Implementing Randomized Algorithms: Text to Matrix Generator Toolbox

Eugenia Maria Kontopoulou, Dimitrios Zeimpekis
and Efstratios Gallopoulos

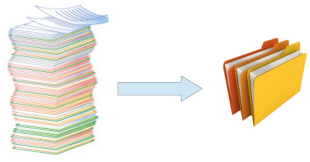
Department of
Compute Engineering and Informatics
University of Patras

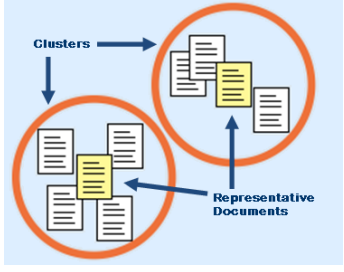
6th Gene Golub SIAM Summer School 2015
G2S3 2015

Delphi, June 2015

Text Mining: Why to use Randomization?

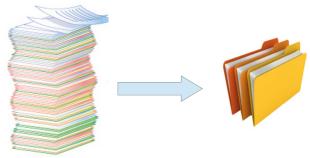
Application

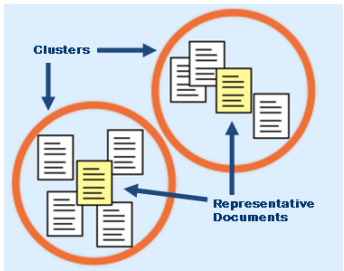




Text Mining: Why to use Randomization?

Application



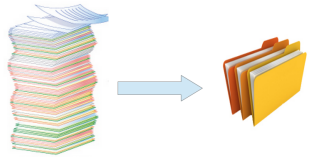


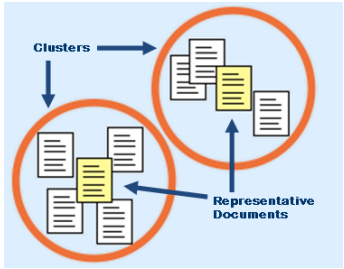
Problem

Big Data { Storage Problems
Interpretation Problems

Text Mining: Why to use Randomization?

Application





Problem

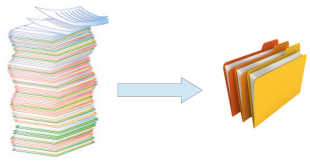
Big Data { Storage Problems
Interpretation Problems

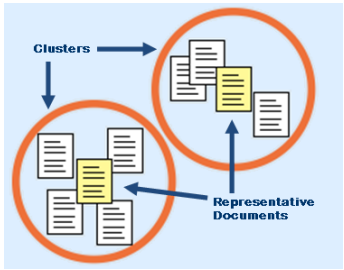
What we need?

- ✓ Few passes to the secondary memory
- ✓ Few blocks in RAM
- ✓ Low dimensionality
- ✓ Interpretability
- ✓ Speed ??

Text Mining: Why to use Randomization?

Application





Problem

Big Data { Storage Problems
Interpretation Problems

What we need?

- ✓ Few passes to the secondary memory
- ✓ Few blocks in RAM
- ✓ Low dimensionality
- ✓ Interpretability
- ✓ Speed ??

Solution

Randomized Techniques

Text-to-Matrix Generator (1)

Zeimpekis+ Kontopoulou + EG '15

What is TMG:

- Toolbox developed in University of Patras for text mining tasks over document collections
- Educational and Research tool

Text-to-Matrix Generator (1)

Zeimpekis+ Kontopoulou + EG ´15

What is TMG:

- Toolbox developed in University of Patras for text mining tasks over document collections
- Educational and Research tool

Implementation:

- over 18.500 lines of `matlab` and `perl`
- takes advantage from sparse technology provided by MATLAB
- first version by Zeimpekis (´06)

Text-to-Matrix Generator (1)

Zeimpekis+ Kontopoulou + EG ´15

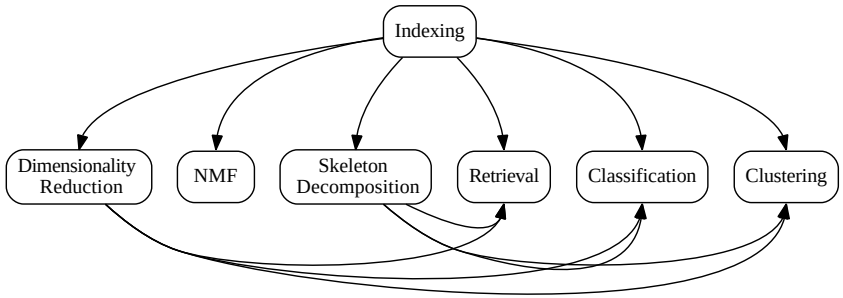
What is TMG:

- Toolbox developed in University of Patras for text mining tasks over document collections
- Educational and Research tool

Implementation:

- over 18.500 lines of `matlab` and `perl`
- takes advantage from sparse technology provided by MATLAB
- first version by Zeimpekis (´06)

Modules



Text-to-Matrix Generator (1)

Zeimpekis+ Kontopoulou + EG ´15

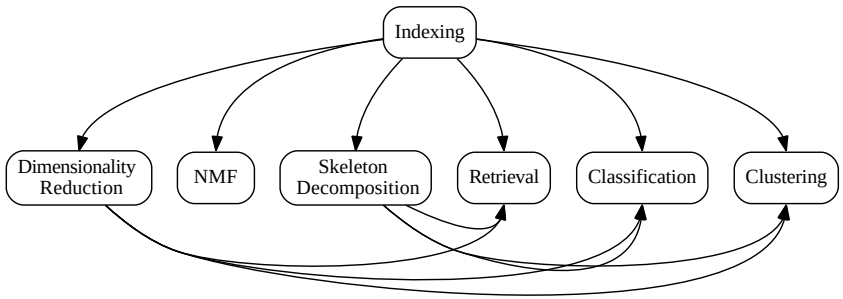
What is TMG:

- Toolbox developed in University of Patras for text mining tasks over document collections
- Educational and Research tool

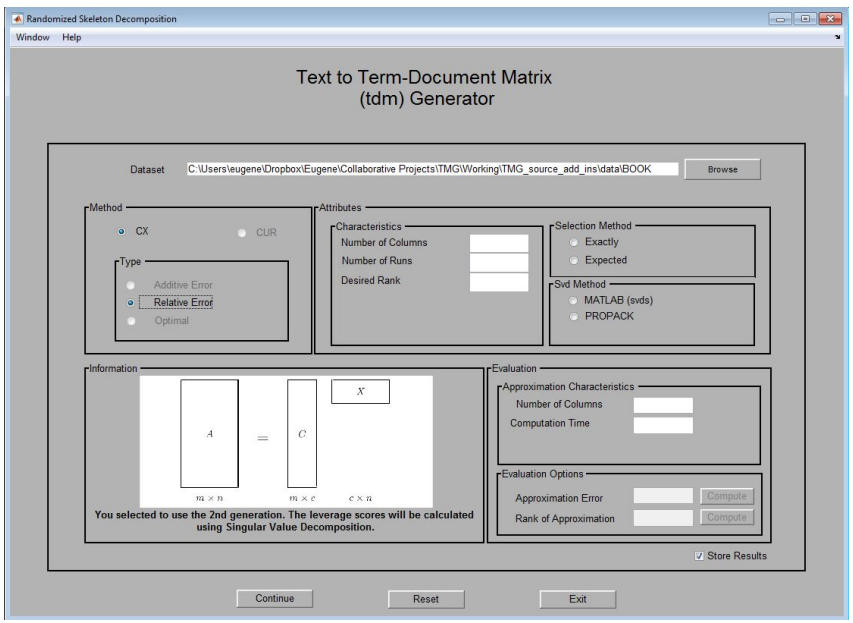
Implementation:

- over 18.500 lines of `matlab` and `perl`
- takes advantage from sparse technology provided by MATLAB
- first version by Zeimpekis (´06)

Modules



Randomized GUIs



A huge thanks to:

Christos Boutsidis

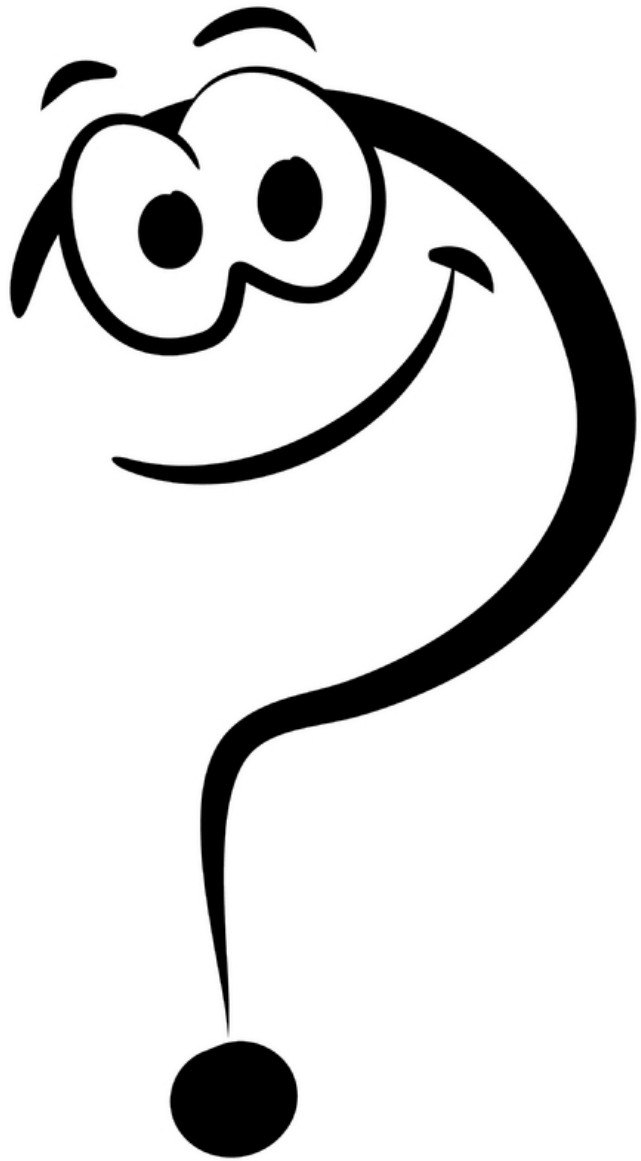
for the really helpful discussions and comments on the implementations :-)

A huge thanks to:

Christos Boutsidis

for the really helpful discussions and comments on the implementations :-)

Questions ?



Bibliography

D. Zeimpekis and E. Gallopoulos. ``TMG: A MATLAB toolbox for generating term document matrices from text collections''. In: *Grouping Multidimensional Data: Recent Advances in Clustering*. Ed. by J. Kogan, C. Nicholas, and M. Teboulle. Berlin: Springer, 2006, 187–210.