NWI-I00041 Information Retrieval
Retrieval Models

Dr. ir. Faegheh Hasibi
f.hasibi@cs.ru.nl

Nijmegen, September 14th, 2020

Classic Retrieval Models

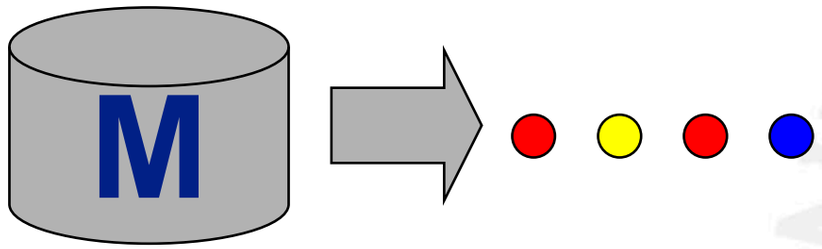
- Vector Space Model (VSM)
  - TF.IDF vector representation

$$TF(w,d) = \text{count}(w,d) \qquad \text{IDF}(w) = \log\left(\frac{M+1}{\text{df}(w)}\right)$$

- Probabilistic Models
  - BM25: Based on estimating *P(R = 1 | d , q)*
  - Language modeling - query likelihood

Language Modeling

- Assigns probability to a sequence of words drawn from some vocabulary

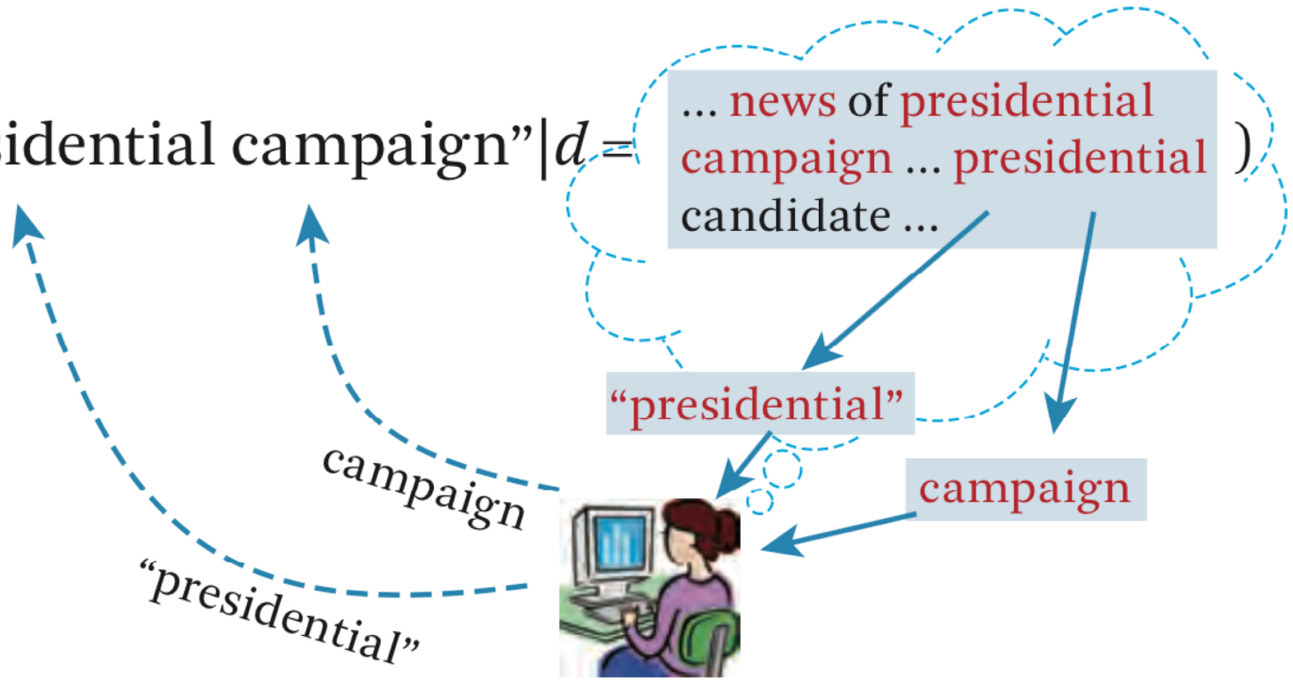


$$P(\bullet\bullet\bullet\bullet \mid M) = P(\bullet \mid M)\, P(\bullet \mid M, \bullet)\, P(\bullet \mid M, \bullet\bullet)\, P(\bullet \mid M, \bullet\bullet\bullet)$$

Query Likelihood Model

- Defines a multinomial probability distribution over all words in the vocabulary of the corpus

- Approximates probability of relevance by P(q | d, R=1)



If the user is **thinking of this doc**,
how likely would she **pose this query**?

Query Likelihood Model

$$P(q|d) = \prod_{i=1}^{n} P(q_i|M_d)$$

i*th* term of query

Language model of document

… but, we do not know the document model

$P(q_i|M_D)$ can be estimated using Maximum Likelihood Estimation (MLE): $\frac{c(q_i, d)}{|d|}$

Query Likelihood Model - Example

q: “Apple Phone”

d1: Apple Samsung
d2: Apple Apple Apple Samsung
d3: Phone Samsung Phone Apple Phone Apple Samsung

Query Likelihood Model - Example

q: “Apple Phone”

d1: Apple Samsung
d2: Apple Apple Apple Samsung
d3: Phone Samsung Phone Apple Phone Apple Samsung

score(q, d1) = 1/2 * 0/2 = 0
score(q, d2) = 3/4 * 0/4 = 0
score(q, d3) = 2/7 * 3/7 = 0.12

Zero probability problem

Smoothing

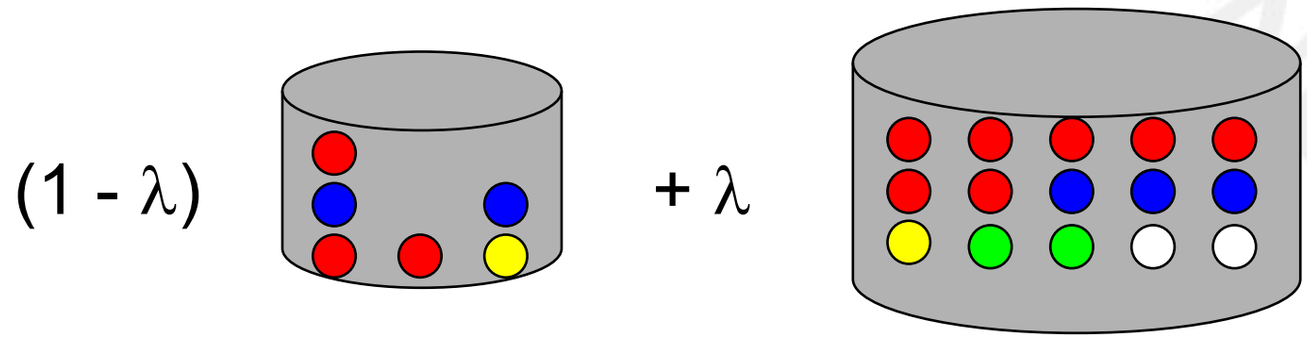
Idea: Assign non-zero probabilities to words that are not observed in the data

Interpolation Method:

- Smoothing the probability coming from the document with the probabilities coming from whole collection

- The interpolation is typically done using a linear fashion

- Behaves like inverse document frequency



Jelinek-Mercer (JM) Smoothing

$$P(q|d) \cong \prod_{i=1}^{n} (1-\lambda) P(q_i|d) + \lambda\, P(q_i|C)$$

$$P(q|d) \cong \prod_{i=1}^{n} (1-\lambda) \frac{c(q_i,d)}{|d|} + \lambda \frac{c(q_i,C)}{|C|}$$

$\lambda \in [0,1]$ *e.g., 0.1*

|*d*|: document length
|*C*|: collection length
c(.): count function

Dirichlet Smoothing

$$P(q|d) \cong \prod_{i=1}^{n} \frac{|d|}{|d|+\mu} \cdot \frac{c(q_i,d)}{|d|} + \frac{\mu}{|d|+\mu} \cdot \frac{c(q_i,C)}{|C|}$$

$$P(q|d) \cong \prod_{i=1}^{n} \frac{c(q_i,d) + \mu\, P(q_i|C)}{|d|+\mu}$$

$$\mu \in [0,\infty)$$

If |*d*| is small, then extra weight is given to the background model, effectively rewarding a short document.

Language Model- Implementation

Multiplying small probabilities may cause underflow:

$$P(q|d) = \prod_{i=1}^{n}(1-\lambda)\frac{c(q_i,d)}{|d|} + \lambda\frac{c(q_i,C)}{|C|})$$

To avoid precision loss, we take the logarithm of the query likelihood:

$$P(q|d) = \sum_{i=1}^{n}\log((1-\lambda)\frac{c(q_i,d)}{|d|} + \lambda\frac{c(q_i,C)}{|C|})$$

Fielded Retrieval Models

- Fielded Language model: **MLM**
- Fielded BM25: **BM25F**

Example Document

Radboud University

Deutsch Nederlands

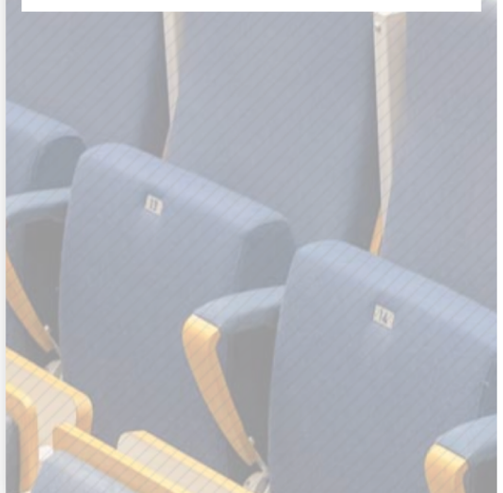
Search

Prospective students • Exchange students • Current Students • Staff • Alumni

HOME EDUCATION RESEARCH NEWS & AGENDA WORKING AT ABOUT US

Radboud University › News & agenda › News › The coronavirus and Radboud University › Corona news › A message from the GGD to students

- All news
- News on research
- News on education
- Columns executive board
- The coronavirus and Radboud University
- Corona news
  - A message from the GGD to students
- Students
- Employees
- Prospective students
- Corona guidelines



A message from the GGD to students: You don't want to receive a call that you have to go into quarantine, do you?

Date of news: 8 September 2020

Following the closure of the Van Rijn student pub in Nijmegen on 1 September, students have now become more conscious of complaints. Over the past week, 820 youngsters in the 20-25 age group have visited the GGD's testing location. This is twice the number of people compared to the number of people in the preceding two weeks.

More than thirty students from Radboud University and HAN University of Applied Sciences are currently infected with the coronavirus. It is partly because of this that around 170 of these students' roommates and close contacts are now in quarantine. The prospect of a second coronavirus outbreak is becoming all the more realistic if we don't continue to comply with the 1.5 metre social distancing rule.

We are therefore asking students to hang in there! Please understand that if you're too close to someone, that that person could test positive. And then, in a few days' time, you yourself could get a phone call from the GGD telling you that you have to go into quarantine for ten days. **So, please act responsibly; this doesn't mean that you have to stop having fun, but you do need to ensure that you continue to do it at a distance of 1.5 metres from each other!**

Many thanks for your cooperation,

The GGD Gelderland-Zuid Coronavirus Team.

Example Document - Structure

<title>
A message from the GGD to students - Radboud University
</title>
<grid-title>
A message from the GGD to students: You don’t want to receive a call that you have to go into quarantine, do you?
</grid-title >
<content>
Following the closure of the Van Rijn student pub in Nijmegen on 1 September, students have now become more conscious of complaints. Over the past week, 820 youngsters in the 20-25 age group have visited the GGD's testing location. This is twice the number of people compared to the number of people in the preceding two weeks.
…
<content>

Mixture of Language Models (MLM)

- Uses mixture model to combine the probability distributions
- Mixture models guides how to combine several language models estimated from different document representations
- Language models of different document representations are linearly interpolated

Mixture of Language Models (MLM)

$$P(q|d) \cong \prod_{i=1}^{n} P(q_i|M_d)$$

$$P(q_i|M_d) = \sum_{f \in F} \mu_f P(q_i|M_{d_f})$$

field weight $\sum_{f \in F} \mu_f = 1$

$$P(q_i|M_{d_f}) = (1-\lambda_f)\frac{c(q_i, d_f)}{|d_f|} + \lambda_f \frac{c(q_i, C_f)}{|C_f|}$$

f: document field
F: all document fields
d_f: field f of document d

BM25F

- Combines the term frequencies of the different fields using a weighted linear combination
  - Weights corresponds to fields weights

- No use of field collection statistics
  - For short fields (such as title) theses statistic may be unstable

- Avoids too many parameter tuning

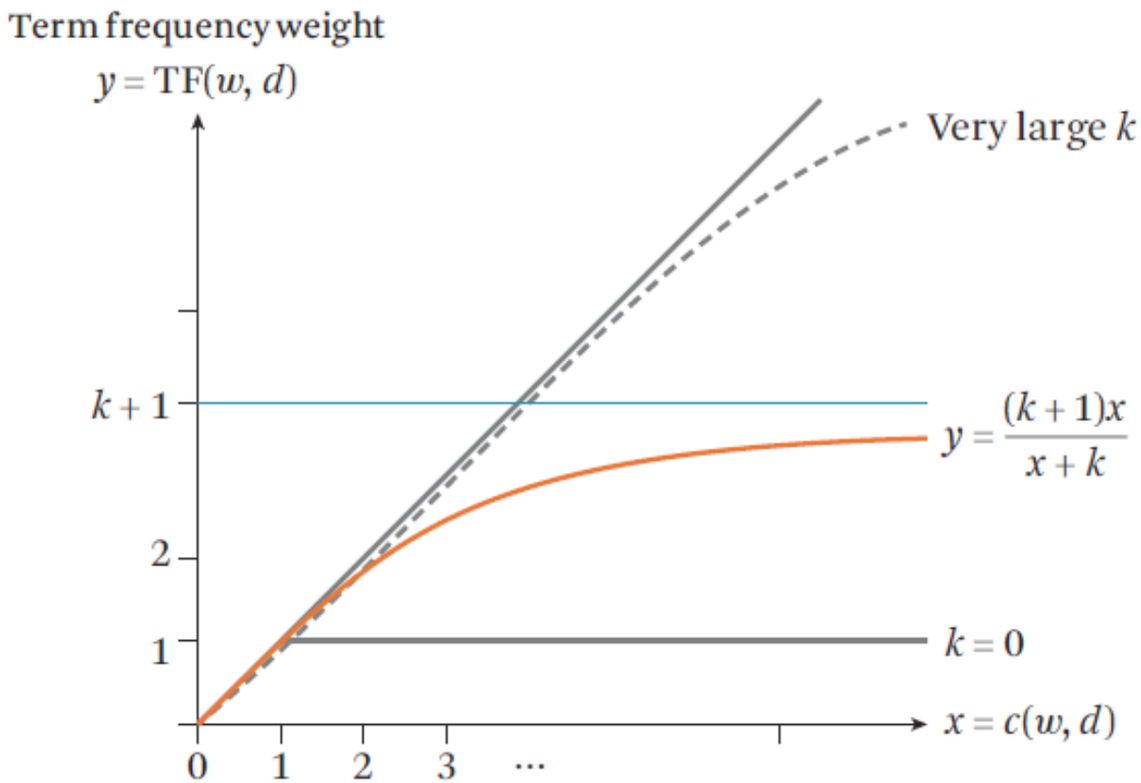
BM25F

BM25

$$score(q,d) = \sum_{i=1}^{n} \frac{(k+1)c(q_i,d)}{k\left(1-b+b\,\frac{|d|}{avdl}\right)+\ c(q_i,d)}\ idf(q_i)$$

BM25F

$$score(q,d) = \sum_{i=1}^{n} \frac{(k+1)c'(q_i,d)}{k+\ c'(q_i,d)}\ idf(q_i)$$



BM25F

BM25

$$score(q,d) = \sum_{i=1}^{n} \frac{(k+1)c(q_i,d)}{k\left(1-b+b\,\frac{|d|}{avdl}\right)+\ c(q_i,d)}\ idf(q_i)$$

BM25F

$$score(q,d) = \sum_{i=1}^{n} \frac{(k+1)c'(q_i,d)}{k+\ c'(q_i,d)}\ idf(q_i)$$

$$\sum_{f\in F} \alpha_f \frac{c(q_i,d_f)}{1-b_f+b_f\,\frac{|d_f|}{avfl})}$$

mean length of field *f*
across the entity catalog

MLM - Example

q: “Apple Phone”

	Title	Content
d1	Apple	Apple phone Apple Samsung phone
d2	Phone Samsung	Phone Samsung Phone Apple Phone Apple Samsung

MLM - Example

	Title	Content
d1	Apple	Apple phone Apple Samsung phone
d2	Phone Samsung	Phone Samsung Phone Apple Phone Apple Samsung

	d1		d2	
	Title-tf	Content-tf	Title-tf	Content-tf
Apple	1	2	0	2
Samsung	0	1	1	2
Phone	0	2	1	3
|d|	1	5	2	7

	Collection	
	Title	Content
Apple	1	4
Samsung	1	3
Phone	1	5
|C|	3	12

MLM - Example

q: "Apple Phone"

$\mu_{title} = 0.2$ $\mu_{content} = 0.8$ $\lambda = 0.1$

	d1		d2	
	Title-tf	Content-tf	Title-tf	Content-tf
Apple	1	2	0	2
Samsung	0	1	1	2
Phone	0	2	1	3
|d|	1	5	2	7

	Collection	
	Title	Content
Apple	1	4
Samsung	1	3
Phone	1	5
|C|	3	12

$$P(Apple|d1) \cong 0.2 \left[0.9 \frac{1}{1} + 0.1 \frac{1}{3}\right] + 0.8 \left[0.9 \frac{2}{5} + 0.1 \frac{4}{12}\right] = 0.5$$

$$P(Phone|d1) \cong 0.2 \left[0.9 \frac{0}{1} + 0.1 \frac{1}{3}\right] + 0.8 \left[0.9 \frac{2}{5} + 0.1 \frac{5}{12}\right] = 0.33$$

$$P(q|d1) \cong 0.5 * 0.33 = 0.165$$

MLM - Example

q: "Apple Phone"

$\mu_{title} = 0.2$ $\mu_{content} = 0.8$ $\lambda = 0.1$

	d1		d2	
	Title-tf	Content-tf	Title-tf	Content-tf
Apple	1	2	0	2
Samsung	0	1	1	2
Phone	0	2	1	3
|d|	1	5	2	7

	Collection	
	Title	Content
Apple	1	4
Samsung	1	3
Phone	1	5
|C|	3	12

$$P(Apple|d2) \cong 0.2 \left[0.9 \frac{0}{2} + 0.1 \frac{1}{3}\right] + 0.8 \left[0.9 \frac{2}{7} + 0.1 \frac{4}{12}\right] = 0.24$$

$$P(Phone|d2) \cong 0.2 \left[0.9 \frac{1}{2} + 0.1 \frac{1}{3}\right] + 0.8 \left[0.9 \frac{3}{7} + 0.1 \frac{5}{12}\right] = 0.44$$

$$P(q|d2) \cong 0.24 * 0.44 = 0.11$$