

1. Motivation and Background

Verizon is interested in gaining insights from cell phone user data such as browsing history [1]. Clustering can be used with this data to perform customer segmentation. Clustering is an **unsupervised machine learning** task that involves grouping data.

Spectral clustering refers to a family of algorithms that use **spectral decomposition** on a **similarity matrix** to reduce the dimensionality for clustering, commonly via **k-means** [2]. This can give good performance but requires multiple **computationally complex** operations, limiting its application to large datasets.

Our team is developing new **landmark-based spectral clustering (LSC)** methods by modifying existing techniques, which can solve this issue. It does so by constructing a **sparse affinity matrix** between the data points and landmark points.

2. Landmark-based Spectral Clustering

LSC is one existing method that tries to improve the **scalability** [3]. The main steps are as follows:

i) Landmark Selection:

By uniform sampling

- observations are sampled randomly
- very fast

By k-means

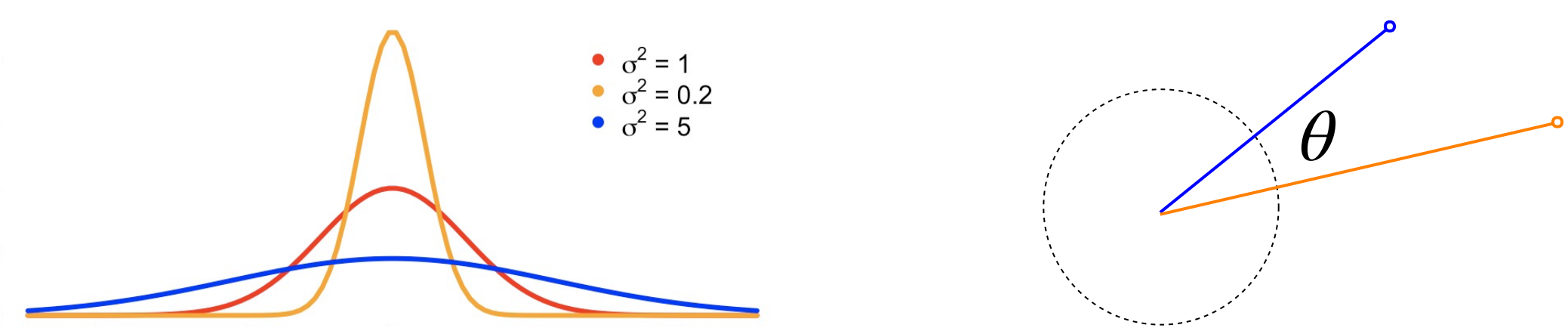
- p centroids
- very slow for larger datasets
- can be more representative

ii) Similarity Computation

- **Gaussian similarity** is computed as: $s(x, y) = e^{-\frac{\|x-y\|^2}{2\beta\sigma^2}}$

σ^2 is estimated with the distance between landmarks and β is a tuning parameter.

- **Cosine similarity** is computed as: $s(x, y) = \cos \theta = \frac{x \cdot y}{\|x\| \cdot \|y\|}$



iii) Nearest Landmarks:

- The largest **r** entries in each row are kept. The rest are set to zero.
- Makes the affinity matrix **sparse**, speeding up computations
- Makes clustering more **robust** to noise

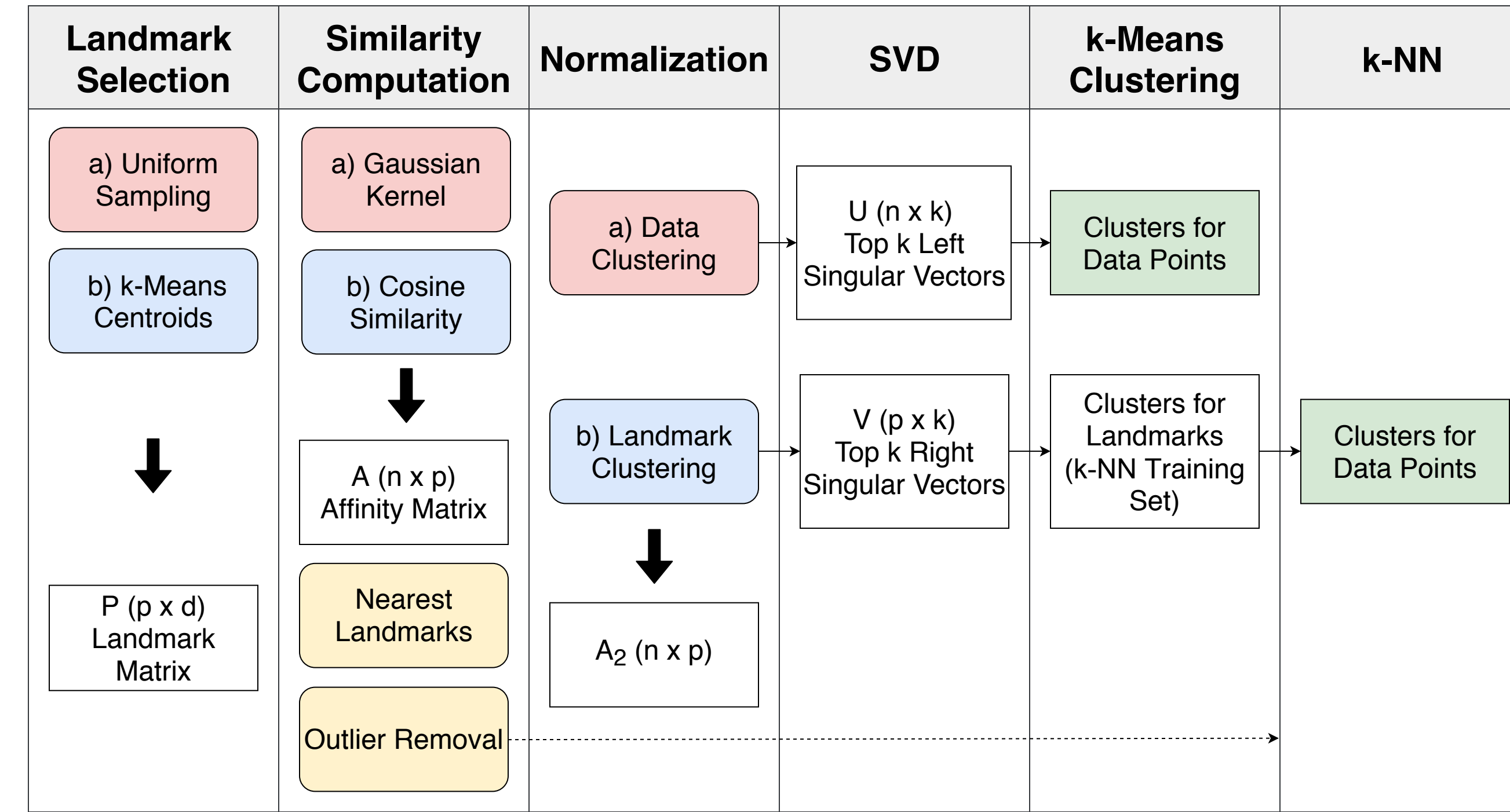
$$A_{ij} = s(x_i, x_j)$$

$$n \times p$$

iv) Data Clustering:

- L_1 row normalization, then $\sqrt{L_1}$ column normalization on A
- Find the top k left singular vectors of A_2
- L_2 row normalization on U
- k-means on U outputs cluster assignments on the data

3. New Algorithm Overview



Input:

- n data points
- **k**: # of desired clusters
- **p**: # of landmarks
- **r**: # of nearest landmarks
- landmark selection method
- similarity measure
- α_1 : proportion of data to treat as outliers
- α_2 : proportion of landmarks to be removed as outliers
- β : the σ^2 tuning parameter for Gaussian similarity
- knn: the number of neighbors considered for **k-nearest neighbor (k-NN)** classification

Output:

- k cluster assignments

4. New Methods

Outlier Removal on the Affinity Matrix

- **Landmark outliers** have low column sums (dissimilar from most data) and are removed.
- **Data outliers** have low row sums (dissimilar from most landmarks) and can be given the zero label or reclassified

Landmark Clustering

- L_1 column normalization, then $\sqrt{L_1}$ row normalization on A
- Find the top k right singular vectors of A_2
- L_2 row normalization on V
- k-means on V outputs cluster assignments on the landmarks
- k-NN outputs cluster assignments on the data

5. Experiments

For evaluating the algorithm and studying **parameter sensitivity**, we tested on three image (handwritten digits) datasets and three text datasets in **R**. These descriptions are recorded after our **preprocessing** steps.

Type	Dataset	# of Instances	# of Features	# of Classes
Text	20Newsgroups	18768	100	20
	Reuters	8067	18933	30
	TDT2	9394	36771	30
Image	USPS	9298	256	10
	Pendigits	10992	16	10
	MNIST	70000	784	10

Compared Algorithms

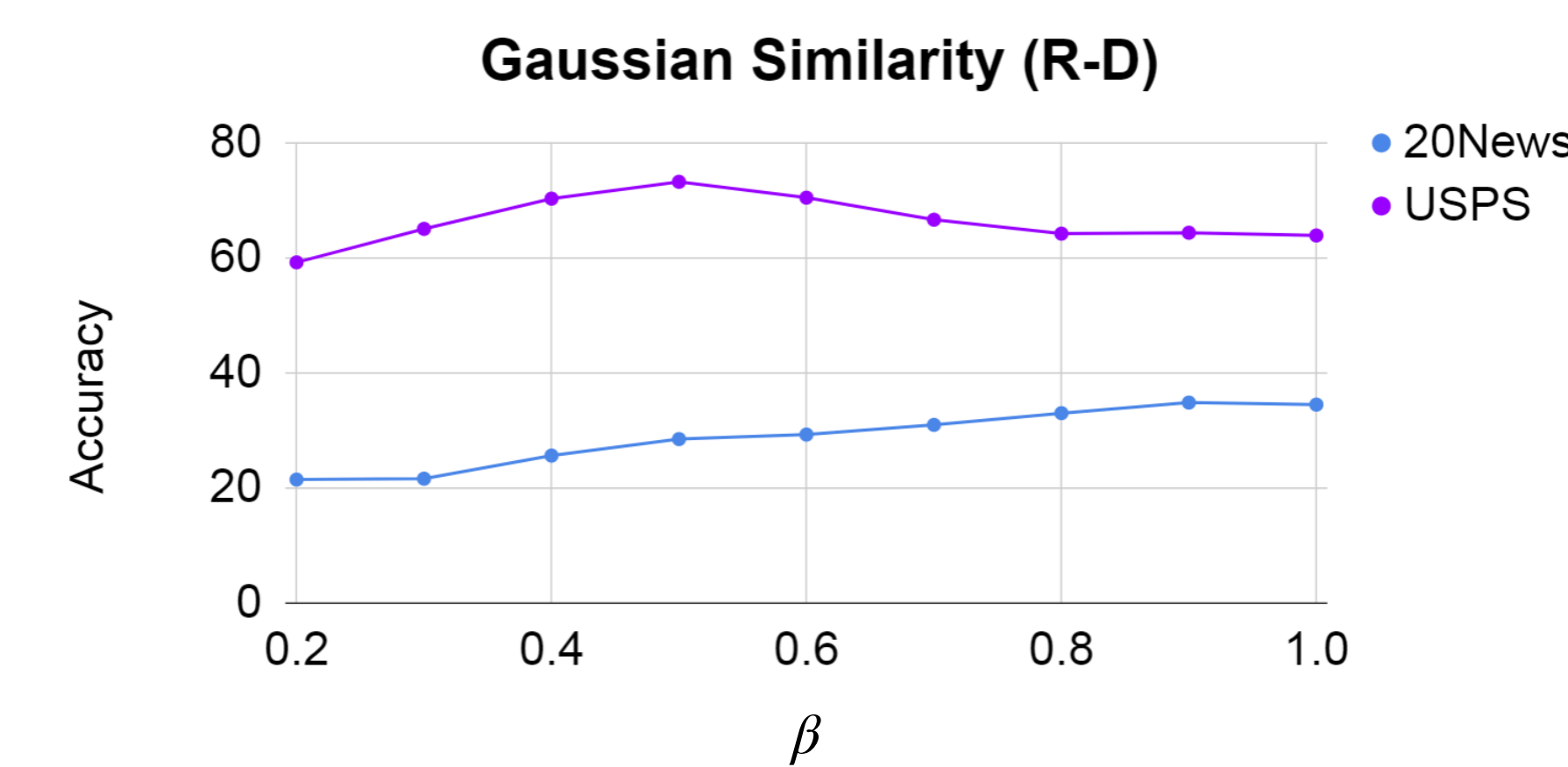
- **R-D**: LSC, random landmark selection, data clustering
- **R-LM**: LSC, random landmark selection, landmark clustering
- **KM-D**: LSC, k-means landmark selection, data clustering
- **KM-LM**: LSC, k-means landmark selection, landmark clustering
- **NJW**: spectral clustering algorithm developed by Ng, Jordan, and Weiss [4]

- **Evaluation metric**: overall classification accuracy
- CPU run-time on the SJSU Golub server is recorded
- All experiments are run for 20 seeds
- Unless specified, the algorithm is run for Cosine similarity, $p = 500$, $r = 6$, $\alpha_1 = 0$, $\alpha_2 = 0$, and $knn = 1$.

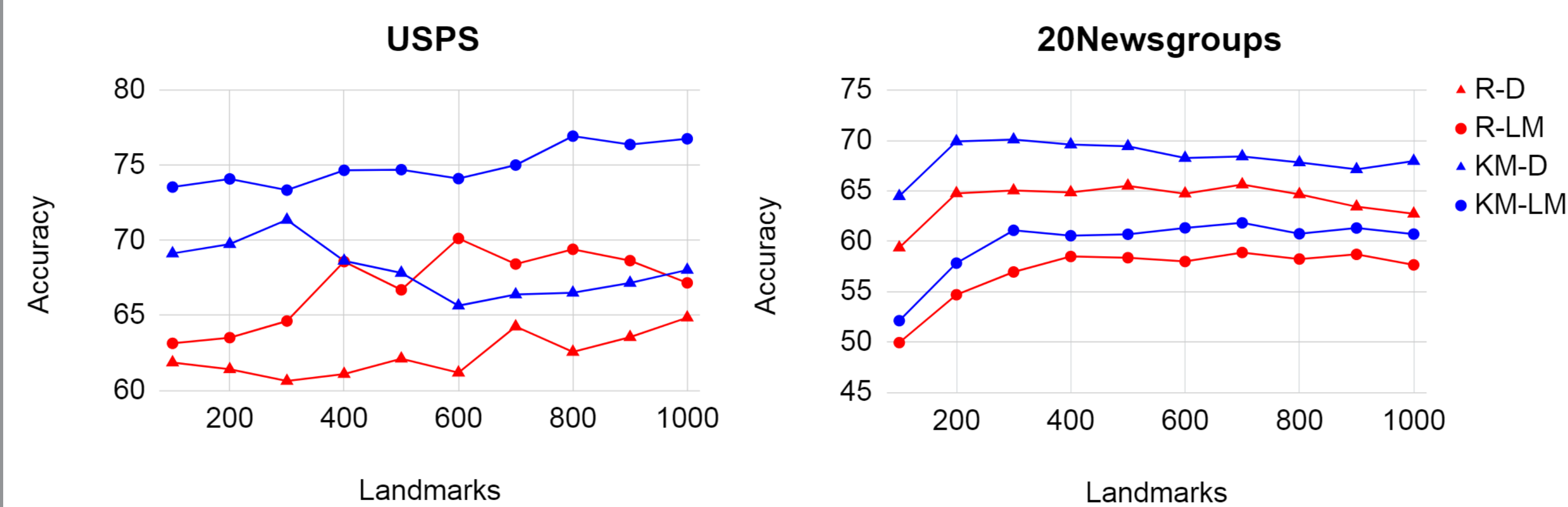
6. Results

Accuracy (%)					
Dataset	k-means LM selection		Random LM Selection		NJW
	Landmark Clustering	Data Clustering	Landmark Clustering	Data Clustering	
20Newsgroups	60.69	69.42	58.37	65.51	63.36
Reuters	31.21	27.38	27.50	25.37	25.68
TDT2	65.69	59.45	64.34	59.85	44.38
USPS	74.70	67.83	66.70	62.12	67.74
Pendigits	81.59	77.94	78.76	78.81	73.75
MNIST	65.10	69.43	59.41	63.32	--

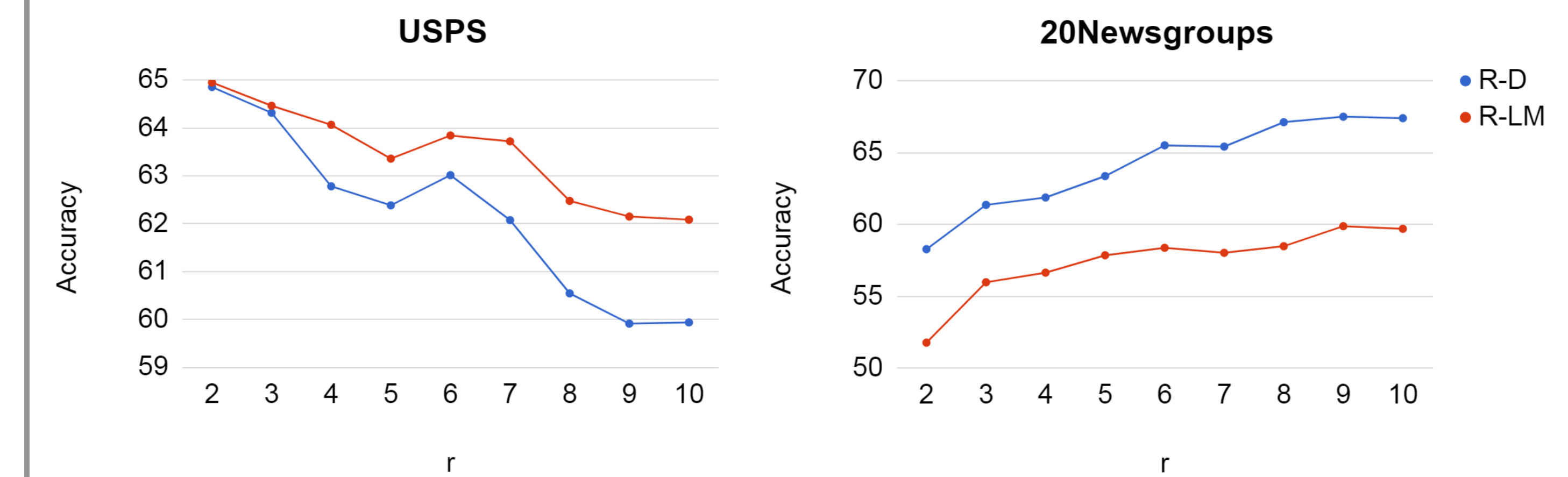
CPU Run-time (s)					
Dataset	k-means LM selection		Random LM Selection		NJW
	Landmark Clustering	Data Clustering	Landmark Clustering	Data Clustering	
20Newsgroups	10.82	12.75	3.94	5.95	150.96
Reuters	503.76	451.88	6.11	7.38	52.31
TDT2	2098.67	1912.68	11.86	12.12	49.46
USPS	11.34	11.65	3.68	3.93	55.46
Pendigits	3.51	3.76	2.15	2.70	95.13
MNIST	597.01	584.06	27.97	31.05	--



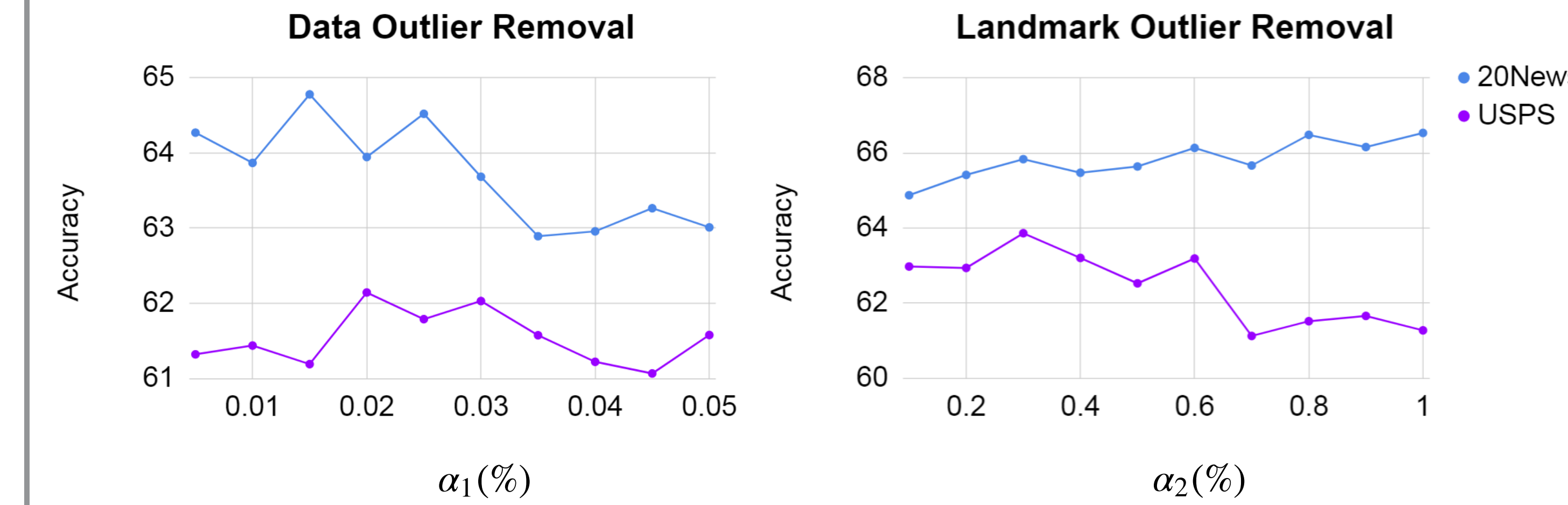
Varying p



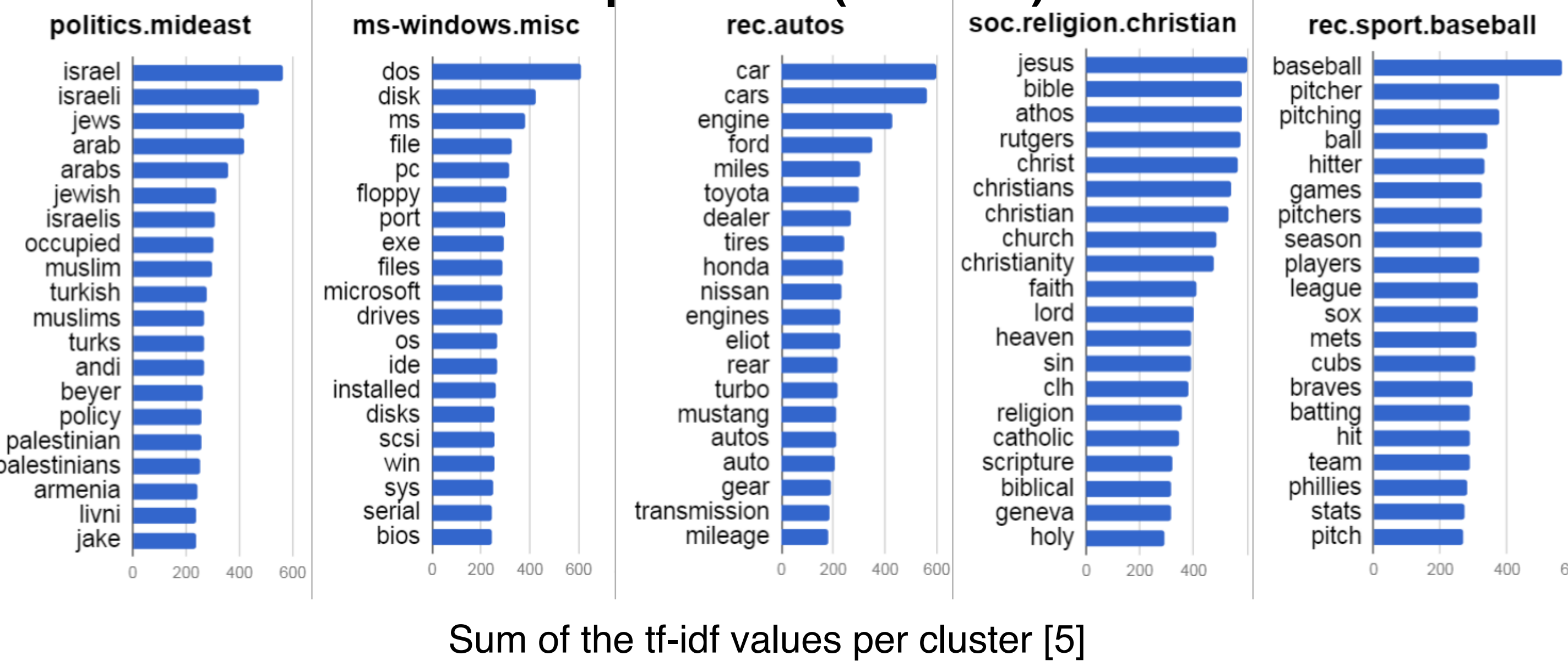
Varying r



Outlier Removal (R-D)



Document Cluster Interpretation (20News)



7. Conclusions

- LSC often provides both speed and accuracy improvements compared to NJW
- Random landmark selection is very efficient compared to k-means, achieving reasonable accuracy in far less time
- Landmark clustering is a promising method, often providing speed and accuracy improvements compared to data clustering
- r and p can be sensitive depending on the dataset

Further Research

- Image segmentation with LSC techniques
- More evaluation metrics
- Further investigation on outlier removal

Acknowledgments

We would like to thank **Guangliang Chen** for his guidance and supervision with this project, **Slobodan Simic** for helping to organize this project, and **Verizon** for their generous sponsorship and support.

References

- [1] J. Fitch et al., "Adaptive Spectral Clustering for High-Dimensional Sparse Count Data", Dept. Math., San Jose State Univ., San Jose, CA, 2017.
- [2] U. Von Luxburg, "A tutorial on spectral clustering", Statistics and computing, 17(4) pp 395–416, 2007.
- [3] D. Cai, X. Chen, "Large Scale Spectral Clustering Via Landmark-Based Sparse Representation", IEEE Trans. Cybernetics, Vol 45 Issue 8, August 2015
- [4] A.Y. Ng, M. I. Jordan, Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm", NIPS Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, pp: 849-856, MIT Press Cambridge, MA, USA, Dec 2001
- [5] C. C. Aggarwal, C. Zhai, "A Survey of Text Clustering Algorithms", ch. 4, pp 77-128, Springer, Boston, MA, doi:10.1007/978-1-4614-3223-4_4