Vocabulary Size and Its Effect on Topic Representation for Informetric and Information Retrieval Data Processing

KUN LU, OKLAHOMA U.

XIN CAI, UW-MILWAUKEE

ISOLA AJIFERUKE, WESTERN U.

DIETMAR WOLFRAM, UW-MILWAUKEE

The Problem

Data processing for large Information Retrieval (IR) & informetric data sets has become computationally intensive

Traditional methods based on vector space and probabilistic models rely on high-dimensional spaces for comparison of entities of interest

Methods to reduce the computational overhead are needed

What is Topic Modeling?

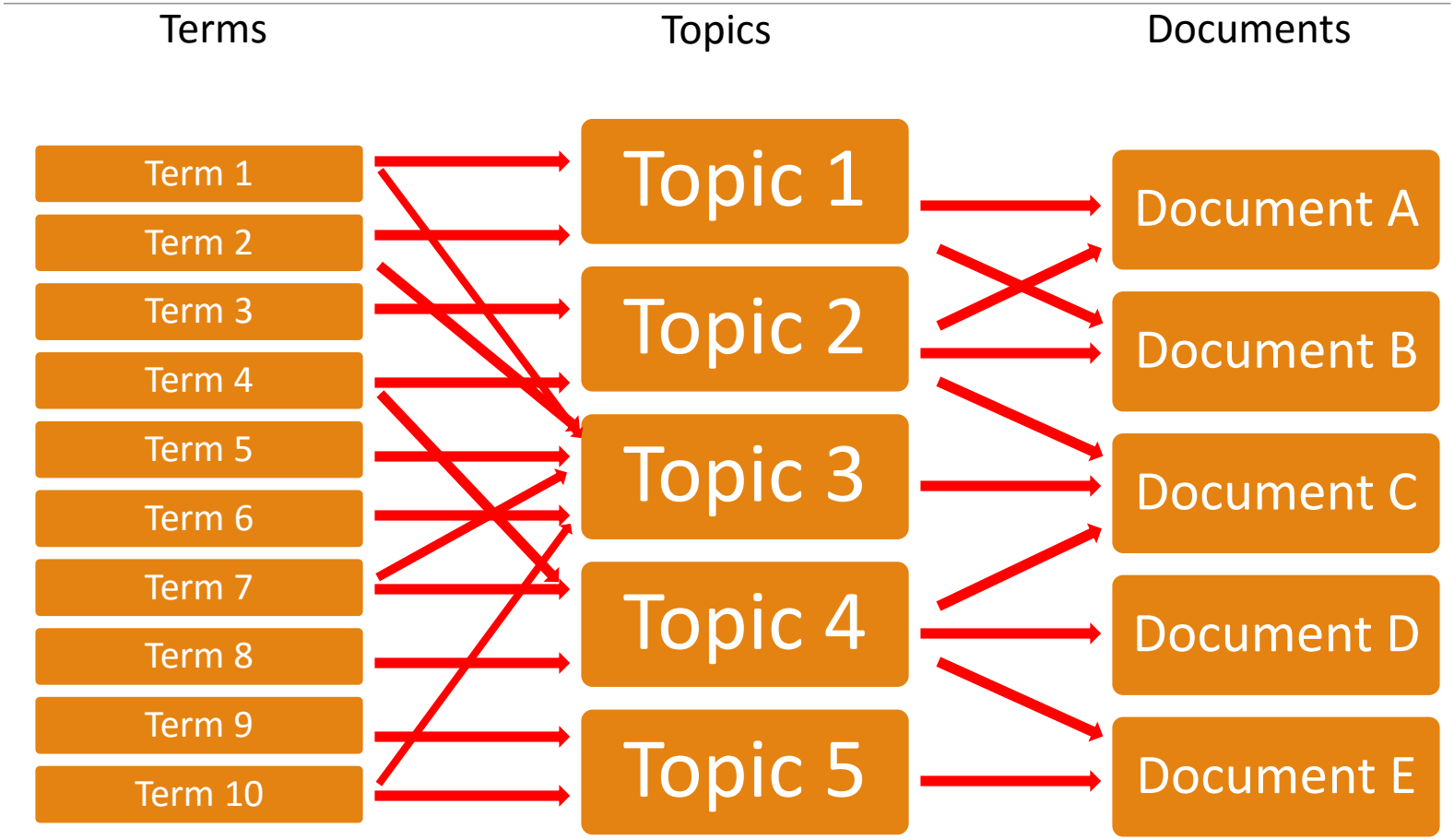
Provides a way to reduce the “dimensionality” of the computing process

Topic modeling reduces many terms down to a manageable number of topics for easier processing

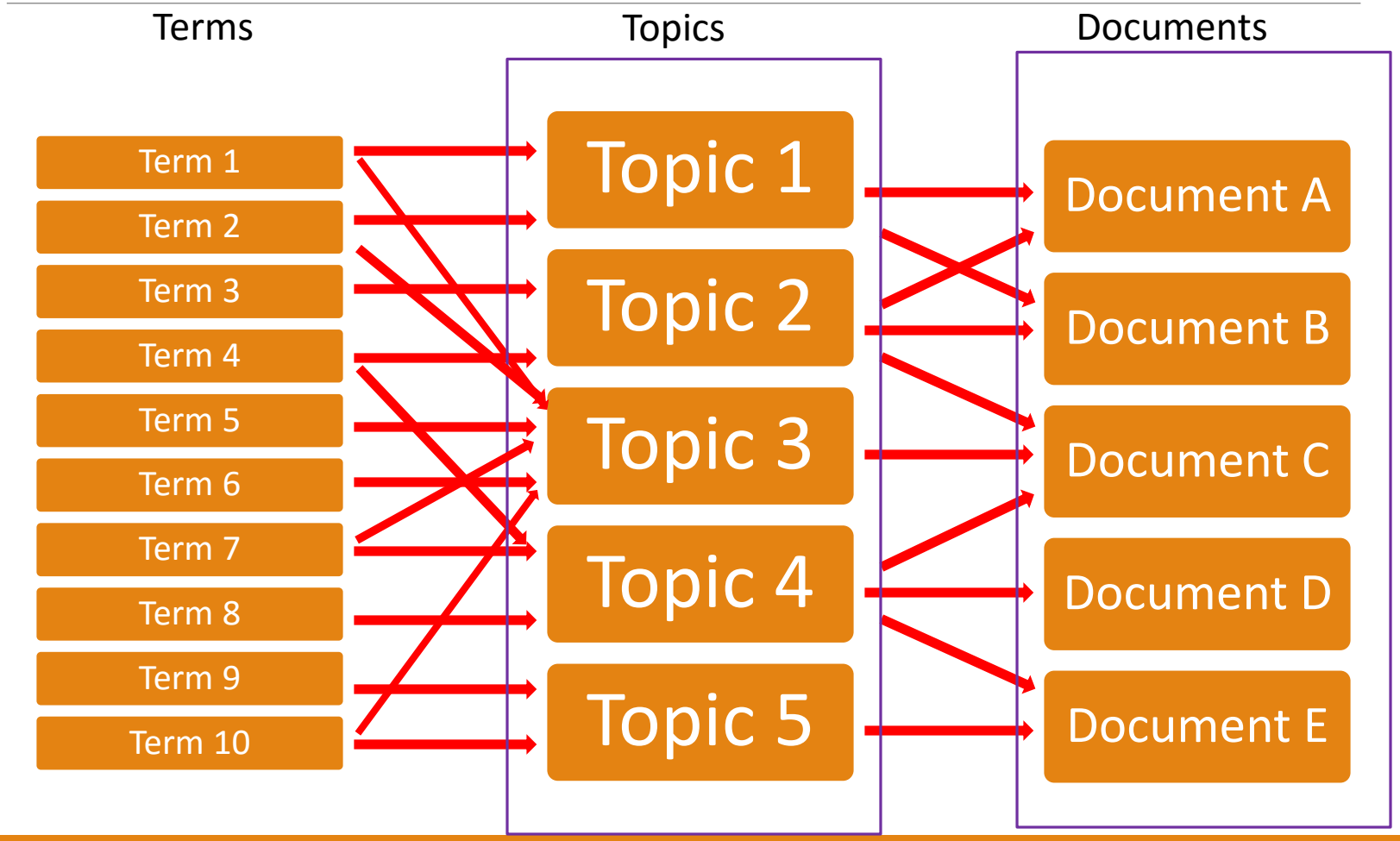
Entities are then compared based on topics and not terms

Useful for metric studies on authors, journals that don’t require citations

Topic Modeling



Research Design



Research Questions

1. What impact does the removal of frequently or infrequently occurring terms for topic training have on the ability to discriminate documents in a text corpus based on the document space density?

2. 2) How does the removal of frequently or infrequently occurring terms affect topic distributions in documents and the distinctiveness of the trained topics using entropy and pairwise topic similarity?

Method

Three datasets

Ohsumed – 34,846 Medline bibliographic records

TREC Genomics Track 2006 – 16,201 full-text articles published in biomedical journals

Elsevier SIGMET - 56,620 bibliographic records from 118 Elsevier Arts & Humanities journals

Method

Stemming & stopword removal applied to raw textual corpora

Latent Dirichlet Allocation (LDA) model implemented for 10, 20, 30, 40, 50 & 100 topics

Topic & document space outcomes assessed for

- Full vocabulary
- Removal of singly occurring terms (infrequent types)
- Removal of top 0.5%, 1% and 5% of most frequently occurring terms (many tokens)

Method

Several measures used to compare resulting corpus characteristics based on topics

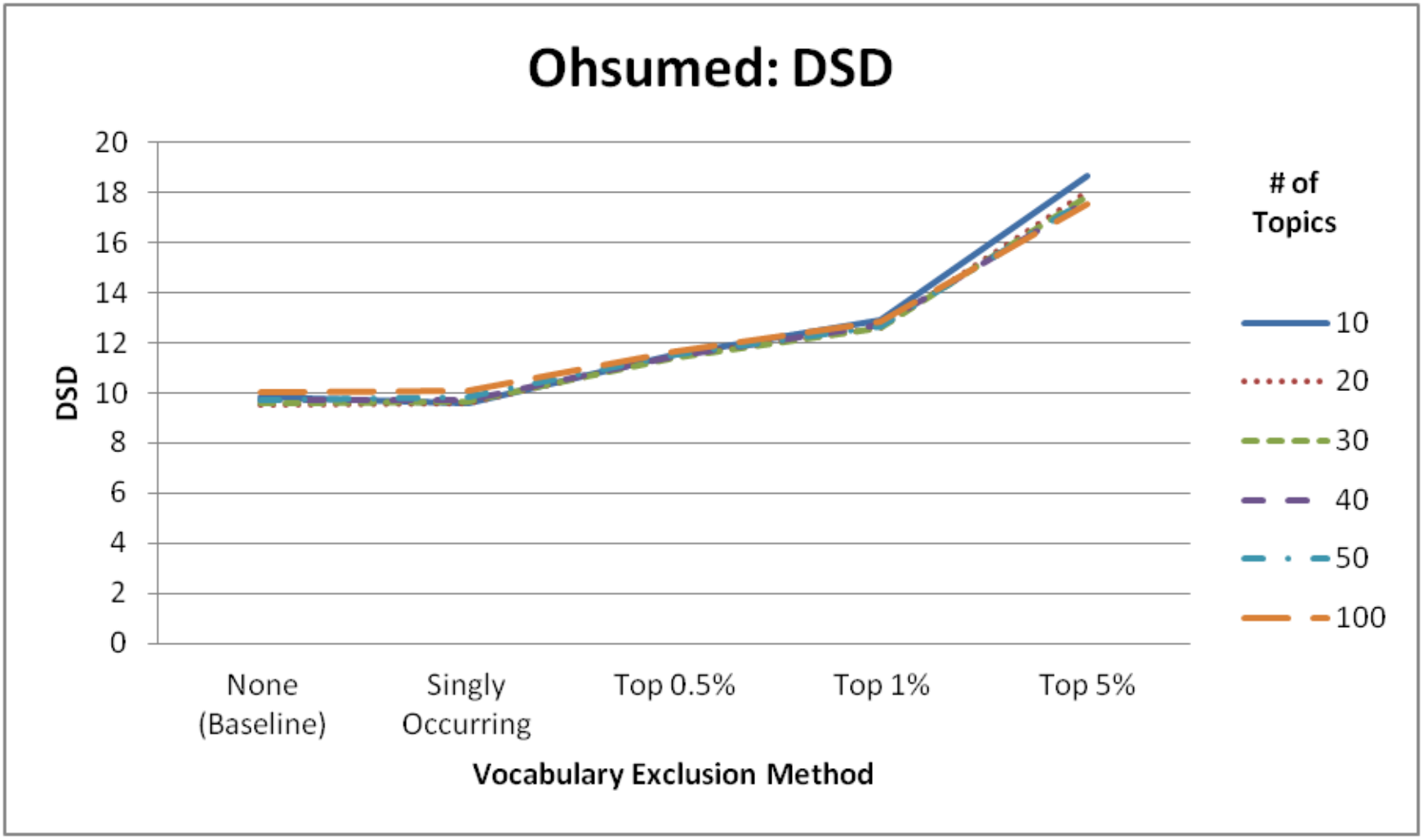
- **Document Space Density (DSD)** – degree of scatter of documents based on topics
- **Information Entropy (IE)** –
  low IE = high concentration on few topics;
  high IE = uniform distribution
- **Pairwise Topic Similarity (PTS)** –
  lower PTS = more distinctive topics

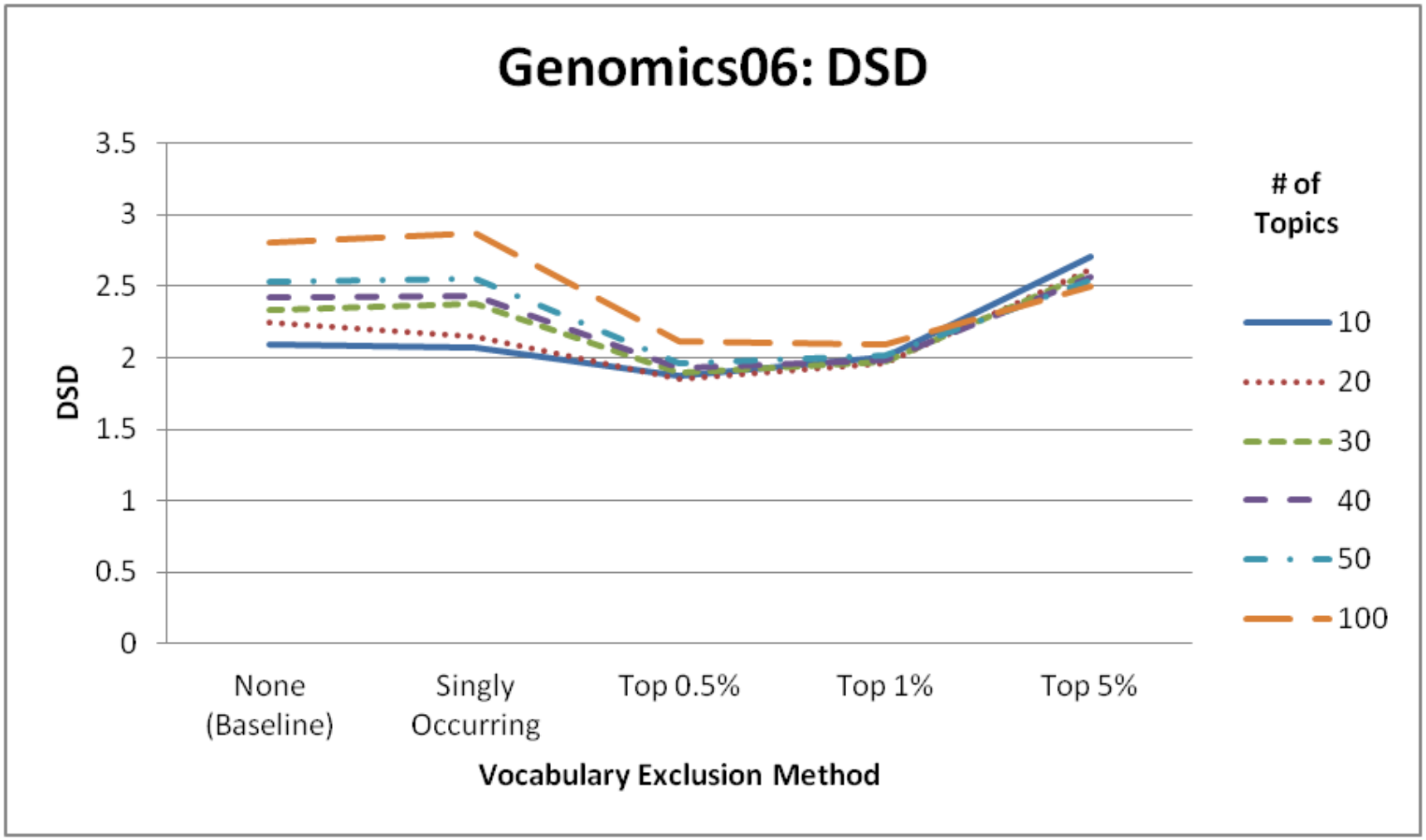
Results

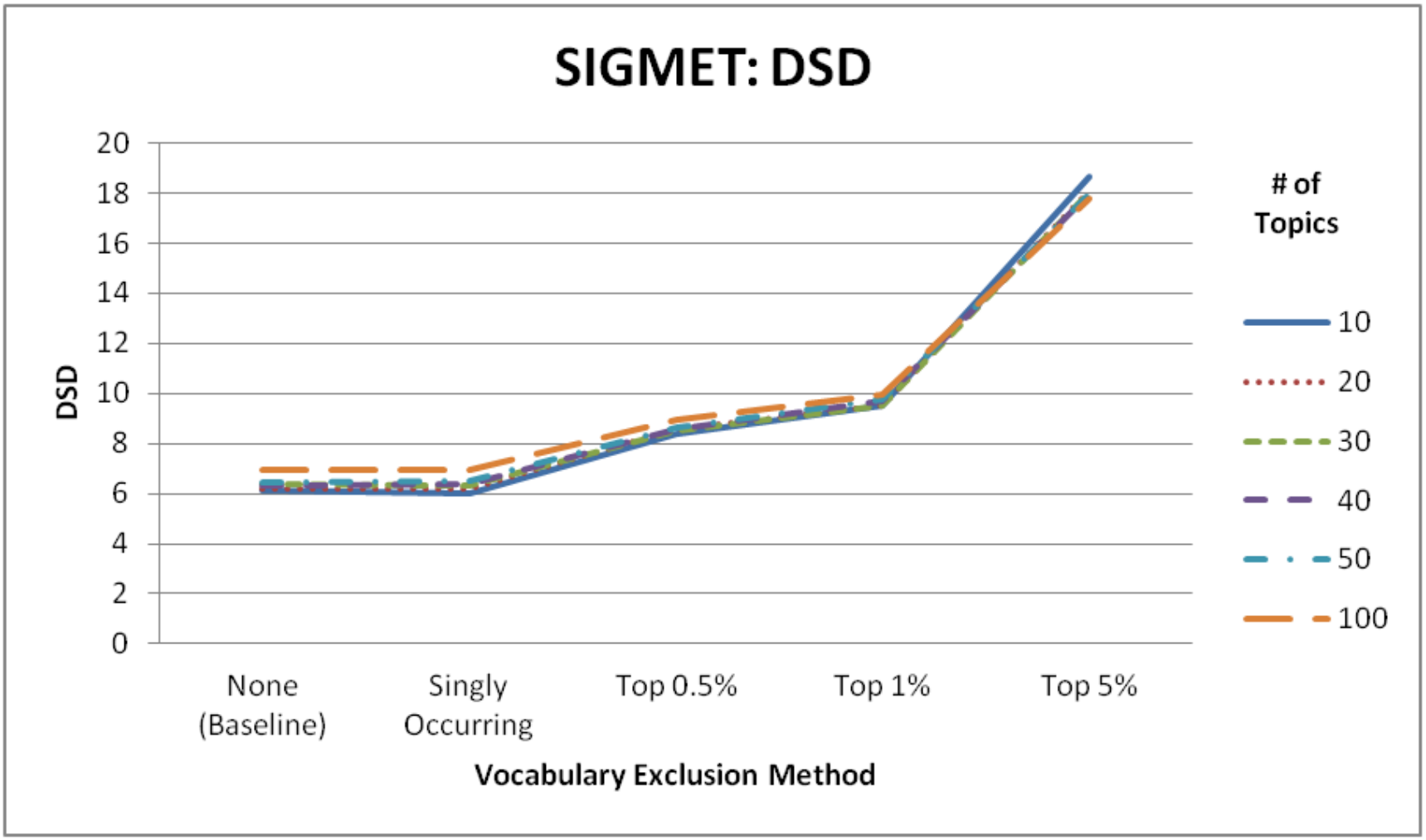
Outcomes reported in graphic format & compared visually

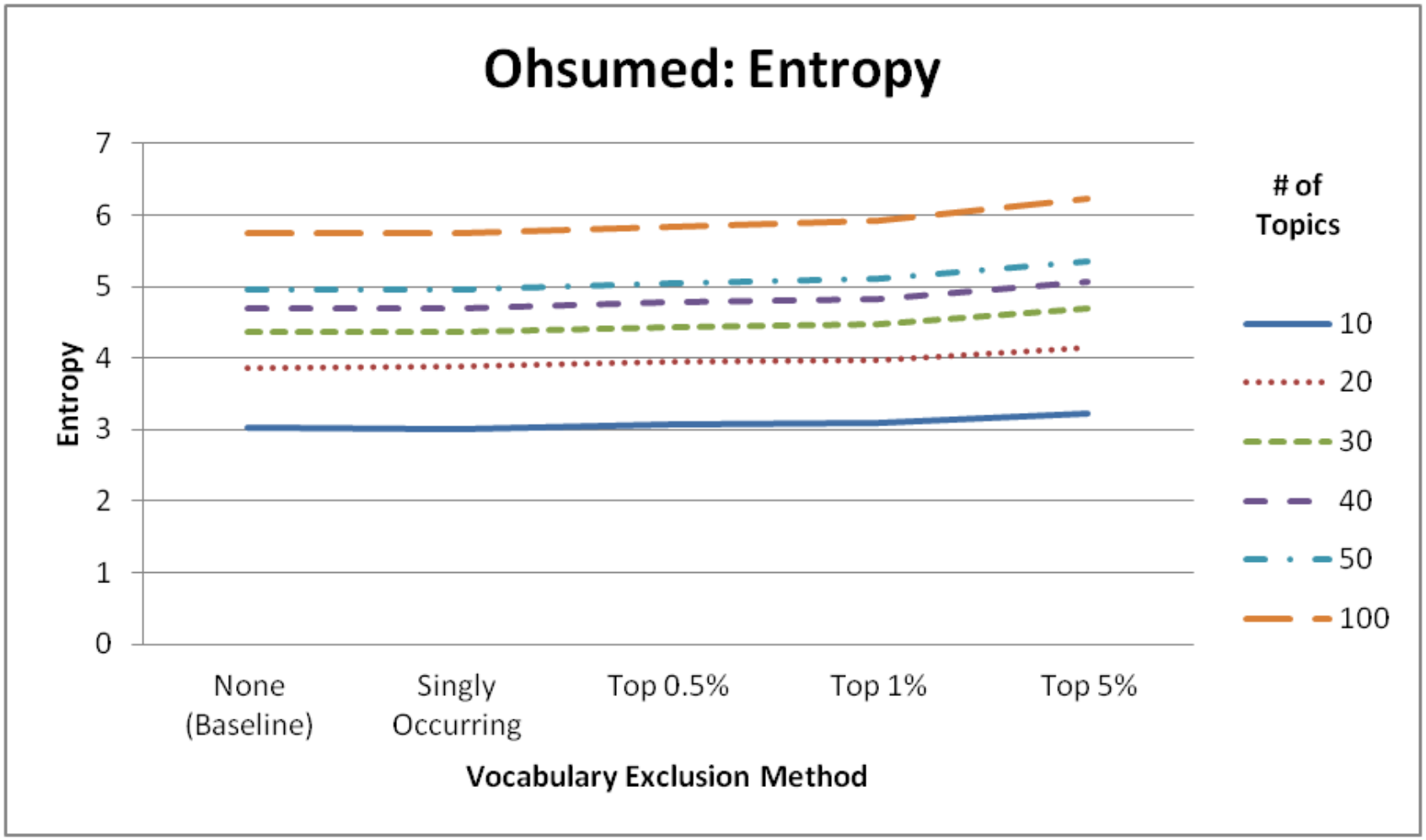
Three comparative measures were compared for each of:

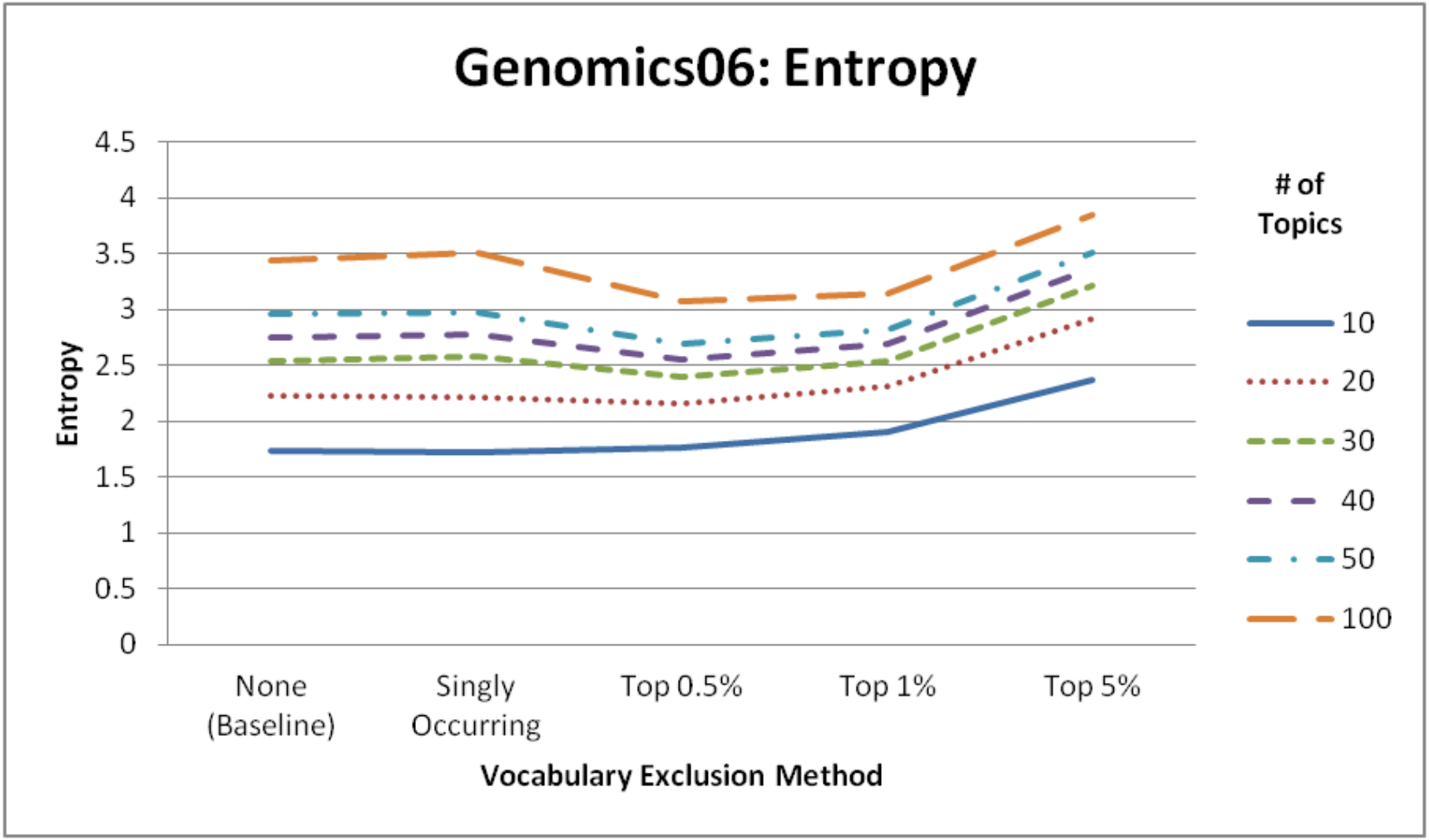
Three datasets X
Six topic totals X
Five vocabulary exclusion methods

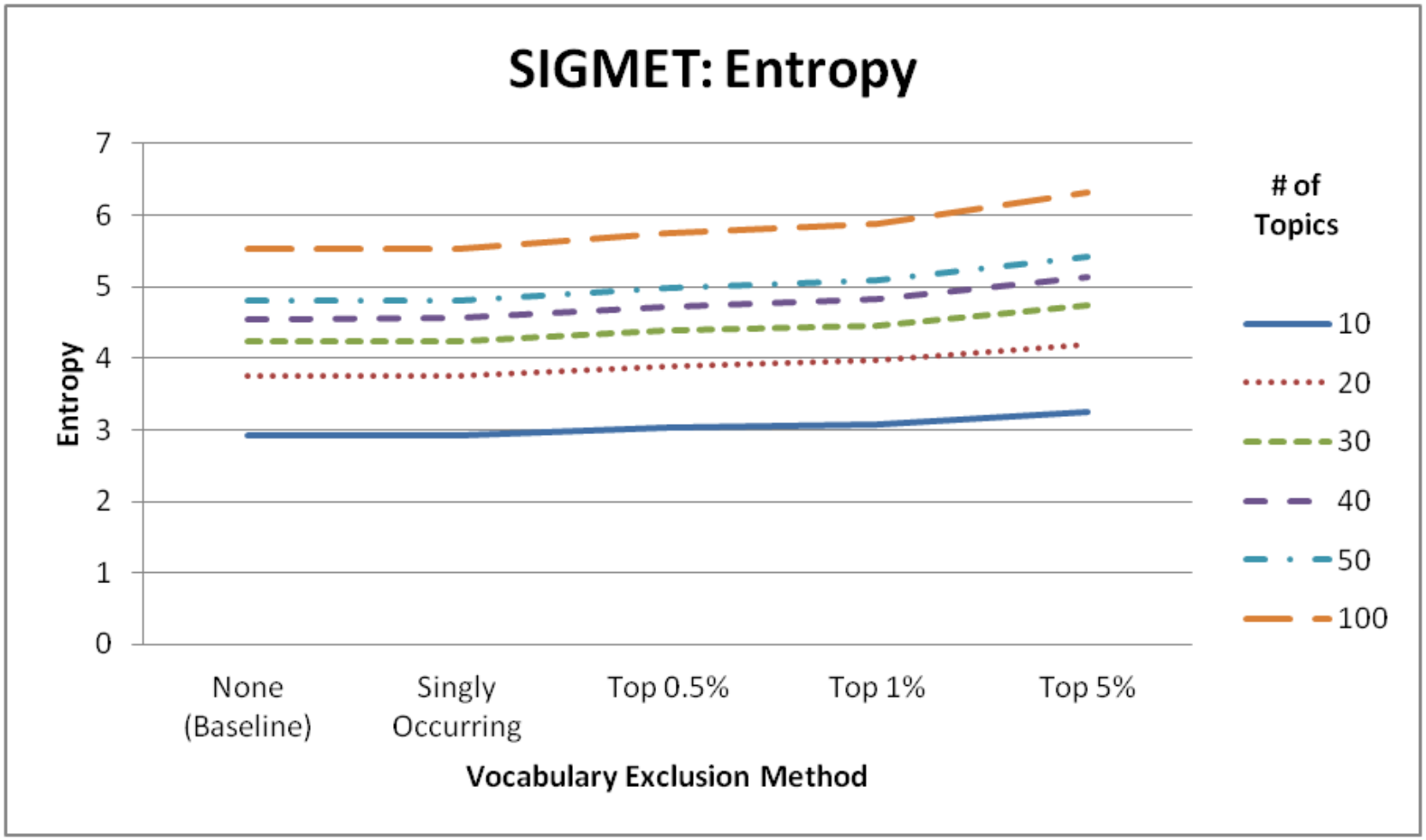

Ohsumed: DSD
DSD
20
18
16
14
12
10
8
6
4
2
0
None (Baseline)
Singly Occurring
Top 0.5%
Top 1%
Top 5%
Vocabulary Exclusion Method
# of Topics
10
20
30
40
50
100

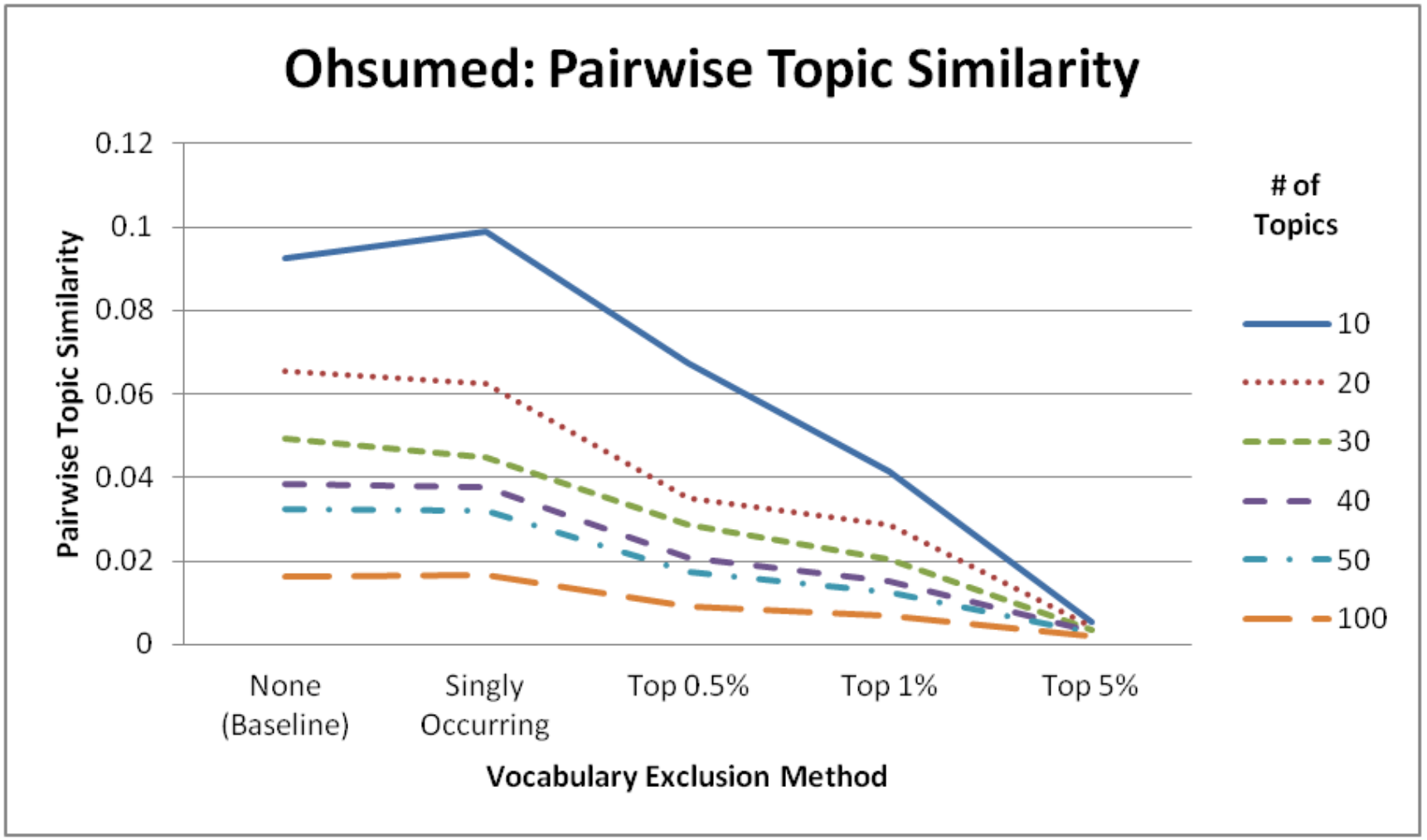

Genomics06: DSD
DSD
3.5
3
2.5
2
1.5
1
0.5
0
None (Baseline)
Singly Occurring
Top 0.5%
Top 1%
Top 5%
Vocabulary Exclusion Method
# of Topics
10
20
30
40
50
100

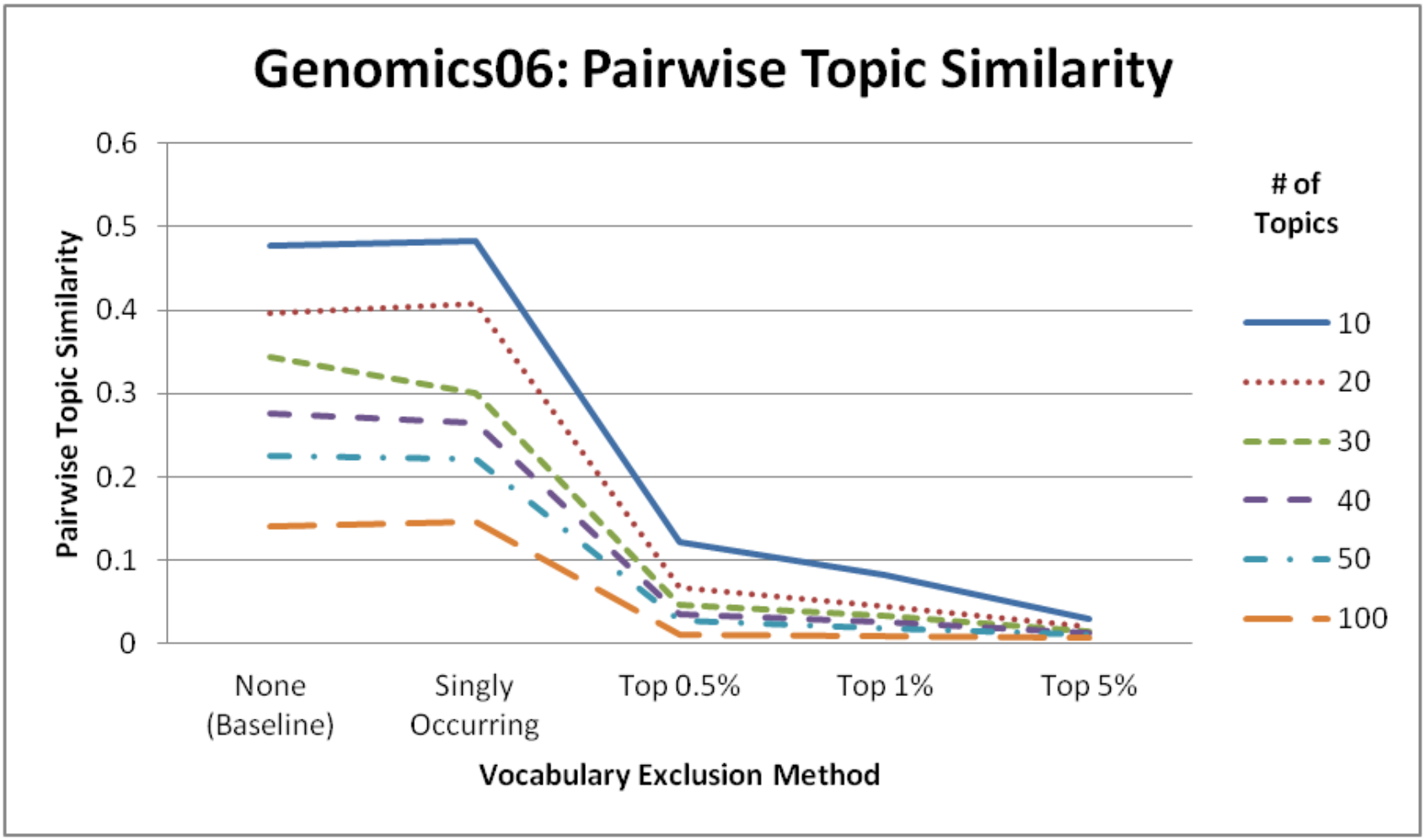

SIGMET: DSD
DSD
20
18
16
14
12
10
8
6
4
2
0
None (Baseline)
Singly Occurring
Top 0.5%
Top 1%
Top 5%
Vocabulary Exclusion Method
# of Topics
10
20
30
40
50
100

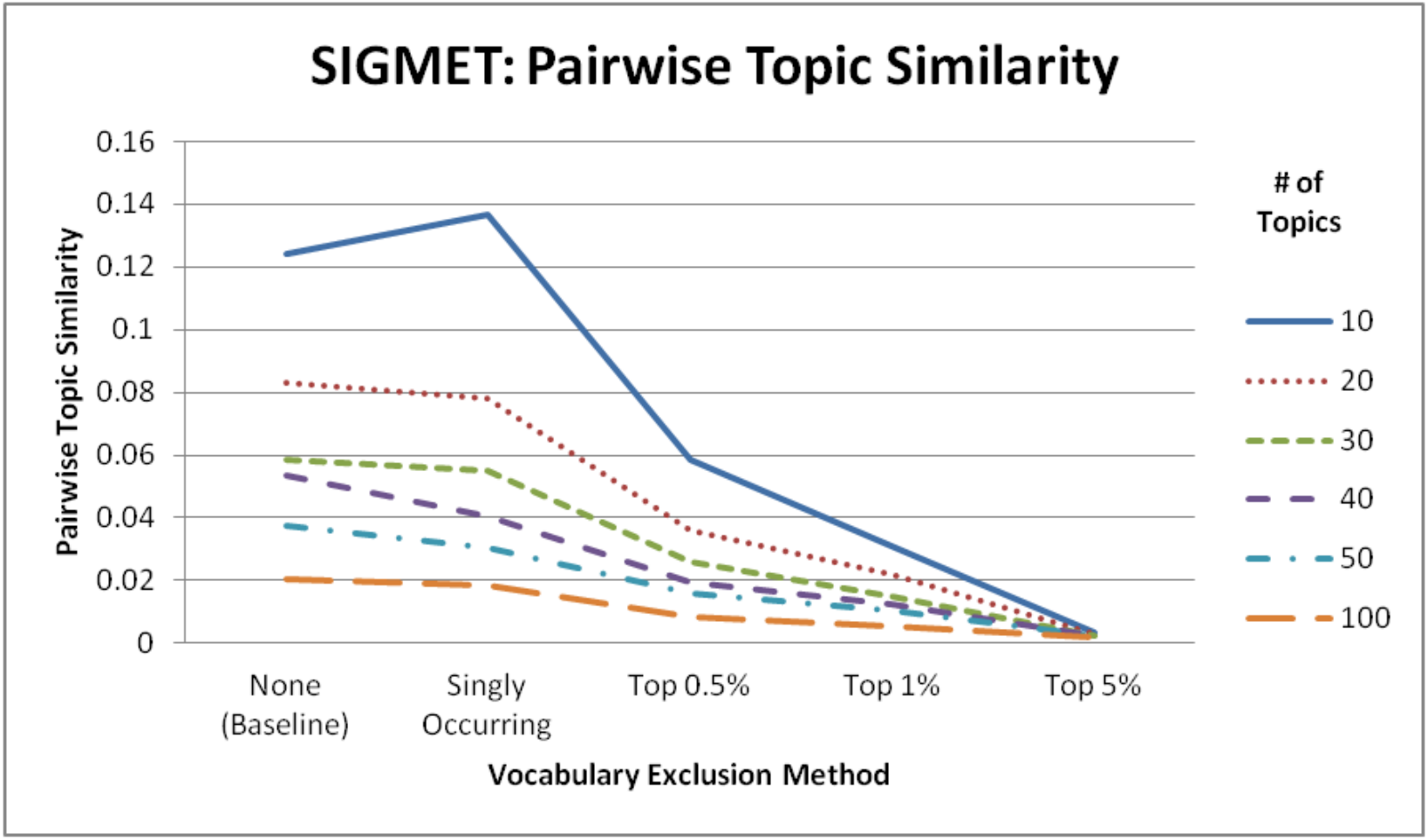

Ohsumed: Entropy
Entropy
7
6
5
4
3
2
1
0
None (Baseline)
Singly Occurring
Top 0.5%
Top 1%
Top 5%
Vocabulary Exclusion Method
# of Topics
10
20
30
40
50
100


Genomics06: Entropy
Entropy
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
None (Baseline)
Singly Occurring
Top 0.5%
Top 1%
Top 5%
Vocabulary Exclusion Method
# of Topics
10
20
30
40
50
100


SIGMET: Entropy
Entropy
7
6
5
4
3
2
1
0
None (Baseline)
Singly Occurring
Top 0.5%
Top 1%
Top 5%
Vocabulary Exclusion Method
# of Topics
10
20
30
40
50
100


Ohsumed: Pairwise Topic Similarity
Pairwise Topic Similarity
0.12
0.1
0.08
0.06
0.04
0.02
0
None (Baseline)
Singly Occurring
Top 0.5%
Top 1%
Top 5%
Vocabulary Exclusion Method
# of Topics
10
20
30
40
50
100


Genomics06: Pairwise Topic Similarity
Pairwise Topic Similarity
0.6
0.5
0.4
0.3
0.2
0.1
0
None (Baseline)
Singly Occurring
Top 0.5%
Top 1%
Top 5%
Vocabulary Exclusion Method
# of Topics
10
20
30
40
50
100


SIGMET: Pairwise Topic Similarity
Pairwise Topic Similarity
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
None (Baseline)
Singly Occurring
Top 0.5%
Top 1%
Top 5%
Vocabulary Exclusion Method
# of Topics
10
20
30
40
50
100

Summary of Findings

DSD

- Topic numbers don't affect DSD … much
- Vocabulary size does (except singly occurring terms)

IE

- Lower #'s of topics produce lower entropy
- Vocabulary size doesn't play a big role

PTS

- #'s of topics and vocabulary size play a role

Implications

Some removal of vocabulary (types and tokens) will reduce computational overhead without affecting outcomes

Metric studies relying on textual analysis can remove singly occurring terms to greatly reduce initial vocabulary size, but removal of many tokens from frequently occurring terms will affect topics

Limitations

Only looked at selected datasets

- More datasets on broader subjects are needed

Measures used indirectly indicate impact on performance, but are not definitive evidence

- Another measures we are trying is Jansen-Shannon Divergence

The content of the topics were not examined

- Content analysis on generated topics is needed

Conclusions

To reduce computational overhead in metric and IR studies involving language modeling, remove singly occurring terms, not just stopwords

Removing more than just frequent stopwords could greatly effect topic composition

Acknowledgement: Thanks to Elsevier, Inc. for providing the SIGMET dataset.