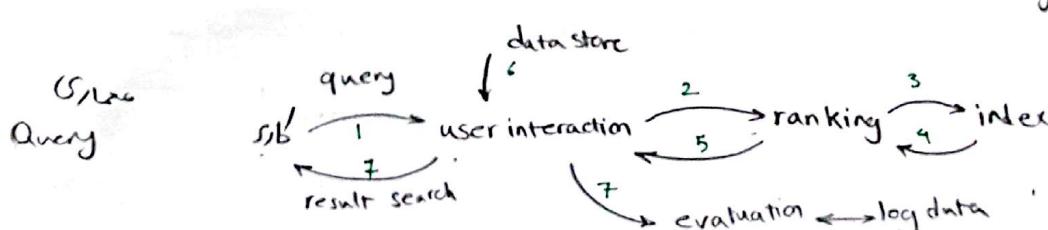
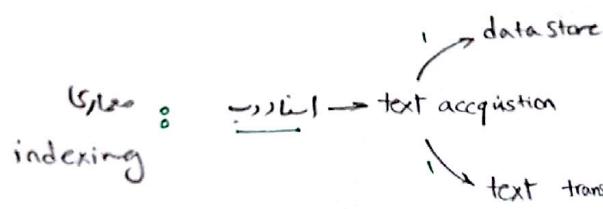
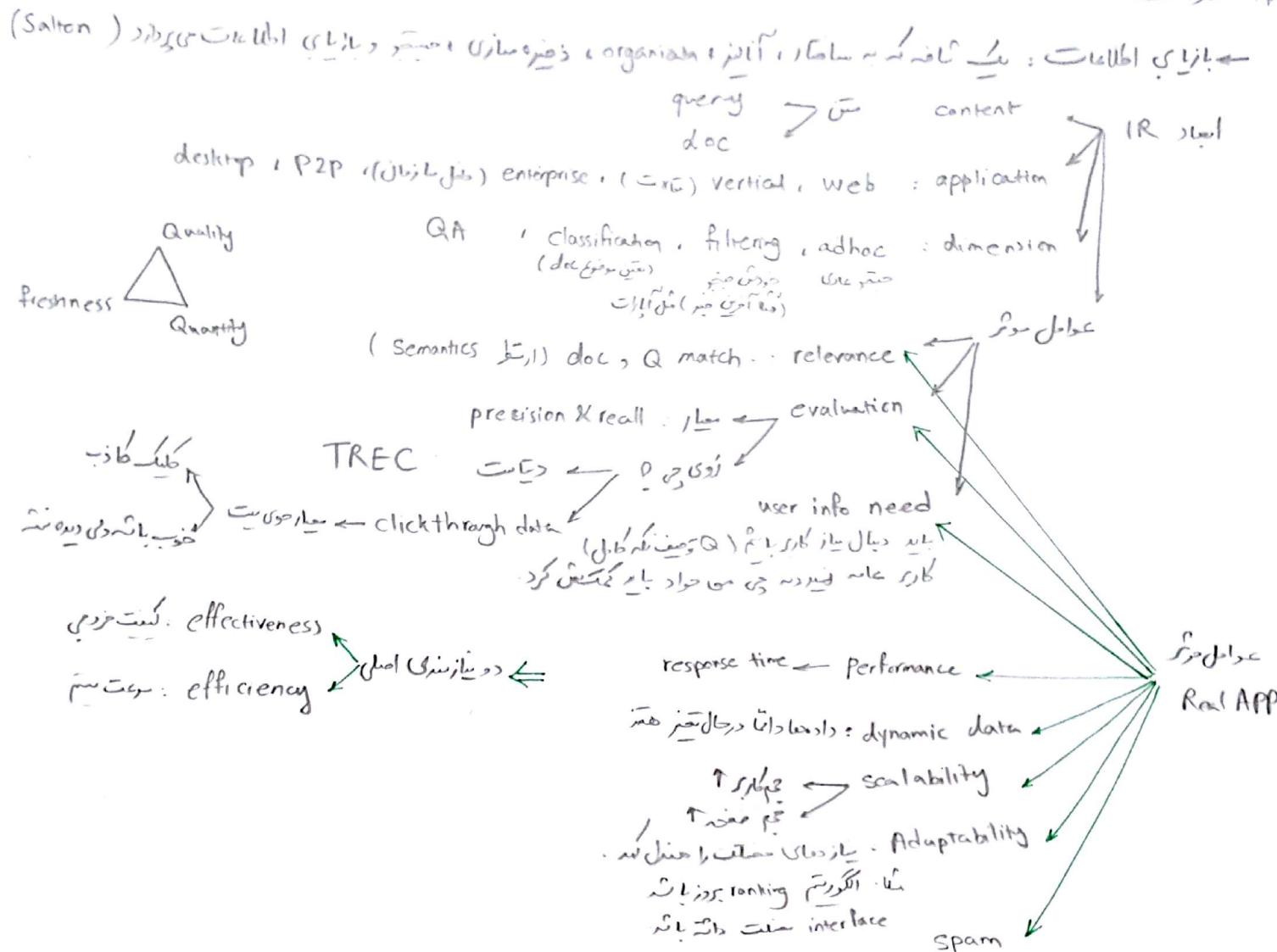


# Information Retrieval





بررسی: (1)  $\text{Q}(\text{Q}_1 \text{Q}_2)$  سیده سنت میکو تووصیل  
 جواب: جواب تا قدر  $\text{Q}_1 \text{Q}_2$  و  $\text{Q}_2 \text{Q}_1$  میتواند  $\text{Q}_1 \text{Q}_2$  باشد

(vector doc units) doc must score  $\leq$  query : Vector space  
 یعنی  $\Delta$  باید این بایاسی خود  $\log$  گرفت

$$\log\left(\frac{N}{df_t}\right) = IDF \quad \text{TF-IDF}$$

نحوه  
نحوه  
نحوه  
نحوه  
نحوه  
نحوه

$$TF_{t,d} \cdot \log\left(\frac{N}{df_t}\right) = w_{T,d}$$

$$cos(q, d) = sim(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| \cdot |\vec{d}|} = \frac{\sum w_{ti,q} \cdot w_{ti,d}}{\sqrt{\sum w_{ti,q}^2} \cdot \sqrt{\sum w_{ti,d}^2}} \quad \text{doc, Q چیزی را بیندازند: } \underline{(t_i)} \text{ را بخواهند}$$

ستاره حضیری: سرعت: حریم‌ساز تباخت می‌توان استاد. کرد

بره: استقال طلب - بندیت مترادف - ارزش سنجی حافظ عنده

$$P(R|D)$$

ترجمات اولیه ← تا دیگری VS : در similarity, معنی، با احتال بودن Relevance

→ PRP میں خوبی دل احتمالات کو دارہ Sort برائیں P(RID) برائیں: اصل

binary independent model : BIM //

$$J, q \in U_1 = P(R=1|d, q) = \frac{P(d|R=1, q) \cdot P(R=1|q)}{P(d|q)}$$

$$\bullet RSV = \sum_{\text{docs} \in S_{t+1}(q)} \log \frac{P_t \cdot (1 - u_t)}{u_t \cdot (1 - P_t)} \cdot \frac{P(x_t=1 | R=0, q)}{P(x_t=1 | R=1, q)} \quad \leftarrow RSV = \text{ranking}$$

$$O(R|d, q) = \frac{P(d|R=1, q)}{P(d|R=0, q)} \cdot \frac{P(R=1|q)}{P(R=0|q)} = O(Nq) \cdot \prod_{x \in d} \frac{P(x|t|R=1, q)}{P(x|t|R=0, q)}$$

$$\frac{1}{\pi} = \prod_{x_t=1} \frac{P(x_t=1|R=1,q)}{P(x_t=1|R=0,q)}, \prod_{x_t=0} = O(R|q), \prod_{x_t=1} \frac{P_t}{U_t}, \prod_{x_t=0} \frac{1-P_t}{1-U_t} \xrightarrow{t \rightarrow \infty} \prod_{x_t=1} \frac{1-P_t}{1-U_t} \times \prod_{x_t=0} \frac{U_t}{1-P_t}$$

$\log \frac{\frac{1}{2}}{\frac{dF}{N}} \cdot \frac{(1-dF)}{\frac{1}{2}} = \log \frac{N}{dF} \leftarrow \text{idf}$   $\leftarrow \frac{dF}{N} = u_t$ ,  $\frac{1}{2} = p_t$   $\rightarrow$   $\text{idf} = \log \frac{N}{dF}$

$$BM25_{q,d} = \sum_{t_i \in q} \log \left[ \frac{N}{df_{t_i}} \right] \cdot \frac{(k+1) \cdot TF_{t_i, d}}{k \cdot (b \times (L_d / L_{Avg}) - b + 1) + TF_{t_i, d}} \quad 1.2 < k < 2$$

$b = 0.75$

(  $P @ 10$  )  
Precision  $\leq P'$

recall میتوشم حساب کنم  

$$q, d \quad t_i \in q \quad \partial [df_{t_i}] \quad k(b \times (L_d / L_{avg}) - b + 1) + 1 \quad t_i, d$$
 از تکمیل  $L_d$  سوداگری Precision  $(P@10)$   
 میتوانم داده precision و recall را محاسبه کنم  
 click through rate = relevance (سینه خوب است، حق است)

نیز  $\rightarrow$  جم اطلاعات  $\uparrow$   $\rightarrow$  نوزٹا  $\leftarrow$  شابه داریای روی حسنه حاصله اندکار است

عوامل دلیلی و موقایعی سنت - rank : درین بازار احتمال

Snippet popularity ↗

مُدَّةً إِقْرَابًاً مُّدَّةً إِقْرَابًاً / Dwell time

3

$\text{AB TCF} = \frac{\text{بـ صورت مولـون}}{\text{ـ ۱۰ زرـست}} + \frac{\text{ـ ۲ صورـت}}{\text{ـ ۱۰ زرـست}} + \frac{\text{ـ ۳ صورـت}}{\text{ـ ۱۰ زرـست}} + \frac{\text{ـ ۴ صورـت}}{\text{ـ ۱۰ زرـست}} + \frac{\text{ـ ۵ صورـت}}{\text{ـ ۱۰ زرـست}} + \frac{\text{ـ ۶ صورـت}}{\text{ـ ۱۰ زرـست}} + \frac{\text{ـ ۷ صورـت}}{\text{ـ ۱۰ زرـست}} + \frac{\text{ـ ۸ صورـت}}{\text{ـ ۱۰ زرـست}} + \frac{\text{ـ ۹ صورـت}}{\text{ـ ۱۰ زرـست}} + \frac{\text{ـ ۱۰ صورـت}}{\text{ـ ۱۰ زرـست}}$

↑ relevance - freshness - UI - response time ← probability

	relevant	not relevant
retrieved	TP	FP
not retrieved	FN	TN

maximize

جودی برآورده شدن  
جودی برآورده شدن

tradeoff

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

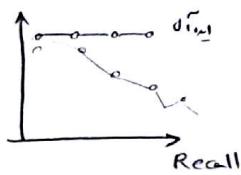
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F\text{-measure} = \frac{(B^2 + 1)P \cdot R}{B^2 P + R}$$

$$\text{harmonic mean} = \frac{2PR}{P+R}$$

$$\text{Accuracy} = \frac{TN + TP}{All}$$

Precision



محلار گلاظه رون ترست باری .

IR ← ↑ جودی برآورده شدن + جودی

↑ P اما مثلاً P را با عبارت عالی نظر نمایم

جودی برآورده شدن داشت انتقال می شود .

Average Precision =  $\frac{\sum AP_i}{n}$

$\frac{1}{n} \sum_{q_i=1}^n AP_i = MAP = \text{Mean Average Precision}$

$P @ k$  = اول k جواب را حساب کنید (جودی

جواب اولین جواب + جواب اولین جواب + ... + جواب اولین جواب)

RR = MRR

$\frac{1}{n} \sum RR_i = MRR$

DCG = discounted cumulative gain = DCG

$DCG_P = rel_1 + \sum_{i=2}^k \frac{rel_i}{log(i)}$

= normalized DCG = NDCG

$NDCG = \frac{DCG}{\text{Perfect ranking DCG}}$

H0: جودی برآورده شدن = Significant Test

$$df = n_1 + n_2 - 2$$

$$T\text{-test} = \frac{\text{signal}}{\text{noise}} = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_1} + \frac{var_C}{n_2}}} \rightarrow$$

اجتنابی کر شنیداده، مسأله باشد

↓ P، df، ↑ T-test: حرجی  $\Delta$ . جون اصل سانی بخوبی می شود

حاجی در یک جمعیت کم و زیاد می تواند باشد! (دست اولی کم می شود)

t-test =  $\frac{\sum d}{N \sum d^2 - (\sum d)^2}$  جمع اختلافها

$$df = N - 1 \rightarrow$$

clue web - TREC - cranfield bench mark

- Al-hoc - بینهایت - داده مفهومی

× recall ← داده مفهومی doc می باشد

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

$$P(E) = P(\text{nonrelevant})^2 + P(\text{relevant})^2$$

$$④ K = \frac{n_{\text{relevant}}}{n_{\text{total}}} + 0.8$$

validity  
relevance  
assessment

$$K = \frac{n_{\text{relevant}}}{n_{\text{total}}} + 0.8$$

$$P(D|Q) = \frac{P(Q|D) \cdot P(D)}{P(Q)} \quad \text{إذا كان } Q \text{ معلوماً باتساعه بين } \leftarrow$$

$$P(Q|D) = \prod_{i=1}^n P(q_i|D) \leftarrow \text{model}$$

Multinomial = unigram

$P(Q|D) = \prod_{w \in q_i} P(w|D) \cdot \prod_{w \notin q_i} (1 - P(w|D))$  multiple beroulli

doc  $\rightarrow$  sequence of  $N$ -gram

$P(Q|D)$

$P(w|D) = \frac{\text{TF}_w}{|D|}$  maximum likelihood estimation

$$\text{Bigram } P(q_1 | D) = P(q_1 | D) \cap P(q_2 | q_1, D)$$

$$\text{Trigram } P(Q|D) = P(q_1|D) \cdot P(q_2|q_1, D) \cdot \pi P(q_t|q_{t-1}, q_{t-2}, D)$$

Skip N-gram  $\rightarrow$  متن سه سویی را در نظر نمایند و فقط بین کلماتی که در نظر نمایند از نظر dependency رابطه گذاری داشته باشند.

تعريف:  $P(Q|D) = 0$  يعنى، حاصل خوب  $\rightarrow$  محسود. اى خلى بد. وان دل =

$$P(w|D) = \frac{TF_w + 1}{|D| + k} : \text{Add 1 / Laplace smoothing}$$

$$P(w|D) = \frac{TF_w + E}{|D| + k \cdot E} \quad \text{Vocab} \quad \text{lindstone correction}$$

مقدمة ملخصه دروس

$$P(w|C) = \frac{CF_w}{|C|}$$

متعدد الكلمات من

Vocab

**تمامیات دارای چه مزایاں - ازاد حاوی کے محتویات** : absolute discounting

$X_{background}$  = مقدار خودش برای این background  
 برای که سایر افعال را backoff :  $X_{background}$   
 تخمین می‌کند .  
 اما  $X_{background}$  =  $\phi$

$$P(w|D) = \lambda \frac{TF_w}{|D|} + (1-\lambda) \frac{CF_w}{|C|}$$

$$\lambda = |D| / (|D| + \mu)$$

$\lambda$  نے  $\Omega$  میں  $u$  کا Dirichlet boundary condition  $u = 0$  کا  $\mathcal{L}$  operator کا inverse ہے۔

$$\lambda = |D| / (|D| + V) \quad \text{smoothing} \leftarrow \text{عمرز داد یوسفونج خاص} - \text{عمرز داد سری اول} \quad \text{crim: } \alpha_1 : \text{witten-bell}$$

$$P(w|D) = \frac{TF_w + \mu}{|D| + V} \cdot \text{CE}_w / |C|$$

$$\text{Bigram : } P(w_i | w_{i-1}, Q) = \lambda_1 \left[ \lambda_2 \frac{\text{TF}_{w_i} - \mu_{i-1}}{\text{TF}_{w_{i-1}}} + (1-\lambda_2) \frac{\text{TF}_{w_i}}{|\mathcal{D}|} \right] + (1-\lambda_1) \cdot \frac{\text{CF}_{w_i}}{|\mathcal{C}|}$$

ایده: کلید تکرار دار که در میان متن - دلیل در خود ادن Q یکسان نیست (امثله موتیال و ترجیح در مورد کلمه ورزشی)

$$P(w|D) = \lambda_1 \left( \lambda_2 \frac{TF_W}{|\mathcal{D}|} + (1-\lambda_2) \frac{TF_{W, \text{cluster}}}{|\text{cluster}|} \right) + (1-\lambda_1) \frac{CF_W}{|\mathcal{C}|}$$

$$P(Q|M_D) = \prod_i P(q_i|M_D)$$

کوئی سر doc میں مل زدگی کی سازنے کے داریں  
کوئی Q مل نہیں کر سکے feed back vs feedback من خود رہے

$$P(D|M_Q) = \prod P(d_i|M_Q) \quad \text{سازه سیستم doc می باشد - Q مجموع کتابه دستوراتی قوای تولید نگذشت}$$

$$KL(M_D || M_Q) = \sum P(w|M_Q) \log \frac{P(w|M_Q)}{P(w|M_D)} = \sum P(w|M_Q) \log P(w|M_Q) - P(w|M_Q) \cdot \log P(w|M_D)$$

model comparison

N-gram	IDF	doc length	TF	
X	✓	cosine ✓	✓	vector space
X	✓	$\sqrt{1 + D}$ ✓	✓	BM25
✓	CFV ✓	(5) $ D  \geq 5$ ✓	✓	language model

$$\overrightarrow{Q}_{opt} = \mu(DC_r) + [\mu(DC_r) - \mu(DC_{nr})]$$

IDF-TF by contrast

$$\overrightarrow{Q}_m = \alpha \overrightarrow{Q}_0 + \beta \mu(DC_r) - \gamma \mu(DC_{nr})$$

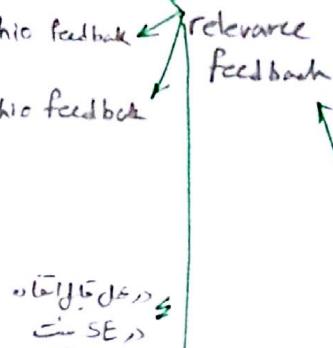
vs result ✓

•  $\mu(DC_r)$  is Q likelihood  $\rightarrow$  result X

• KL divergence  $\rightarrow$  model comparison  $\rightarrow$  result ✓

$$KL(M_D || M_Q) = -\sum P(w|M_Q) \cdot \log(P(w|M_D))$$

Model



$$M_Q' = (1-\alpha)M_Q + \alpha M_F$$

$\arg \max_m \sum \sum C(WD_i) \cdot \log P(w|M)$

residual collection  $\overrightarrow{Q}_m$  is evaluated against  $Q_0$  to get  $Q_m$  = evaluation

ادن هایی که طبیعت متن مغایر دارند در درنظر گرفته شوند = residual collection

خوبی توجه کنید: عبارتی که کامپوننتهای این سیستم

سیستم  $\leftarrow$  modified query - این کامپوننت را بعنوان ضربه + درنظر نگیرید.  $\leftarrow$  اول (کوچک) را بعنوان ضربه + درنظر نگیرید.  $\leftarrow$  این دست  $\uparrow$

word mismatch  
vocab gap

حدف:  
recall  $\uparrow$

Vector space  $\leftarrow$  Q

$Q_s$ : vocab  $\leftarrow$  مابین طبقاتی  $\leftarrow$  جزو

مانند  $\leftarrow$  Q expands

مانند  $\leftarrow$  دوستانه داشته باشند

statistical: تحلیل تاریخی از متن داده دارند

synonyms: similar  $\leftarrow$  دوستانه داشته باشند

thesaurus: دوستانه داشته باشند

manual: دستی تراویح از متن داده دارند

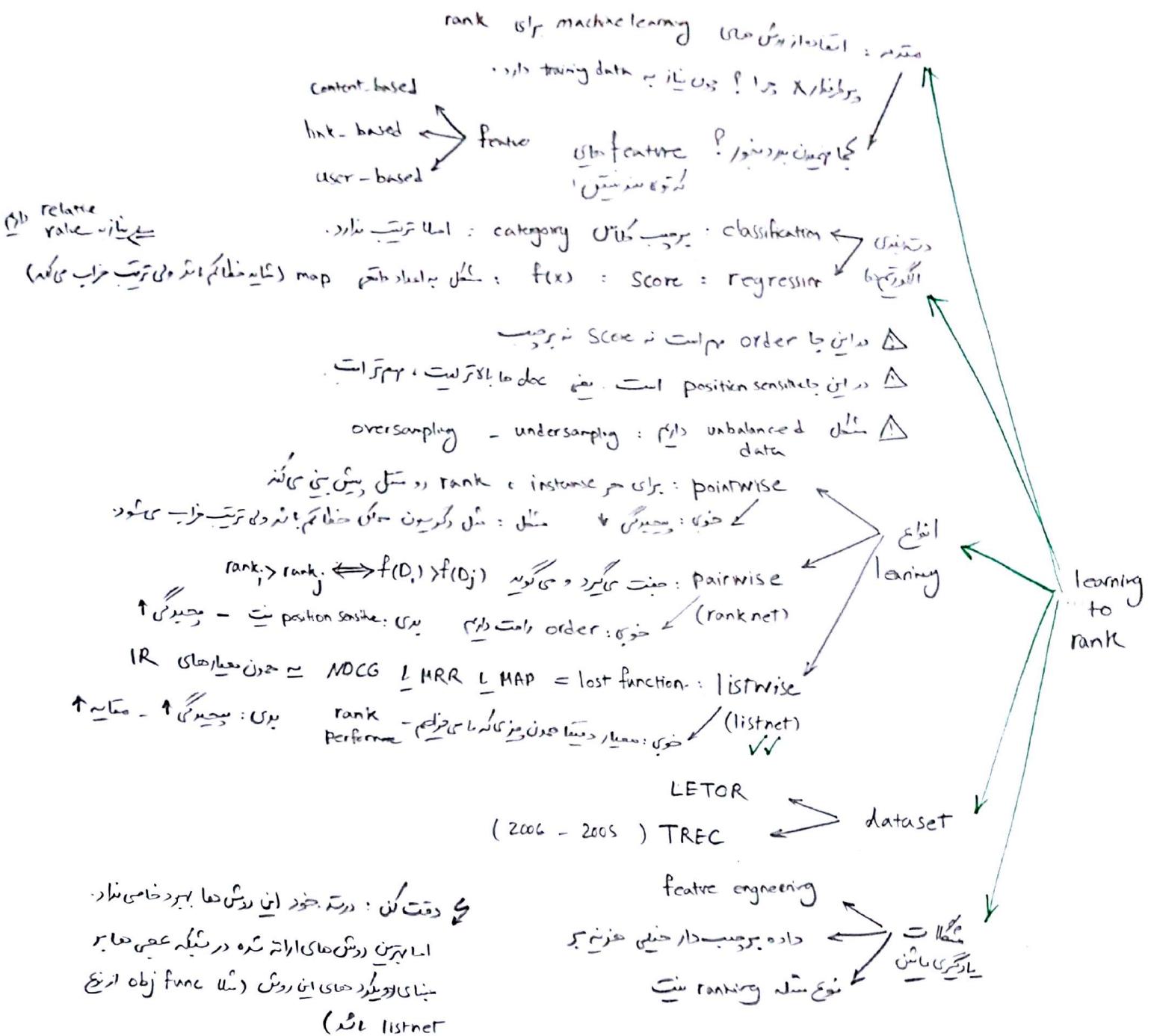
co-occur

او را که با هم-نموده باشد  
است similar  $\leftarrow$  دوستانه داشته باشند

اضافه کردن مطالبی که در دیگر Q session داشته باشند

- وردیلد - Query logs  $\leftarrow$  url خود را می‌دانند و تغیرات پنهان

وریژن



wikipedia > info-box (eg . nlp, IR, DBpedia: min,

در قالب مدل Subject-predict-object (SPO) در دست زانی ذخیره شود.

