

## PRP PROBABILISTIC MODEL

La probabilità che un documento  $d$  è rilevante per la query  $q$ :

$$P(z | d, q) = \frac{P(d | z, q) P(z)}{P(d)}$$

d'altra parte abbiamo la non rilevanza

$$P(\bar{z} | d, q) = \frac{P(d | \bar{z}, q) P(\bar{z})}{P(d)}$$

$\delta$  deve valere la normalizzazione delle probabilità

$$P(z | d, q) + P(\bar{z} | d, q) = 1$$

## BINARY INDEPENDENCE MODEL (BIN)

Rappresentiamo i documenti come vettori binari grandi  $|V|$ , dove 1 indica che il termine  $i$ -esimo è presente nel documento, 0 altrimenti.

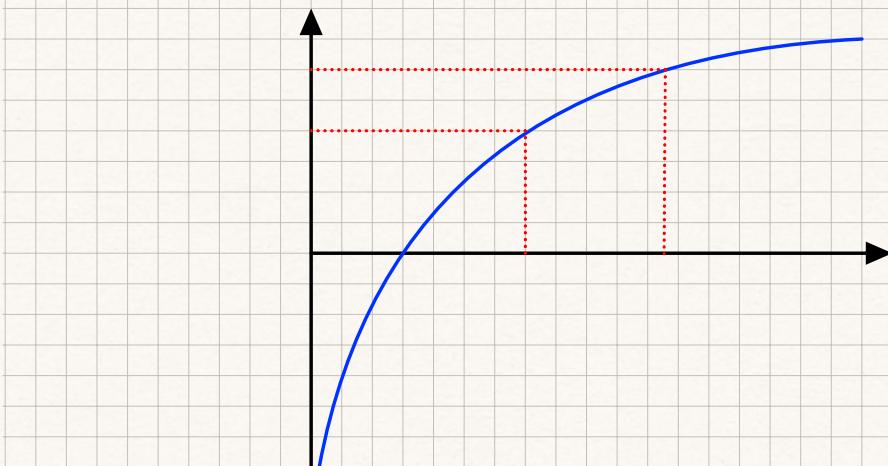
$$d = \{x_1 \dots x_{|V|}\} \text{ con } x_i \in \{0, 1\}$$

dunque  $d$  è rappresentato dal vettore  $\vec{x}$ , per cui ci serve trovare:

$$P(z | \vec{x}, q)$$

Il nostro obiettivo è di stilare una classifica, per farlo si usano funzioni di tipo **ranking preserving**.

Esempio funzione RPF:  $\text{Log}(\cdot)$

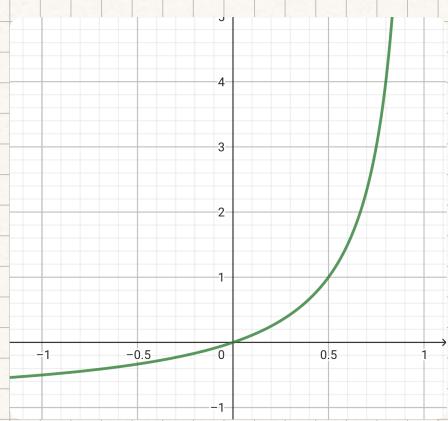


## ODD: Quotazione

Definiamo la quota di un certo evento  $\delta$ :

$$O(\delta) = \frac{P(\delta)}{1 - P(\delta)}$$

Che è il rapporto fra la probabilità che l'evento succede su la probabilità che non succede. Ne possiamo fare il grafico



Per  $P(\delta) < 1$ , questa funzione è rank-preserving

ODD e BIM

Calcoliamo la quota di  $P(\varepsilon|x, q)$ :

$$O(\varepsilon|q, d) = \frac{P(\varepsilon|x, q)}{P(\bar{\varepsilon}|x, q)}$$

Usando il Teorema di Bayes:

$$\frac{\cancel{P(x|\varepsilon, q)} P(\varepsilon|q)}{\cancel{P(x)}} \cdot \frac{\cancel{P(x)}}{P(x|\bar{\varepsilon}, q) P(\bar{\varepsilon}|q)} =$$

$$= \frac{\cancel{P(x|\varepsilon, q)} P(\varepsilon|q)}{\cancel{P(x|\bar{\varepsilon}, q)} P(\bar{\varepsilon}|q)}$$

Assumendo l'indipendenza dei termini che appaiono in un documento:

$$P(x) = P([0, 1, 0, \dots, 1, 0]) = P(0) \cdot P(1) \cdots P(0)$$

Possiamo quindi scrivere

$$O(\varepsilon|q, x) = O(\varepsilon|q) \cdot \prod_{i=1}^{|V|} \frac{P(x_i|\varepsilon, q)}{P(x_i|\bar{\varepsilon}, q)} =$$

Consideriamo ora che ogni  $x_i$  può essere 0 o 1:

$$= O(\varepsilon|q) \cdot \prod_{i=1}^{|V|} \frac{P(x_i=1|\varepsilon, q)}{P(x_i=1|\bar{\varepsilon}, q)} \cdot \prod_{i=1}^{|V|} \frac{P(x_i=0|\varepsilon, q)}{P(x_i=0|\bar{\varepsilon}, q)}$$

Per leggibilità sostituiamo:

$$P(x_{i=1} | \tau, q) = \tau_i$$

$$P(x_{i=0} | \tau, q) = \bar{\tau}_i$$

Quindi abbiamo:

$$= O(\tau | q) \cdot \prod_{x_{i=1}} \frac{\tau_i}{\bar{\tau}_i} \cdot \prod_{x_{i=0}} \frac{1 - \tau_i}{1 - \bar{\tau}_i}$$

Ora mettiamo nel conto la query.

I termini che non appaiono nella query, e che quindi nel vettore binario della query hanno valore  $q_i = 0$ , sono equiprobabili nell'apprice sia nei documenti rilevanti sia in quelli non rilevanti:  $\tau_i = P_i$ , per cui possiamo non considerarli e scrivere:

$$O(\tau | x, q) = O(\tau | q) \cdot \prod_{\substack{x_{i=1} \\ q_i=1}} \frac{\tau_i}{\bar{\tau}_i} \cdot \prod_{\substack{x_{i=0} \\ q_i=1}} \frac{1 - \tau_i}{1 - \bar{\tau}_i}$$

fanno match  
coi termini di q    non fanno match  
coi termini di q

Moltiplicando e dividendo per la stessa quantità:

$$= O(\tau | q) \cdot \prod_{\substack{x_{i=1} \\ q_i=1}} \frac{\tau_i}{\bar{\tau}_i} \cdot \prod_{\substack{x_{i=1} \\ q_i=1}} \frac{1 - \bar{\tau}_i}{1 - \tau_i} \cdot \frac{1 - \tau_i}{1 - \bar{\tau}_i} \prod_{\substack{x_{i=0} \\ q_i=1}} \frac{1 - \tau_i}{1 - \bar{\tau}_i}$$

Possiamo ora unire, come indicano i rettangoli blu le tre produttivitie. Considerando gli indici, la seconda produttività diventa su tutti i termini del documento e non ( $x_i = 0$  e  $x_i = 1$ ), ma presenti nella query ( $q_i = 1$ )

$$O(\varepsilon |x, q) = O(\varepsilon |q) \prod_{\substack{x_i=1 \\ q_i=1}} \frac{P_i \cdot (1 - \varepsilon_i)}{\varepsilon_i \cdot (1 - P_i)} \cdot \prod_{\substack{q_i=1}} \frac{1 - P_i}{1 - \varepsilon_i}$$

Ora:

1)  $O(\varepsilon |q)$ : costante data la query

2)  $\prod_{\substack{i=1 \\ q_i=1}} \frac{1 - P_i}{1 - \varepsilon_i}$ : costante data la query

3)  $\prod_{\substack{i=1 \\ q_i=1}} \frac{P_i \cdot (1 - \varepsilon_i)}{\varepsilon_i \cdot (1 - P_i)}$ : l'unico non costante data la query.  
Bisogna stimarlo.

## RETRIEVAL STATUS VALUE (RSV)

Prendiamo il termine variabile rispetto alla query e ne facciamo il logaritmo:

$$RSV = \log \prod_{\substack{x_i=1 \\ q_i=1}} \frac{P_i \cdot (1 - \varepsilon_i)}{\varepsilon_i \cdot (1 - P_i)} = \sum_{\substack{x_i=1 \\ q_i=1}} \log \frac{P_i \cdot (1 - \varepsilon_i)}{\varepsilon_i \cdot (1 - P_i)}$$

## STIMA DI $\hat{P}_i$ e $\hat{\epsilon}_i$

	relevant	not-relevant	total
term present	$r_i$	$n_i - r_i$	$n_i$
term not present	$R - r_i$	$(N - n_i) - (R - r_i)$	$N - n_i$
total	$R$	$N - R$	$N$

Dove:

$N$ : numero di docs nel campione

$n_i$ : numero di docs nel campione contenenti  $t_i$

$R$ : numero di docs rilevanti nel campione

$\epsilon_i$ : numero di docs rilevanti nel campione contenenti  $t_i$

Perc un discorso di stabilità numerica, aggiungiamo un fattore di 0.5

	relevant	not-relevant	total
term present	$r_i + 0.5$	$n_i - r_i + 0.5$	$n_i + 1$
term not present	$R - r_i + 0.5$	$(N - n_i) - (R - r_i) + 0.5$	$N - n_i + 1$
total	$R + 1$	$N - R + 1$	$N + 2$

Ricordando che  $\hat{P}_i = \hat{P}(x_i=1 | \Sigma, q)$ , calcoliamo questa probabilità con la tabella:

$$\hat{P}_i = \frac{\epsilon_i + 0.5}{R + 1}$$

Stessa cosa per  $\hat{\epsilon}_i = \hat{P}(x_i=1 | \bar{\Sigma}, q)$

$$\hat{\epsilon}_i = \frac{n_i - \epsilon_i + 0.5}{N - R + 1}$$

Attenzione che queste due  $\epsilon_i$  sono diverse.

Chiamiamo ora  $C_i$  il seguente rapporto:

$$C_i = \frac{P_i (1 - \varepsilon_i)}{\varepsilon_i (1 - P_i)}$$

Da cui:

$$C_i^{\text{BIM}} = W_i^{\text{RSS}} = \log \frac{(\varepsilon_i + 0.5) \cdot (N - R - m_i + \varepsilon_i + 0.5)}{(m_i - \varepsilon_i + 0.5) \cdot (R - \varepsilon_i + 0.5)}$$

Ovvero i **Robert/Sparck-Jones Weights**.

Possiamo facilmente affermare, a questo punto, che:

$$N \gg R \quad e \quad m_i \gg \varepsilon_i$$

da cui deriva l'approssimazione:

$$\varepsilon_i \approx \frac{m_i}{N}$$

Questo perché, data la query, il numero di documenti non rilevanti sarà vicino al numero totale di documenti nello collezione.

$$\varepsilon_i = \frac{m_i - \varepsilon_i + 0.5}{N - R + 1} \approx \frac{m_i}{N}$$

Ma per  $\varepsilon_i \approx 0$  e  $R \approx 0$  abbiamo

$$P_i = \frac{\varepsilon_i + 0.5}{R + 1} \approx 0.5$$

Ottieniamo così quello che si chiama **BEST MATCH L weights**:

$$w_i^{\text{BM1}} = \log \frac{N - n_i + 0.5}{n_i + 0.5} \approx \log \frac{N - n_i}{n_i} \approx \log \frac{N}{n_i} = \text{IDF}_i$$

# POISSON DISTRIBUTION

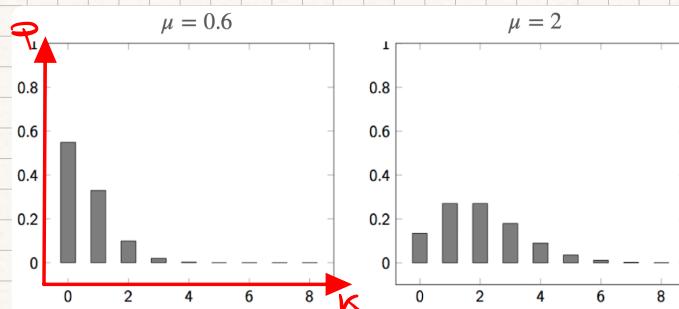
I documenti vengono disegnati indipendentemente dal vocabolario usando una **distribuzione multinomiale**. Un processo multinomiale è una sequenza di variabili aleatorie indipendenti e identicamente distribuite  $X = (X_1, X_2, \dots)$  ognuna delle quali può avere  $K$  possibili valori.

La distribuzione delle **term frequency** è la **distribuzione di Poisson**.

## Modello di Poisson

La distribuzione di Poisson cerca di modellare la probabilità che  $K$  eventi si verifichino in un intervallo di tempo/spazio fisso. Questi eventi si succedono e sono indipendenti. Ciò che sappiamo è che in media si verificano  $\mu$  eventi indipendenti.

$$f(K | \mu) = P(X = K | \mu) = \frac{\mu^K e^{-\mu}}{K!}$$



Noi useremo questa distribuzione per le  $t_f$ , in un intervallo fisso, che per noi vuol dire assumere che le doc-length siano fisse.

Il modello di Poisson è ragionevole se usato con termini molto generali, molto meno se usato con termini così specifici a un certo topic.

Guardando il seguente esempio, appare evidente

Freq	Word	Word occurrences											
		0	1	2	3	4	5	6	7	8	9	10	11
53	expected	599	49	2									
52	based	600	48	2									
53	conditions	604	39	7									
55	cathexis	619	22	3	2	1	2	0	1				
51	comic	642	3	0	1	0	0	0	0	0	1	1	2

Number of documents containing  $k$  word occurrences

La tabella è da leggersi, per il termine "expected" ci sono 599 docs che non hanno nessuna occorrenza di questo termine, 49 con 1 occorrenza e 2 con 2 occorrenze.

Dalla tabella si evince che i primi tre termini seguono Poisson, perché andando avanti nell'intervallo i tf scendono. Gli ultimi due termini, a un certo punto hanno un incremento.

Diffatti l'assunzione fatta è sbagliata, non tutti le parole seguono lo stesso modello. Questo è anche detto **1-Poisson Model**.

Se le parole vengono generate da due modelli diversi allora si parla di **2-Poisson Model**. In particolare i due modi differenziano le parole di élite e quelle non d'élite.

## ELITNESS MODEL

Bisogna capire come i termini sono distribuiti nei docs.

Assumiamo che esista una variabile aleatoria nascosta, detta **elitness**, per ogni coppia documento-termine.

Questa variabile aleatoria, è anche binaria. Il concetto di elite è che un termine è d'elite se il documento che lo contiene ha a che fare col concetto che trasporta quel termine.

Proprietà elitness  $\delta$ :

$$1) t_{fi} \text{ dipende da } \delta_i : t_{fi} = f(\delta_i)$$

$$2) \delta_i \text{ dipende dalla rilevanza } R : \delta_i = f(R)$$

Se rappresentiamo i documenti con vettori di numeri interi ognuno rappresentante  $t_{fi}$ , con la stessa procedura con cui siamo arrivati a BM:

$$RSV^{\text{elite}} = \sum_{i \in q, t_{fi} > 0} c_i^{\text{elite}}(t_{fi})$$

Da cui:

$$c_i^{\text{elite}}(t_{fi}) = \log \frac{P(t_{fi} | \epsilon) P(\epsilon | \bar{\epsilon})}{P(\bar{\epsilon} | \epsilon) P(t_{fi} | \bar{\epsilon})}$$

Usando il concetto di elitness, dividiamo i tipi di termini usando 2-Poisson Model:

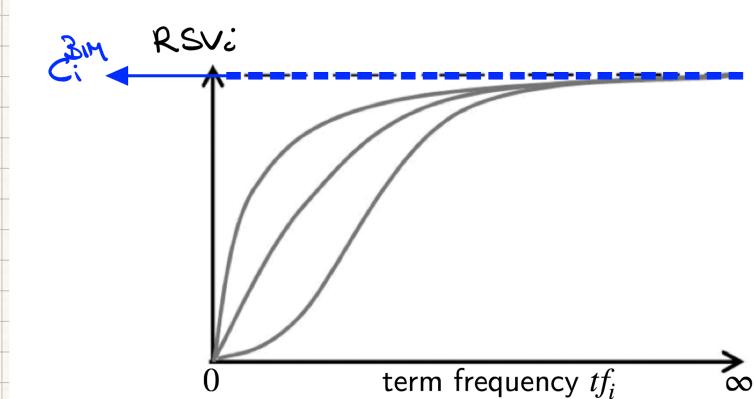
$$P(t_{fi} | \epsilon) = P(t_{f1} | e_1) \cdot P(e_1 | \epsilon) + P(t_{f1} | \bar{e}_1) \cdot P(\bar{e}_1 | \epsilon) =$$

$$= \underbrace{\mathbb{P}(tf_i | e_i)}_{\text{Poisson}} \cdot \underbrace{\pi}_{\text{TC}} + \underbrace{\mathbb{P}(tf_i | \bar{e}_i)}_{\text{Poisson}} \cdot \underbrace{[1 - \mathbb{P}(e_i | \varepsilon)]}_{\pi}$$

Quindi assumiamo di avere due differenti distribuzioni, una per i termini d'élite e una per quelli non d'élite:

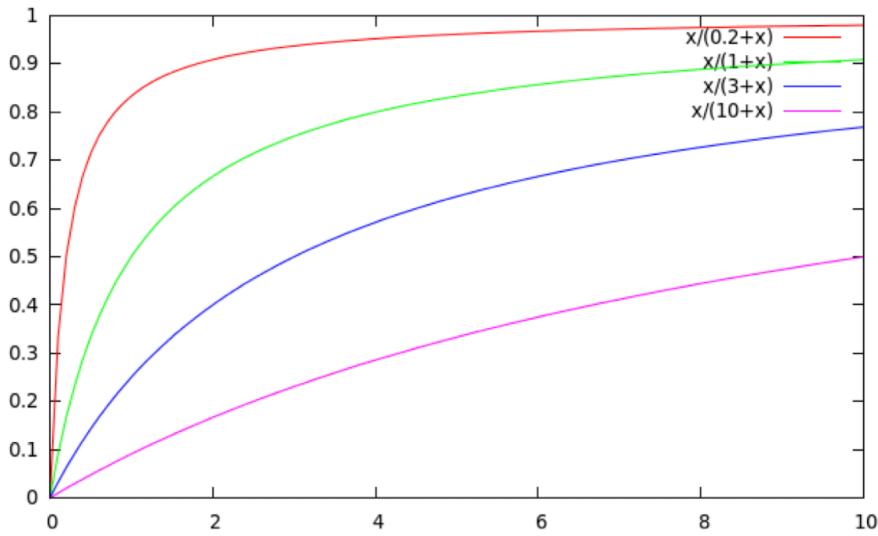
$$\mathbb{P}(k | \varepsilon) = \pi \cdot \frac{\lambda^k}{k!} e^{-\lambda} + (1 - \pi) \frac{\mu^k}{k!} e^{-\mu}$$

Bisogna stimare  $\pi, \lambda$  e  $\mu$ . **Troppo difficile.**



Tutte queste curve RSV<sub>i</sub> sono complesse da calcolare, e noi siamo interessati alla media di tutti i risultati, e le singole curve sono meno importanti, perciò possiamo approssimarle, a una semplice curva parametrica che ha le stesse proprietà:

$$\frac{tf_i}{K_1 + tf_i} \quad \text{oppure, per avere } \zeta \text{ con } tf_i = 1: \quad \frac{(K_1 + 1) \cdot tf_i}{K_1 + tf_i}$$



Da qui ricaviamo  $C_i^{BM15}$ , ovvero una semplice approssimazione di RSV con 2-Poisson Model. **BEST MATCH 15:**

$$C_i^{BM15}(t_{fi}^f) = C_i^{BM1} \cdot \frac{t_{fi}^f}{K_1 + t_f^f}$$

Usando ora la semplificazione di BM1 con IDF:

$$C_i^{BM15}(t_{fi}^f) \approx \log \frac{N}{n_i} \cdot \frac{t_{fi}^f}{K_1 + t_f^f}$$

Che è simile a TFIDF ma con  $t_{fi}^f$  delimitata

## DOCUMENT LENGTH NORMALIZATION

Fin ora, per usare la distribuzione di Poisson, abbiamo assunto che tutti i documenti avessero lunghezza costante.

Document Length:

$$dl_i = \sum_{i \in V} tf_i$$

Average Document Length:

$$avdl = \frac{1}{N} \sum_{i=1}^N dl_i$$

Rappresentiamo ora  $tf_i$ , normalizzato rispetto a  $avdl$ , utilizzando la seguente:

$$tf_i(d_i) : dl_i = tf'_i(d_i) : avdl$$

$$tf'_i(d_i) = tf_i(d_i) \cdot \frac{avdl}{dl_i}$$

Se mettiamo questa normalizzazione dentro BM15 otteniamo BM11:

$$c_i^{BM11}(tf_i, dl_i) = c_i \frac{BM1 tf'_i(d_i)}{K_1 + tf'_i(d_i)} = c_i \cdot \frac{tf'_i(d_i)}{K_1 \frac{dl_i}{avdl} + tf'_i(d_i)}$$

BM11 applica una normalizzazione totale rispetto alle doc len, ma questa cosa è ottimale per i documenti che hanno  $dl_i$  molto lontano da  $avdl$ . Quindi introduciamo una **normalizzazione parziale**:

$$B_f = \left( (1-b) + b \frac{de}{\text{avde}} \right)$$

Dove quando  $b=1$  otteniamo una full-normalization, mentre per  $b=0$  nessuna normalizzazione.



## BEST MATCH 25 (BM25)

Normalizziamo  $tf_i$  con  $B_f$ :

$$tf'_i(d_f) = \frac{tf_i(d_f)}{B_f}$$

Lo mettiamo in BM15 per ottenere BM25

$$C_i^{BM25}(tf_i, d_f) = C_i^{BM15} \frac{tf'_i(d_f)}{K_1 + tf'_i(d_f)} = C_i^{BM15} \frac{tf_i(d_f)}{K_1 B_f + tf_i(d_f)}$$

$$RSV^{BM25}(q, d) = \sum_{t \in q} \frac{tf_i(d)}{K_1 \left( (b-1) + b \frac{de}{\text{avde}} \right) + tf_i(d)} \cdot \log \frac{N}{n_i}$$

## RANKING WITH FIELDS

Se i documenti hanno campi diversi, e ben distinti (Titolo, body, URL, Anchor text, ...)?

Bisogna applicare lo score a ogni campo del documento, e poi farne la somma.

Così facendo stiamo trattando i vari campi in modo indipendente, anche se dipendono dello stesso documento.

Il modo per superare questa cosa è trattare l'elitness come condivisa fra i vari campi del documento. Ma la relazione fra elitness e tf è dipendente anche del campo di riferimento.

Vogliamo applicare una sorta di normalizzazione di tf rispetto al campo di appartenenza.

$$\tilde{tf}_i(d_j) = \sum_{f \in F} w^f \frac{tf_i^f(d_j)}{B_j^f}$$

$$B_j^f = \left( (1 - b^f) + b^f \frac{dl_j^f}{avdl^f} \right) \text{ with } 0 \leq b^f \leq 1$$

$$c_i^{\text{BM25F}}(tf_i, d_j) = \frac{\tilde{tf}_i(d_j)}{k_1 + \tilde{tf}_i(d_j)} \log \frac{N}{n_i}$$

$$RSV^{\text{BM25F}}(q, d) = \sum_{t_i \in q} \frac{\tilde{tf}_i(d)}{k_1 + \tilde{tf}_i(d)} \log \frac{N}{n_i}$$