E\}$

$V(s_3) = \{F\}$

หาระยะห่างของ kw จากเอกสารที่ระยะทางออกมาโดยยึดตัวที่น้อยสุด

สมมติ return

| | A,B distance |
|---|---|
| $d_5$ | 5 ✓ |
| $d_6$ | 8 |

② แปลงค่าในตาราง (x) ด้วย $1/x$

| ระยะห่าง | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | 0 | 5 | ∞ | ∞ | 1 | 2 |
| B | 5 | 0 | 3 | 2 | 1 | 1 |
| C | ∞ | 3 | 0 | 3 | 4 | ∞ |
| D | ∞ | 2 | 3 | 0 | ∞ | 5 |
| E | 1 | 1 | 4 | ∞ | 0 | 1 |
| F | 2 | 1 | ∞ | 5 | 1 | 0 |

ex. $0 \rightarrow 1/0 = -$

$5 \rightarrow 1/5 = 0.2$

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | - | 0.20 | 0 | 0 | 1 | 0.50 |
| B | 0.20 | - | 0.33 | 0.50 | 1 | 1 |
| C | 0 | 0.33 | - | 0.33 | 0.25 | 0 |
| D | 0 | 0.50 | 0.33 | - | 0 | 0.20 |
| E | 1 | 1 | 0.25 | 0 | - | 1 |
| F | 0.50 | 1 | 0 | 0.20 | 1 | - |

③ หาค่า $C_{u,v}$ (คล้าย.ของ stem)

ex. $C_{1,2} = C(A,D) + C(A,E) + C(B,D) + C(B,E) + C(C,D) + C(C,E)$

$= 0 + 1 + 0.5 + 1 + 0.33 + 0.25$

$= 3.08$

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| $S_1$ | 0 | 3.08 | 1.50 |
| $S_2$ | 3.08 | 0 | 1.20 |
| $S_3$ | 1.50 | 1.20 | 0 |

④ Normalized Corr Matrix

ex. $S_{2,3} = \dfrac{C_{2,3}}{|V(s_2)| \times |V(s_3)|} = \dfrac{1.2}{2\times1} = 0.6$

ขนาดของ S

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| $S_1$ | 0 | 0.51 | 0.50 |
| $S_2$ | 0.51 | 0 | 0.60 |
| $S_3$ | 0.50 | 0.60 | 0 |

⑤ หา Query ใหม่

ex. Stem Relation

$S_1 \rightarrow \{S_1, S_2\}$
$S_2 \rightarrow \{S_1, S_3\}$
$S_3 \rightarrow \{S_3, S_2\}$

$q = A+2D \rightarrow q' = (S_1 + 0.51S_2) + 2(S_2 + 0.6S_3)$

$= S_1 + 2.51S_2 + 1.2S_3$

นายภัทรชัย วงรณสันทัด 65010815

## Scalar Clusters → หาคำที่สอดคล้อง

vector คลัส. ของ stem : $\vec{s}_u = (c_{u,1}, c_{u,2}, \ldots, c_{u,t})$

ขนาดของ stem : $|s_u| = \sqrt{c_{u,1}^2 + c_{u,2}^2 + c_{u,3}^2}$

Normalized Corr. Matrix : $s_{u,v} = \dfrac{\vec{s}_u \cdot \vec{s}_v}{|\vec{s}_u| \times |\vec{s}_v|}$

ex. Query $= 3S_1 + S_3$

① เทียบ vector คลัส. ของ stem

ex. $\vec{s}_1 = (c_{1,1}, c_{1,2}, c_{1,3}) = (5, 6, 1)$

Database $\vec{s}_3 = (c_{3,1}, c_{3,2}, c_{3,3}) = (1, 0, 2)$

$c_{Network, Algebra}$, $c_{Network, Math}$, $c_{Network, Set}$

stem Network = {Set(2), Algebra(1), Math(0)}

② หาขนาดของแต่ละ stem

ex. $|s_3| = \sqrt{1+0+4} = 2.236$

$|s_1| = \sqrt{25+36+1} = 7.874$

③ หา Normalized Corr. Matrix คำนวณให้ครบทุก stem

ex. $s_{1,3} = \dfrac{\vec{s}_1 \cdot \vec{s}_3}{|\vec{s}_1| \times |\vec{s}_3|} = \dfrac{(5\times1)+(6\times0)+(1\times2)}{7.874 + 2.236} = 0.398$

| $S$ | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| $S_1$ | 1 | 0.986 | 0.398 |
| $S_2$ | 0.986 | 1 | 0.248 |
| $S_3$ | 0.398 | 0.248 | 1 |

④ หา New Query

Stem Relation

$S_1 \rightarrow \{S_1, S_2\}$

$S_2 \rightarrow \{S_2, S_1\}$

$S_3 \rightarrow \{S_3, S_1\}$

$q = 3S_1 + S_3$

$q' = 3(S_1 + 0.986S_2) + (0.398S_1 + S_3)$

$= 3.398S_1 + 2.958S_2 + S_3$

## Text Operations → จัดการข้อความให้เหมาะกับการเปรียบเทียบ (doc, Query)

① Lexical Analysis บางคำต้องเอาเครื่องหมายบางอย่างออก

② Elimination of Stopwords

③ stemming → มี 4 วิธี

↳ affix removal, table lookup, successor variety, N-gram

### successor variety

ex. corpus : able, axle, accident, ape, about

test word : apple

| Prefix | Successor Variety | Letters |
|---|---|---|
| a | 4 | b, x, c, p |
| b | 2 | l, o |
| o | 1 | u |
| u | 1 | t |
| t | 1 | blank |

### n-gram → $S = 2C / A+B$

ex. statistics ⇒ st ta at ti is st ti ic cs

unique diagrams ⇒ at cs ic is st ta ti

statistical ⇒ st ta at ti is st ti ic ca al

unique diagrams ⇒ al at ca ic is st ta ti

$S = 2(6)/7+8$

### Affix Removal Stemmers

ex. ถ้าลงท้าย ies และไม่ใช่ eies, aies เปลี่ยน ies → y

④ Index Terms Selection → คำเดียว = ความ

## Text Compression

### Huffman Coding → สร้าง Tree

#### case 1

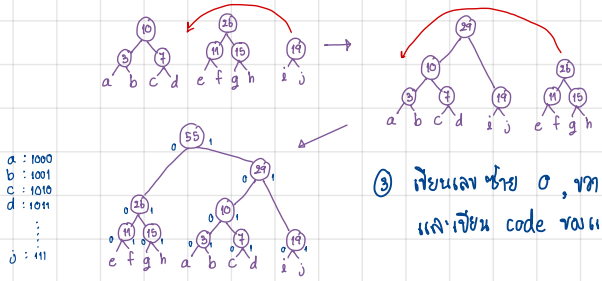
ex. ① รวมความถี่เป็นคู่ 2 ชิ้น

เรียงน้อยไปมากก่อนทำ

| a : 1 | b : 2 | c : 3 | d : 4 | e : 5 | f : 6 | g : 7 | h : 8 | i : 9 | j : 10 |
|---|---|---|---|---|---|---|---|---|---|

3, 7, 11, 15, 19 → 10, 26, 19 } 2 ชิ้น

② ทดกิ่งที่ได้และเรียงลำดับน้อย → มาก และค่อยสร้าง Root เพิ่ม loop ②





a : 1000
b : 1001
c : 1010
d : 1011
⋮
j : 111

③ เขียนเลข ซ้าย 0, ขวา 1 และเขียน code ของแต่ละตัว

#### case 2 → save กว่า case 1

ex. ① รวมความถี่ทีละคู่ และเรียงทันทีหลังจากรวม ทำซ้ำจนเหลือ 2 กลุ่ม

| a : 1 | b : 2 | c : 3 | d : 4 | e : 5 | f : 6 | g : 7 | h : 8 | i : 9 | j : 10 |
|---|---|---|---|---|---|---|---|---|---|

3, 6

ตัวนี้คนเดียว เพราะอยู่ตรงเดิม

| d : 4 | e : 5 | abc : 6 | f : 6 | g : 7 | h : 8 | i : 9 | j : 10 |
|---|---|---|---|---|---|---|---|

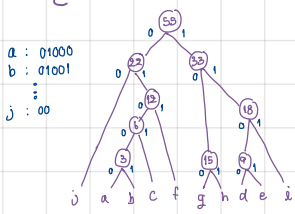
9

de : 9

ผลลัพธ์

| jabcf : 22 | ghdei : 33 |
|---|---|

② ค่อยๆแตกเป็น tree และเขียน code ให้แต่ละตัว



a : 01000
b : 01001
⋮
j : 00

encode : สมมติต้องส่ง e a b a d c e d

01 000 11 000 001 10 01 001 ต่อกัน

decode : ทำตลายกัน

### Ziv Lempel Welch

ex. ① เทียบ P, C, P+C ไล่ input ทีละตัว เทียบ character แต่ละตัวใน Dict

Dictionary

| | |
|---|---|
| 1 | A |
| 2 | B |
| 3 | W |

Encode Text : WABBAWABBA

P : W → P : W, A → P : W, A, B

C : W → C : W, A → C : W, A, B

P+C : W → P+C : W, WA → P+C : W, WA, AB

② ถ้า P+C มีใน Dict : P+C → P ไม่มีใช้ C → P และเพิ่มลง Dict

Dictionary

| | |
|---|---|
| 1 | A |
| 2 | B |
| 3 | W |
| 4 | WA |
| 5 | AB |

③ ไล่ไปเรื่อยๆ จนเสร็จ

P : W, A, B, B, A, W, WA, B, BB, A

C : W, A, B, B, A, W, A, B, B, A

P+C : W, WA, AB, BB, BA, AW, WA, WAB, BB, BBA

มากลงเมื่อเข้า Dict

Dictionary

| | |
|---|---|
| 1 | A |
| 2 | B |
| 3 | W |
| 4 | WA |
| 5 | AB |
| 6 | BB |
| 7 | BA |
| 8 | AW |
| 9 | WAB |
| 10 | BBA |

④ เทียบ Encode output จาก P เทียบใน Dict Encode : 31221461

⑤ ทำ Decode

Decode : 31221461

Dictionary

| | |
|---|---|
| 1 | A |
| 2 | B |
| 3 | W |

Pw : 3, Cw : 1, StrP : W, StrC : A

P : W, C : A, P+C : WA

Output : WA

Decode : 31221461

Dictionary

| | |
|---|---|
| 1 | A |
| 2 | B |
| 3 | W |
| 4 | WA |
| 5 | AB |
| 6 | BB |
| 7 | BA |
| 8 | AW |

Pw : 1, Cw : 4, StrP : A, StrC : WA

P : A, C : W, P+C : AW

Output : WABBAWA

ไล่ C แต่ตัวหน้า

Decode : 31221461

Dictionary

| | |
|---|---|
| 1 | A |
| 2 | B |
| 3 | W |
| 4 | WA |
| 5 | AB |
| 6 | BB |
| 7 | BA |
| 8 | AW |
| 9 | WAB |
| 10 | BBA |

Pw : 6, Cw : 1, StrP : BB, StrC : A

P : BB, C : A, P+C : BBA

Output : WABBAWABBA

นายภัทรชัย วรรณสินทัต 65010815

# Indexing and Searching

→ ทำให้ค้นหาได้เร็วขึ้น , 3 วิธี

**Inverted files** → เก็บ word ไว้ doc/pages (ตน.) ไหน

ex. "abacus" (WORD) : 3, 19, 22 (DOCUMENT)

↳ Block Addressing → ก็ต้องทราบตน. แน่นอน ห้ามใช้

| Block 1 | Block 2 | Block 3 | Block 4 |
|---|---|---|---|
| That house has a | garden. The garden has | many flowers. The flowers | are beautiful |

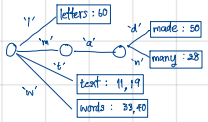
↳ Searching

↳ Construction → ถ้ามีคำใหม่ ⇒ ต่อท้าย (add new position)

เรียงตาม Dictionary (แต่ มีแค่ 5 ตัว) จากซ้าย → ขวา ไม่นับ "."

1 6 9 11 17 19 24 28 33 40 46 50 55 60

This is a text. A text has many words. Words are made from letters. text




Vocabulary trie

↳ เหมาะกับข้อมูลที่นิ่งเดิม ไม่เปลี่ยนแปลง (≠ ตลาดหุ้น)

# Suffix Trees and Suffix Arrays

ex. เรียงตาม Dict มี '.' อยู่หลัง

This is a text. A text has many words. Words are made from letters.

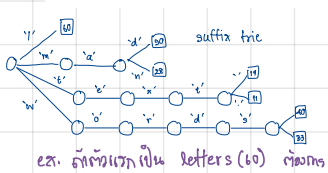
Text

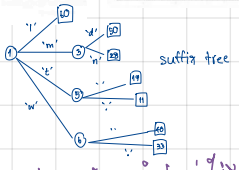
text. A text has many words. Words are made from letters.
text has many words. Words are made from letters.
many words. Words are made from letters.
words. Words are made from letters.
Words are made from letters.
made from letters.
letters.

Suffixes

| 60 | 50 | 28 | 19 | 11 | 40 | 33 |
|---|---|---|---|---|---|---|

Suffix Array




ex. ถ้าตัวแรกเป็น letters (60) ต้องการ หา h... หาตัวแรกเสร็จรู้เลยว่าไม่มี

↳ สร้าง Array เพื่อให้สามารถเริ่มเข้าคำจุดอื่นๆ ได้ (มีหลายชั้นได้)

| left | text | word |
|---|---|---|
| 60 | 50 | 28 | 19 | 11 | 40 | 33 |

Supra-Index

Suffix Array

# Signature Files

→ แบ่งข้อมูลเป็น Block คน. ตัวเท่ากัน

↳ เอา hw ของแต่ละ Block ไปเข้า Hashing function เพื่อให้ได้ code ของ block

มี 2 ตัว เข้า hashing func. แล้วเอามา OR

| Block 1 | Block 2 | Block 3 | Block 4 |
|---|---|---|---|
| This is a text. | A text has many | words. Words are | made from letters. |

Text

| 000101 | 110101 | 100100 | 101101 |
|---|---|---|---|

Text signature

h(text) = 000101
h(many) = 110000
h(words) = 100100
h(made) = 001100
h(letters) = 100001

Signature function

ex. เมื่อมี Query เข้ามาเอา Query ไปเข้า hash แล้วเอาไป AND แต่ละ Block

ถ้าได้เท่าเดิม Block นั้นมี hw ที่ต้องการ (ถ้าได้ตัวเดิม แล้ว Block นั้นไม่มี → hash func. แย่ ต้องเปลี่ยน)

# Boolean Diagram

- adj ต่อท้ายที่ ex. abacus adj actor
- near 4 → อยู่ห่างไม่เกิน 4 คำ ex. abacus near 4 actor

Priority
1. adj, near
2. and, not
3. or

# Brute Force

→ ง่าย, พื้นฐาน, ใช้พลังงานเยอะ, hard search

ex Query = "for"

i: 0 1 2 3 4 5 6

b r u t e f o r c e ไล่ไปเรื่อยๆ จนเจอ

f o r

f o r

for

# Boyer Moore

↳ เลื่อนตามตาราง , เลื่อนตามความยาว (length)

ex. Query = THOTH

① สร้างตาราง กฎคือ index เริ่มที่ 0 , ใช้ตัวอักษรนั้นที่อยู่ท้ายๆ แทวก่อน , ไม่นำตัวสุดท้ายมาคำนวณ (ใช้ length แทน)

ex. THOTH

Length = 5

Value = length - index - 1

T = 5 - 3 - 1 = 1
H = 5 - 1 - 1 = 3
O = 5 - 2 - 1 = 2

| Letter | T | H | O | * |
|---|---|---|---|---|
| Value | 1 | 3 | 2 | 5 |

② ลองเลื่อนไปทางขวา doc เป็นตัวเทียบกับตาราง

doc : T R U S Z H A T T H T H O T H

Q : T H O T H
T H O T H
T H O T H
T H O T H

③ ถ้าเจอแล้ว doc ยังมีต่อให้หาต่อ

# Kruth - Morris - Pratt

→ ดีน้อยกว่า Boyer

คล้าย Boyer แต่เช็คตัวหน้า และ สร้าง ตารางจาก Prefix suffix

ex. Query = ACAGAT

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Pattern[i] | A | C | A | C | A | G | T |
| Prefix[i] | 0 | 0 | 1 | 2 | 3 | | |

① ไล่ทีละ Index แล้วดูว่ามี Duplicate มั้ย ถ้ามีเอามากสุด

| | | |
|---|---|---|
| A | Preffix : | Suffix : |
| C | Preffix : A | Suffix : C |
| A | Preffix : A, AC | Suffix : A, CA |
| C | Preffix : A, AC, ACA | Suffix : C, AC, CAC |
| A | Preffix : A, AC, ACA, ACAC | Suffix : A, CA, ACA, CACA |

Dupe 1 , Length 1
" 1 , " 2
" 2 , " 3

⋮

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Pattern[i] | A | C | A | C | A | G | T |
| Prefix[i] | 0 | 0 | 1 | 2 | 3 | 0 | 0 |

② เช็คแบบ Boyer แต่ดูตัวแรกสุด

# Google Tech

① Crawling → ใช้ตัว google bot/spider รวมข้อมูลบนหน้า web ใหม่ๆ , index update

มี 3 ขั้นตอน → กรองด้วย filter, แยก/ตัดคำ , วัดดัชนี

② Page Rank → วัดค.สำคัญหน้า web

ตั้งเลขบอกค.สำคัญ , 0-10 ยิ่งมาก = แสดงก่อน , ตั้ง page rank คำนวณจาก คน.อ้างอิงหน้าอื่น

③ Google Query

↳ คำที่ค้นหาจะถูกนำมาใช้หมด

↳ ผลที่ได้จากการค้นหาจะตรงกับคำที่ใช้

↳ ขยายคำย่อให้ ex. → example

↳ ค้นคำใกล้เคียง child + children

↳ stopwords ไม่ใช้เป็น the, on, where, how

↳ จำกัด 32 คำ

↳ สำดับคำที่จะใช้ ค.สำคัญกับหน้าที่มีคำเรียงตามลำดับตรงกับที่ใช้ค้น

↳ ไม่สนอักษรใหญ่เล็ก

↳ อักขระพิเศษบางตัวจะไม่ใช้ แต่บางกรณีใช้

↳ Apostrophes("'") จะตัด

↳ ถ้า "-" จะตัดคำค้นเพิ่มให้ครอบคลุม

ex. E-mail, Email, E mail