Integrating News Article Metadata into Topic Models

June 15, 2021

Rasmus Engesgaard Christensen, Peter Langballe Erichsen, and Dennis Højbjerg Rose

Department of Computer Science
Aalborg University
Denmark



Agenda



Query Generation

Introduction

Information Retrieval Methods

Experiment

Results

Conclusion

TF-IDF BM25

....,

Introduction

Information Retrieval Methods

_xpciiiici

Results

Language Model Information retrieval goal



 We want to find relevant documents based on words in a query

$$P(q|d) = \prod_{w \in S} P(w|d)$$

Language Model

BM25

Combination of method

Query Generation

Introduction

Information Retrieva Methods

Results



$$P(w|d) = \frac{N_d}{N_d + \lambda} \cdot \frac{tf(w, d)}{N_d} + \left(1 - \frac{N_d}{N_d + \lambda}\right) \cdot \frac{tf(w, D)}{N_D}$$

- \triangleright d = document
- $ightharpoonup N_d = \text{Number of words in } d$
- \blacktriangleright λ is the average document length
- \triangleright D = Corpus

2 Language Model

TF-IDF BM25

Query Generation

ntroduction

Information Retrieva Methods

_ протито

icouito



$$P(w|d) = \underbrace{\frac{N_d}{N_d + \lambda} \cdot \frac{tf(w, d)}{N_d}}_{\text{weight term}} + \underbrace{\left(1 - \frac{N_d}{N_d + \lambda}\right) \cdot \frac{tf(w, D)}{N_D}}_{\text{inverse weight term}}$$

- \rightarrow d = document
- $ightharpoonup N_d = \text{Number of words in } d$
- $ightharpoonup \lambda$ is the average document length
- \triangleright D = Corpus

2 Language Model

BM25

Query Generation

Introduction

Information Retrieval Methods

широппп

Results



$$P(w|d) = \underbrace{\frac{N_d}{N_d + \lambda}}_{\text{weight term}} \cdot \underbrace{\frac{tf(w,d)}{N_d}}_{\text{% of w in d}} + \underbrace{\left(1 - \frac{N_d}{N_d + \lambda}\right)}_{\text{inverse weight term}} \cdot \underbrace{\frac{tf(w,D)}{N_D}}_{\text{% w in D}}$$

2 Language Model

BM25

Query Generation

ntroduction

Information Retrieval Methods

_xperime

Results

- ightharpoonup d = document
- $ightharpoonup N_d = \text{Number of words in } d$
- $ightharpoonup \lambda$ is the average document length
- ightharpoonup D = Corpus



$$P(w|d) = \underbrace{\frac{N_d}{N_d + \lambda}}_{\text{weight term}} \cdot \underbrace{\frac{tf(w,d)}{N_d}}_{\text{% of w in d}} + \underbrace{\left(1 - \frac{N_d}{N_d + \lambda}\right)}_{\text{inverse weight term}} \cdot \underbrace{\frac{tf(w,D)}{N_D}}_{\text{% w in D}}$$

2 Language Model

BM25

Query Generation

ntroduction

Information Retrieval Methods

_xperime

Results

- ightharpoonup d = document
- $ightharpoonup N_d = \text{Number of words in } d$
- $ightharpoonup \lambda$ is the average document length
- ightharpoonup D = Corpus





) Language Model

BM25

Combination of methods

Query Generation

Introduction

Information Retrieval Methods

Results

Conclusion

► Favors high percentage of a word in a document or corpus

TF-IDF Explanation



$$tf-idf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

- ightharpoonup t = Term
- ightharpoonup d = Document
- ightharpoonup D = Corpus

Language Model
TF-IDF

BM25

Combination of method

Query Generation

Introduction

Information Retriev Methods

Results

TF-IDF Explanation



 $\mathsf{tf\text{-}idf}(t,d,D) = \mathsf{tf}(t,d) \cdot log \frac{|\{d \in D\}|}{|\{d \in D : t \in d\}|}$

Language Model

) TF-IDF BM25

Combination of methods

Query General

Introduction

Information Retrieva Methods

Exponiii

Results

TF-IDF Summary



► Favors high usage of unique word(s) in a document or corpus

Language Model

) TF-I

Combination of method

Query Generatio

Introduction

Information Retrieval Methods

Results



bm25 $(d, q) = \sum_{i=1}^{n} idf(q_i) \cdot \frac{tf(q_i, d) \cdot (k_1 + 1)}{tf(q_i, d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})}$

- ▶ b adjust the sensitivity of varying document lengths
- $ightharpoonup k_1$ adjust how quickly a term is saturated

Language Model

) BM25

Combination of methods

Query Generation

Introduction

Information Retrieva Methods

Ехротипо

....

BM25 Explanation



Language Model

) BM25

Combination of methods

Query Generation

Introduction

Information Retrieva Methods

Experiment

Results

$$bm25(d,q) = \sum_{i=1}^{n} idf(q_i) \cdot \frac{tf(q_i,d) \cdot (1.5+1)}{tf(q_i,d) + 1.5 \cdot (1 - 0.75 + 0.75 \cdot \frac{|d|}{avgdl})}$$

BM25 Summary



Language Model

) BM25

Combination of methods

Query Generation

Introduction

Information Retrieval Methods

Expoiiii

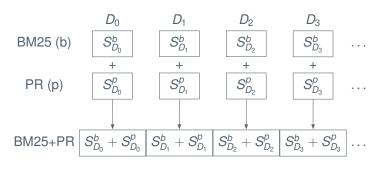
Results

- ► Similar to tf-idf but with some other focus points
 - ► Document length
 - Word saturation

Combination of methods

How to combine?





TF-IDF BM25

Combination of methods

Query Generation

ntroduction

Information Retrieva Methods

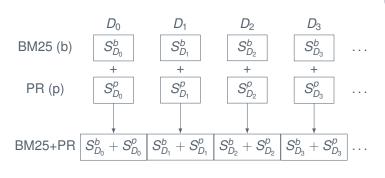
Lybellille

Results

Combination of methods

How to combine?





TF-IDF BM25

Combination of methods

Query Generation

Information Retrieval Methods

Experime

Results

Query Generation Types of queries



- Document query
 - Specificity Finding a specific document
- ► Topic query
 - Generality Finding topic relevant documents

8 Query Generation

Document Query Topic Query

Introducti

Information Retrieva Methods

Results

Document Query Generation



Corpus

Query Generation

Document Query

Introducti

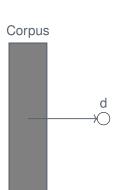
Information Retrieval Methods

LAPOIII

Results

Document Query Generation





Query Generation

Document Query Topic Query

Introducti

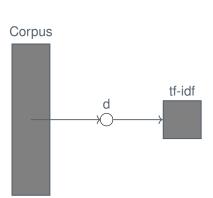
Information Retrieval Methods

Lxpeiiii

Results

Document Query Generation





Query Generation

9 Document Query

Topic Query

Introducti

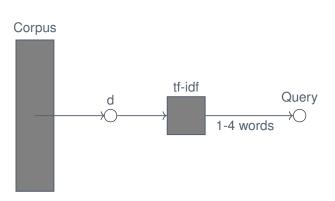
Information Retrieval Methods

LAPOIIII

Results

Document Query Generation





Query Generation

9 Document Query

Topic Query

IIIIIOGGCIIOII

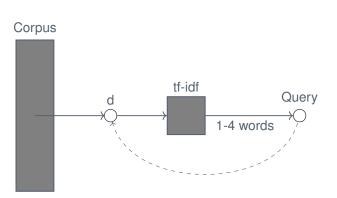
Information Retrieva Methods

LAPOIIII

Results

Document Query Generation





Document Query

Methods

Results



Topics

Query Generation



Introducti

Information Retrieval Methods

Experin

Results



Topics

t

Query Generation

Document Que 10 Topic Query

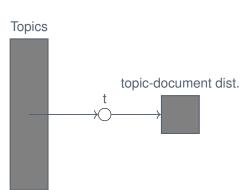
Introduction

Information Retrieval Methods

Lxpeiii

Results





Query Generation

Document Que 10 Topic Query

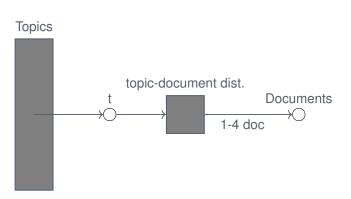
Introducti

Information Retrieval Methods

LAPCIIII

Results





Query Generation

Document Query

Topic Query

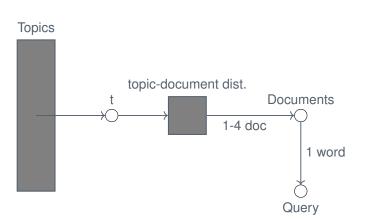
Introduction

Information Retrieval Methods

Lybeilli

Results





Query Generation

Topic Query

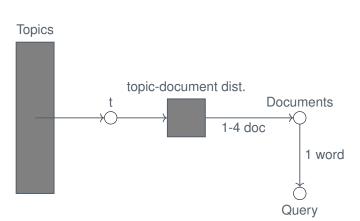
Introduction

Information Retrieva Methods

Lybeiiiii

Results





► Sample the topic-word distribution instead

Query Generatio

Topic Query

Introducti

Information Retrieva Methods

Lybeiiiii

Results

Introduction Motivation



Query Generation

1) Introduction

Information Retrieval Methods

Experiment Results

riesuits

- Query based search of documents
 - ► Google Scholar
- ► Encourage abstractions of underlying topics
 - ► Rather than word frequency

Information Retrieval Methods



Query Generation

2 Information Retrieval

Latent Dirichlet Allocation

Experiment

Methods

Results

- ► Latent Dirichlet Allocation (LDA)
- ► PageRank (PR)
- ► Language Model (LM)
- ► Term Frequency Inverse Document Frequency (TF-IDF)
- ► Best Match 25 (BM25)



Introduction

Information Re

(13) Latent Dirichlet Allocation

Lxpeiiiii

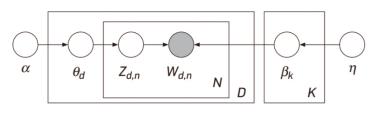
....

► Create a generative process to produce documents, based on topics

- ► Fine-tune to maximize probability of generating the original documents
- ► Use generated topics for calculating similarity

Plate Notation





Query Generation

Introduct

Information Retriev

14 Latent Dirichlet Allocation

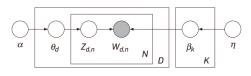
Experime Results

0----

- $ightharpoonup \alpha, \eta$ dirichlet distributions
- \blacktriangleright θ , β multinomial distributions
- ► *Z*, *W* sampled topics and words

Dirichlet Distributions





Alpha of 0.1



Alpha of 1



Alpha of 4



1

Query Generation

Introduction

Information Retrieva Methods

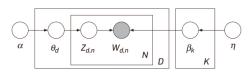
Latent Dirichlet Allocation

Results

¹https://mollermara.com/blog/lda/

Dirichlet Distributions











- ► typical sample based on low alpha = {1,0,0}
- ► typical sample based on high alpha = {0.333, 0.333, 0.333}

uery Generation

Introduction

Information Retrie Methods

Latent Dirichlet Allocation

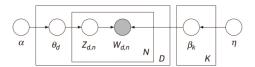
Experiment

Results

https://mollermara.com/blog/lda/

Multinomial Distributions





▶ Sample *N* topics (Z) based on θ

▶ Sample *N* words (*W*) based on *Z* and β

Query Generation

Introd

Information Retrie



Lxpciiiii

Latent Dirichlet Allocation Generation Probability



Query Generation

Introdu

Information Retrie

17 Latent Dirichlet Allocation

Lxpeiiiii

....

$$P(W, Z, \theta, \beta; \alpha, \eta) = \prod_{d=1}^{D} P(\theta_d; \alpha) \prod_{k=1}^{K} P(\beta_k; \eta) \prod_{n=1}^{N} P(Z_{d,n} | \theta_d) P(W_{d,n} | \beta, Z_{d,n})$$

PageRank Overview



Query Generation

Introduction

Methods

18 PageRank

Expe

Results

Conclusion

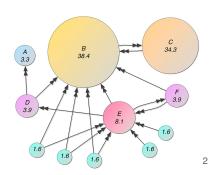
► Used to rank nodes in a graph

- ► Underlying assumption: important nodes have in-going connections from other important nodes
- ▶ Based on the 'random surfer' model

PageRank Overview



- ► Used to rank nodes in a graph
- Underlying assumption: important nodes have in-going connections from other important nodes
- ► Based on the 'random surfer' model



Query Generation

Introduction

Information Retrieva Methods

18 PageRank

Evporimo

Results

²https://en.wikipedia.org/wiki/PageRank

PageRank Graph Construction



. . . .

Introductio

Methods

19 PageRank

- 7

► Used on adjacency matrix

- ▶ Similarity between documents based on θ
 - Calculated using Jensen-Shannon similarity
- While fully connected each edge has a value which will influence the ranking

Experime

Results



Grid-search

Parameter	Tested Values
K ₁	10, 50, 100, 200, 300
K_2	5, 10, 15, 20, 25, 30, 35, 40, 45, 50
α	0.5, 0.1, 0.01, 0.001
η	0.1, 0.01, 0.001, 0.0001

Query Generation

Introduction

Information Retrieval Methods

Experiment 20 Hyperparameters

Results



Grid-search

Parameter	Tested Values
K ₁	10 , 50 , 100, 200, 300
K_2	5, 10, 15, 20, 25, 30, 35, 40, 45, 50
α	0.5, 0.1, 0.01, 0.001
η	0.1, 0.01, 0.001, 0.0001

Query Generation

Introduction

Information Retrieval Methods

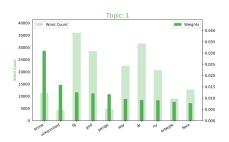
Experiment 20 Hyperparameters

Results



Grid-search

Parameter	Tested Values
K ₁	10, 50, 100, 200, 300
K_2	5, 10, 15, 20, 25, 30, 35, 40, 45, 50
α	0.5, 0.1, 0.01, 0.001
η	0.1, 0.01, 0.001, 0.0001



Query Generation

Introduction

Information Retrieval Methods

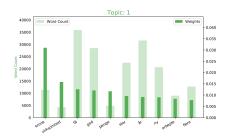
Experiment

) Hyperparameters Results



Grid-search

Parameter	Tested Values
	10, 50, 100, 200, 300
K_2	5, 10, 15, 20, 25, 30 , 35, 40, 45, 50
α	0.5, 0.1 , 0.01, 0.001
η	0.1 , 0.01, 0.001, 0.0001



► K = 30, $\alpha = 0.1$, and $\eta = 0.1$

Query Generation

Introduction

Information Retrieval Methods

Experiment

Hyperparameters

Results



Mean Average Precision Results

	Do	cument Qu	th	Т	opic Que	ery Lengt	th	
IR methods / MAP	1	2	3	4	1	2	3	4
LDA-IR	0.00457	0.00527	0.0429	0.0538	0.155	0.186	0.168	0.178
LM	0.198	0.152	0.291	0.260	0.126	0.130	0.128	0.129
BM25	0.270	0.656	0.866	0.908	0.155	0.158	0.155	0.161
tf-idf	0.210	0.621	0.799	0.897	0.155	0.157	0.155	0.161
LDA-IR + PR	0.00458	0.00526	0.0429	0.0538	0.162	0.195	0.177	0.187
LDA-IR * PR	0.00781	0.00569	0.0410	0.0537	0.156	0.186	0.168	0.179
LM + LDA-IR	0.0419	0.0214	0.0602	0.120	0.147	0.163	0.145	0.146
LM * LDA-IR	0.0931	0.0462	0.175	0.177	0.150	0.175	0.155	0.166
LM + PR	0.170	0.153	0.283	0.256	0.130	0.132	0.130	0.131
LM * PR	0.163	0.138	0.259	0.236	0.130	0.133	0.129	0.130
LM + LDA-IR + PR	0.0499	0.0214	0.0601	0.120	0.148	0.164	0.146	0.147
LM * LDA-IR * PR	0.0911	0.0459	0.157	0.170	0.150	0.175	0.155	0.166
BM25 + LDA-IR	0.276	0.524	0.588	0.365	0.155	0.184	0.168	0.176
BM25 * LDA-IR	0.139	0.270	0.412	0.365	0.154	0.159	0.156	0.162
BM25 + PR	0.269	0.656	0.866	0.902	0.192	0.193	0.175	0.183
BM25 * PR	0.267	0.663	0.884	0.904	0.155	0.159	0.155	0.161
BM25 + LDA-IR + PR	0.276	0.525	0.589	0.366	0.162	0.192	0.176	0.184
BM25 * LDA-IR * PR	0.150	0.266	0.446	0.381	0.155	0.159	0.156	0.163

Query Generation

Introduct

Information Retrieval Methods

Experiment



Results



Precision Results

	Topic Query Length (P@10)				Topic Query Length (P@100)			
IR methods	1	2	3	4	1	2	3	4
LDA-IR	0.02	0.103	0.13	0.203	0.062	0.131	0.164	0.191
LM	0.126	0.118	0.116	0.069	0.090	0.092	0.087	0.093
BM25	0.136	0.161	0.164	0.174	0.142	0.165	0.175	0.151
tf-idf	0.160	0.125	0.200	0.148	0.163	0.169	0.188	0.170
LDA-IR + PR	0.0188	0.103	0.141	0.211	0.062	0.131	0.175	0.198
LDA-IR * PR	0.0125	0.100	0.133	0.200	0.062	0.132	0.167	0.192
LM + LDA-IR	0.02	0.101	0.136	0.196	0.060	0.129	0.161	0.188
LM * LDA-IR	0.02	0.085	0.109	0.161	0.055	0.114	0.129	0.152
LM + PR	0.138	0.130	0.116	0.0763	0.110	0.108	0.097	0.098
LM * PR	0.148	0.128	0.116	0.0963	0.110	0.112	0.101	0.101
LM + LDA-IR + PR	0.019	0.101	0.134	0.195	0.061	0.129	0.163	0.187
LM * LDA-IR * PR	0.026	0.090	0.110	0.161	0.059	0.115	0.130	0.152
BM25 + LDA-IR	0.124	0.160	0.154	0.204	0.113	0.155	0.168	0.198
BM25 * LDA-IR	0.09	0.139	0.175	0.206	0.134	0.170	0.174	0.187
BM25 + PR	0.135	0.163	0.165	0.173	0.155	0.165	0.176	0.151
BM25 * PR	0.165	0.169	0.184	0.170	0.148	0.177	0.186	0.161
BM25 + LDA-IR + PR	0.124	0.160	0.154	0.204	0.113	0.155	0.17	0.198
BM25 * LDA-IR * PR	0.095	0.141	0.176	0.206	0.135	0.174	0.174	0.188

Query Generation

Introduction

Information Retrieval Methods

Experiment

(22) Results

Results



Average Rank Results

	Document Query Length						
IR methods / Avg. Rank	1	2	3	4			
LDA-IR	2287.93	1599.95	1241.18	1926.78			
LM	7120.04	9082.9	6501.85	7782.65			
BM25	19.58	7.94	1.78	1.41			
tf-idf	30.0	9.3	2.03	1.29			
LDA-IR + PR	2491.31	1342.53	1126.23	1906.76			
LDA-IR * PR	2305.04	1600.93	1223.14	1920.175			
LM + LDA-IR	1971.19	1192.91	1027.95	1482.69			
LM * LDA-IR	1874.81	1456.21	954.66	1853.44			
LM + PR	7299.85	9134.81	6429.24	7725.36			
LM * PR	7328.7625	9137.23	6504.85	7772.4			
LM + LDA-IR + PR	1978.74	1179.21	994.91	1438.88			
LM * LDA-IR * PR	1892.12	1453.56	941.4	1850.43			
BM25 + LDA-IR	30.45	28.59	17.7	23.76			
BM25 * LDA-IR	163.76	557.13	297.48	1159.33			
BM25 + PR	19.69	7.88	1.79	1.43			
BM25 * PR	23.96	8.45	1.61	1.24			
BM25 + LDA-IR + PR	30.35	28.44	17.65	23.76			
BM25 * LDA-IR * PR	163.5	555.35	295.69	1158.08			

Query Generation

Introducti

Information Retrieval Methods

Experiment

23 Results



MAP Random Document Results

ID mode of AAAD	Do	cument Qu	, , ,		Topic Query Length			
IR methods / MAP	l l	2	3	4		2	3	4
LDA-IR	0.00457	0.00527	0.0429	0.0538	0.155	0.186	0.168	0.178
LM	0.198	0.152	0.291	0.260	0.126	0.130	0.128	0.129
BM25	0.270	0.656	0.866	0.908	0.155	0.158	0.155	0.161
tf-idf	0.210	0.621	0.799	0.897	0.155	0.157	0.155	0.161
LDA-IR + PR	0.00458	0.00526	0.0429	0.0538	0.162	0.195	0.177	0.187
LDA-IR * PR	0.00781	0.00569	0.0410	0.0537	0.156	0.186	0.168	0.179
LM + LDA-IR	0.0419	0.0214	0.0602	0.120	0.147	0.163	0.145	0.146
LM * LDA-IR	0.0931	0.0462	0.175	0.177	0.150	0.175	0.155	0.166
LM + PR	0.170	0.153	0.283	0.256	0.130	0.132	0.130	0.131
LM * PR	0.163	0.138	0.259	0.236	0.130	0.133	0.129	0.130
LM + LDA-IR + PR	0.0499	0.0214	0.0601	0.120	0.148	0.164	0.146	0.147
LM * LDA-IR * PR	0.0911	0.0459	0.157	0.170	0.150	0.175	0.155	0.166
BM25 + LDA-IR	0.276	0.524	0.588	0.365	0.155	0.184	0.168	0.176
BM25 * LDA-IR	0.139	0.270	0.412	0.365	0.154	0.159	0.156	0.162
BM25 + PR	0.269	0.656	0.866	0.902	0.192	0.193	0.175	0.183
BM25 * PR	0.267	0.663	0.884	0.904	0.155	0.159	0.155	0.161
BM25 + LDA-IR + PR	0.276	0.525	0.589	0.366	0.162	0.192	0.176	0.184
BM25 * LDA-IR * PR	0.150	0.266	0.446	0.381	0.155	0.159	0.156	0.163
Random 0.000357						0.1	44	

Query Generation

Introdu

Information Retrieval Methods

Experiment

Results



Precision Random Document Results

	Topic	Topic Query Length (P@10)			Topic	Query Le	ength (Po	@100)
IR methods	1	2	3	4	1	2	3	4
LDA-IR	0.02	0.103	0.13	0.203	0.062	0.131	0.164	0.191
LM	0.126	0.118	0.116	0.069	0.090	0.092	0.087	0.093
BM25	0.136	0.161	0.164	0.174	0.142	0.165	0.175	0.151
tf-idf	0.160	0.125	0.200	0.148	0.163	0.169	0.188	0.170
LDA-IR + PR	0.0188	0.103	0.141	0.211	0.062	0.131	0.175	0.198
LDA-IR * PR	0.0125	0.100	0.133	0.200	0.062	0.132	0.167	0.192
LM + LDA-IR	0.02	0.101	0.136	0.196	0.060	0.129	0.161	0.188
LM * LDA-IR	0.02	0.085	0.109	0.161	0.055	0.114	0.129	0.152
LM + PR	0.138	0.130	0.116	0.0763	0.110	0.108	0.097	0.098
LM * PR	0.148	0.128	0.116	0.0963	0.110	0.112	0.101	0.101
LM + LDA-IR + PR	0.019	0.101	0.134	0.195	0.061	0.129	0.163	0.187
LM * LDA-IR * PR	0.026	0.090	0.110	0.161	0.059	0.115	0.130	0.152
BM25 + LDA-IR	0.124	0.160	0.154	0.204	0.113	0.155	0.168	0.198
BM25 * LDA-IR	0.09	0.139	0.175	0.206	0.134	0.170	0.174	0.187
BM25 + PR	0.135	0.163	0.165	0.173	0.155	0.165	0.176	0.151
BM25 * PR	0.165	0.169	0.184	0.170	0.148	0.177	0.186	0.161
BM25 + LDA-IR + PR	0.124	0.160	0.154	0.204	0.113	0.155	0.17	0.198
BM25 * LDA-IR * PR	0.095	0.141	0.176	0.206	0.135	0.174	0.174	0.188
Random		0.1	142			0.1	52	

Query Generation

Introduction

Information Retrieval Methods

Experiment

Results





Average Rank Random Document Results

	Document Query Length							
IR methods / Avg. Rank	1	2	3	4				
LDA-IR	2287.93	1599.95	1241.18	1926.78				
LM	7120.04	9082.9	6501.85	7782.65				
BM25	19.58	7.94	1.78	1.41				
tf-idf	30.0	9.3	2.03	1.29				
LDA-IR + PR	2491.31	1342.53	1126.23	1906.76				
LDA-IR * PR	2305.04	1600.93	1223.14	1920.175				
LM + LDA-IR	1971.19	1192.91	1027.95	1482.69				
LM * LDA-IR	1874.81	1456.21	954.66	1853.44				
LM + PR	7299.85	9134.81	6429.24	7725.36				
LM * PR	7328.7625	9137.23	6504.85	7772.4				
LM + LDA-IR + PR	1978.74	1179.21	994.91	1438.88				
LM * LDA-IR * PR	1892.12	1453.56	941.4	1850.43				
BM25 + LDA-IR	30.45	28.59	17.7	23.76				
BM25 * LDA-IR	163.76	557.13	297.48	1159.33				
BM25 + PR	19.69	7.88	1.79	1.43				
BM25 * PR	23.96	8.45	1.61	1.24				
BM25 + LDA-IR + PR	30.35	28.44	17.65	23.76				
BM25 * LDA-IR * PR	163.5	555.35	295.69	1158.08				
Random	16080							

Query Generation

Introdu

Information Retrieval Methods

Experiment

Results

26 New Experiments



- ► Looking for a specific document (EU og klimaet til debat)
 - Debate about EU and converting to green energy

Information Retrieval Methods

Experiment

Results



- Looking for a specific document (EU og klimaet til debat)
 - Debate about EU and converting to green energy
- ► "EU", "grøn" (green), "omstilling" (conversion)

Methods

Experiment



- Looking for a specific document (EU og klimaet til debat)
 - Debate about EU and converting to green energy
- ► "EU", "grøn" (green), "omstilling" (conversion)
- ► IR method: BM25 + PR

Methods

Experiment



- ► Looking for a specific document (EU og klimaet til debat)
 - ► Debate about EU and converting to green energy
- ► "EU", "grøn" (green), "omstilling" (conversion)
- ► IR method: BM25 + PR
- ► 17717, 2657, 18245, 9213, **30000**, 18809, 13197, 15307, 20145, 19180

Query Generation

Introduction

Information Retrieval Methods

experiment

Result



- ► Looking for a specific document (EU og klimaet til debat)
 - ► Debate about EU and converting to green energy
- ► "EU", "grøn" (green), "omstilling" (conversion)
- ► IR method: BM25 + PR
- ► 17717, 2657, 18245, 9213, **30000**, 18809, 13197, 15307, 20145, 19180
- Article 17717 is also about green growth and protecting the environment

Query Generation

Introduction

Information Retrieval Methods

xperiment

Resu

27 New Experiments



► How can we generate queries for a dataset to use for IR?

► How can we evaluate IR methods in a way that favors abstraction, rather than word frequency?

Can PR be used on a document dataset with no explicit connections to improve IR methods? Query Generation

Introducti

Information Retrieval Methods

Experiment

Results



- ► How can we generate queries for a dataset to use for IR?
 - Generate two types of queries
 - ► Important words from a specific document or topic
- ► How can we evaluate IR methods in a way that favors abstraction, rather than word frequency?

Can PR be used on a document dataset with no explicit connections to improve IR methods? Query Generation

Introduc

Information Retrieval Methods

Experiment

Results



- ► How can we generate queries for a dataset to use for IR?
 - ► Generate two types of queries
 - ► Important words from a specific document or topic
- ► How can we evaluate IR methods in a way that favors abstraction, rather than word frequency?
 - Query types favor both specificity and abstraction
 - ► Evaluated using MAP, P@n, and average rank
- Can PR be used on a document dataset with no explicit connections to improve IR methods?

Query Generation

Introduc

Information Retrieval Methods

Experimen

Result



- ► How can we generate queries for a dataset to use for IR?
 - Generate two types of queries
 - ► Important words from a specific document or topic
- ► How can we evaluate IR methods in a way that favors abstraction, rather than word frequency?
 - Query types favor both specificity and abstraction
 - ► Evaluated using MAP, P@n, and average rank
- Can PR be used on a document dataset with no explicit connections to improve IR methods?
 - Creating the PR adjacency matrix using the similarity between document topic distributions
 - Highly effective

Query Generation

Introduc

Information Retrieval Methods

Experiment

Result

(28) Conclusion

Thank you

