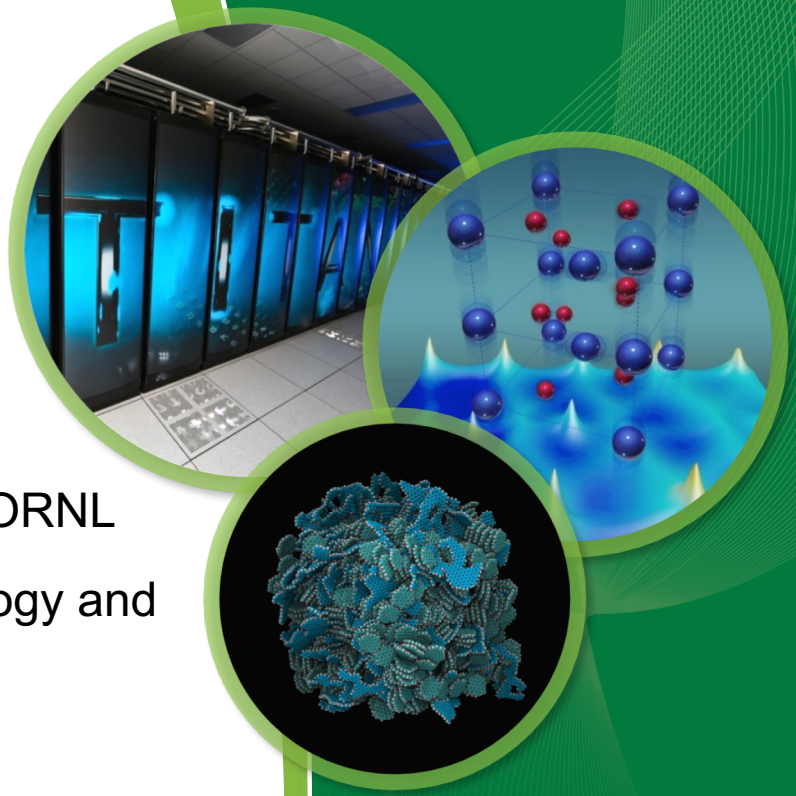
Outlier Detection for Text Data



Ramakrishnan Kannan, kannanr@ornl.gov, ORNL

Hyenkyun Woo, Korea University of Technology and Education

Charu Aggarwal, IBM

Haesun Park, GATech

https://github.com/ramkikannan/outliernmf

Acknowledgements

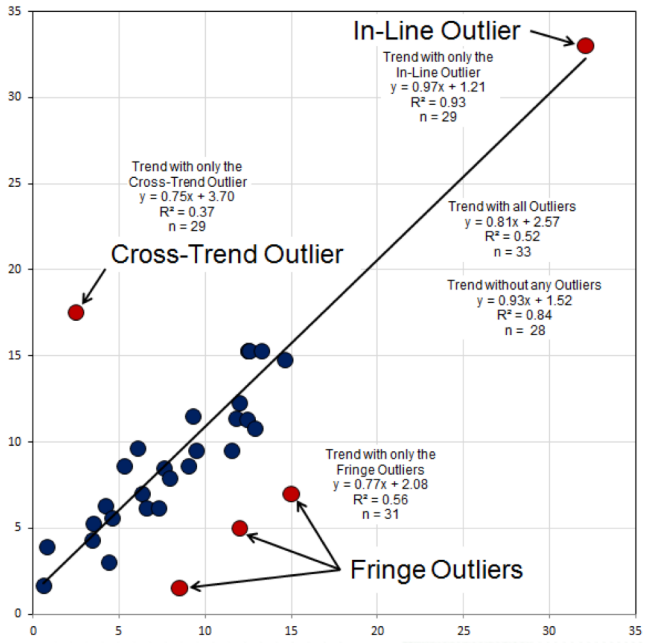
This manuscript has been co-authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. This project was partially funded by the Laboratory Director's Research and Development fund. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy.
This project was partially funded by the Laboratory Director's Research and Development fund and also sponsored by the Army Research Laboratory (ARL) and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. Also, H. Woo is supported by NRF-2015R101A1A01061261.

Agenda

- Introduction to Outliers
- Related work
- Matrix Factorization Model
- Text Outliers using NMF (TONMF)
- Algorithm
- Experiments
  - Baselines and Datasets
  - Performance comparison

What Are Outliers?

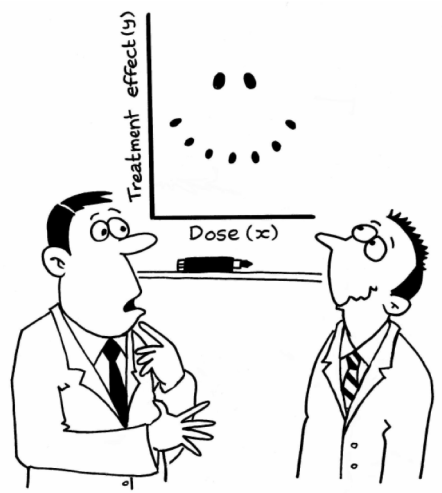
- **Outlier**: An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism
- Applications:
  - Web Site Management
  - Sparse High dimensional data
  - News Article Management
  - Credit card fraud
  - Medical analysis



*https://statswithcats.wordpress.com/

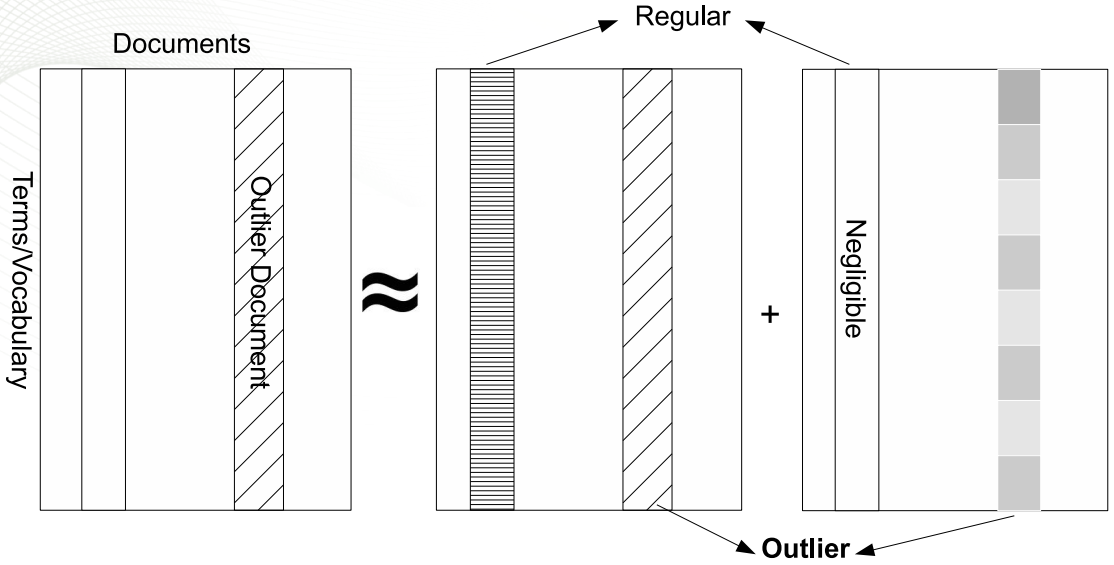
Challenges of Outlier Detection

- Modeling normal objects and outliers properly
  - Hard to enumerate all possible normal behaviors in an application
  - The border between normal and outlier objects is often a gray area
- Application-specific outlier detection
  - E.g., clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations
- Text Specific Problems
  - Very sparse high dimensional data
  - Context - word “Jaguar” may correspond to a car or a cat

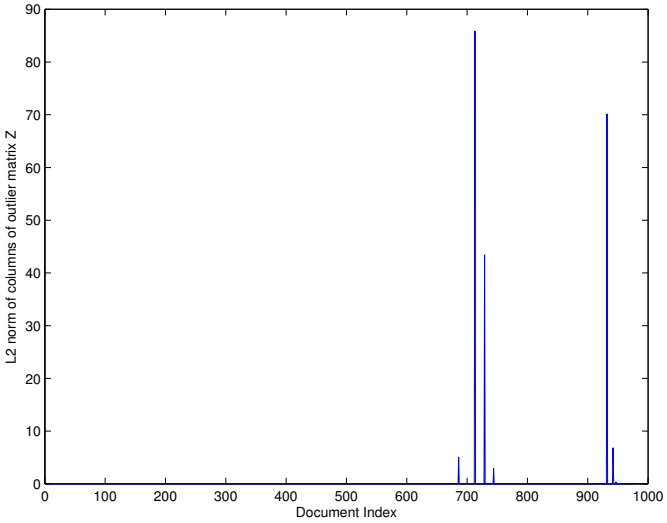


*http://jacobjwalker.effectiveeducation.org/

Matrix Factorization Model



$$\arg\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0;\mathbf{Z}} \frac{1}{2}\|\mathbf{A}-\mathbf{W}\mathbf{H}-\mathbf{Z}\|_F^2+\alpha\|\mathbf{Z}\|_{1,2}$$



BBC News Dataset : http://mlg.ucd.ie/datasets/bbc.html

All the documents from business and politics and 50 documents from tech labeled as outliers.

- A pragmatic approach
  - Understand why these are outliers: Justification of the detection
  - Specify the degree of an outlier: the unlikelihood of the object being generated by a normal mechanism

Text Outliers using NMF (TONMF)

$$\min_{\mathbf{W}\geq 0,\mathbf{H}\geq 0;\mathbf{Z}} \frac{1}{2}\|\mathbf{A}-\mathbf{W}\mathbf{H}-\mathbf{Z}\|_F^2 + \alpha\|\mathbf{Z}\|_{1,2} + \beta\|\mathbf{H}\|_1$$

Outlier Sparsity

3 Blocks - Block Coordinate Descent (BCD)

Block1

$$\mathbf{Z}^{(k+1)} \leftarrow \arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{A}-\mathbf{Z}-\mathbf{W}^{(k)}\mathbf{H}^{(k)}\|_F^2 + \alpha\|\mathbf{Z}\|_{1,2}$$

Block 2 and 3 Sparse NMF

$$(\mathbf{W}^{(k+1)},\mathbf{H}^{(k+1)}) \leftarrow \arg\min_{\mathbf{W}\geq\mathbf{0},\mathbf{H}\geq\mathbf{0}} \frac{1}{2}\|\mathbf{A}-\mathbf{W}\mathbf{H}-\mathbf{Z}^{(k+1)}\| + \beta\|\mathbf{H}\|_1$$

TONMF Algorithm

input : Matrix $\mathbf{A} \in \mathbb{R}_+^{m \times n}$,reduced rank r, α, β

output: Matrix $\mathbf{W} \in \mathbb{R}_+^{m \times r}$,$\mathbf{H} \in \mathbb{R}_+^{r \times n}$,$\mathbf{Z} \in \mathbb{R}^{m \times n}$

`// Rand initialization of W, H, Z`

Initialize $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{Z}$ *as a nonnegative random matrix* ;

while *stopping criteria* $\mathfrak{C}_1$ *not met* **do**

  `// Compute Z for the given` $\mathbf{A}, \mathbf{W}, \mathbf{H}, \alpha, \beta$ `based on Theorem 2`

  for $i \leftarrow 1$ **to** n **do**

    $\mathbf{z_i} \leftarrow max(\|\mathbf{a}_i\|_2 - \frac{\alpha}{\gamma}, 0) \frac{\mathbf{a}_i}{\|\mathbf{a}_i\|_2}$

  $\bar{\mathbf{A}} = \mathbf{A} - \mathbf{Z}$;

  while *stopping criteria* $\mathfrak{C}_2$ *not met* **do**

    for $j \leftarrow 1$ **to** r **do**

      $\mathbf{h}_j^{(k+1)} = \underset{\mathbf{h}_j \geq 0}{\operatorname{argmin}} \frac{\alpha}{2} \|\mathbf{w}_j^{(k)} \mathbf{h}_j^T - (\bar{\mathbf{A}} - \tilde{\mathbf{W}}_j^{(k)})\|_F^2 + g(\mathbf{h}_1^{(k+1)}, \cdots, \mathbf{h}_j, \cdots, \mathbf{h}_r^{(k)})$;

      where, $\tilde{\mathbf{W}}_j^{(k)} = \sum_{i=1}^{j-1} \mathbf{w}_i^{(k)} (\mathbf{h}_i^{(k+1)})^T + \sum_{i=j+1}^{r} \mathbf{w}_i^{(k)} (\mathbf{h}_i^{(k)})^T$

    for $j \leftarrow 1$ **to** r **do**

      $\mathbf{w}_j^{(k+1)} = \underset{\mathbf{w}_j \geq 0}{\operatorname{argmin}} \|\mathbf{w}_j (\mathbf{h}_j^{(k+1)})^T - (\bar{\mathbf{A}} - \tilde{\mathbf{H}}_j^{(k+1)})\|_F^2$;

      where, $\tilde{\mathbf{H}}_j^{(k+1)} = \sum_{i=1}^{j-1} \mathbf{w}_i^{(k+1)} (\mathbf{h}_i^{(k+1)})^T + \sum_{i=j+1}^{r} \mathbf{w}_i^{(k)} (\mathbf{h}_i^{(k+1)})^T$.

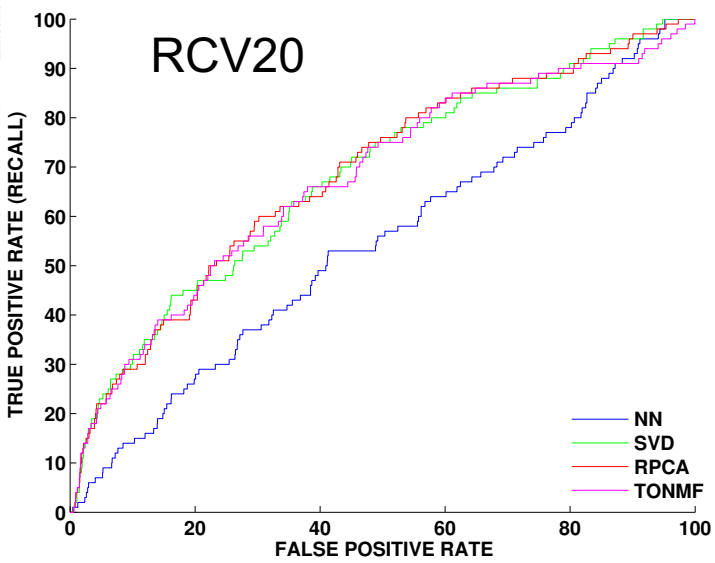
Datasets

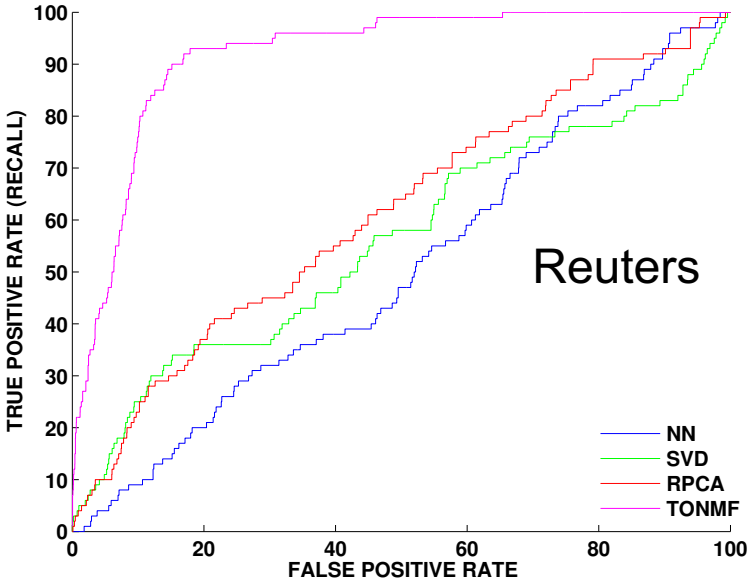
Datasets	# Docs	#Words	Outliers
RCV20 http://qwone.com/~jason/20Newsgroups/	4025	61188	All from IBM and Mac. 50 from Windows OS
Reuters-21578 http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection	5768	18933	All from *earn* and *acq*. 100 from *interest*
Wiki People http://en.wikipedia.org/wiki/Category:Lists_of_politicians	9593	18834	Sections career and life were regular classes. Section Death is outlier
Market Basket Data	10000	50000	2500 data points from four different seeds and 250 as outliers

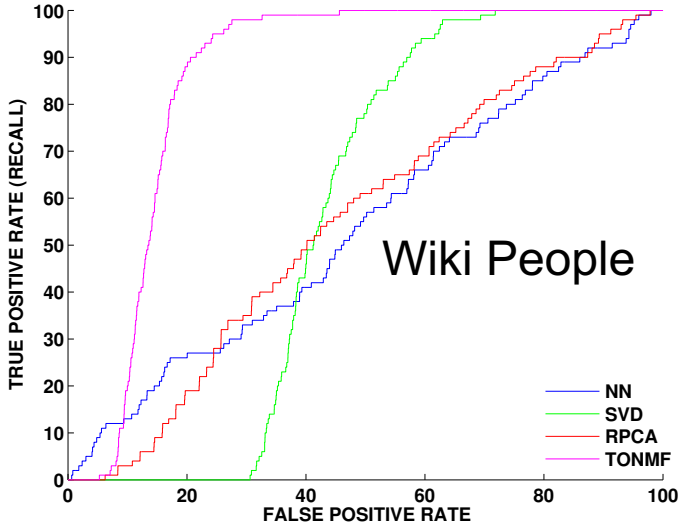
Baselines and Metrics

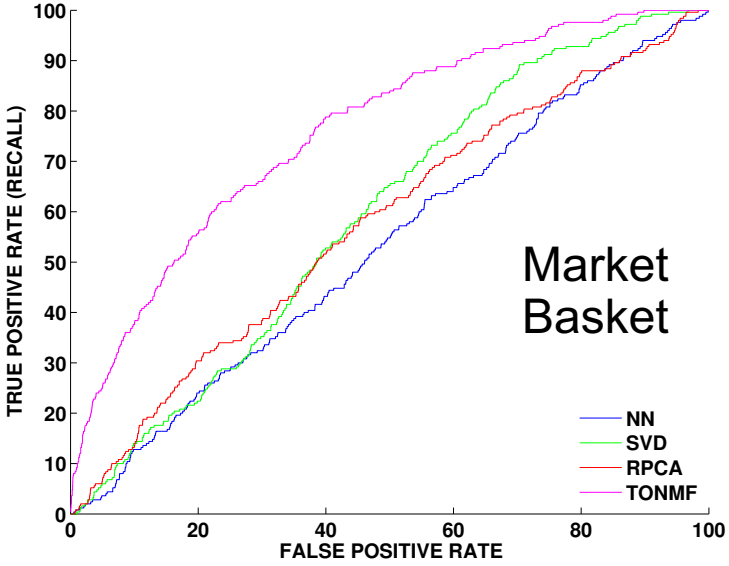
- Metrics – Area under Receiver Operating Characteristics (ROC) Curve
- Baselines
  - Distance-based kNN Algorithm – Sweeping k from 1 to 50.
  - Singular Value Decomposition (SVD)
  - Robust Principal Component Analysis (RPCA)

ROC Curves

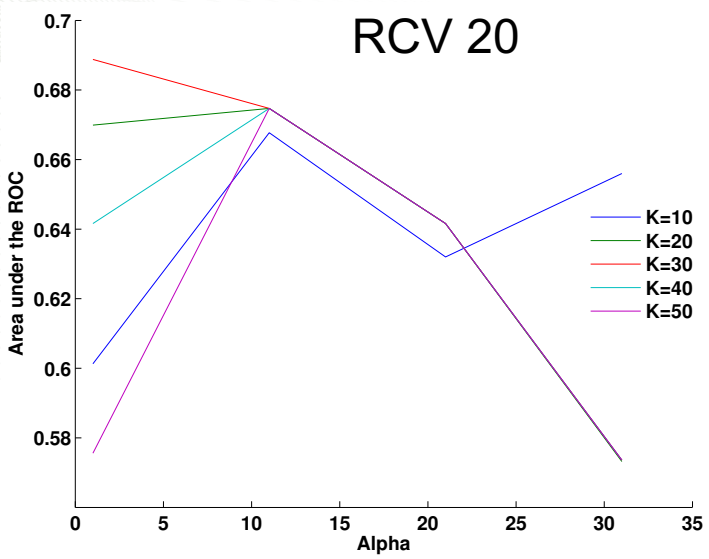


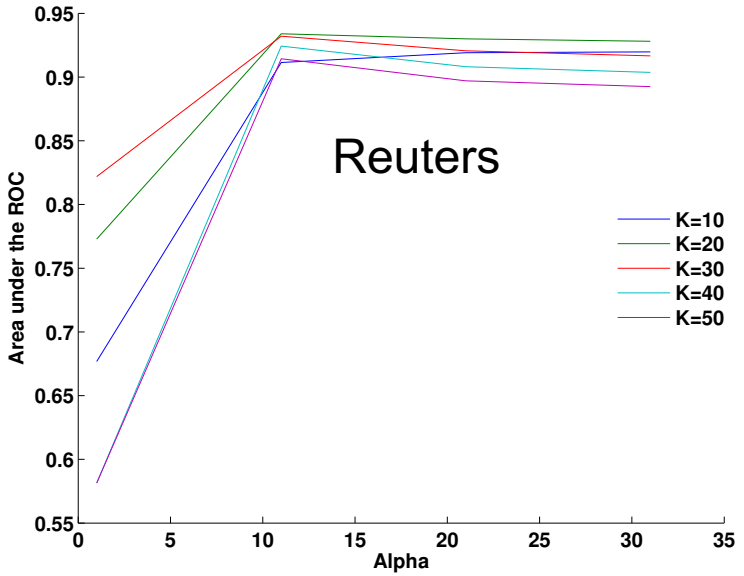


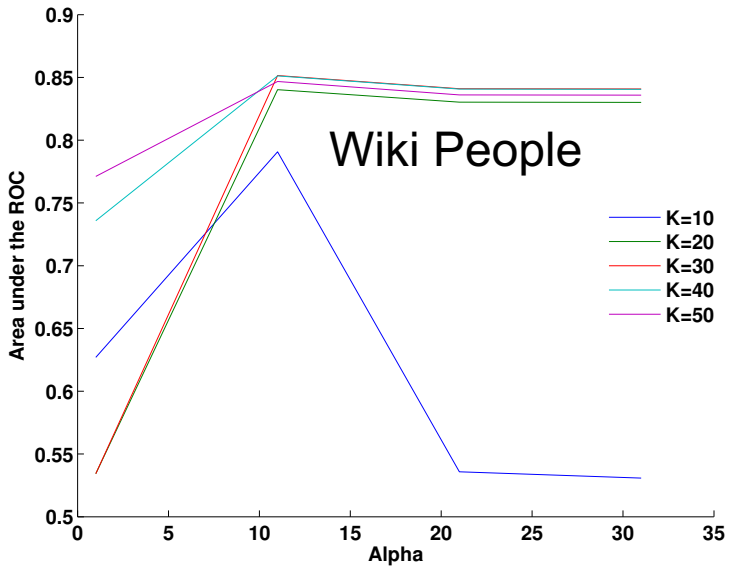


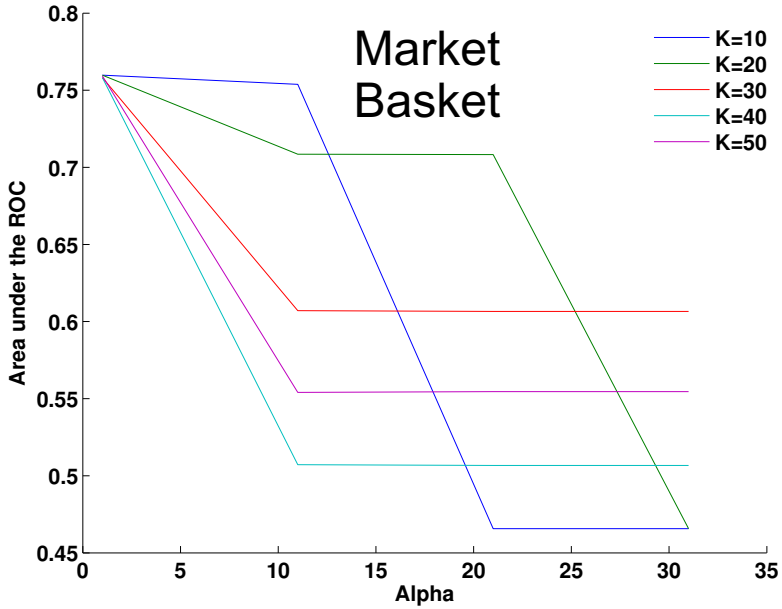


Parameter Sensitivity









Conclusion

- Matrix Factorization based approach to text outlier analysis
- Different representation other than bag of words
- Distributed implementation
- Temporal and Spatial aspects
- Topic Detection and streaming data

Questions