Recommendations

Philipp Maschke, Matthias Pohl

Agenda

- Infrastructure
- Algorithms
  - Clustering, Cosine Distance, Simhash
- Quality Comparisons
  - Demo
- Runtime Comparisons
- Conclusion
- Additional graphs

Infrastructure

- Considered domain:
    - Articles containing ‘german’
- Different table sizes:
    - 1.000 -122.263 articles
- Keyword retrieval
    - All links within the article
    - Used parser:
        - JWPL (University of Darmstadt)

KMeans/KMedoids

- Weighted keyword-maps as sparse dataset
- Compute clusters
  - JavaML
- Find 10 nearest articles from cluster
- **But**: Bad performance

Cosine Distance

- Computes distance between vectors
- Weighted keyword map as vector
  - Per article

SimHash

- Algorithm for finding near-duplicates
  - We are looking for “topical duplicates”
- Used hash value sizes:
  - 32 Bit
  - 512 Bit

Sample Recommendations of ABBA

List of commonly misused English words
Snippet for article #309724

Fort McPherson
Snippet for article #901292

Ken Kavanagh
Snippet for article #1353892

French ship Dupleix
Snippet for article #2486078

Metriacanthosaurus
Snippet for article #3509867

LÃiszlÃ³ Krasznahorkai
Snippet for article #3794578

Norbert Schultze
Snippet for article #8113362

Jason Rowe
Snippet for article #8633945

Margherita Zimmermann
Snippet for article #16544411

Joan Fontcuberta
Snippet for article #18382108

SimHash 32 Bit

SimHash 512 Bit

Omaha Beach
Snippet for article #60120

Mieszko III the Old
Snippet for article #378258

Swedish iron ore during World War II
Snippet for article #585230

Matthias Platzeck
Snippet for article #3057569

Money, Money, Money
Snippet for article #3590021

Pressath
Snippet for article #5960687

Sarah Natochenny
Snippet for article #6048021

Hartlepool's Maritime Experience
Snippet for article #9896017

Schulzendorf
Snippet for article #10356834

SchÃ¶nwald, Brandenburg
Snippet for article #11653762

Scandinavia
Snippet for article #26740

Mamma Mia (song)
Snippet for article #409496

Chiquitita
Snippet for article #2501546

Summer Night City
Snippet for article #2524933

Znaps Vodka
Snippet for article #3537987

Money, Money, Money
Snippet for article #3590021

AC/DC discography
Snippet for article #4354534

Hasta MaÃ±ana
Snippet for article #7063215

Archduke Karl Albrecht of Austria
Snippet for article #7707981

Nordische Gesellschaft
Snippet for article #17324174

Cosine Distance

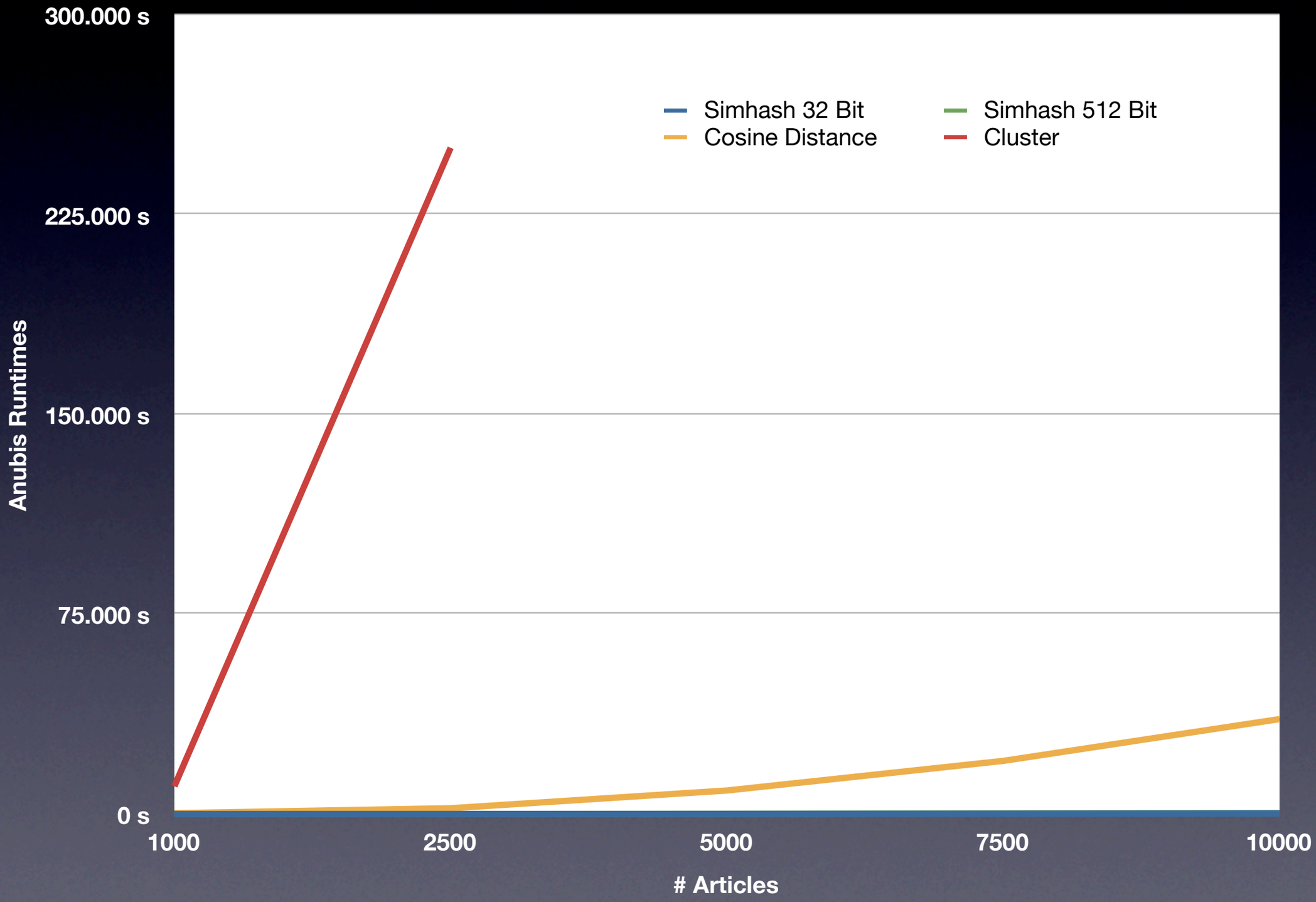
Table size:
10.000 articles

except Clustering:
2500 articles

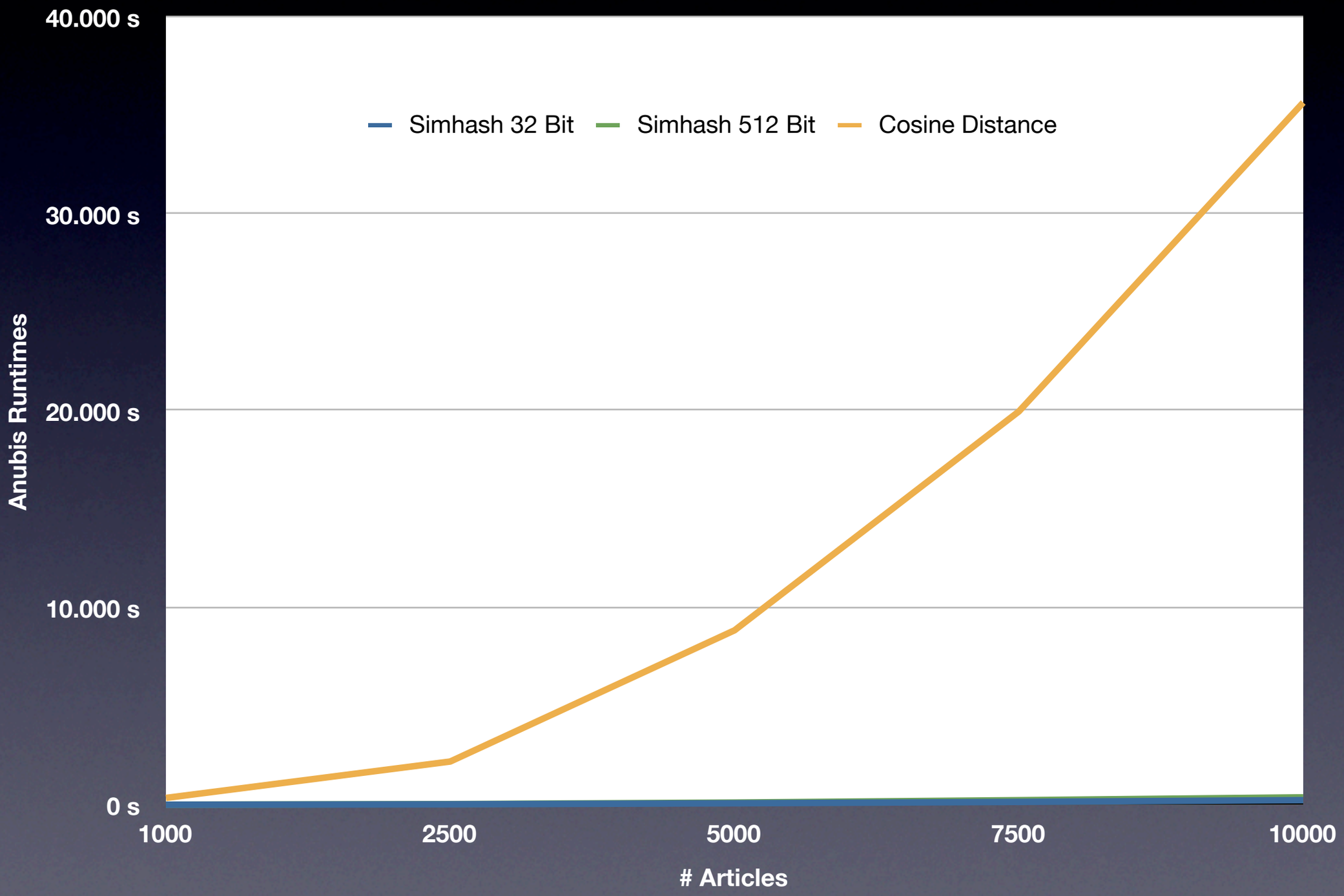
Clustering

Dancing Queen
Snippet for article #504209

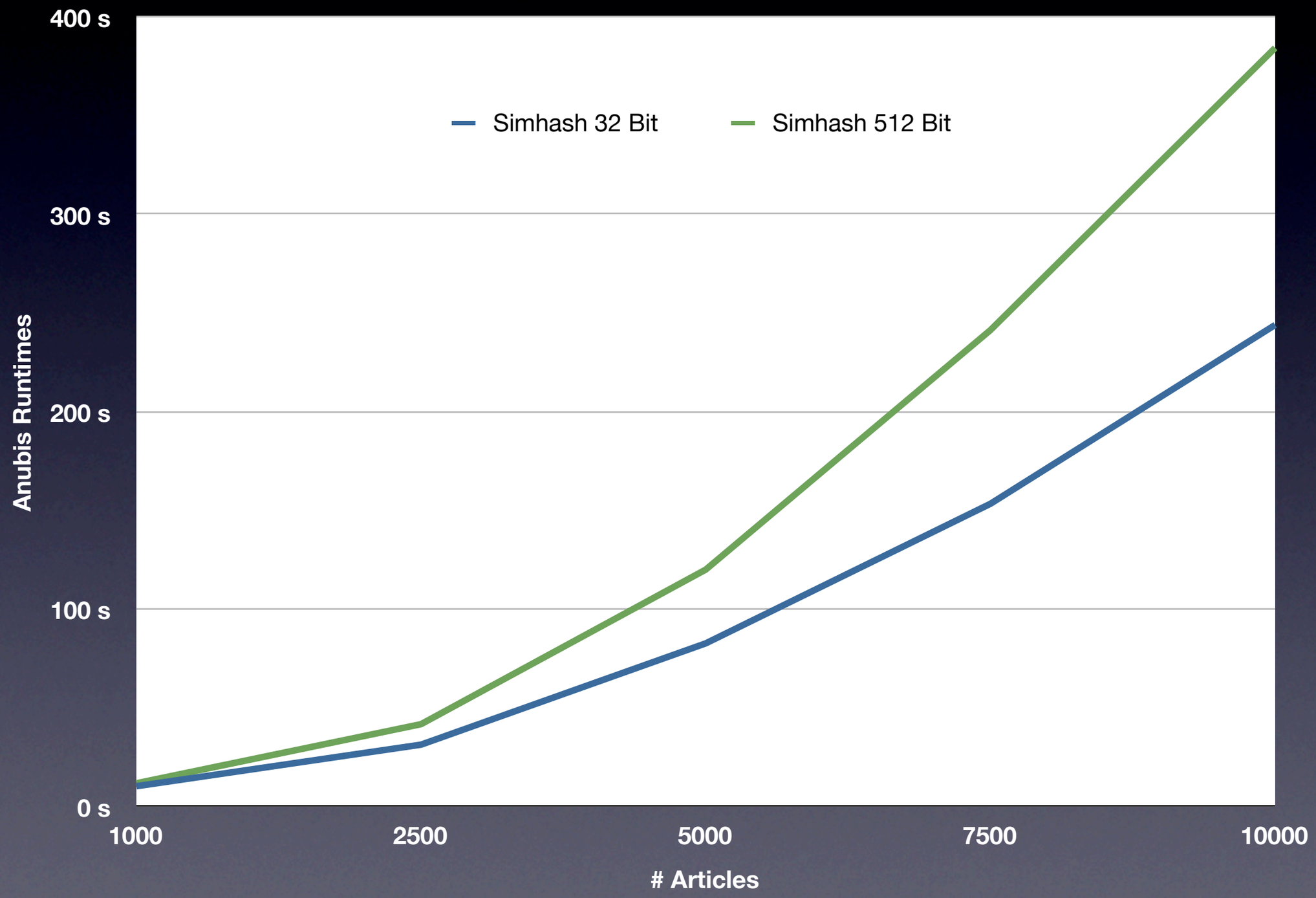
Runtime Evaluation



Runtime Evaluation



Runtime Evaluation



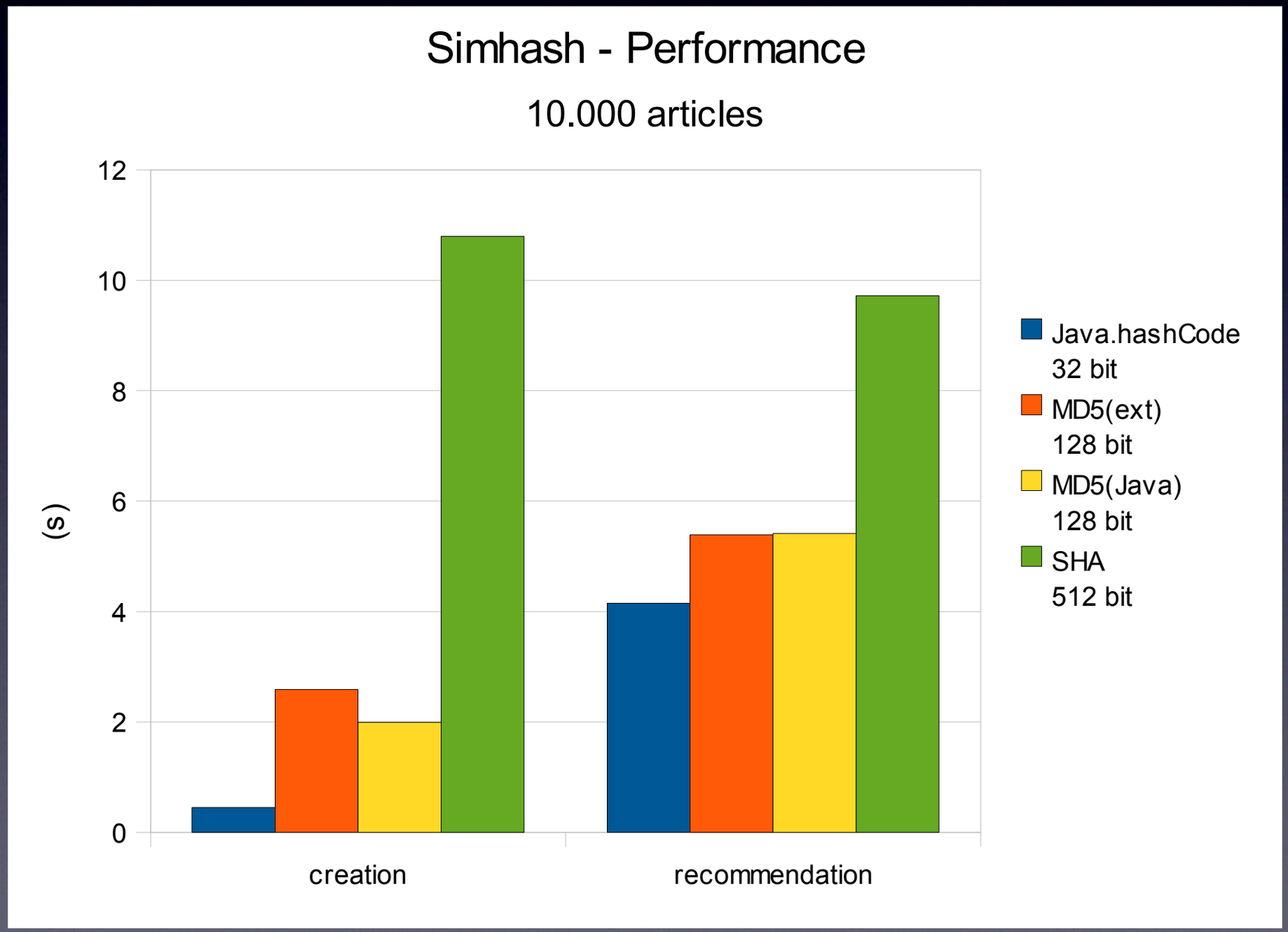
Demo

- 122263 articles
  - All articles including the phrase ‘german’
- Runtime:
  - 45.061,462 s

Conclusion

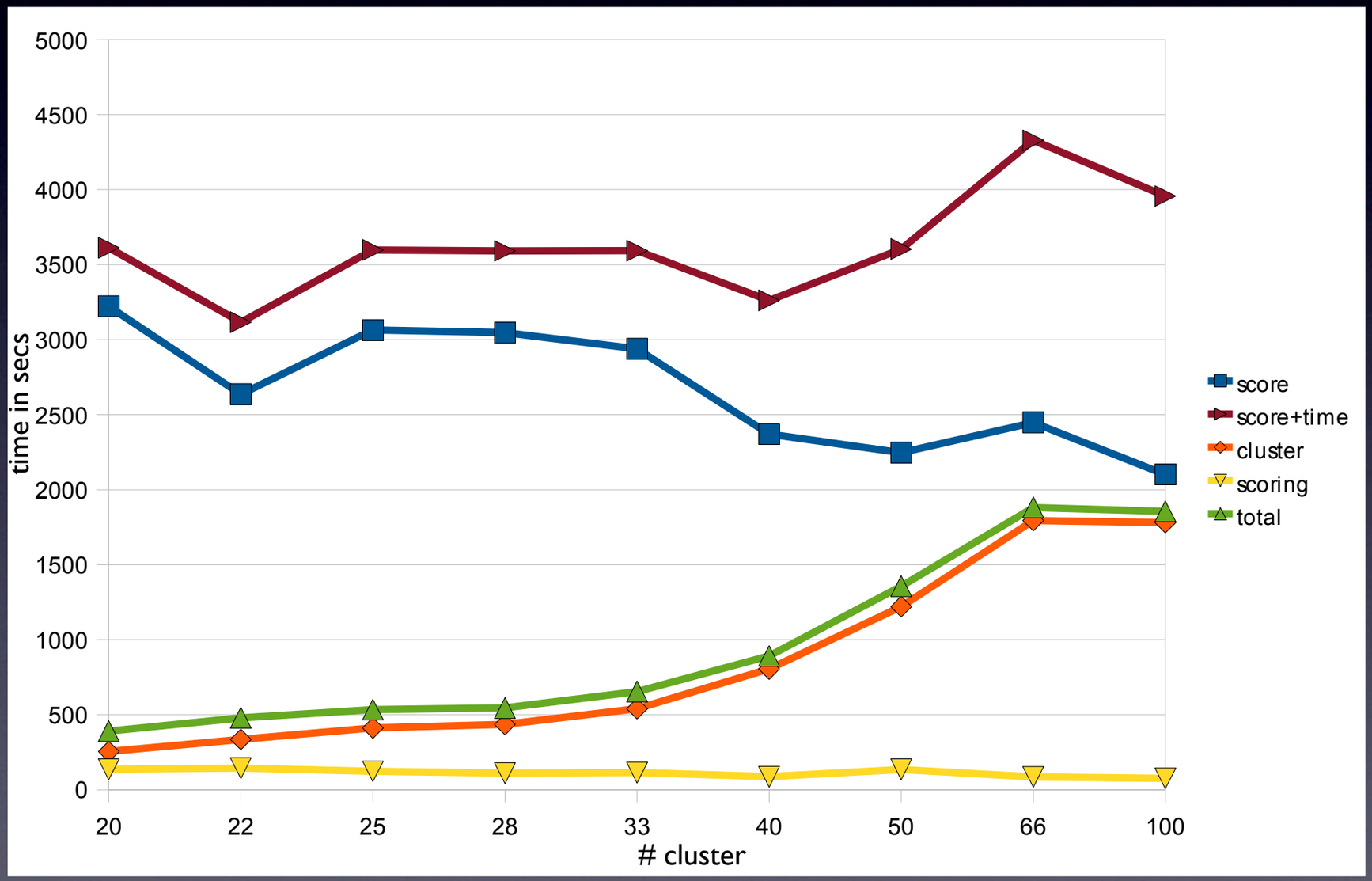
- Clustering is good but slooooooow on huge data sets
- Cosine Distance is better but still too slow
- SimHash is the only feasible approach

Simhash Performance



Runtime of the Simhash algorithm's two phases depending on the bitset size

Cluster Count (1000 articles)



Iterations (1000 articles)