Faster Learned Sparse Retrieval with Guided Traversal

Antonio Mallia, Joel Mackenzie, Torsten Suel, and Nicola Tonellotto

Introduction and Preliminaries

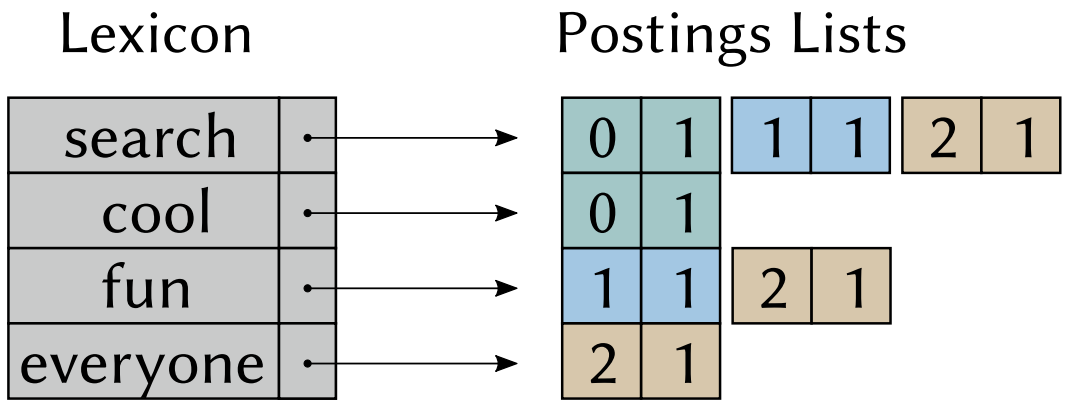
Context: Recently, a number of neural-based techniques have been explored for improving inverted index based retrieval.

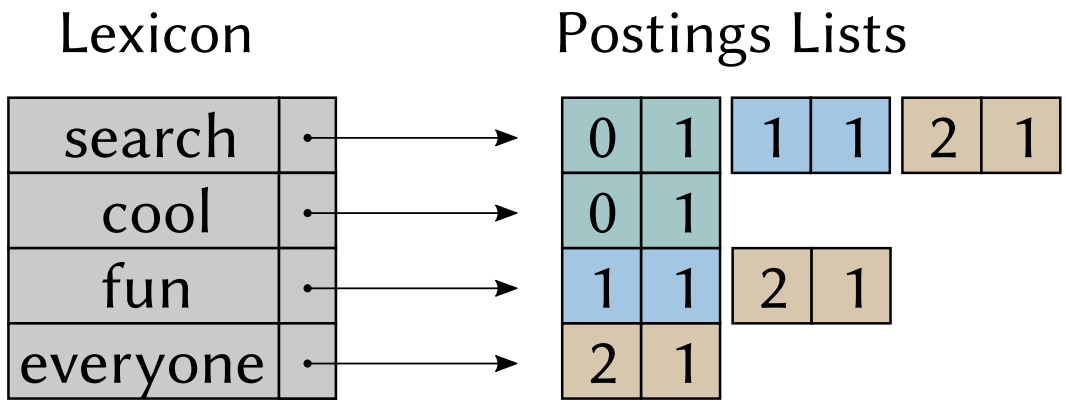
- Term Expansion: Both offline (applied to documents) and online (applied to queries).
- Term (Re)weighting: Both offline (reweighted term frequencies, or learned impacts) and online (query term weighting).

search is cool

search is fun

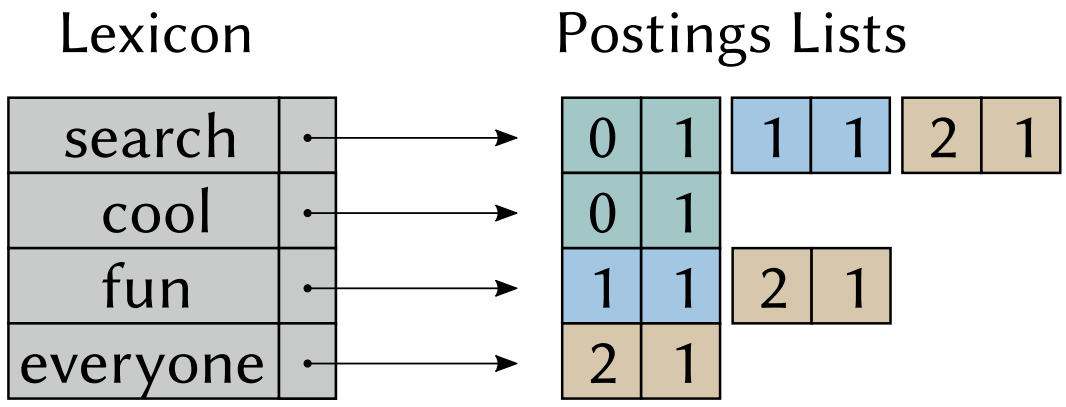
search is fun for everyone

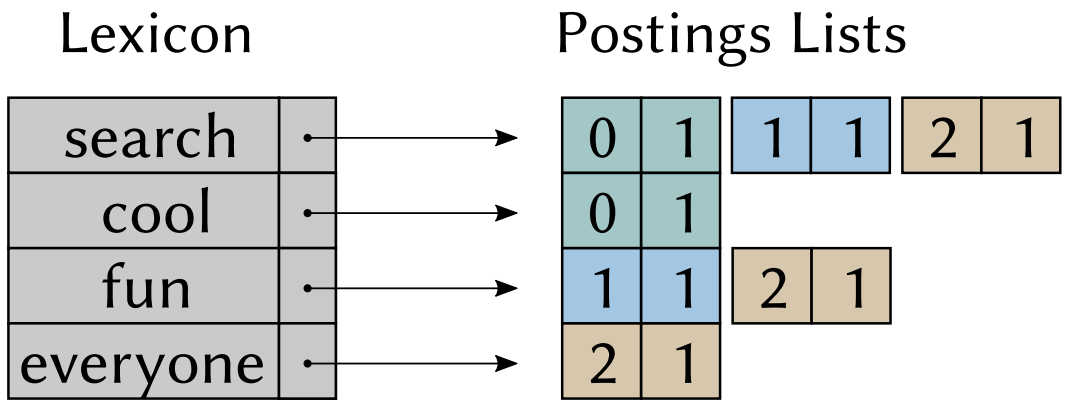

Lexicon
Postings Lists
search
cool
fun
everyone
0 1 1 1 2 1
0 1
1 1 2 1
2 1



Score documents with a bag-of-words ranker

-BM25, Language Model, DPH, etc


Lexicon
Postings Lists
search
cool
fun
everyone
0 1 1 1 2 1
0 1
1 1 2 1
2 1

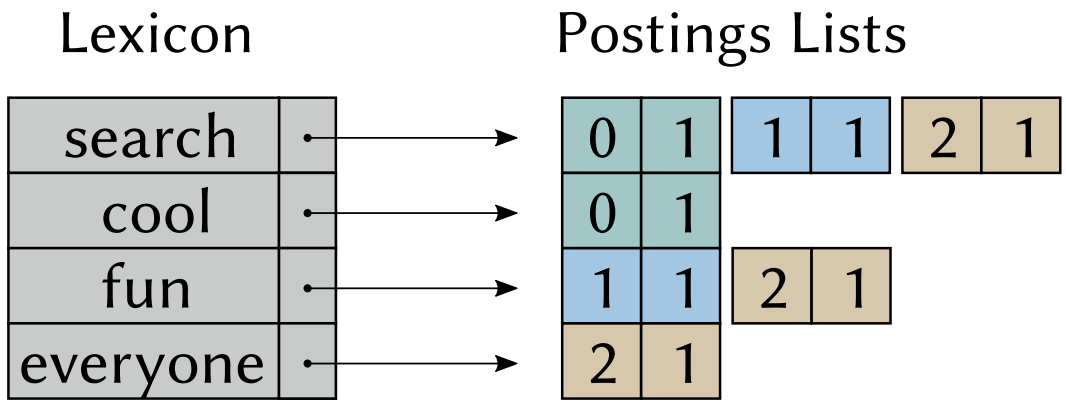


Idea: Add questions to documents to avoid vocabulary mismatch!

search is cool

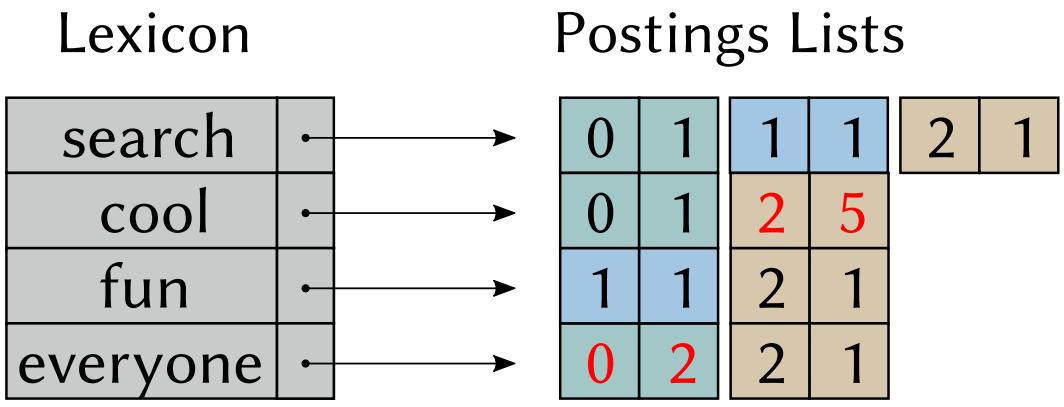
search is cool

+ *why is search cool?*
+ *who thinks search is cool?*
+...



Idea: Add questions to documents to avoid vocabulary mismatch!

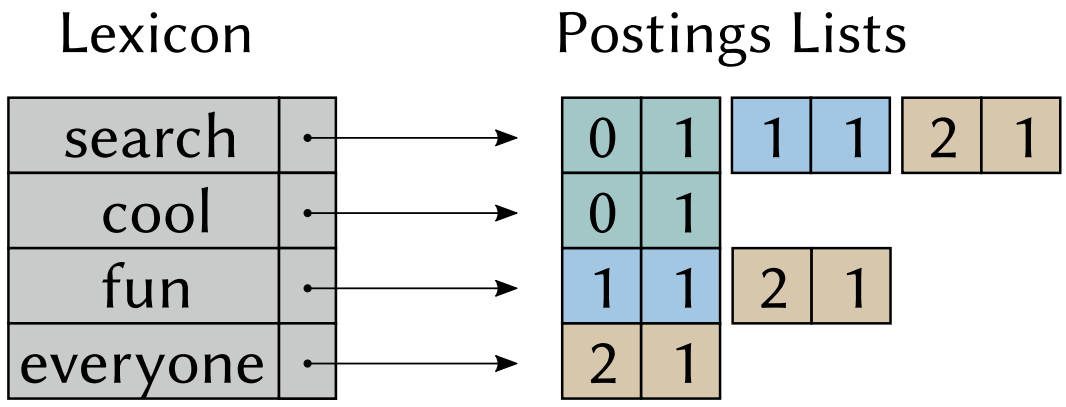
- Doc2Query, DocTTTTTQuery, TILDE

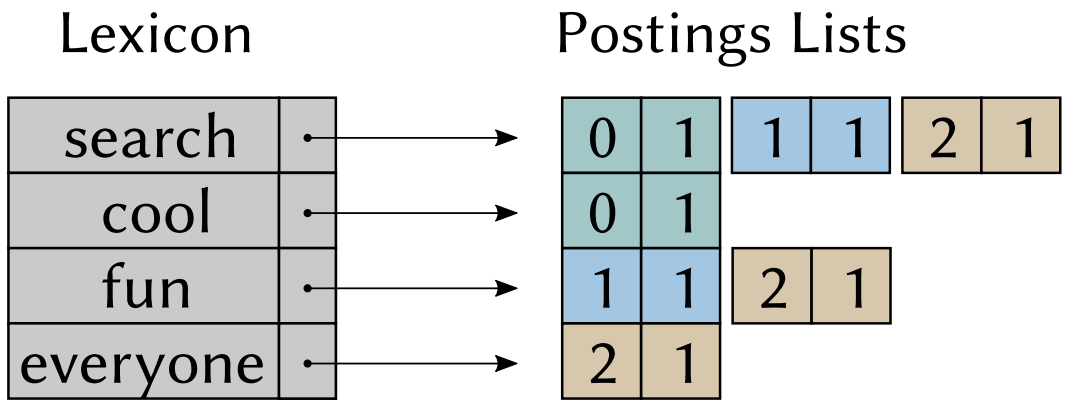


Idea: Add questions to documents to avoid vocabulary mismatch!

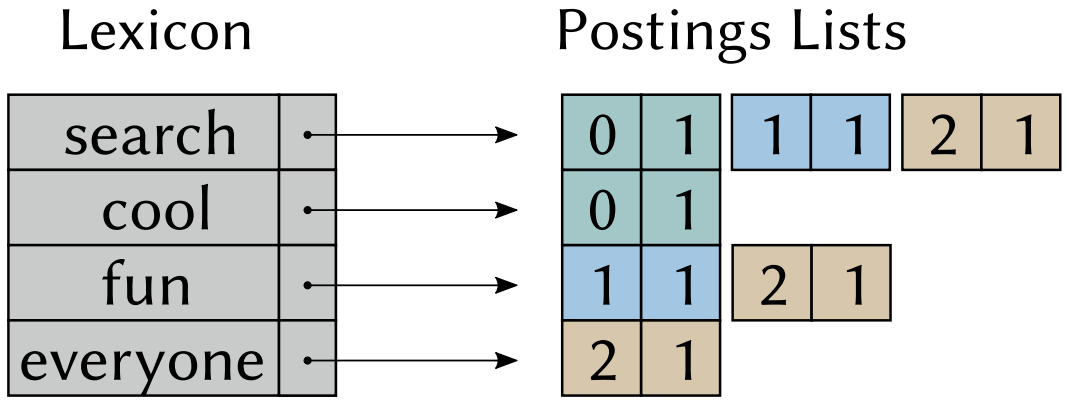
- Doc2Query, DocTTTTTQuery, TILDE

Expanded documents $\rightarrow$ longer postings lists


Lexicon
Postings Lists
search
cool
fun
everyone
0 1 1 1 2 1
0 1
1 1 2 1
2 1

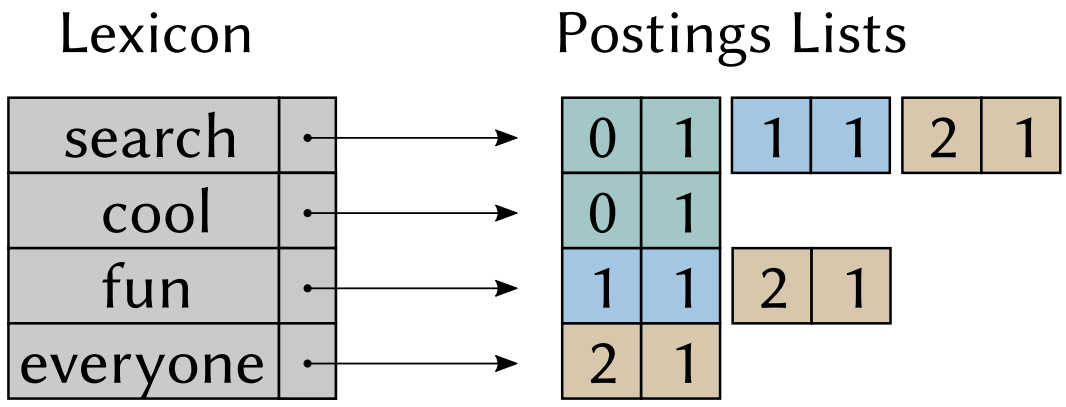


Idea: Why don't we **learn the term frequencies?**

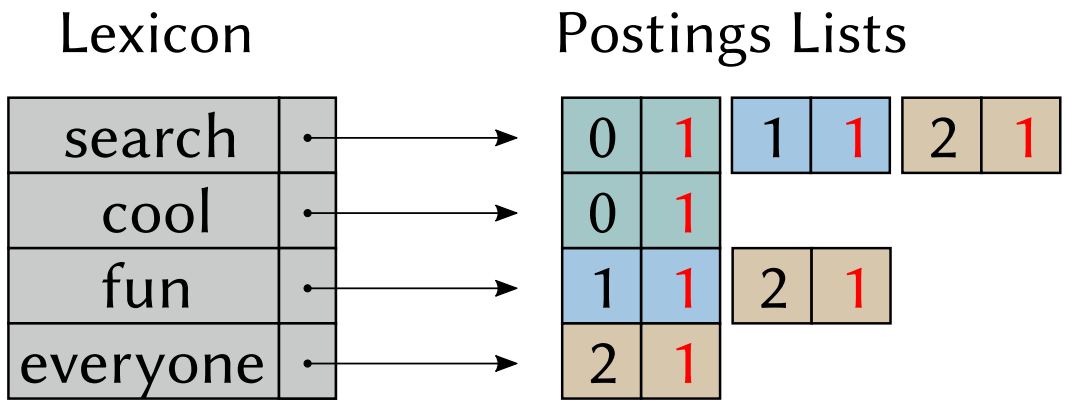


Idea: Why don't we **learn the term frequencies?**

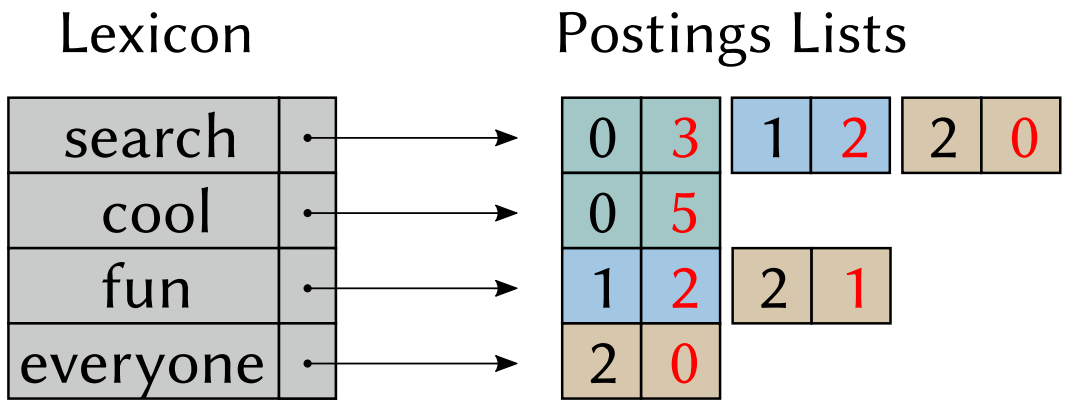
- DeepCT, HDCT
- Use contextual language models to re-weight tf's
- Still score with traditional models!



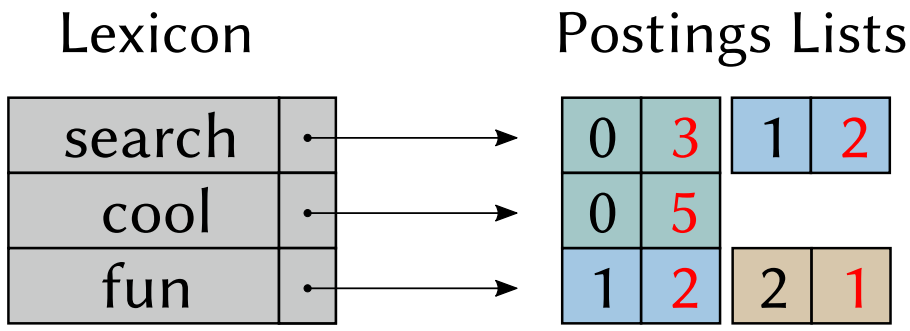
Idea: Why don't we **learn the term frequencies?**



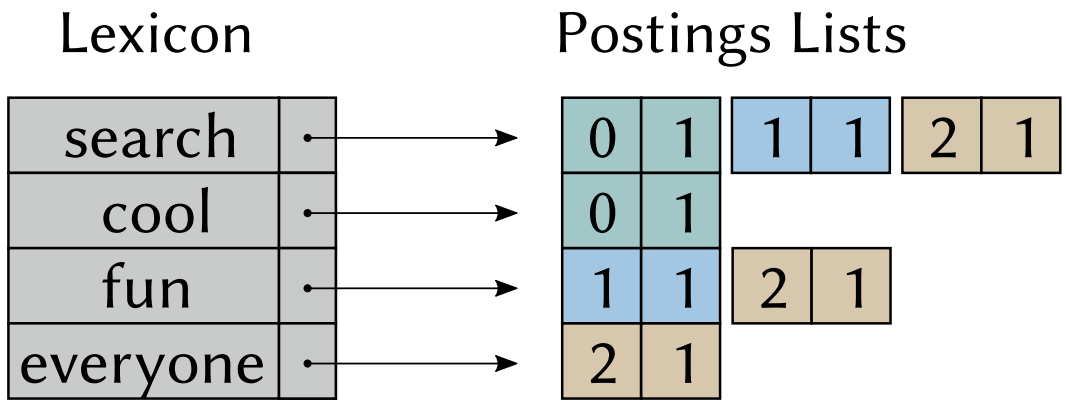
Idea: Why don't we **learn the term frequencies?**

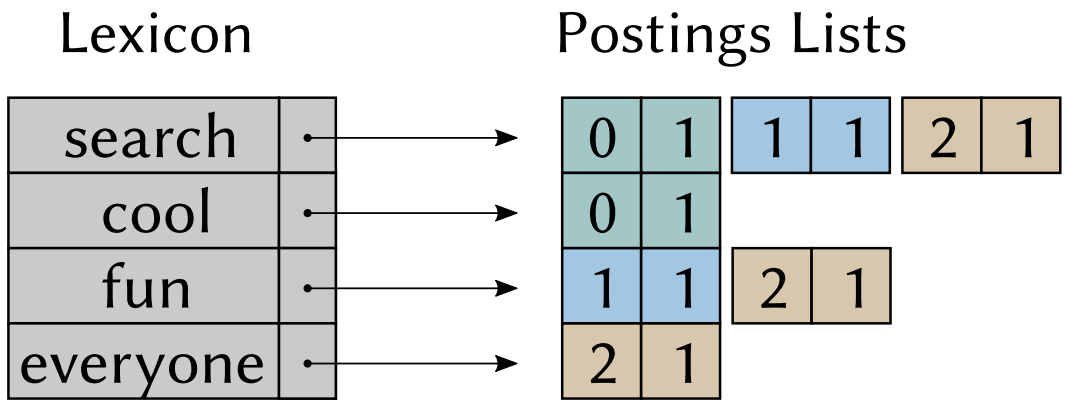


Idea: Why don't we **learn the term frequencies?**

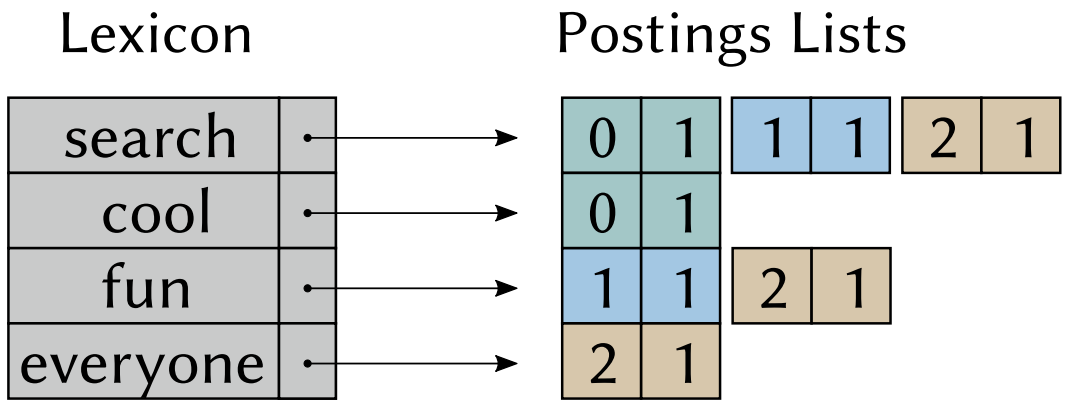


Idea: Why don't we **learn the term frequencies?**


Lexicon
Postings Lists
search
cool
fun
everyone
0 1 1 1 2 1
0 1
1 1 2 1
2 1

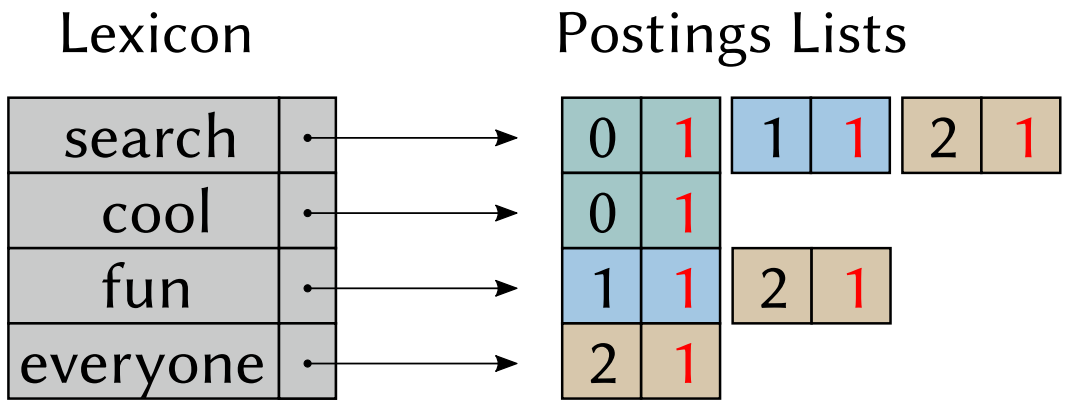


Idea: Learn the **impact** (not the term frequency)



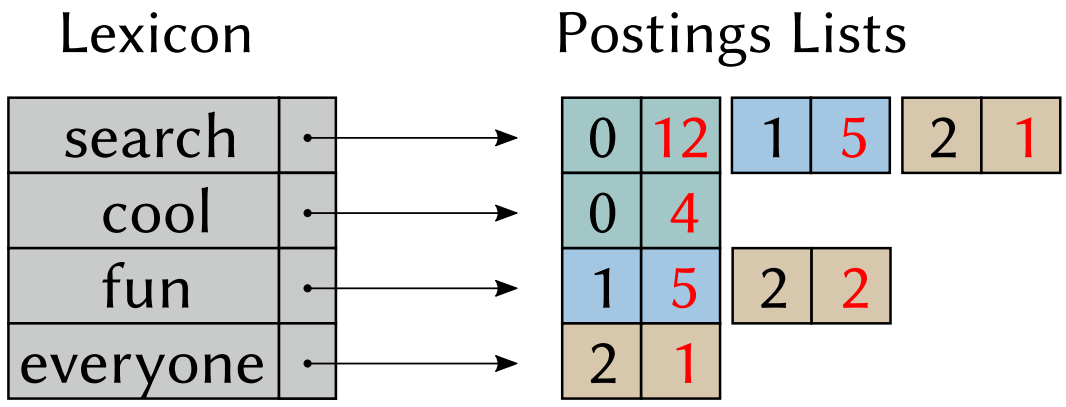
Idea: Learn the **impact** (not the term frequency)

- DeepImpact
- TILDE(v2)



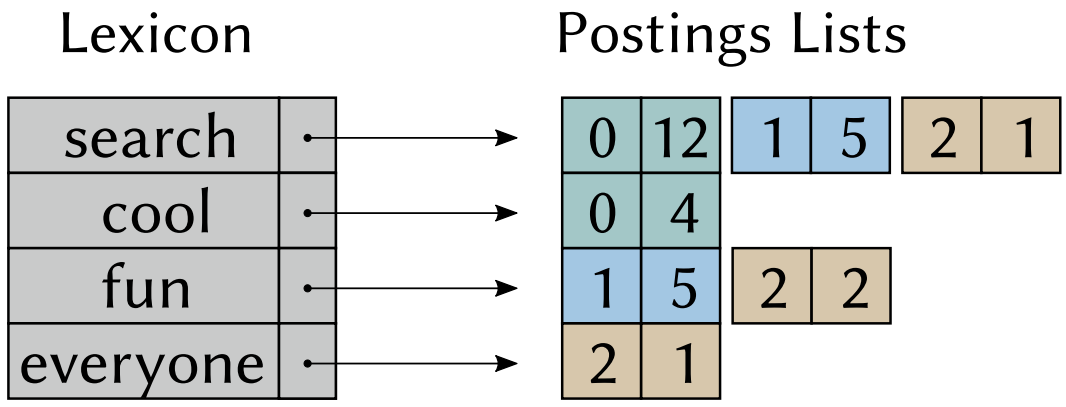
Idea: Learn the **impact** (not the term frequency)

- DeepImpact
- TILDE(v2)



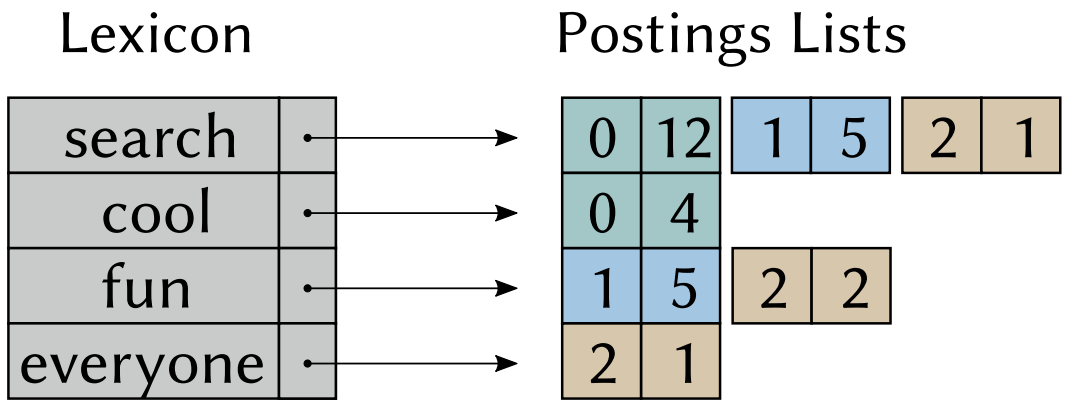
Idea: Learn the **impact** (not the term frequency)

- DeepImpact
- TILDE(v2)

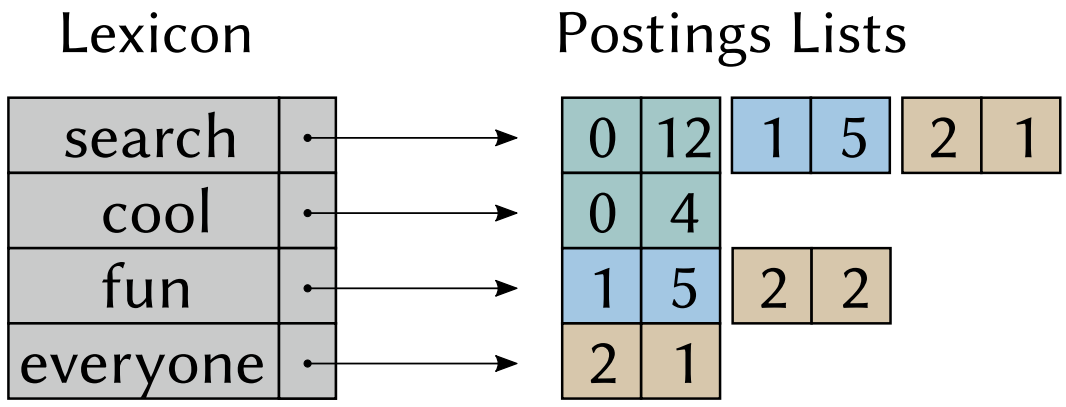


Idea: Learn the **impact** (not the term frequency)

- DeepImpact
- TILDE(v2)

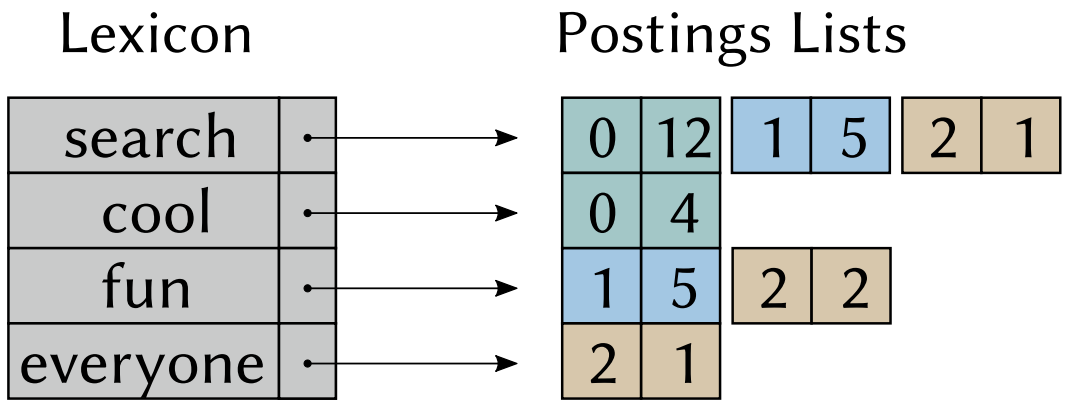


Idea: Learn the **impact** (not the term frequency)



Idea: Learn the **impact** (not the term frequency)

\+ Learn weights for each query



Idea: Learn the **impact** (not the term frequency)

\+ Learn weights for each query

\- uniCOIL, SPLADEv2

Default

999391 where is bulli creek queensland australia

uniCOIL-TILDE

999391 where where where where where where where where where where where where where where where where where
where where where where where where where where where where where where where where where where where where
where where where where where where where is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is is
is is is is is is is is is is bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull
bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull
bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull
bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull
bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull
bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull
bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull bull
bull bull bull ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i ##i
....

[where 42][is 44][bull 194][##i 116][creek 165]
[queensland 149][australia 70]

SPLADEv2 [Adds query expansion]

999391 it it it it it it it it it it it it it it it it it it it it it it it it it it it it it it it it it from from from from from from are are are are are are are him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him him ##i where ....

[where 268] [bull 237] [creek 219] [queensland 211] [australia 183] [##i 132] [headquarters 92] [australian 86] [habitat 72] [stream 71] [brisbane 64] [him 40] [lake 37] [scotland 31] [it 29] [are 7] [from 6] [river 5] [italy 3]

Problem: Learned sparse models negatively impact processing latency.

Purpose: Make learned sparse models as fast as traditional models.

Approach: Heuristics.

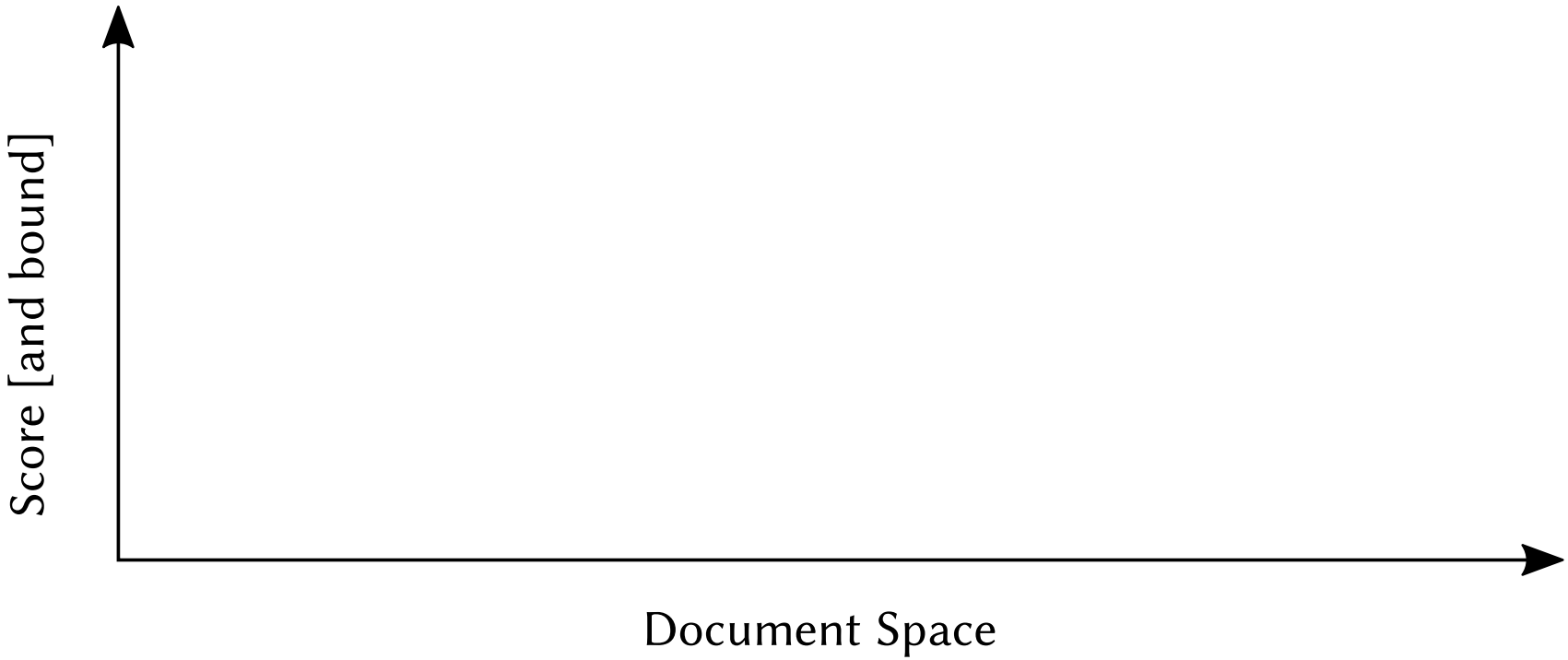
Why are learned sparse models slower?

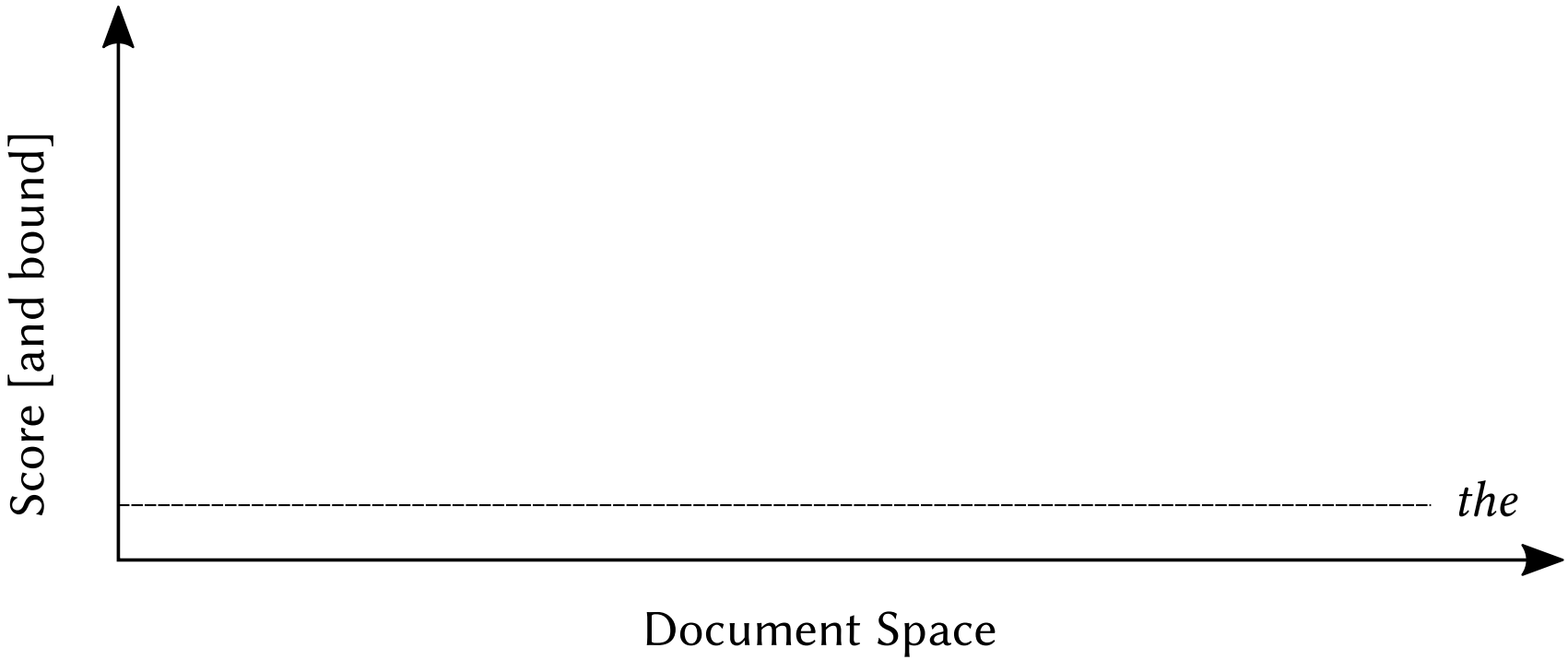
Dynamic Pruning at a glance

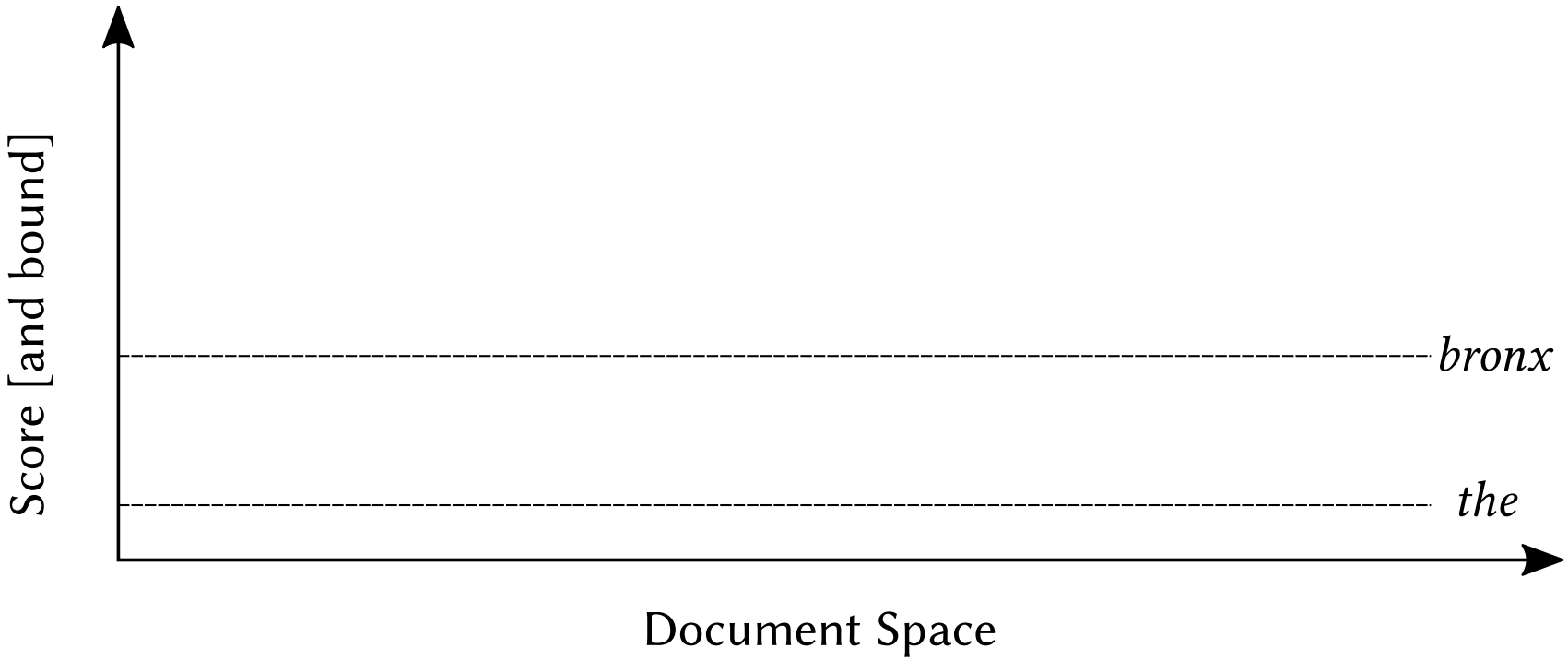
Assumption: Ranking is *additive*.

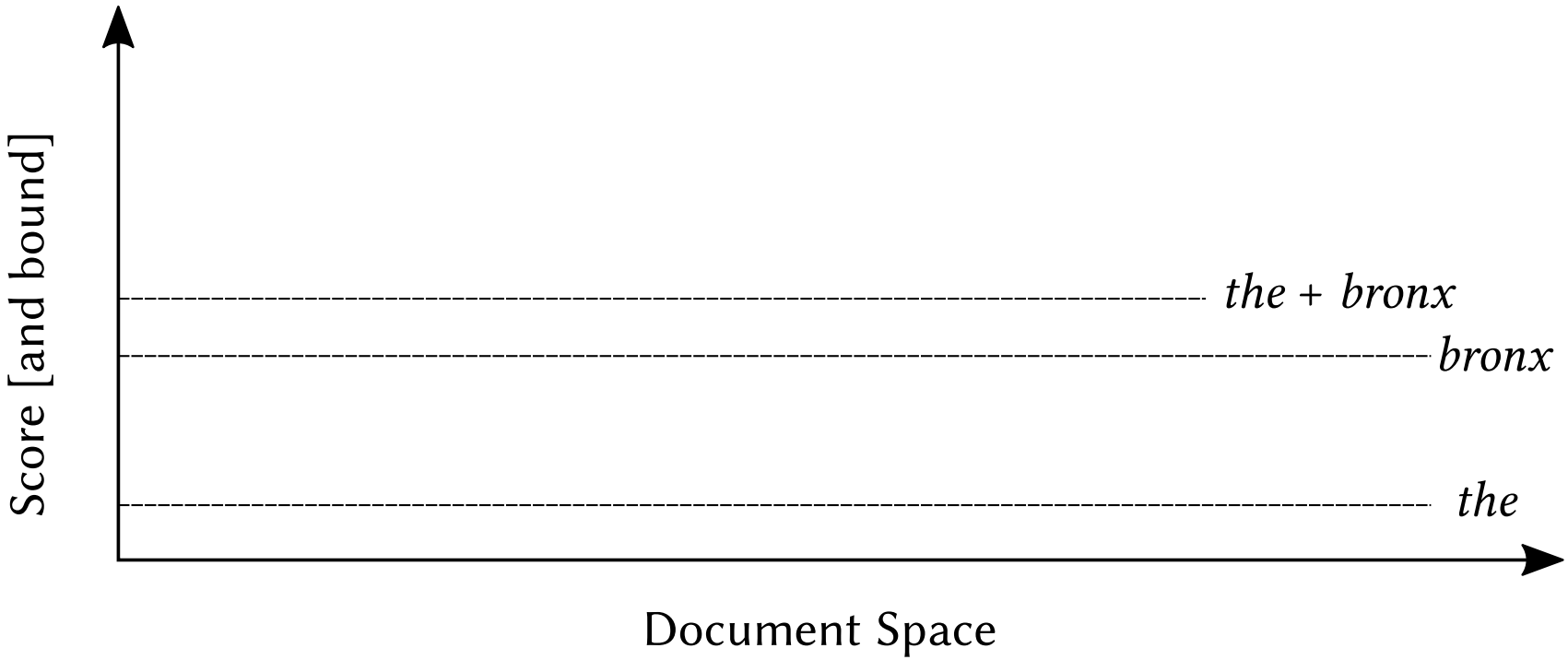
Requirement: We have pre-computed the maximum impact for each term (postings list) and stored it.

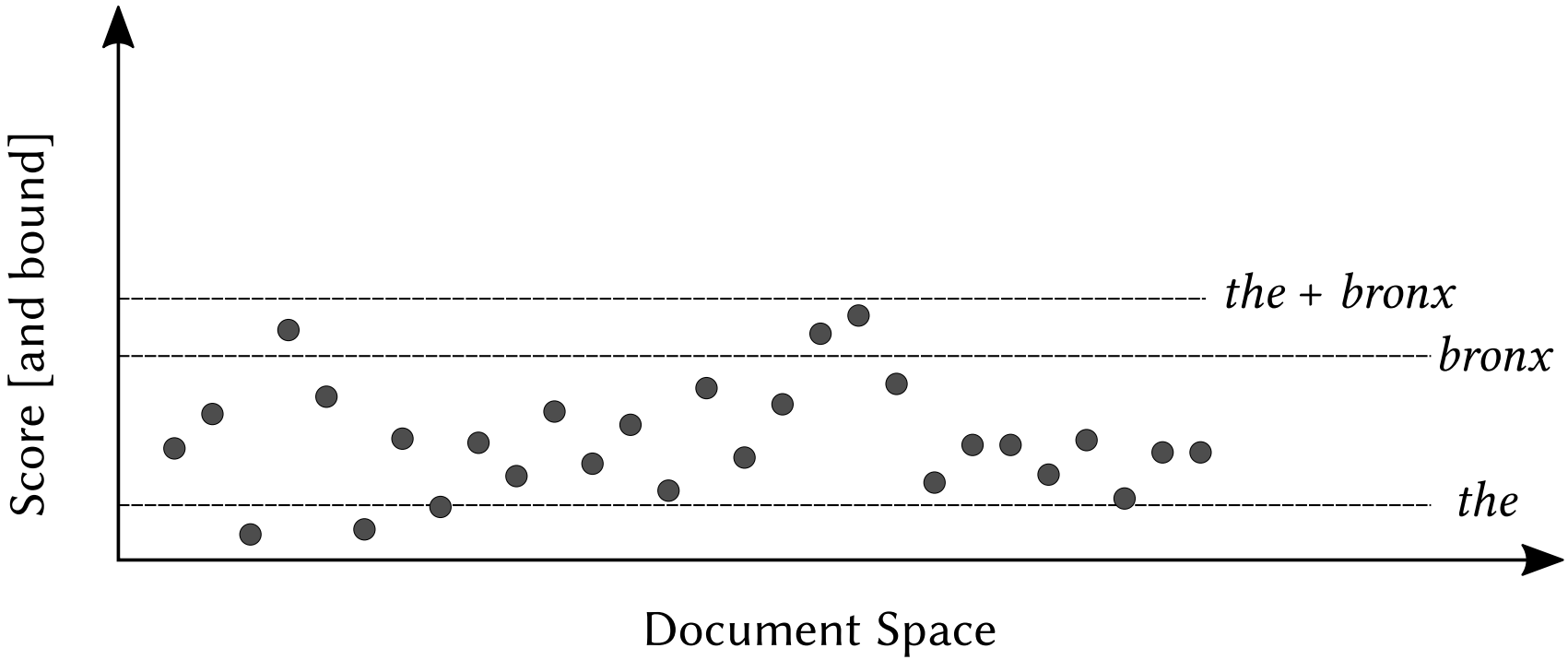
Requirement: During processing, we have access to the kth highest score "seen so far" (a threshold θ).

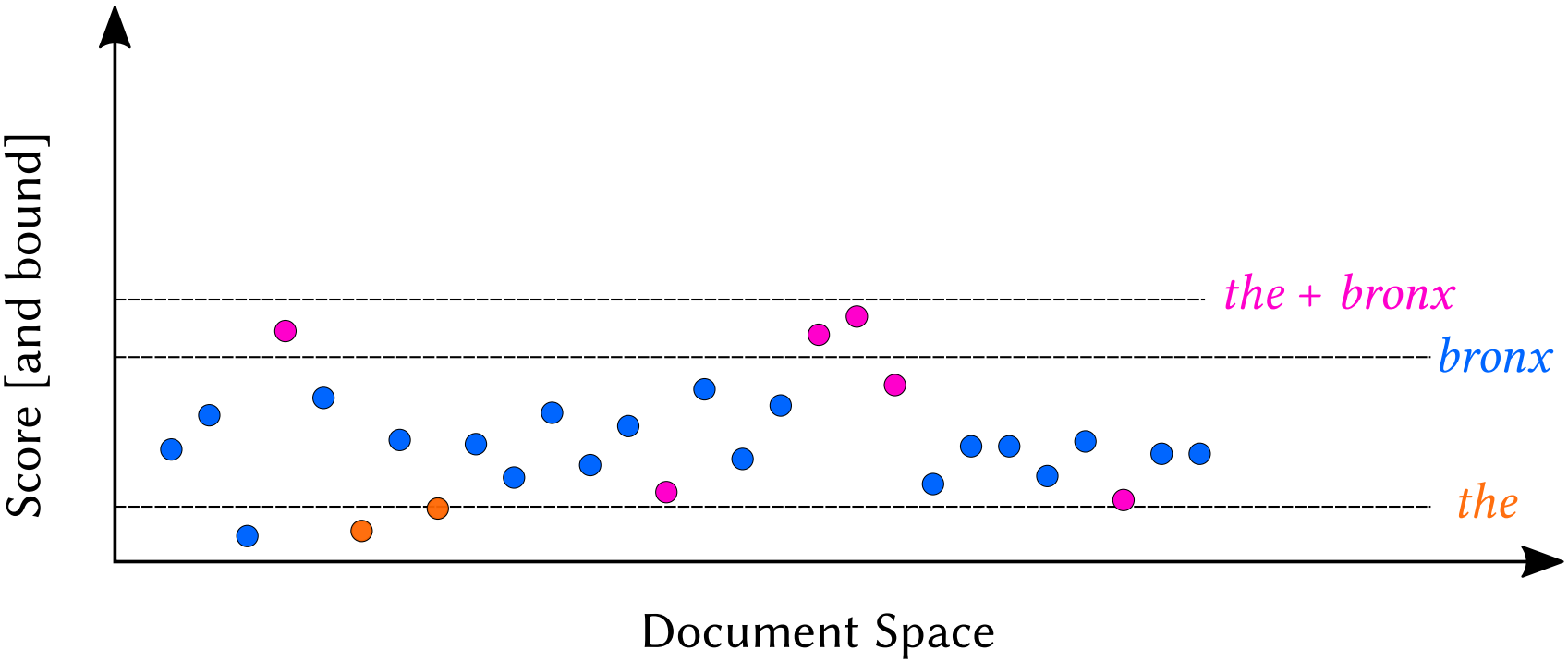

Score [and bound]
Document Space

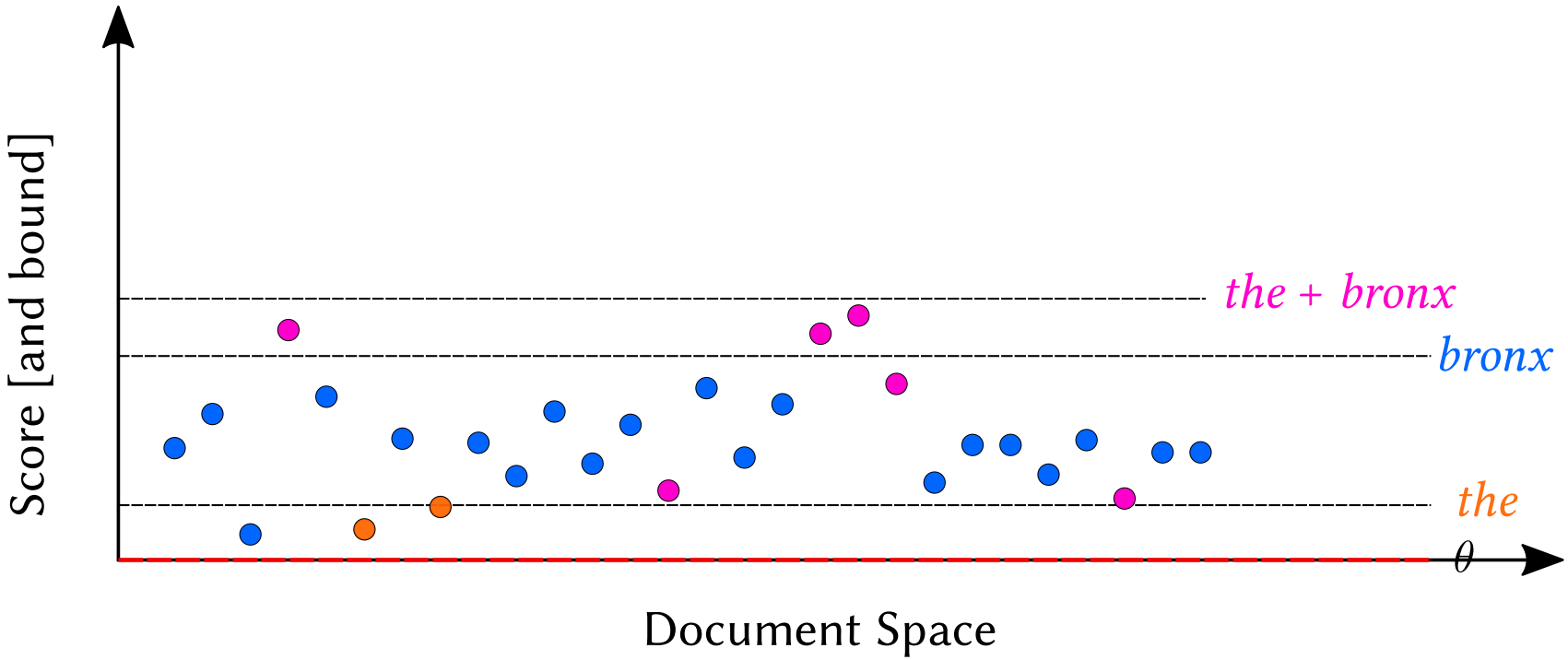

Score [and bound]
the
Document Space

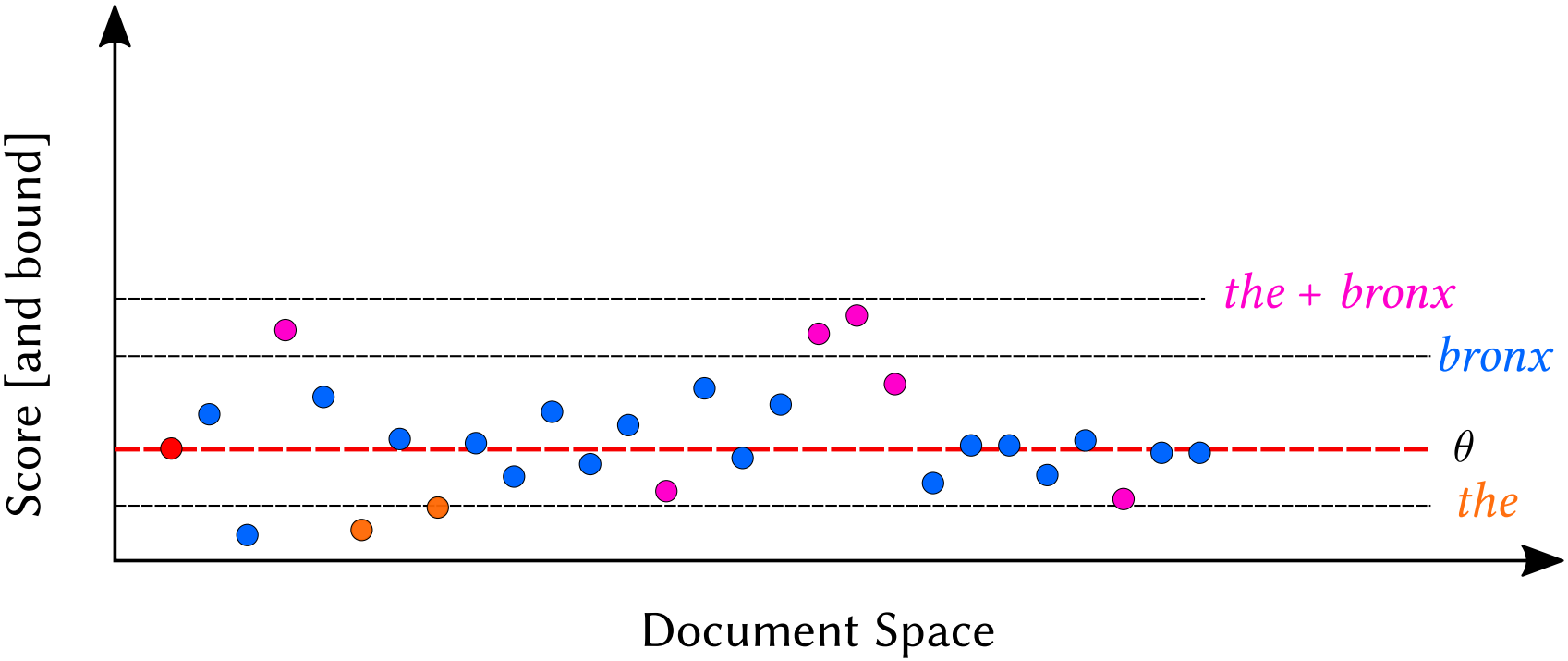

Score [and bound]
bronx
the
Document Space

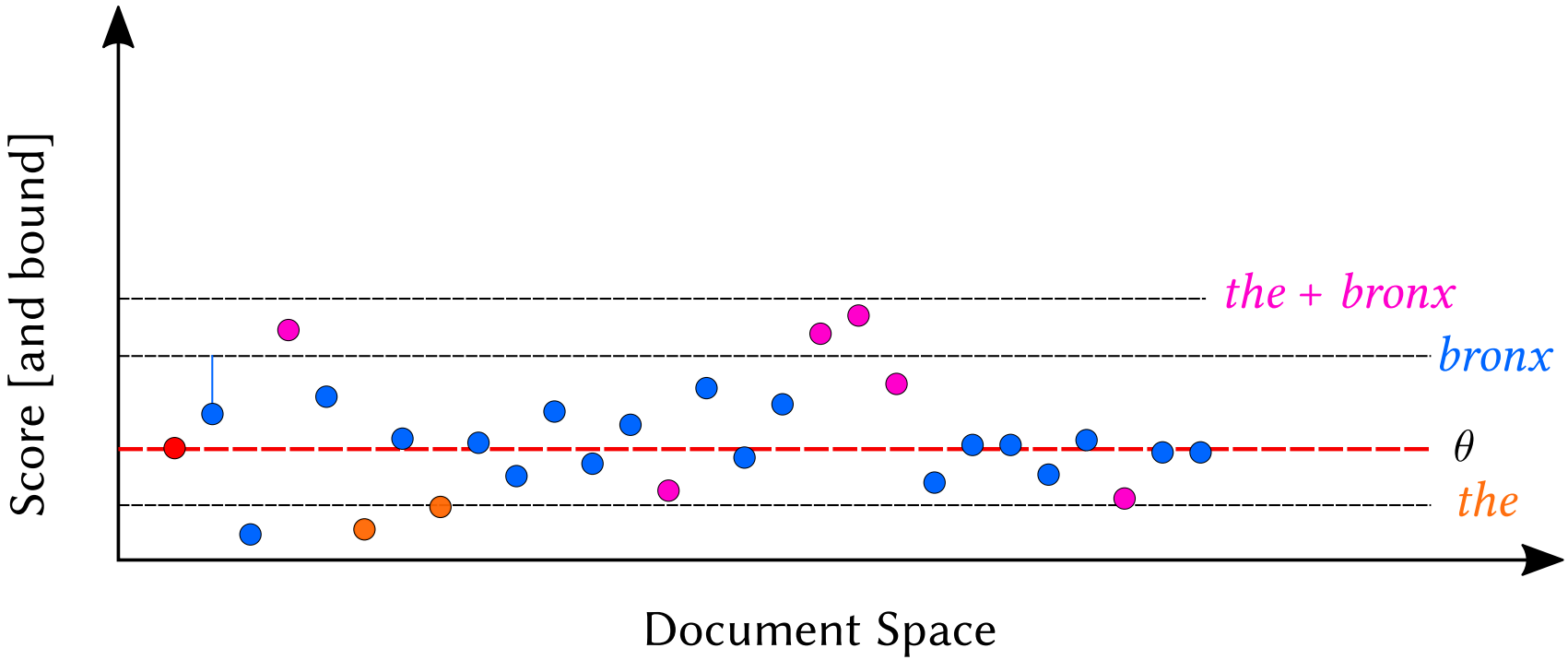

Score [and bound]
the + bronx
bronx
the
Document Space

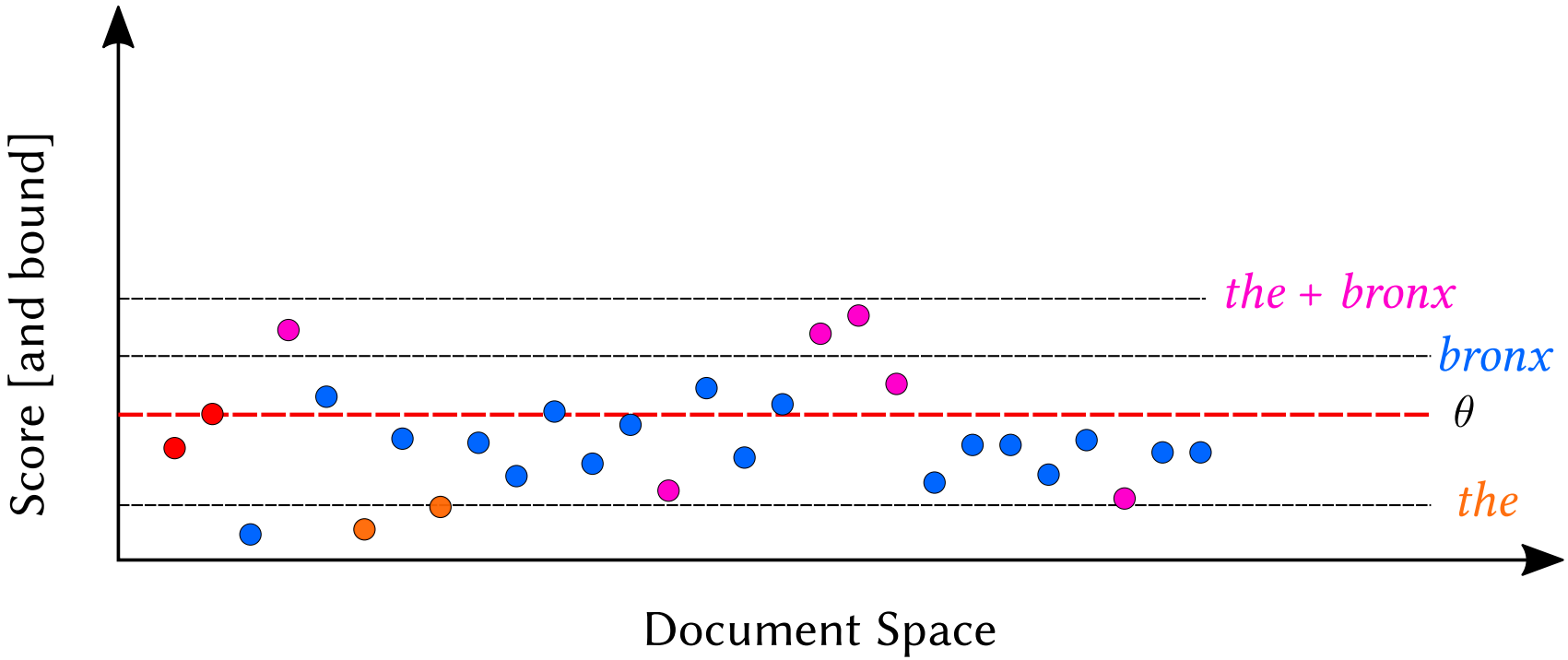

Score [and bound]
the + bronx
bronx
the
Document Space

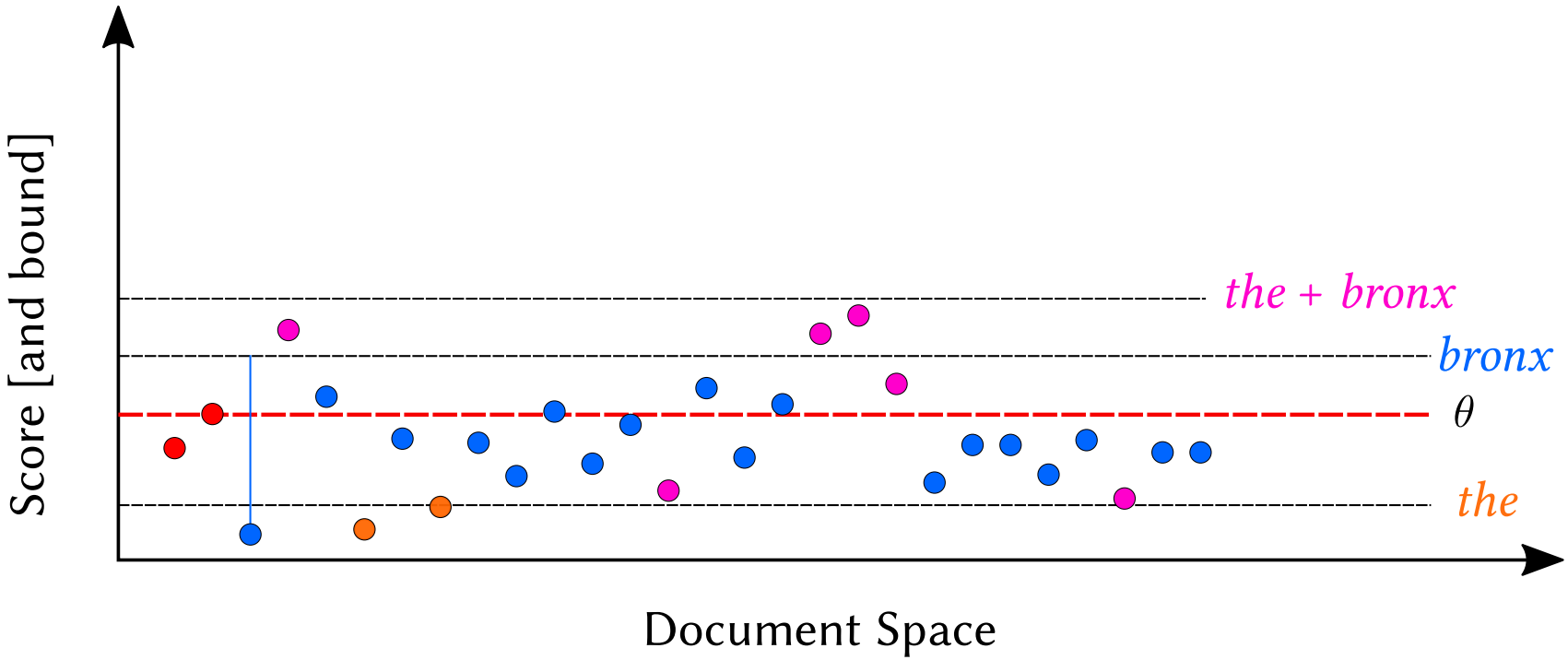

Score [and bound]
the + bronx
bronx
the
Document Space

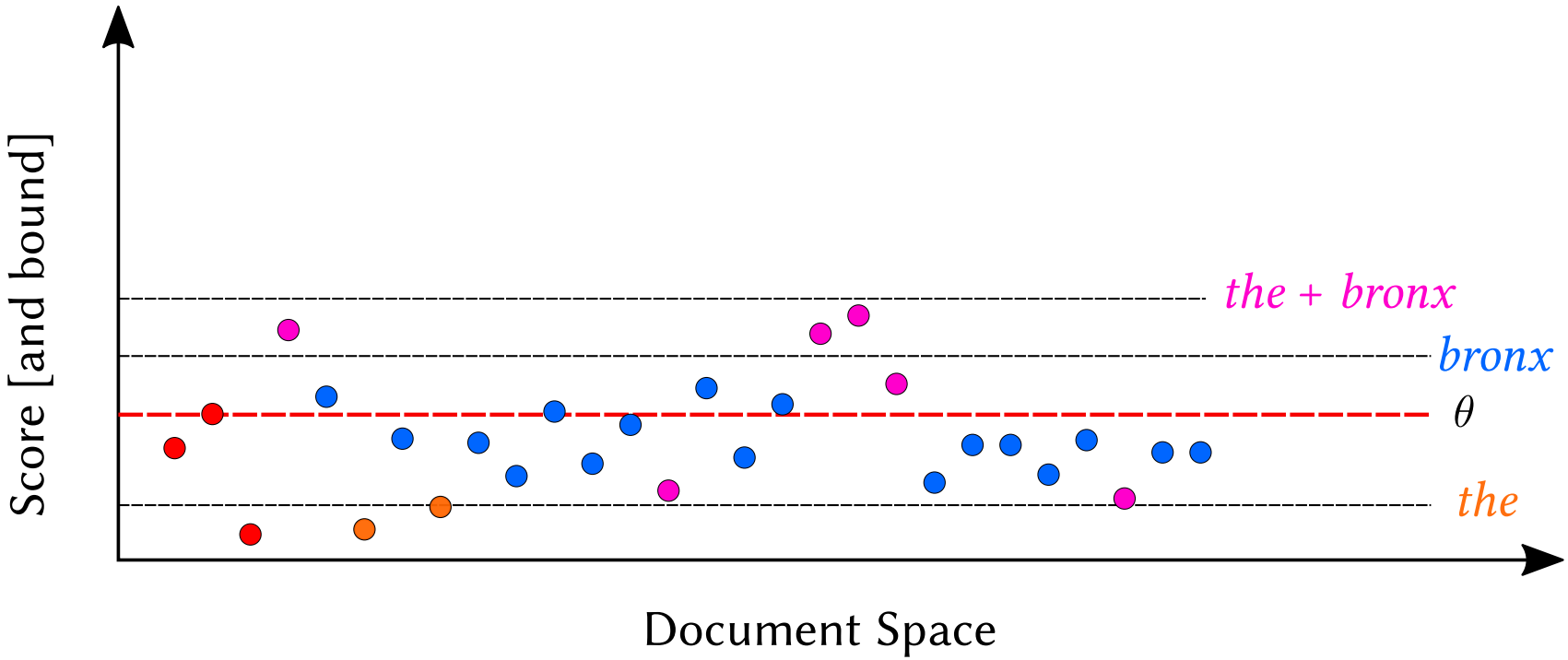

Score [and bound]
the + bronx
bronx
the
0
Document Space

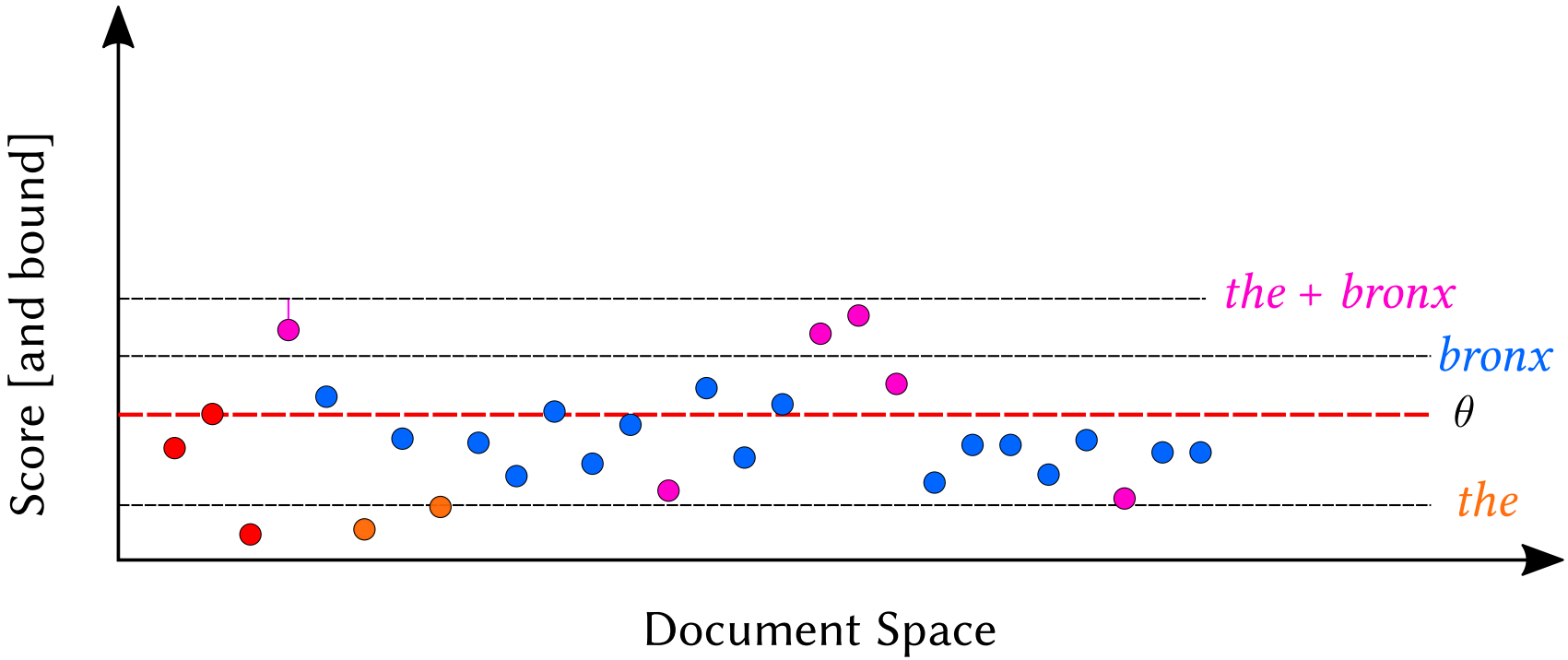

Score [and bound]
the + bronx
bronx
θ
the
Document Space

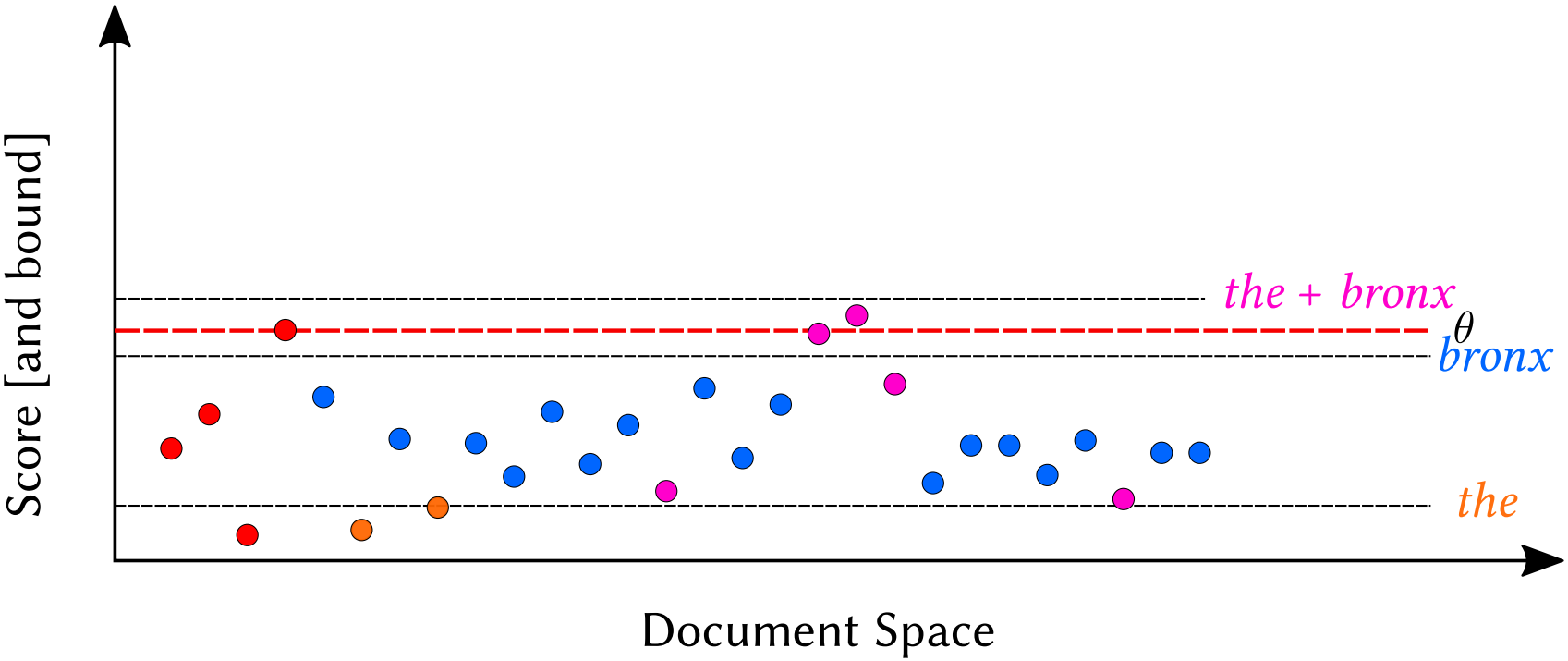

Score [and bound]
the + bronx
bronx
θ
the
Document Space

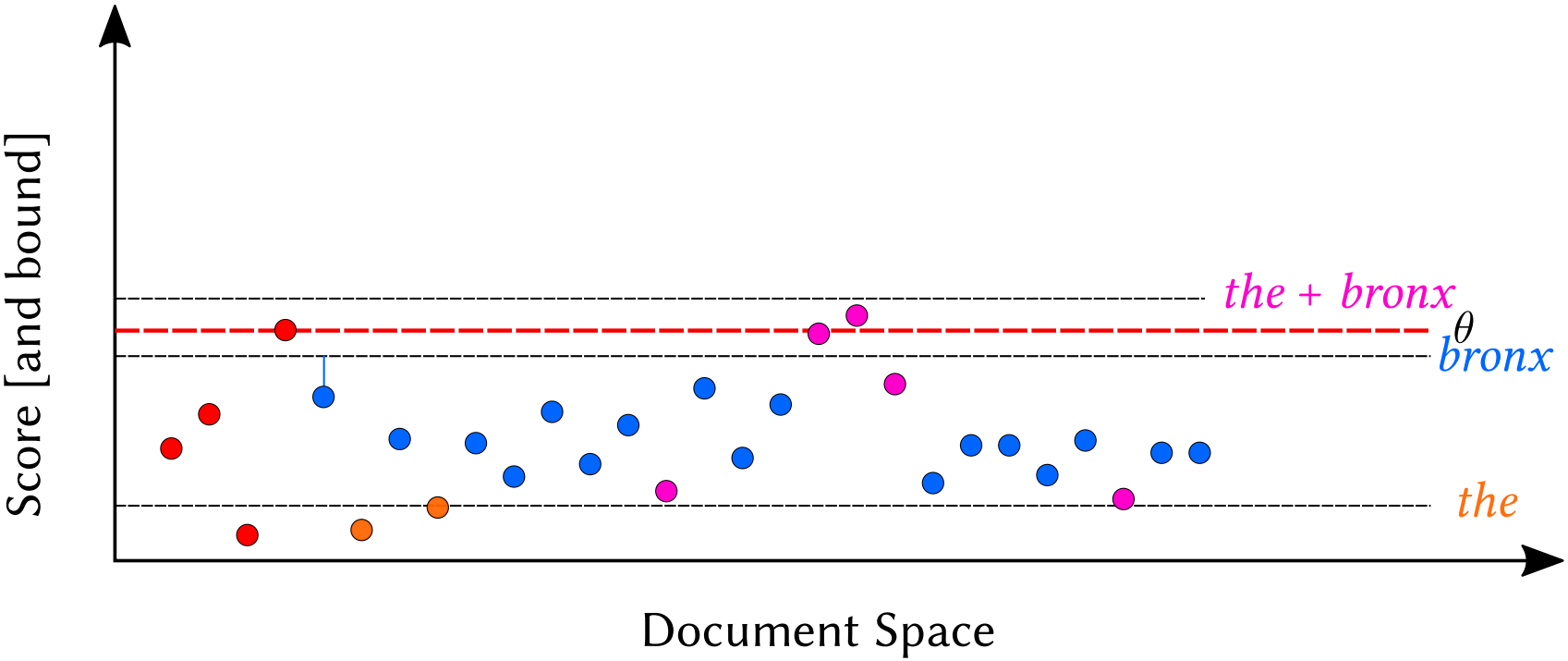

Score [and bound]
the + bronx
bronx
θ
the
Document Space

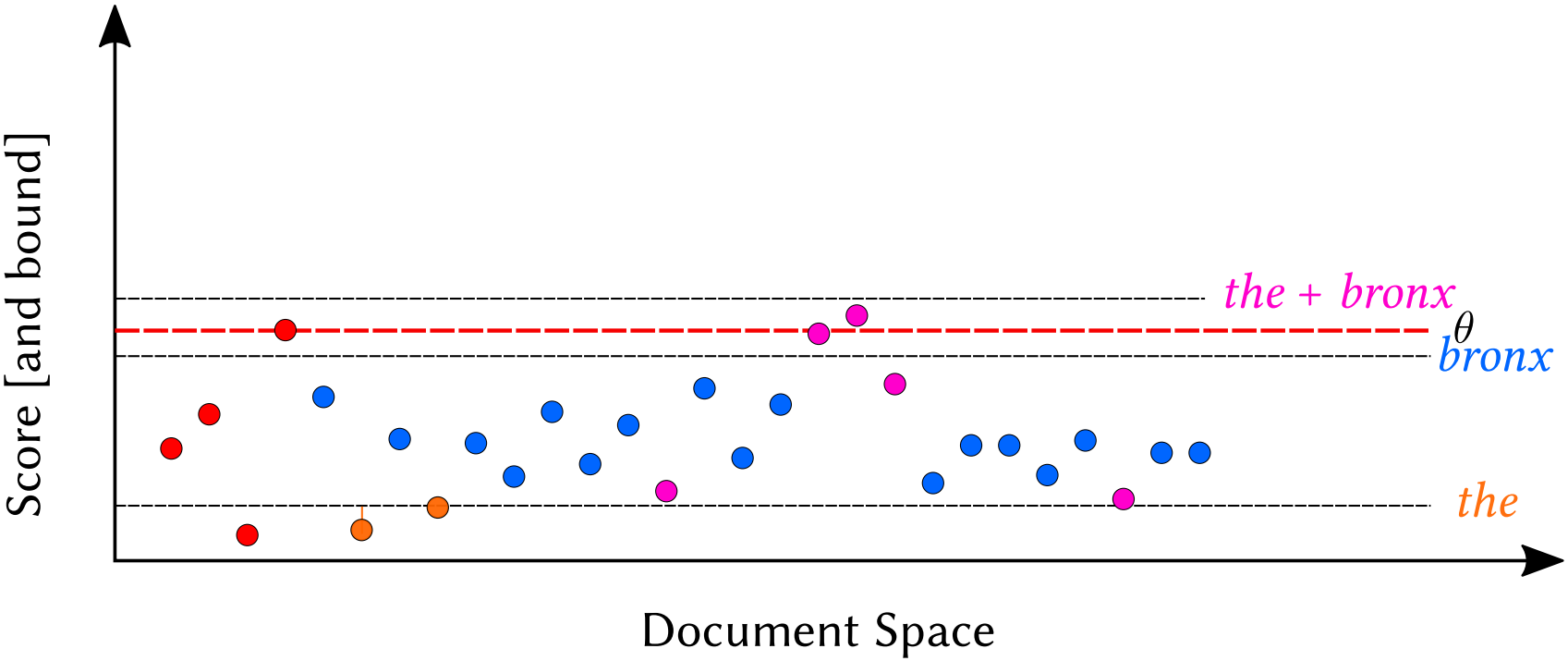

Score [and bound]
the + bronx
bronx
θ
the
Document Space

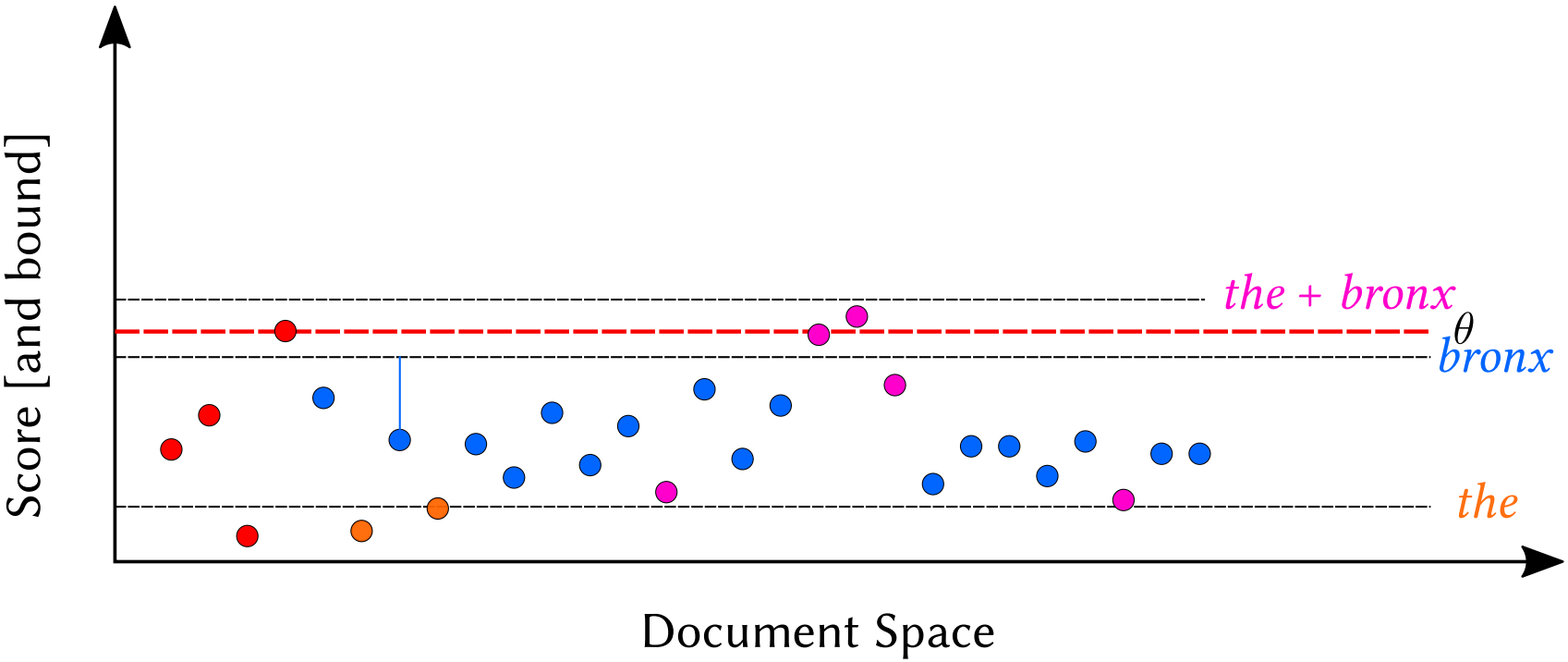

Score [and bound]
the + bronx
bronx
θ
the
Document Space

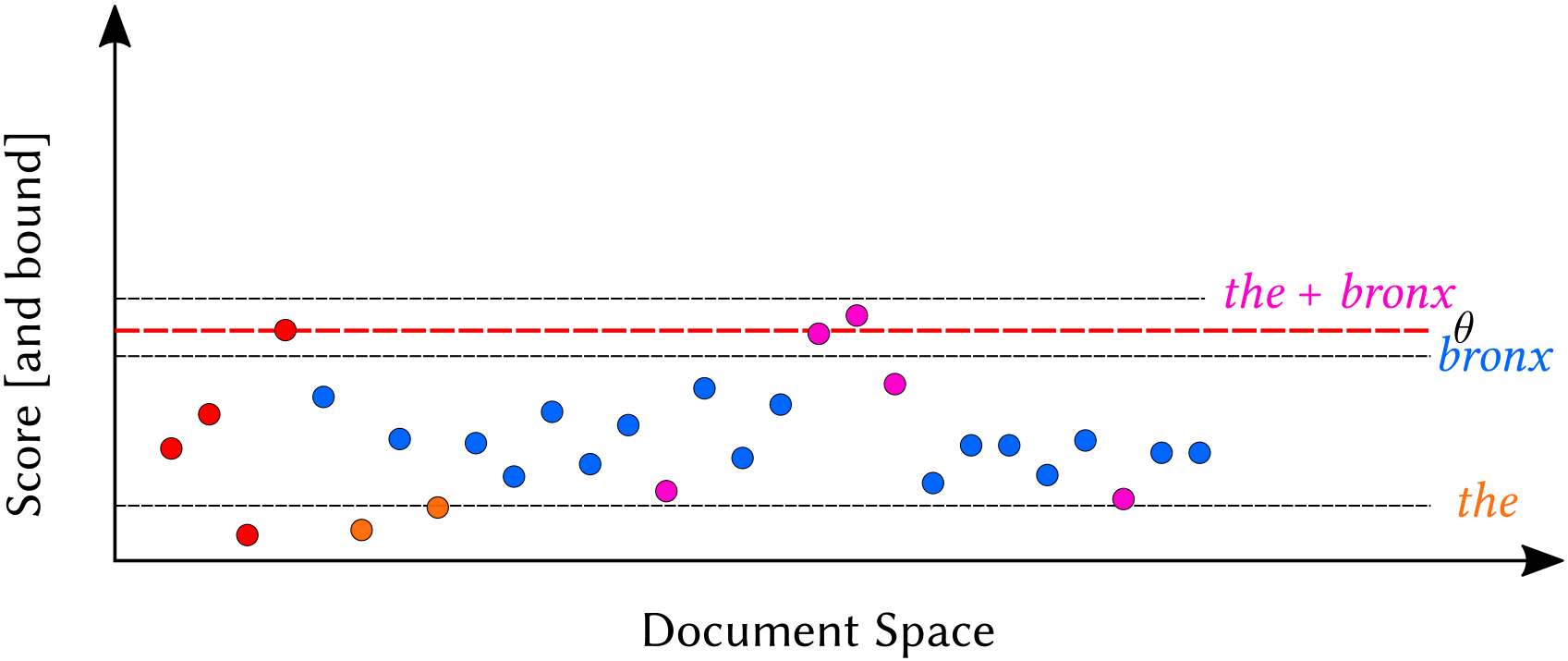

Score [and bound]
the + bronx
bronx
θ
the
Document Space

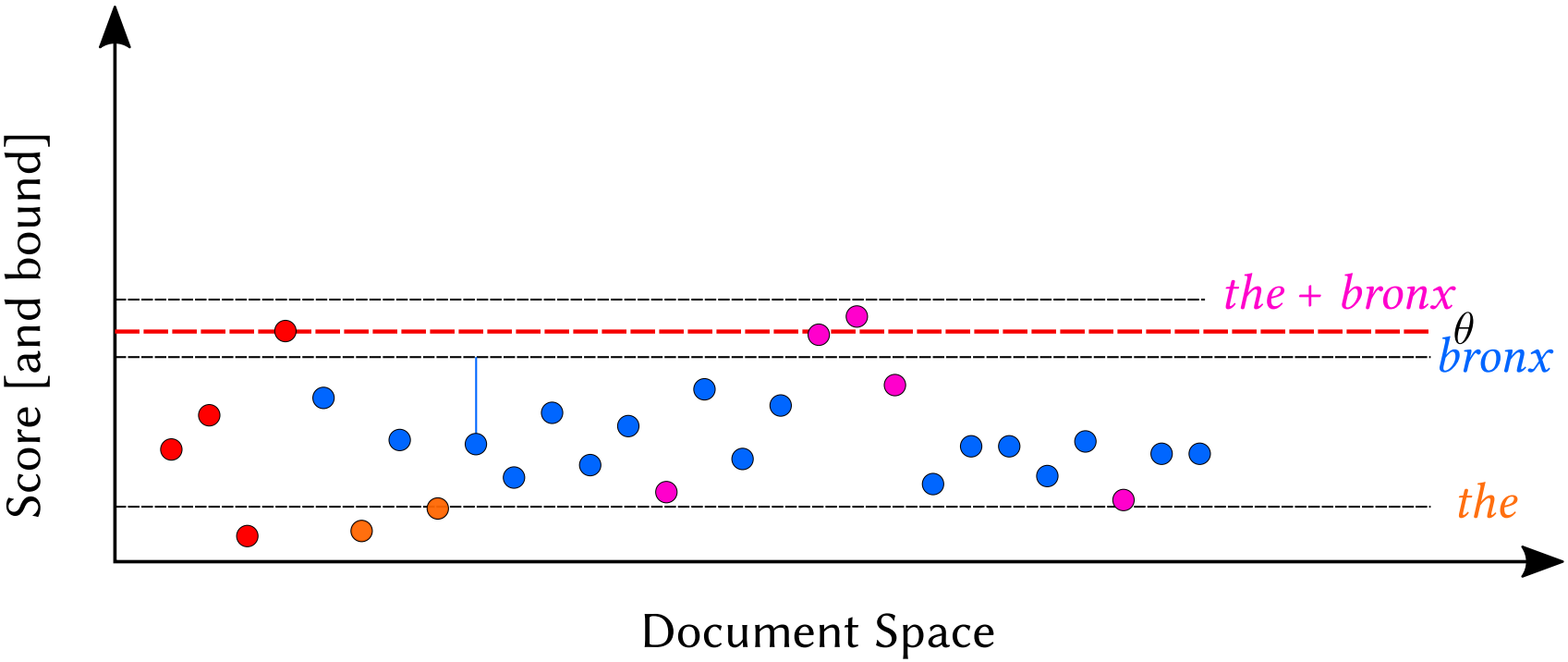

Score [and bound]
the + bronx
θ
bronx
the
Document Space

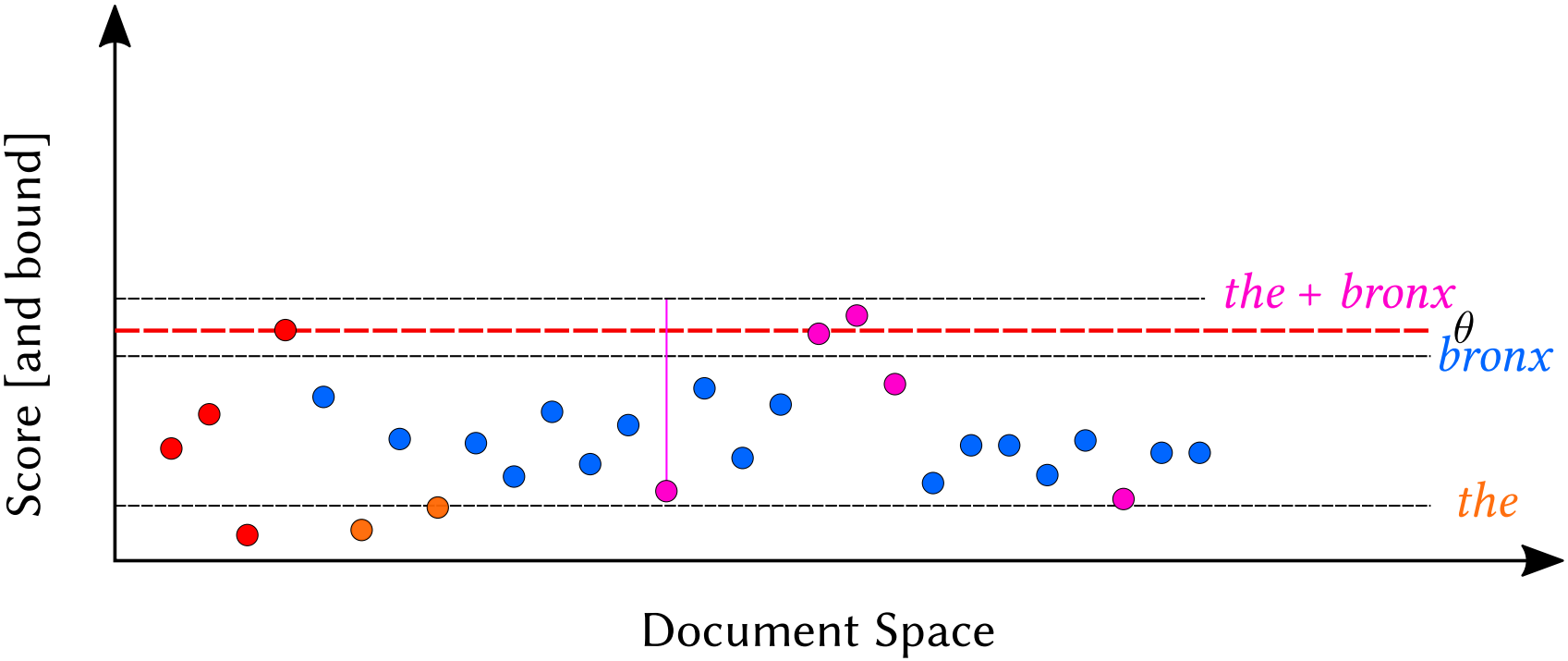

Score [and bound]
the + bronx
θ
bronx
the
Document Space

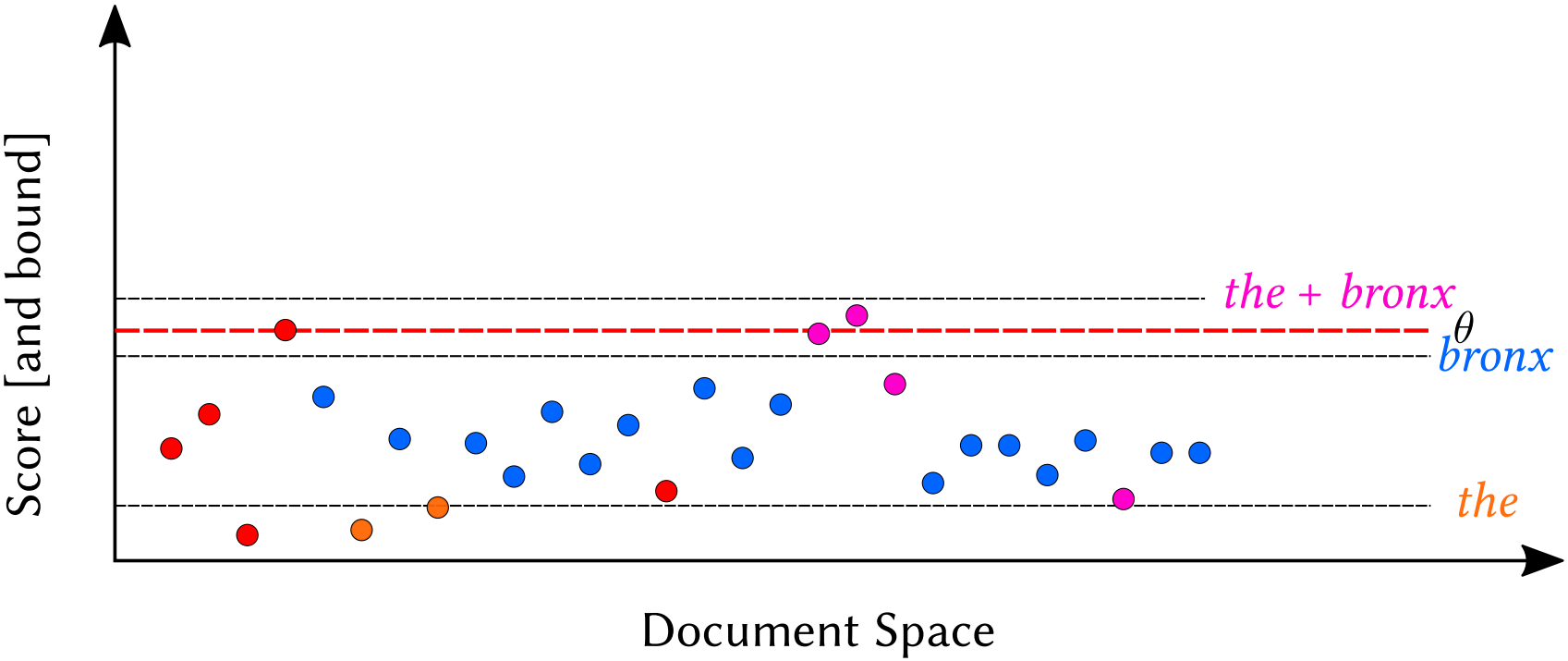

Score [and bound]
the + bronx
θ
bronx
the
Document Space

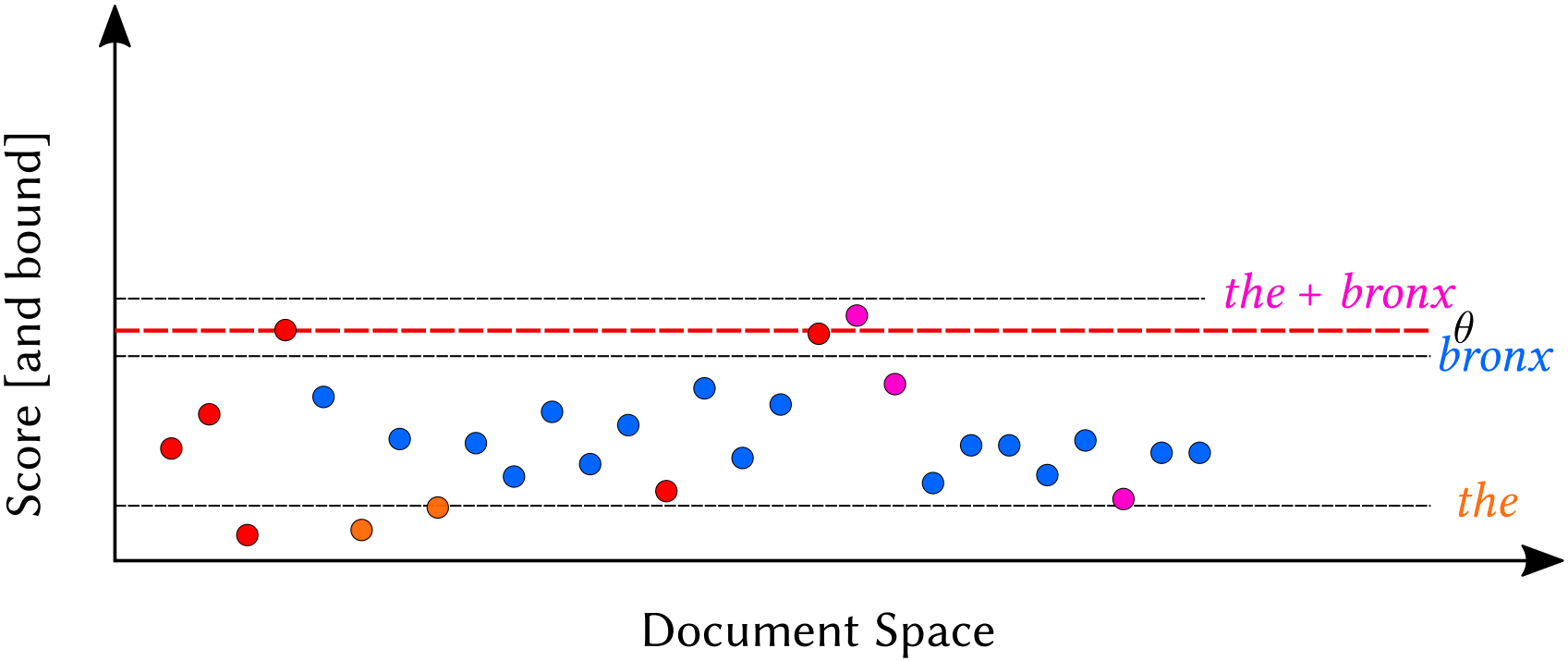

Score [and bound]
the + bronx
θ
bronx
the
Document Space

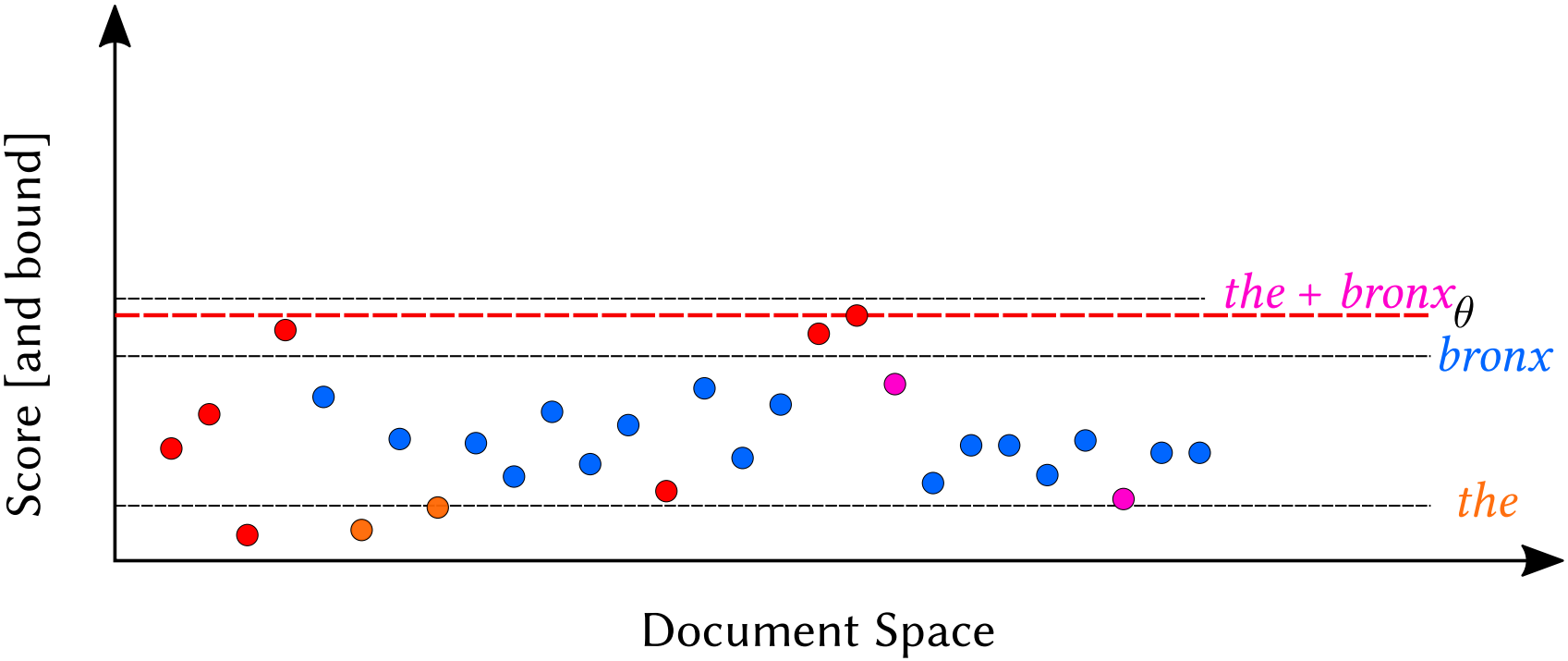

Score [and bound]
the + bronx
θ
bronx
the
Document Space

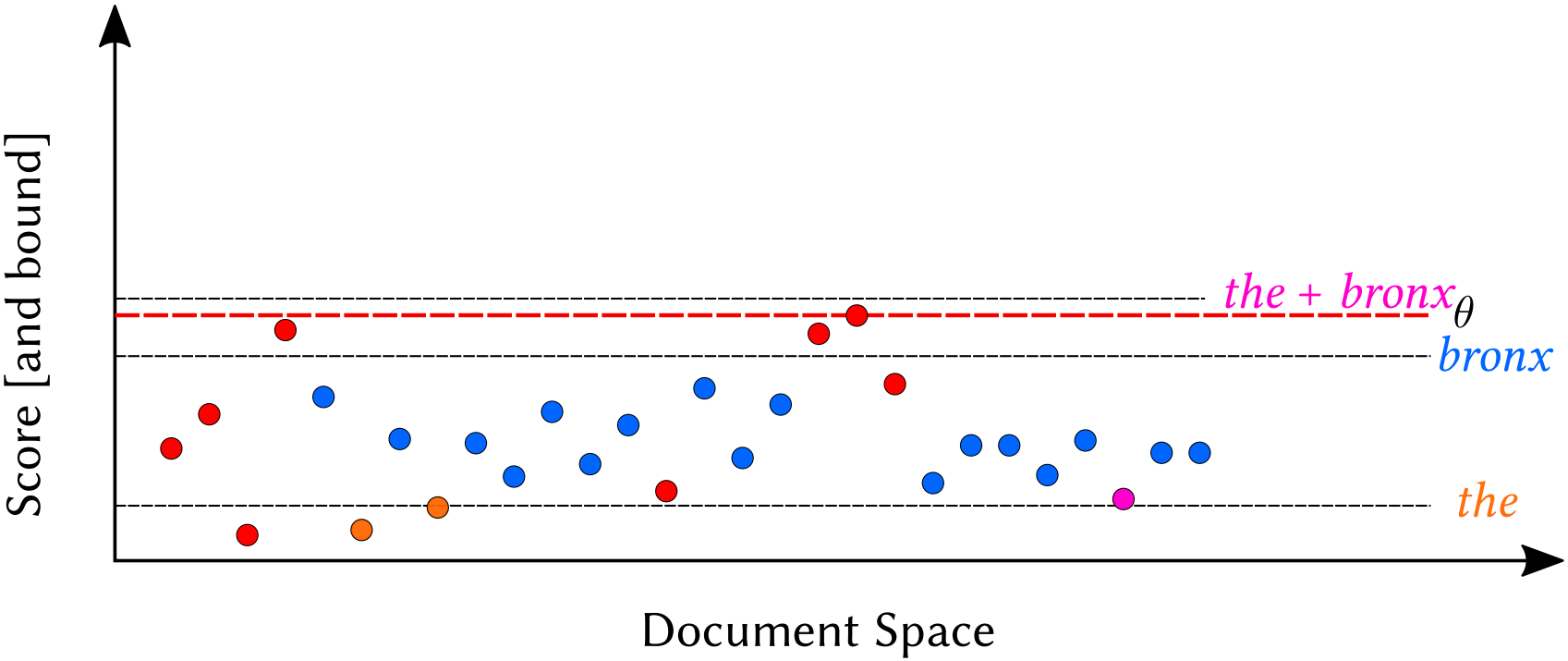

Score [and bound]
the + bronx
θ
bronx
the
Document Space

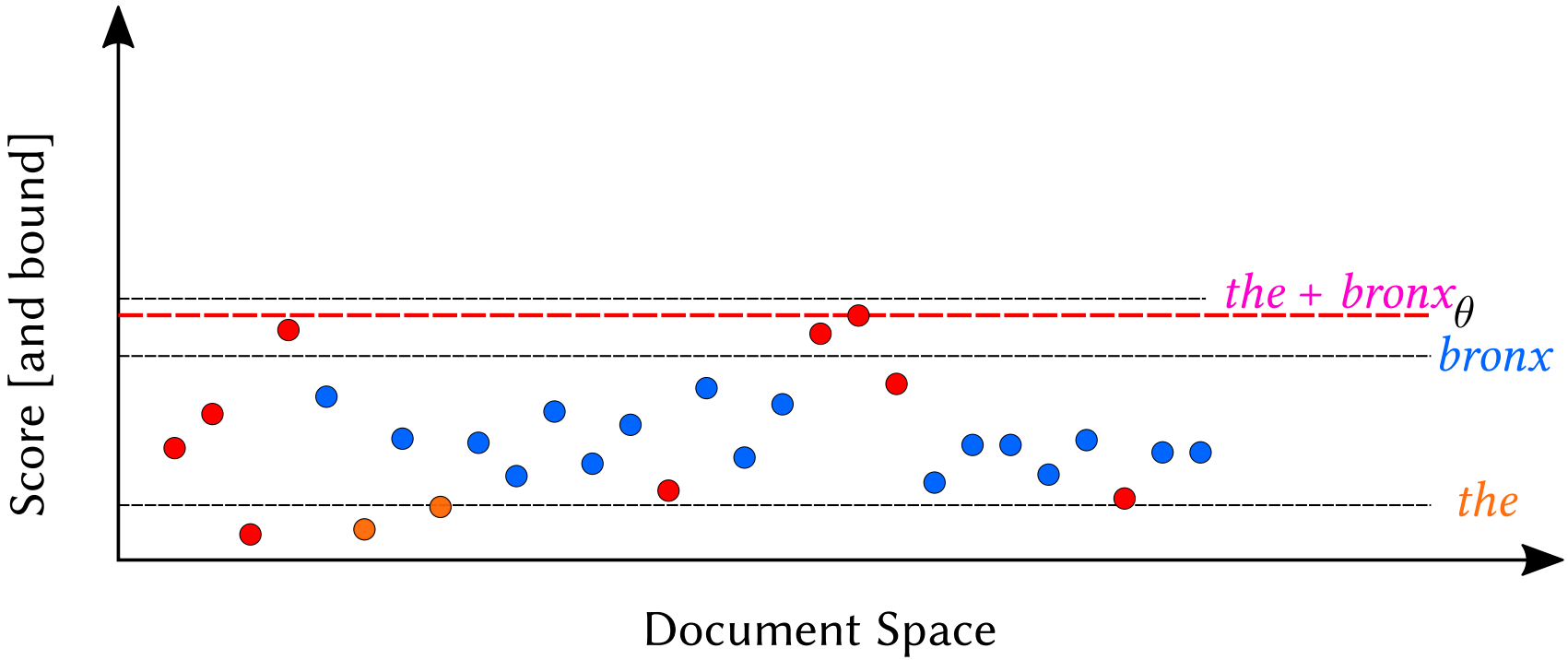

Score [and bound]
the + bronx
θ
bronx
the
Document Space


Score [and bound]
the + bronx
θ
bronx
the
Document Space

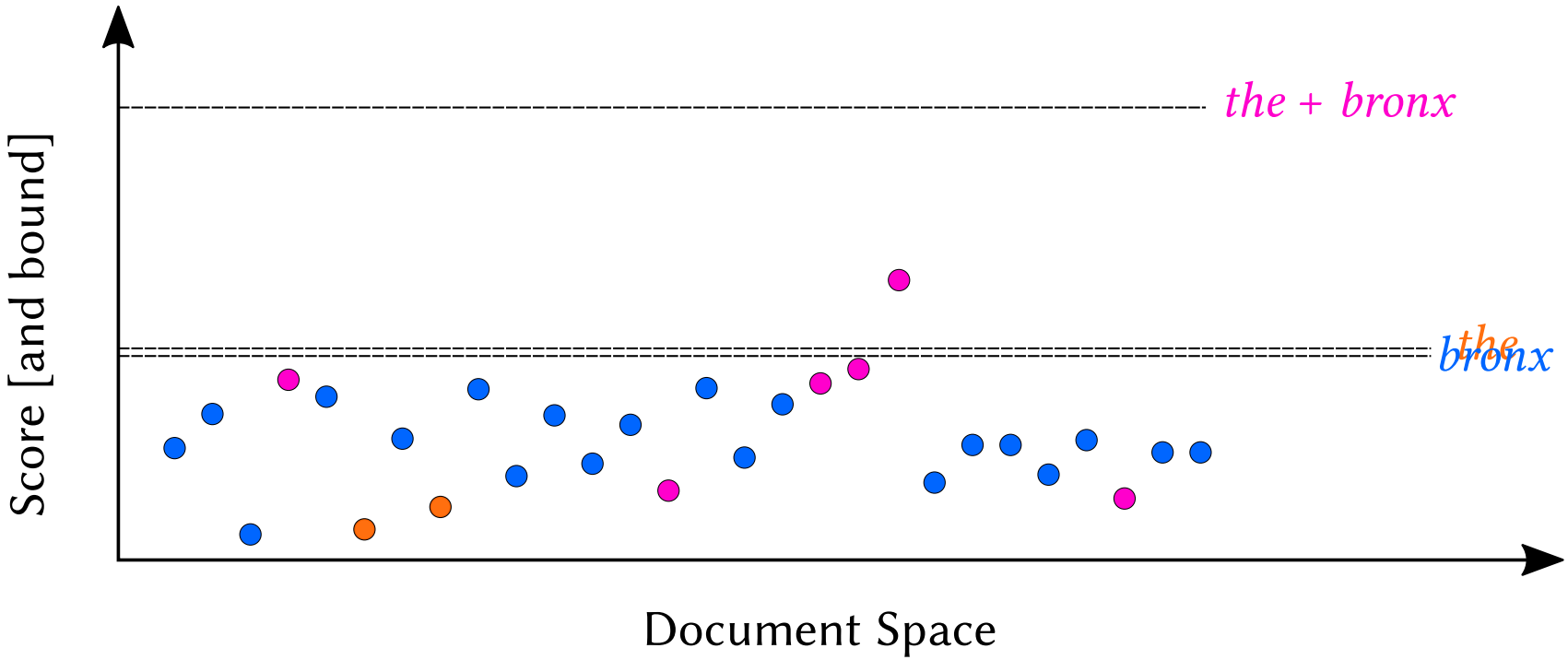

Score [and bound]
the + bronx
θ
bronx
the
Document Space


Score [and bound]
the + bronx$_\theta$
bronx
the
Document Space

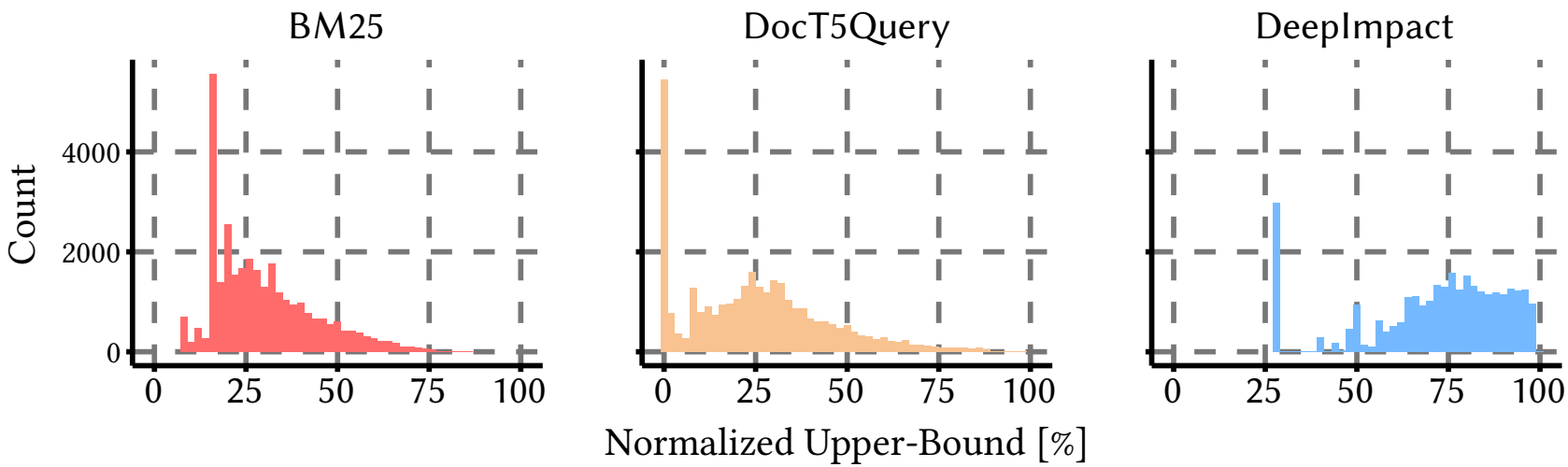

Score [and bound]
the + bronx$_\theta$
bronx
the
Document Space

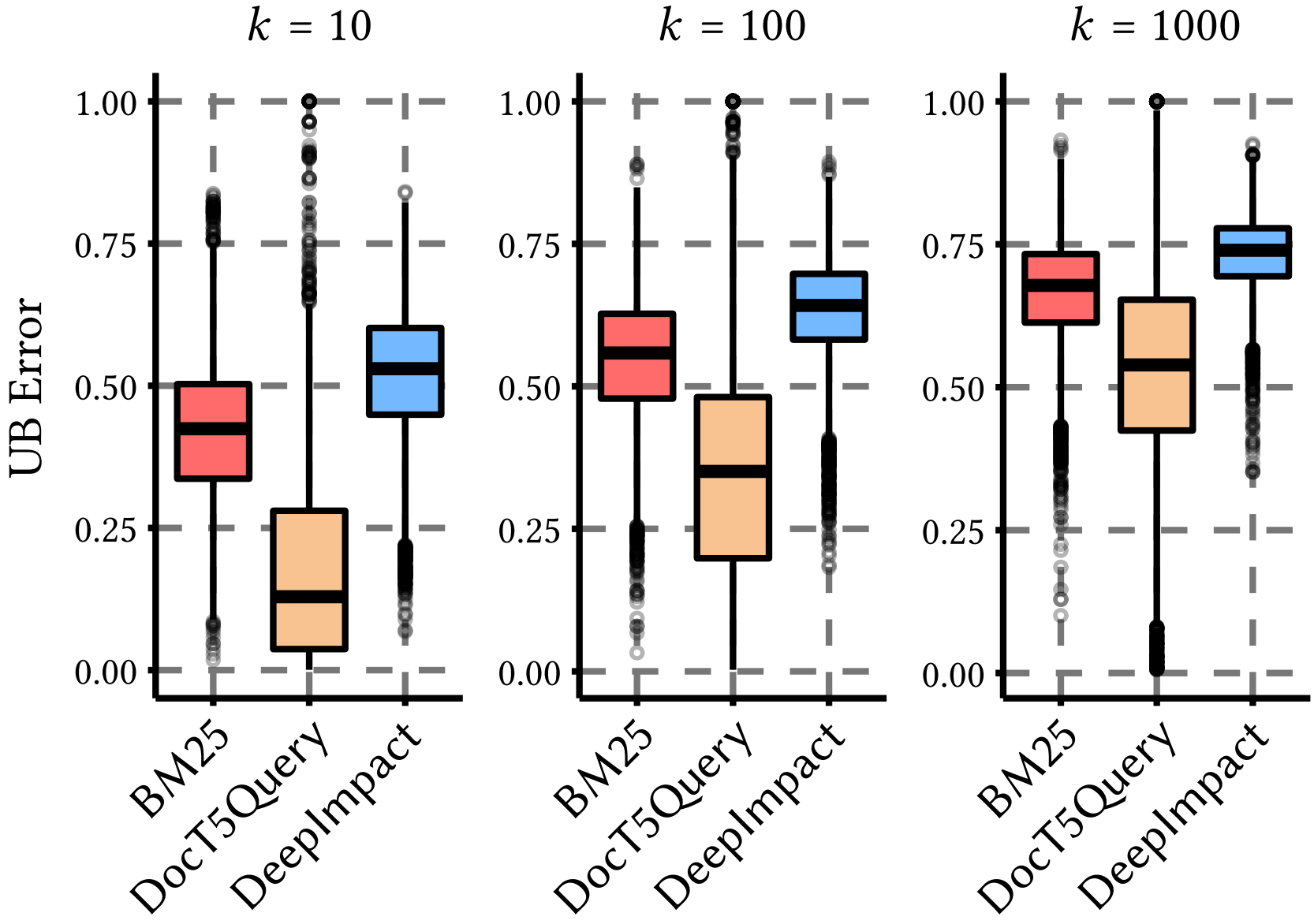

Score [and bound]
the + bronx$_\theta$
bronx
the
Document Space

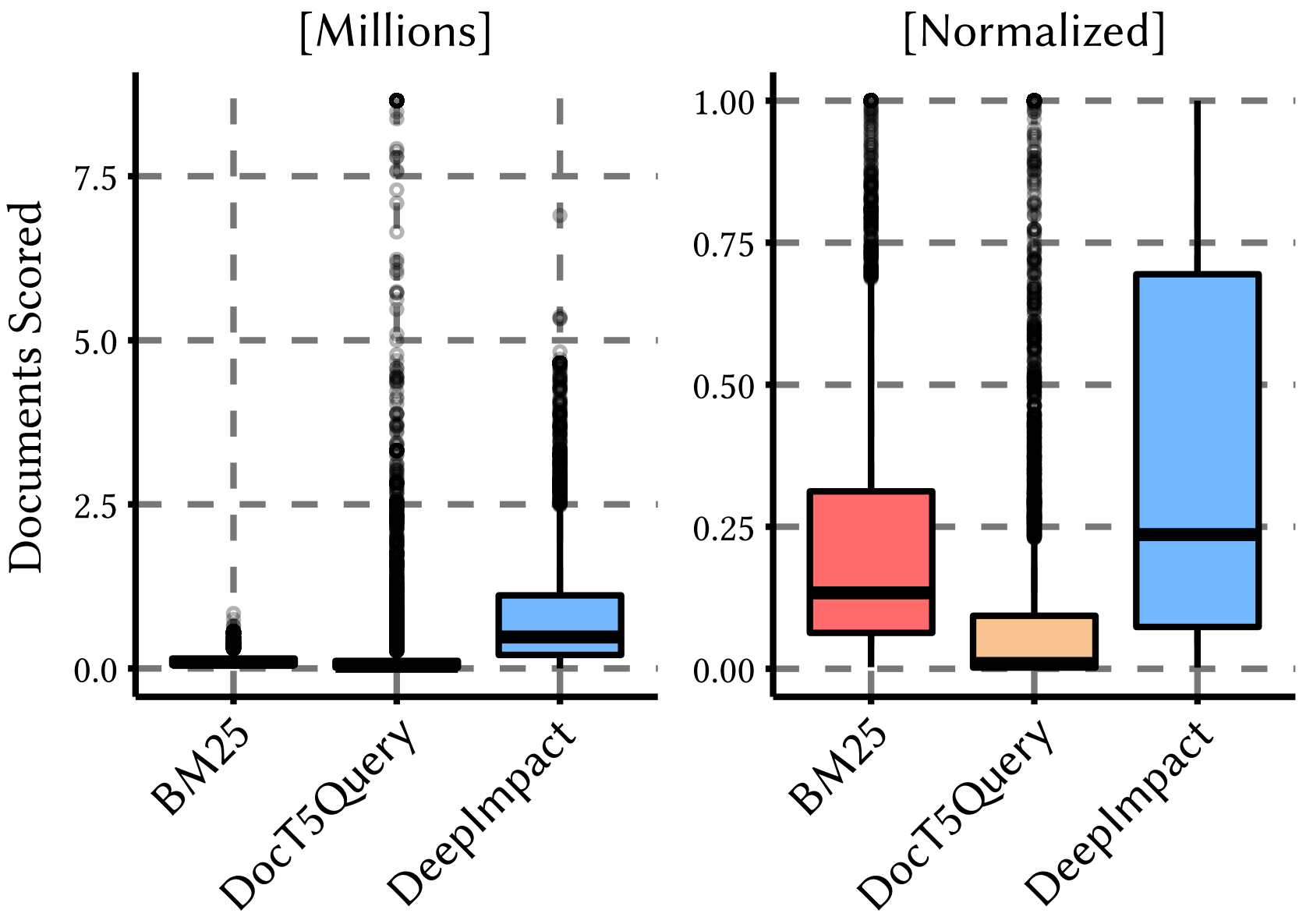
But what if the distributions change?


Score [and bound]
the + bronx
the
bronx
Document Space

Learned Models violate the embedded assumption of IDF: The more rare a term, the more highly it should be weighted with respect to the other terms.

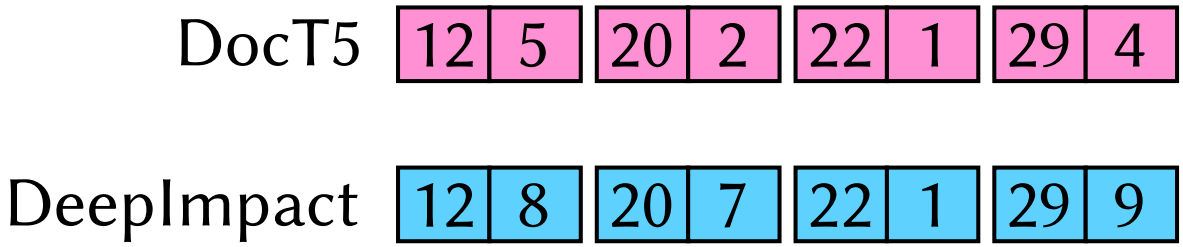

BM25
DocT5Query
DeepImpact
Count
4000
2000
0
0
25
50
75
100
Normalized Upper-Bound [%]

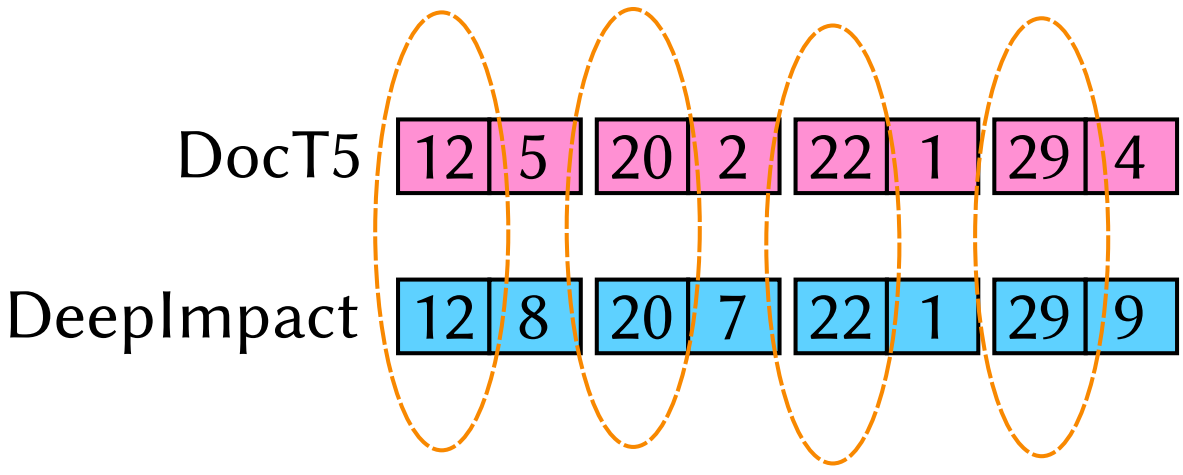

k = 10
k = 100
k = 1000
UB Error
1.00
0.75
0.50
0.25
0.00
BM25
DocT5Query
DeepImpact

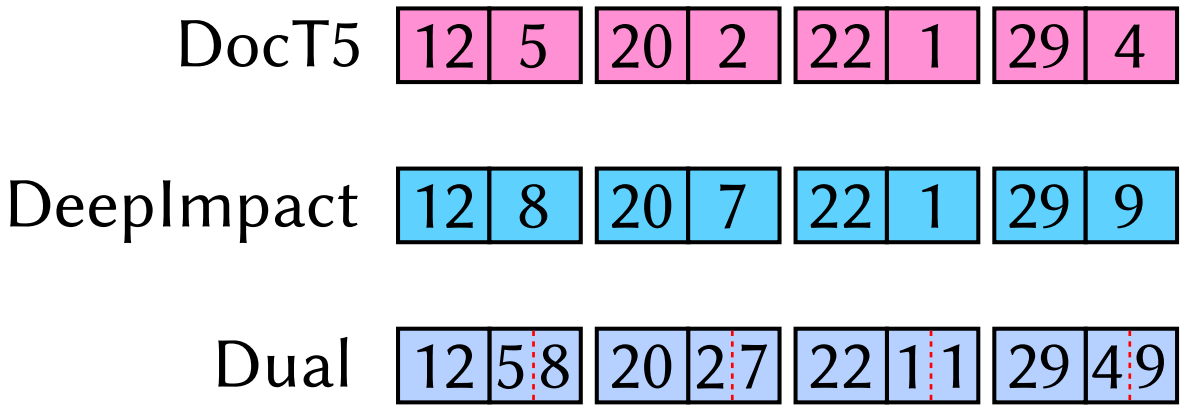

[Millions]
[Normalized]
Documents Scored
7.5
5.0
2.5
0.0
1.00
0.75
0.50
0.25
0.00
BM25
DocT5Query
DeepImpact
BM25
DocT5Query
DeepImpact

Our Heuristic

Observation: DeepImpact is built over a DocT5Query Index... Postings lists are (almost) one-to-one.

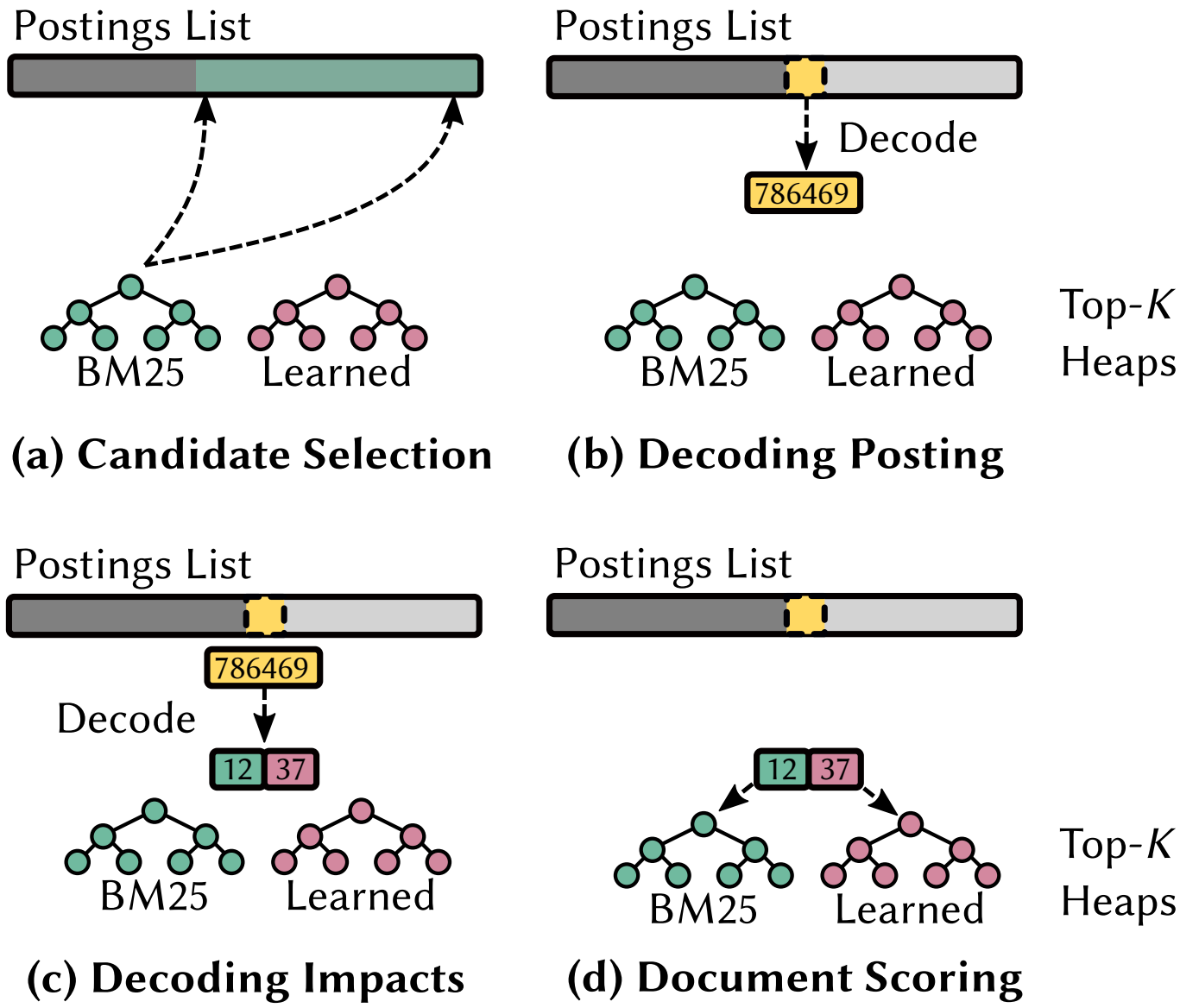

DocT5
12 5 20 2 22 1 29 4
DeepImpact
12 8 20 7 22 1 29 9


DocT5
12
5
20
2
22
1
29
4
DeepImpact
12
8
20
7
22
1
29
9


DocT5
12 5 20 2 22 1 29 4
DeepImpact
12 8 20 7 22 1 29 9
Dual
12 5 8 20 2 7 22 1 1 29 4 9

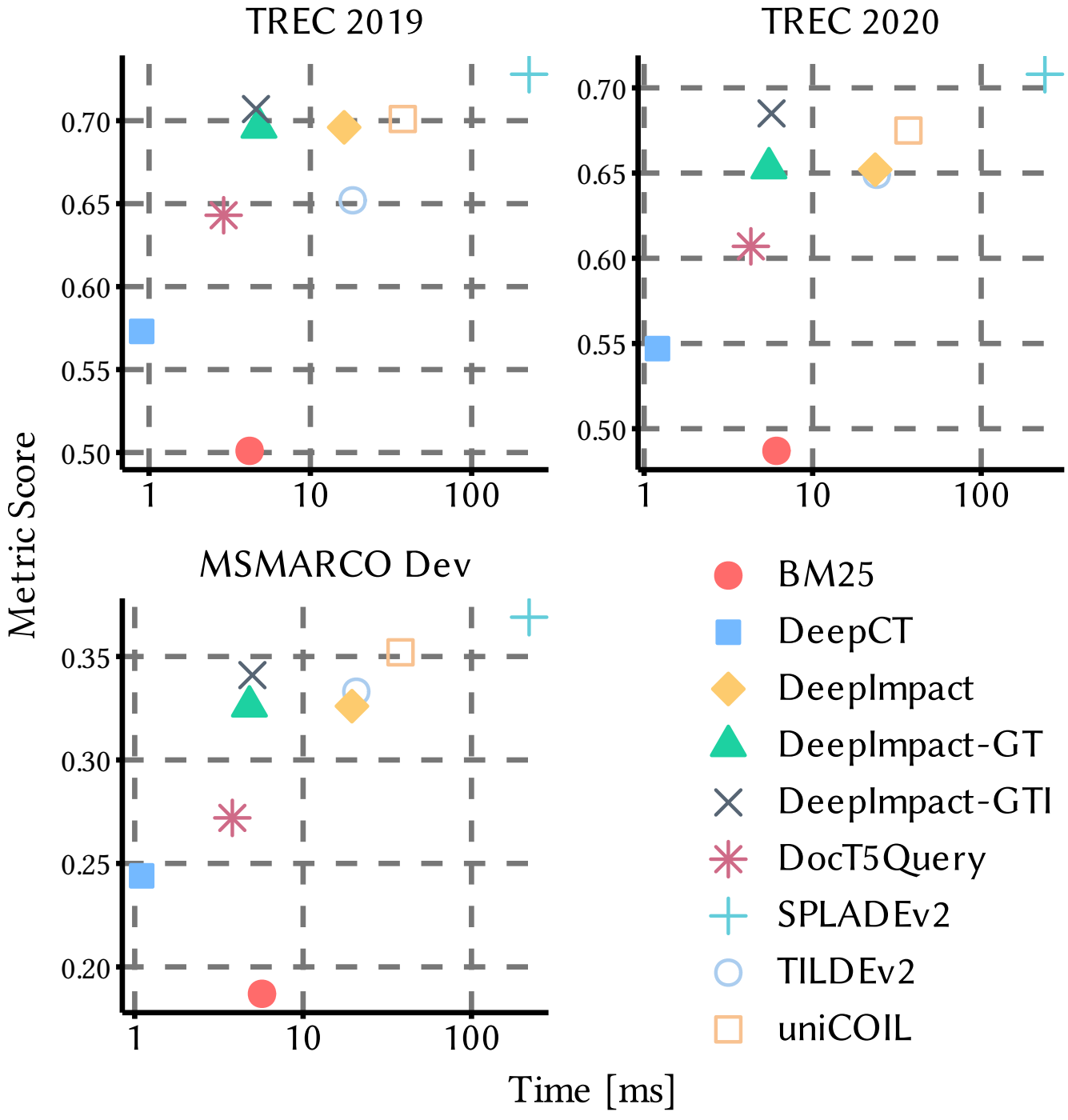
Hypothesis: BM25 visits the right documents; it just doesn't rank them well.

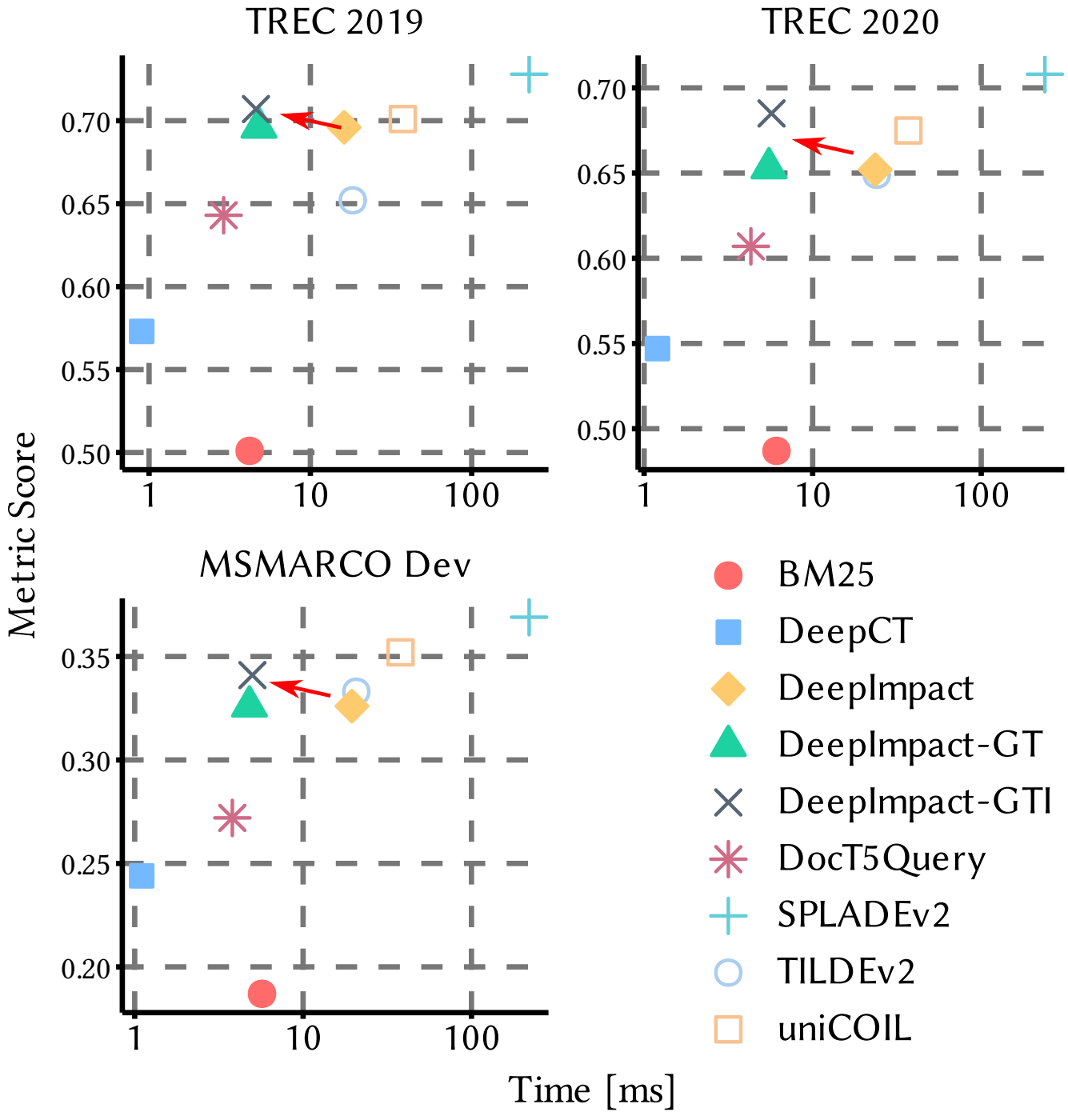
Guided Traversal (GT)



Interpolation is easy/fast (we call this GTI).

In our experiments, we just do an unweighted linear interpolation.


TREC 2019
TREC 2020
MSMARCO Dev
Metric Score
Time [ms]
0.70
0.65
0.60
0.55
0.50
0.35
0.30
0.25
0.20
1
10
100
BM25
DeepCT
DeepImpact
DeepImpact-GT
DeepImpact-GTI
DocT5Query
SPLADEv2
TILDEv2
uniCOIL


TREC 2019
TREC 2020
MSMARCO Dev
Metric Score
Time [ms]
0.70
0.65
0.60
0.55
0.50
0.35
0.30
0.25
0.20
1
10
100
BM25
DeepCT
DeepImpact
DeepImpact-GT
DeepImpact-GTI
DocT5Query
SPLADEv2
TILDEv2
uniCOIL

Limitations and Future Work

Packing method is naïve and difficult to compress.

Specialized solution; works only when postings align.

Only works in contexts where BM25 (over an expanded index) *visits* the "good" documents.

Concluding Notes: Benchmarking is hard!

Improvements That Don't Add Up:
Ad-Hoc Retrieval Results Since 1998

Timothy G. Armstrong, Alistair Moffat, William Webber, Justin Zobel

Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia

{tgar,alistair,wew,jz}@csse.unimelb.edu.au

CIKM 2009: *"How confident are we that a technique that yields an improvement over a weak baseline would also give an improvement over a strong one?"*

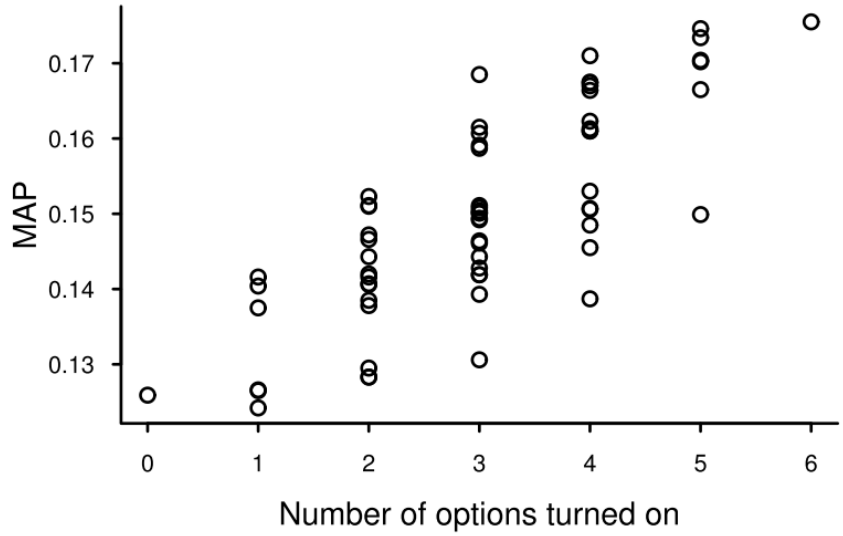


Figure 5: MAP as a function of number of options turned on, for Indri running against the TREC-5 Ad-Hoc test collection.

"There are instances where an option creates a significant improvement when added to certain configurations, but has no overall effect when added to others."

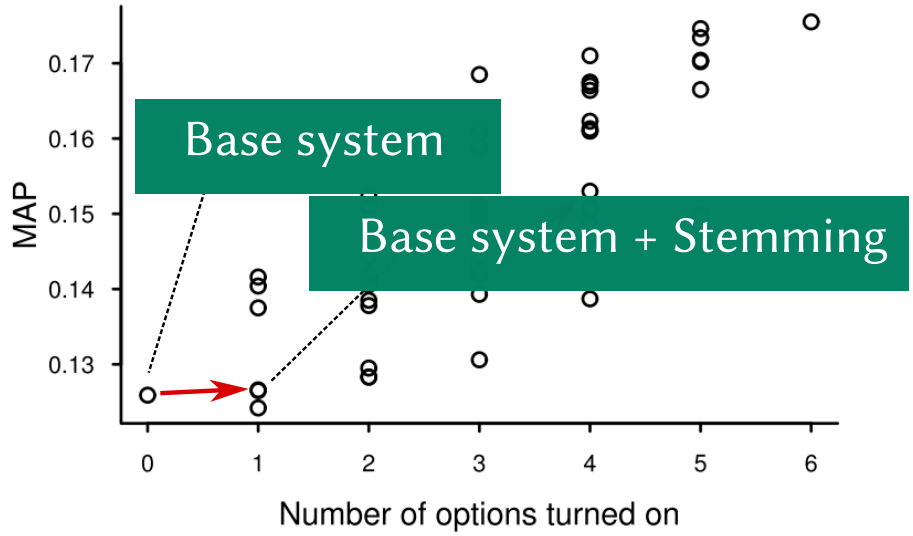


Figure 5: MAP as a function of number of options turned on, for Indri running against the TREC-5 Ad-Hoc test collection.

"There are instances where an option creates a significant improvement when added to certain configurations, but has no overall effect when added to others."

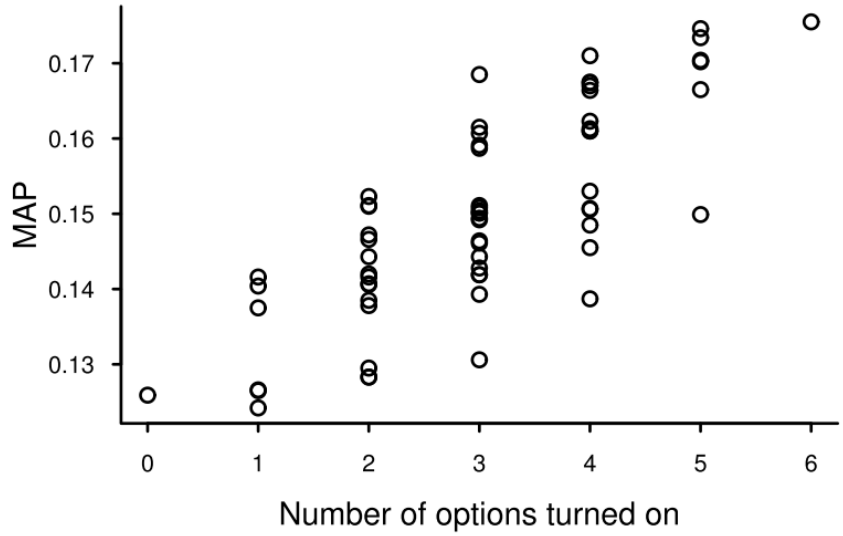


Figure 5: MAP as a function of number of options turned on, for Indri running against the TREC-5 Ad-Hoc test collection.

"There are instances where an option creates a significant improvement when added to certain configurations, but has no overall effect when added to others."

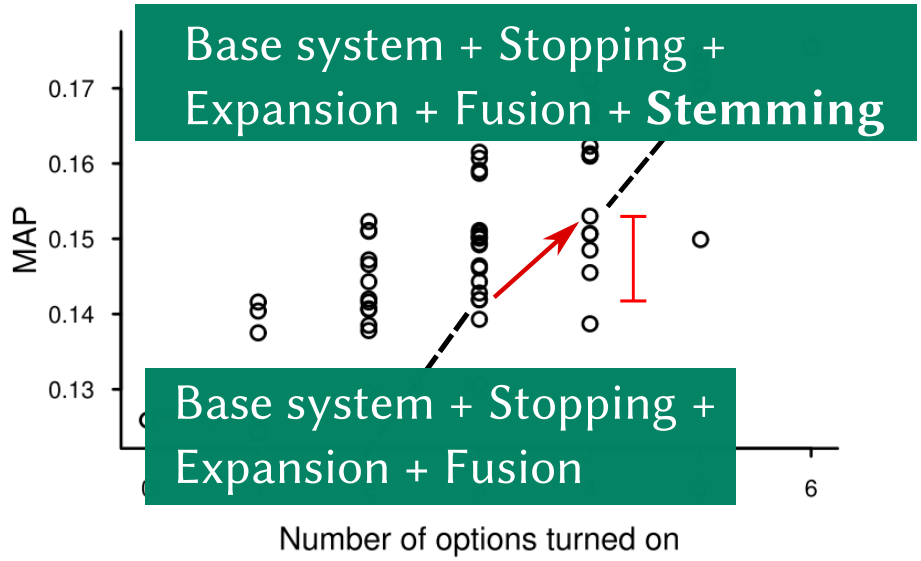
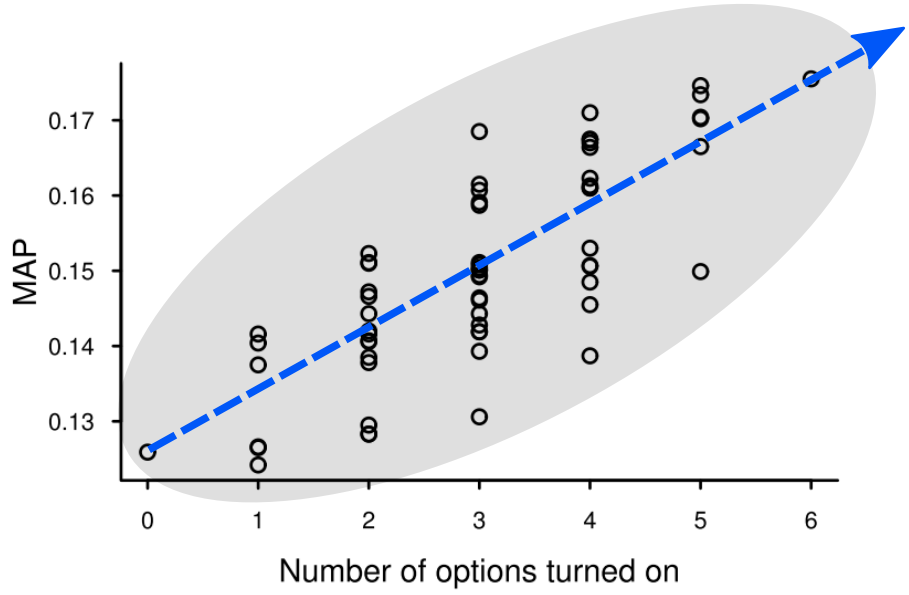


Figure 5: MAP as a function of number of options turned on, for Indri running against the TREC-5 Ad-Hoc test collection.

"There are instances where an option creates a significant improvement when added to certain configurations, but has no overall effect when added to others."



Improvements are broadly additive

"There are instances where an option creates a significant improvement when added to certain configurations, but has no overall effect when added to others."

Examining Additivity and Weak Baselines

SADEGH KHARAZMI, RMIT University & NICTA
FALK SCHOLER, RMIT University
DAVID VALLET, Google
MARK SANDERSON, RMIT University

On the Additivity and Weak Baselines for Search Result Diversification Research

Mehmet Akcay
Middle East Technical University & ASELSAN
Ankara, Turkey
meakcay@aselsan.com.tr

Ismail Sengor Altingovde
Middle East Technical University
Ankara, Turkey
altingovde@ceng.metu.edu.tr

Craig Macdonald
University of Glasgow
Glasgow, Scotland, UK
craig.macdonald@glasgow.ac.uk

Iadh Ounis
University of Glasgow
Glasgow, Scotland, UK
iadh.ounis@glasgow.ac.uk

Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models

Wei Yang,[1] Kuang Lu,[2] Peilin Yang, and Jimmy Lin[1]
[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] Department of Electrical and Computer Engineering, University of Delaware

Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches

Maurizio Ferrari Dacrema
Politecnico di Milano, Italy
maurizio.ferrari@polimi.it

Paolo Cremonesi
Politecnico di Milano, Italy
paolo.cremonesi@polimi.it

Dietmar Jannach
University of Klagenfurt, Austria
dietmar.jannach@aau.at

Examining the Additivity of Top-k Query Processing Innovations

Joel Mackenzie
The University of Melbourne
Melbourne, Australia

Alistair Moffat
The University of Melbourne
Melbourne, Australia

This also applies to efficiency studies!

Benchmarking is difficult...

CPU vs GPU? Which CPU and which GPU??

Optimized codebase? Research code? How much do we know (and need to know) about the underlying codebase?

Context of measurement; Scale of collections, hardware, etc... tension between latency, throughput, space occupancy, scalability, computing power and cost $\rightarrow$ CO2