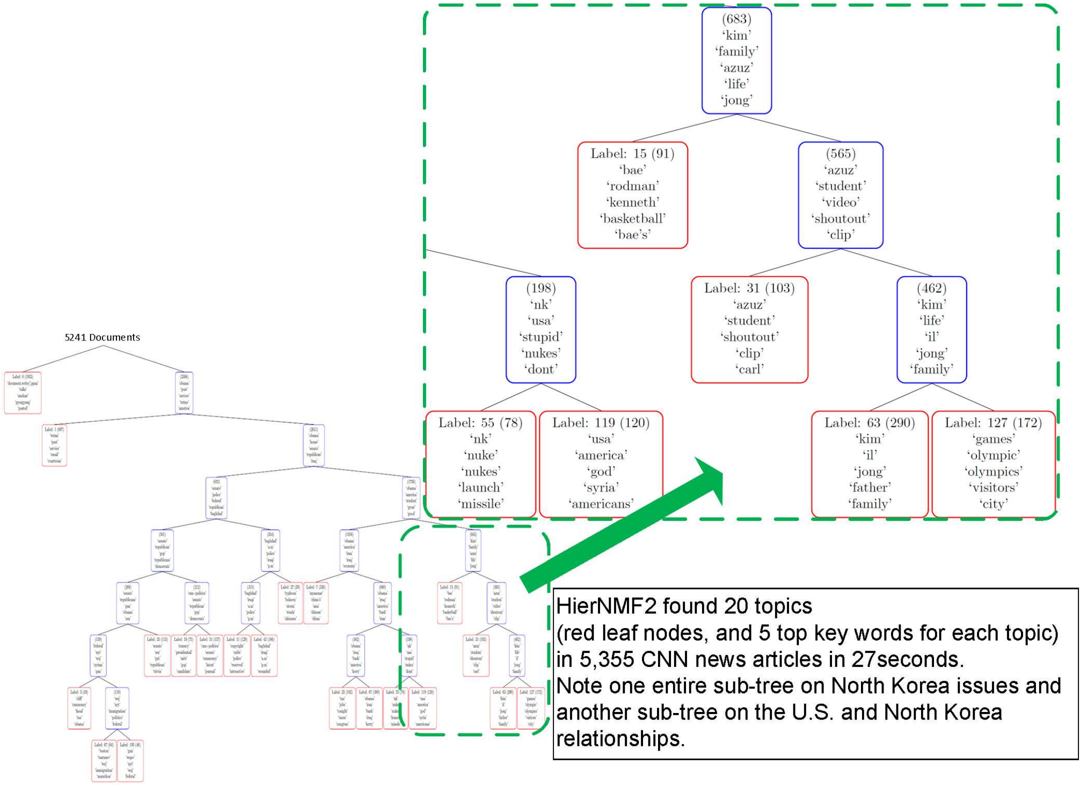
DARPA XDATA: Fast Automatic Topic and Keyword Discovery for Large Document Collections (SmallK Software : smallk.github.io)

Georgia Institute of Technology and Georgia Tech Research Institute

Key Contributions

- Fast, scalable, and effective algorithms for automatically generating topics and keywords in large scale text data sets based on nonnegative matrix factorization (NMF), called HierNMF2.
- HierNMF2 can discover two topics extremely fast in each step, and traverse down the tree deciding the next best node to further split into two topics
- HierNMF2 is currently the fastest and most accurate method for discovery of topics and top keyword on variety of computing systems such as commodity laptop hardware, GPU, and distributed environment

HierNMF2 for Discovery of Topics: Results on CNN News Data: North Korea

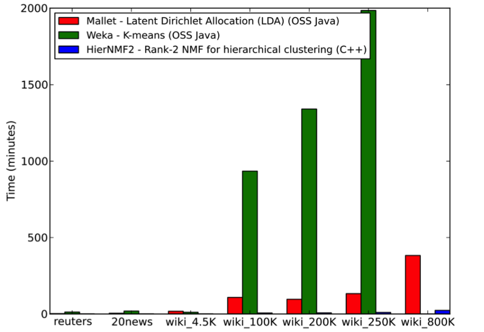

5241 Documents
(683) 'kim' 'family' 'azuz' 'life' 'jong'
Label: 15 (91) 'bae' 'rodman' 'kenneth' 'basketball' 'bae's'
(565) 'azuz' 'student' 'video' 'shoutout' 'clip'
(198) 'nk' 'usa' 'stupid' 'nukes' 'dont'
Label: 31 (103) 'azuz' 'student' 'shoutout' 'clip' 'carl'
(462) 'kim' 'life' 'il' 'jong' 'family'
Label: 55 (78) 'nk' 'nuke' 'nukes' 'launch' 'missile'
Label: 119 (120) 'usa' 'america' 'god' 'syria' 'americans'
Label: 63 (290) 'kim' 'il' 'jong' 'father' 'family'
Label: 127 (172) 'games' 'olympic' 'olympics' 'visitors' 'city'

HierNMF2 found 20 topics (red leaf nodes, and 5 top key words for each topic) in 5,355 CNN news articles in 27seconds. Note one entire sub-tree on North Korea issues and another sub-tree on the U.S. and North Korea relationships.

Computing Time: much faster than competing methods

Number of Documents (and topics)

- Reuters: 10,377 (80)
- 20news: 11,314 (80)
- Wiki4.5K: 4,673 (80)
- Wiki100K: 92,899 (120)
- Wiki200K: 212,980 (120)
- Wiki250K: 272,750 (120)
- Wiki800K: 810,454 (160)
  (Weka did not finish on Wiki800K)
  Preprocessing: 11 sec
  Total time for HierNMF2
  k=40: 6.75 minutes k=80: 9.5 minutes
  I/O for loading input files, writing results: ≈ 5 min.



*** HierNMF2: 4.5M Wikipedia doc. found 80 topics in 43.1 min; MacbookPro laptop, Intel Core i7 2.6 GHz, 4 cores, 16 GB memory

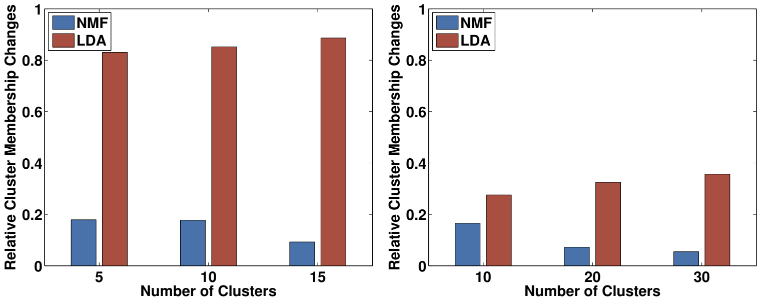
*** LDA (Latent Dirichlet Allocation) and k-means have been the most commonly used methods for topic modeling and clustering. However, HierNMF2 is far superior in speed and quality of solutions.

Topic Consistency: more consistent than generative models

Documents topic membership may change among multiple runs (due to initializations): NMF more consistent than LDA

Infovis-Vast Data: 515 academic papers in visualization area

20Newsgroups Data: 20,000 newsgroup documents



NMF Variants and Applications

We have developed other variants of NMF that can be applied to many important large scale data analytics problems. Some examples are *Robust NMF* for outlier detection and moving object detection *Symmetric NMF* for graph clustering, *AdapNMF* for Adaptive NMF for changing data, *DynNMF* for Dynamic NMF for rank updating, and *Distributed NMF* for distributed computer systems.

DARPA XDATA Open Source Software (smallk.github.io)

SmallK provides fast and efficient software for variations of NMF with usability and extensibility as key design features. SmallK has a wide range of applications to real-world large-scale data analytics problems.

Documentation and Tutorials

- Step-by-step procedures for installation and execution, test case inputs and outputs documented for comparison, and tutorials provide example use-cases.

Implementation

- C++ codes: fast NMF; hierarchical, and flat clustering. All based on Elemental: numerically robust, distributed matrix computations.
- High level Python code in addition to command line interface. Linux and Mac OS X supported. Will expand to Windows.
- Virtual Machine (platform-agnostic) installation option: Vagrant installation based on Ubuntu minimal installation

Reference: Da Kuang and Haesun Park, Fast rank-2 nonnegative matrix factorization for hierarchical document clustering, KDD 2013.

da.kuang@cc.gatech.edu