Kernel Spectral Document Clustering Using Unsupervised Precision-Recall Metrics

Raghvendra Mall[1] & Johan A.K. Suykens

[1]ESAT-STADIUS
KU Leuven, Belgium



IJCNN 2015

Introduction

- Clustering is widely used tool in data mining, machine learning, graph compression and textual analysis.
- Aim of clustering is to locate natural groups in data.
- Groups have high intra-cluster similarity and low inter-cluster similarity.
- **Focus on the problem of clustering complex heterogeneous textual data**.
- Document clustering finds applications in domains like document organization and browsing [1], corpus summarization [2], document classification [3] etc.
- Brief survey of text clustering algorithms is provided in [4].

Spectral Clustering - Background

- Create a Laplacian matrix from a pairwise similarity matrix [5, 6, 7].
- Performs eigen-decomposition of the Laplacian matrix.
- Perform k-means top k eigenvectors to obtain localized clusters in eigenspace.
- A non-negative matrix factorization (NMF) method based on eigenspectra was utilized for document clustering in [8].
- **Computationally expensive ($O(N^3)$) and memory inefficient ($O(N^2)$)**.

Kernel Spectral Clustering (KSC)

- KSC is based on weighted kernel PCA formulation as provided in [9].
- It is setup in a primal-dual optimization based learning framework with a proper training, validation and test phase.
- The clustering model is built on a small representative subset of data.
- It has a powerful out-of-sample extensions property which allows to infer cluster affiliation for previously unseen data.
- Applied extensively on large scale networks [10, 11] which are both **sparse and high-dimensional** for clustering/community detection.

Motivations

- Cluster quality metrics are considered good if they possess low intra-cluster distances compared to their inter-cluster distances.
- **Distance based indices like Davies-Bouldin & Silhouette are strongly biased and dependent on clustering method [12]**.
- These indices are often reported on unrealistic low-dimensional text corpora [12].
- Fails to obtain good clusters for sparse & high-dimensional complex textual data.
- Quality metrics like Precision and Recall based on IR principles were shown to be efficient for unsupervised learning in case of heterogeneous textual data [13].
- An advantage is its independence of clustering method & their operation mode.
- Goal is to locate optimal number of homogeneous clusters which is well supported by the unsupervised precision-recall quality metrics proposed in [13].

Contributions

- Owing to success of KSC for community affiliation in large scale networks, we investigate its applicability on **sparse & high-dimensional** textual data.
- Propose a new quality metric Q using the principles and metrics proposed in [13].
- Optimize the Q quality metric in combination with the validation stage of KSC resulting in **Kernel Spectral Document Clustering (KSDC)** method.
- Show effectiveness of KSDC against k-means and neural gas (NG), which have performed well in combination with unsupervised precision-recall metrics.

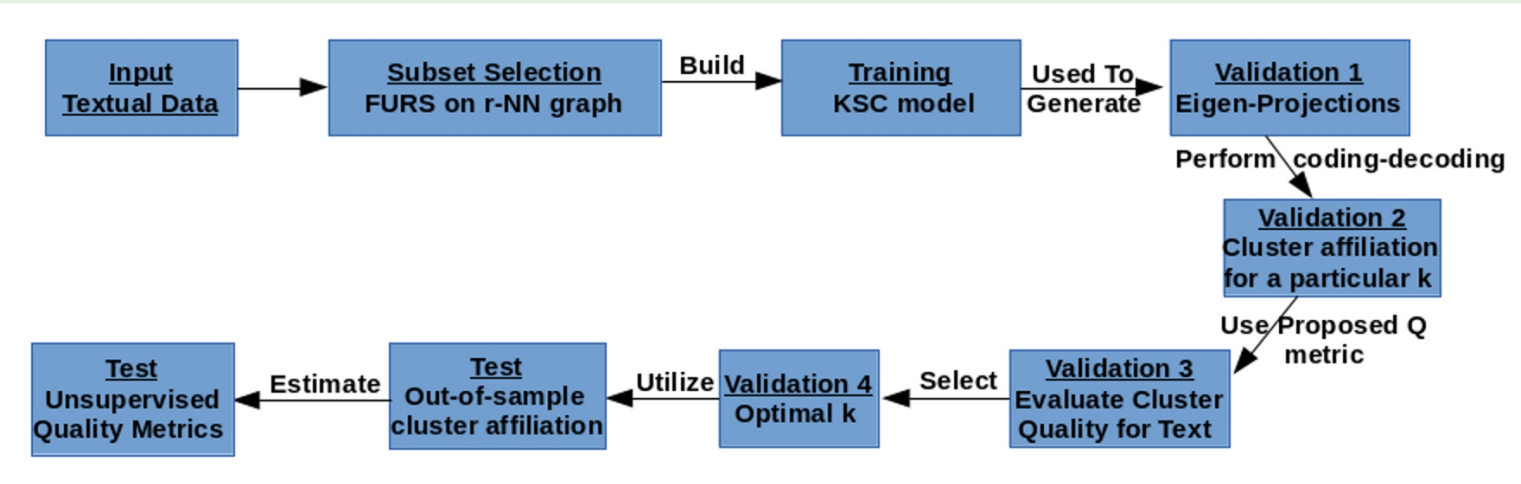


Figure: Steps involved in Proposed Method

KSC Formulation

Given $\mathcal{D} = \{x_i\}_{i=1}^{N_{tr}}$ and *maxk*, the primal formulation of the weighted kPCA [9] is:

$$\min_{w^{(l)}, e^{(l)}, b_l} \quad \frac{1}{2} \sum_{l=1}^{maxk-1} w^{(l)\intercal} w^{(l)} - \frac{1}{2N_{tr}} \sum_{l=1}^{maxk-1} \gamma_l e^{(l)\intercal} D_\Omega^{-1} e^{(l)} \qquad (1)$$

$$\text{such that} \quad e^{(l)} = \Phi w^{(l)} + b_l 1_{N_{tr}}, l = 1, \ldots, maxk - 1$$

KSC Primal-Dual Model

- Primal clustering model is: $e_i^{(l)} = w^{(l)\intercal} \phi(x_i) + b_l, i = 1, \ldots, N_{tr}$.
- Corresponding dual: $D_\Omega^{-1} M_D \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)}$
- Dual is closely related to **Random Walk** problem.
- Dual predictive model is: $\hat{e}^{(l)}(x) = \sum_{i=1}^{N_{tr}} \alpha_i^{(l)} K(x, x_i) + b_l$.
- $\Omega_{ij} = K(x_i, x_j) = \phi(x_i)^\intercal \phi(x_j)$ is obtained by using normalized linear kernel ($\frac{x_i^\intercal x_j}{\|x_i\| \|x_j\|}$) for textual data.
- Predictive model allows cluster affiliation for previously unseen data.

KSC Steps

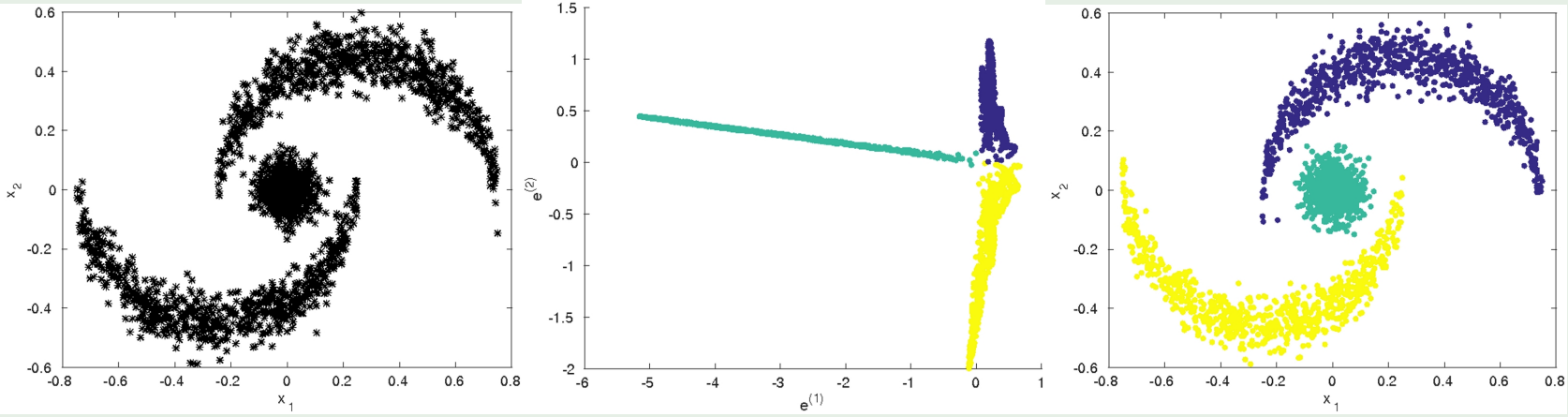


Figure: Illustration of steps undertaken by KSC technique.

KSC Training & Validation Requirements

- Requires a representative subset of data to build the KSC model.
- FURS [14] selection technique used to generate subsets.
- It converts the text data into an r-NN graph and then applies FURS [14] as depicted in [15].
- **A good model selection criterion to estimate k homogeneous clusters in complex textual data.**

Comparison of Model Selection Criteria for KSC

Table: Comparison of different model selection criteria for KSC exploiting the projection structures in the eigenspace.

Criterion	Strengths
BLF [9]	Works best in case of well-separated, non-overlapping clusters in input space with clear line-structure in eigenspace
BAF [10]	Works well in case of $\geq$ 3 overlapping clusters by exploiting angular similarity of eigen-projections to mean projection vector in each cluster.
AMS [16]	Based on the same principle as **BAF** but able to identify $k = 2$ clusters.
SIL [12]	Based on principle of intra and inter cluster euclidean distance in the eigenspace.

Unsupervised Precision-Recall based Metrics

- After cluster affiliation for a given k using coding-decoding scheme of KSC, we evaluate quality of clusters using unsupervised IR metrics.
- For validation set $\mathcal{V}$, Recall (Rec) and Precision ($Prec$) metrics for a particular word p of cluster c is defined as [13]:

$$Rec_c(p) = \frac{\sum_{d \in c} W_d^p}{\sum_{c' \in \mathcal{C}} \sum_{d \in c'} W_d^p},$$

$$Prec_c(p) = \frac{\sum_{d \in c} W_d^p}{\max_{p'} (\sum_{d \in c} W_d^{p'})}$$

- W_x^p is weight of property p in document x & $\mathcal{C}$ represent the set of clusters.
- To estimate overall clustering quality, average Macro-Recall (R_M) & average Macro-Precision (P_M) are used:

$$R_M = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|S_c|} \sum_{p \in S_c} Rec_c(p),$$

$$P_M = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|S_c|} \sum_{p \in S_c} Prec_c(p)$$

- S_c is the set of peculiar properties of the cluster c and $\bar{C}$ represents the peculiar set of clusters extracted from the set $\mathcal{C}$.

Unsupervised Quality Metrics

- Macro-Precision and Macro-Recall have antagonistic behavior [13] & optimum k is obtained when both values are high.
- A Macro F-measure is defined as: $F_M = \frac{2 \times R_M \times P_M}{R_M + P_M}$.
- F-measure cannot detect degenerate heterogeneous clusters [13].
- Property oriented indices like Micro-Recall (R_m) and Micro-Precision (P_m) overcome this issue:

$$R_m = \frac{1}{d} \sum_{c \in \bar{C}, p \in S_c} Rec_c(p),$$

$$P_m = \frac{1}{d} \sum_{c \in \bar{C}, p \in S_c} Prec_c(p).$$

- A Cumulative Micro-Precision (CP_m) was proposed in [13] to efficiently identify heterogeneous clusters of large size:

$$CP_m = \frac{\sum_{i=|c_{inf}|,|c_{sup}|} \frac{1}{|C_{i+}^2|} \sum_{c \in C_{i+}, p \in S_c} Prec_c(p)}{\sum_{i=|c_{inf}|,|c_{sup}|} \frac{1}{C_{i+}}},$$

- C_{i+} is the subset of clusters of $\mathcal{C}$ for which associated data is greater than i, and: $inf = \text{argmin}_{c_i \in \mathcal{C}} |c_i|$, $sup = \text{argmax}_{c_i \in \mathcal{C}} |c_i|$.

Proposed Q metric

- Define a Micro F-measure (F_m) as: $F_m = \frac{2 \times R_m \times CP_m}{R_m + CP_m}$.
- Optimizes clustering quality for a particular k w.r.t. R_m and CP_m.
- Goal is to identify optimal k s.t. quality of obtained clusters is maximum.
- Obtained clusters should be homogeneous.
- Quality metric should be able to distinguish them from degenerate ones.
- Need to optimize both F_M and F_m.
- F_M larger for smaller values of k and decreases monotonically.
- F_m smaller for lower values of k and increases monotonically.
- Propose a Q metric which minimizes $|F_M - F_m|$.
- Q metric retains both macro and micro cluster qualities.
- Q is defined as: $1 - |F_M - F_m|$.
- Takes values between $[0, 1]$ and higher values represent better quality clusters.

KSDC Algorithm

Input: Given dataset $\mathcal{T} = \{x_i\}_{i=1}^{N}$.
Output: Categorize documents into k clusters.

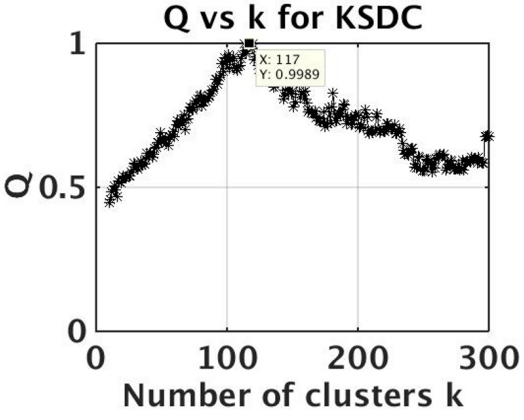
1. Converts documents to bag-of-words using TF-IDF weighted scheme.
2. Generate r-NN graph [15] and perform FURS [14] to get training & validation set.
3. Calculate Ω using normalized linear kernel on sparse format data.
4. Eigen-decompose Ω & use out-of-sample extensions to generate e_{valid}.
5. **Foreach** k **do**
   - Use coding-decoding scheme to obtain cluster affiliations.
   - Calculate Macro F-measure (F_M) & Micro F-measure (F_m).
   - Estimate $Q = 1 - |F_M - F_m|$.
6. Select k corresponding to which Q is maximum.
7. Perform out-of-sample extensions by using the top k-1 eigenvectors to assign clusters to unseen test data.
8. Estimate F_M and F_m for entire corpus.

Datasets

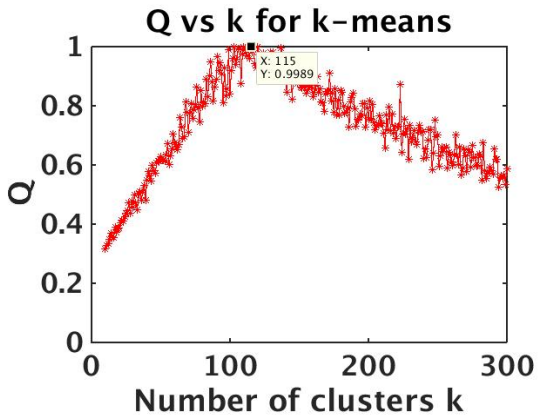
Table: Text data used in the experiments. Here *N* represents number of documents, *d* represents the number of words obtained after performing TF-IDF weighting scheme and *nnz* represents the number of non-zero counts in the bag-of-words.

Dataset	N	d	nnz
Nips	1,500	115,60	217,900
Kos	3,430	5,631	58,146
R8	6,743	6,737	77,582
Classic4	7,095	5,896	247,158
Reuters	8,293	18,933	389,455
TDT2	9,394	36,771	1,224,135
TDT2all	10,212	36,771	1,323,869
20NewsGroups	18,774	61,188	2,435,219
Enron	36,439	24,311	1,009,003
NYTimes	300,000	82,358	90,180,242

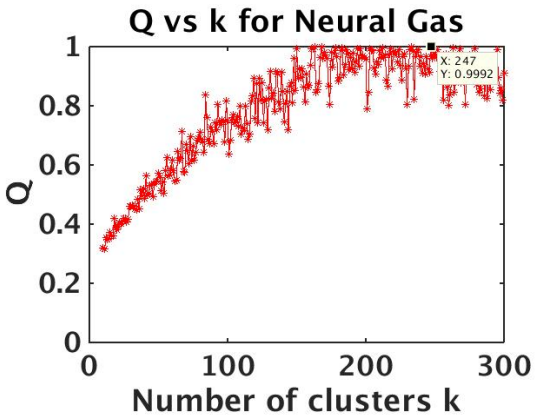
Experimental Results



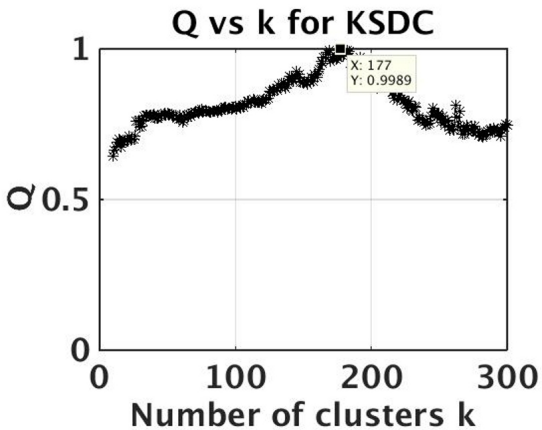
(a) $k = 117$ - KSDC)



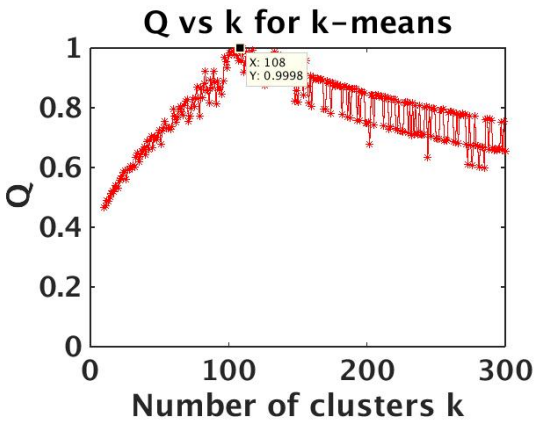
(b) $k = 115$ - k-means



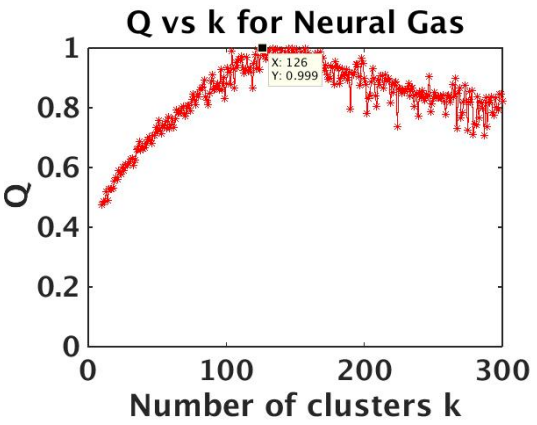
(c) $k = 247$ - NG



(d) $k = 177$ - KSDC



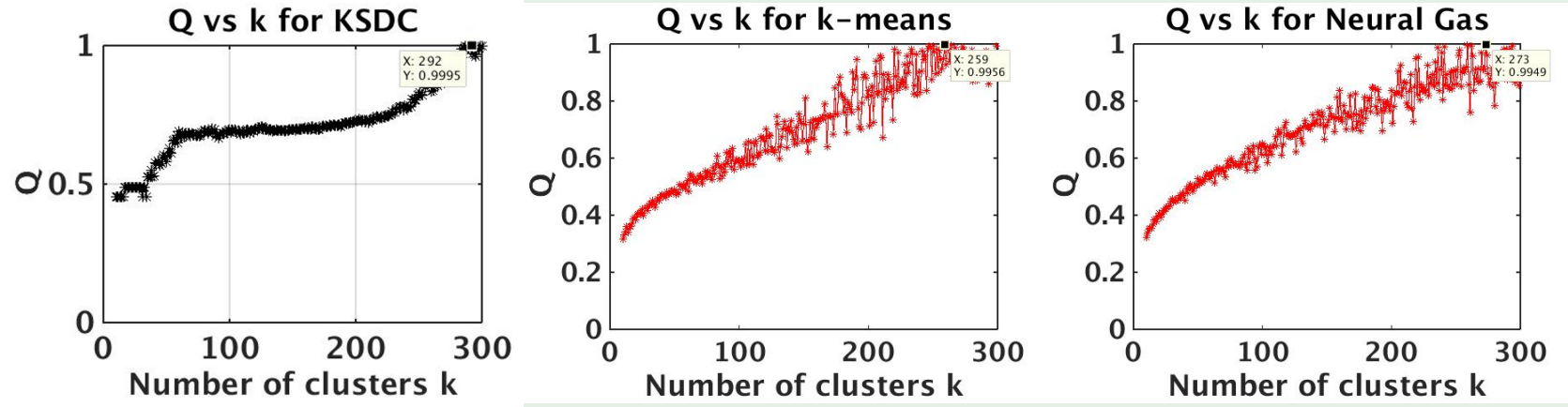
(e) $k = 108$ - k-means



(f) $k = 126$ - NG

Figure: Comparison of proposed Q quality metric for KSDC, k-means and NG algorithms on Kos and Classic4 datasets respectively.

Reuters Experimental Result



(a) $k = 292$ - KSDC (b) $k = 259$ - k-means (c) $k = 273$ - NG

Figure: Q metric comparison for KSDC, k-means and NG algorithms on Reuters. For fair comparison the optimal k search is performed on the validation set for all the methods.

Observations

- Final clustering quality of entire corpus is usually different from validation subset.
- Trends for KSDC are relatively smoother in comparison to k-means and NG.
- Attributed to their dependence on initial randomization.

Experiments on All Datasets

Table: Comparison of quality using F_M, F_m and Q metrics for KSDC, k-means and NG algorithms. Highlighted results are maximum for that metric not necessarily the best.

Dataset	KSDC				k-means				Neural Gas (NG)			
	k	F_M	F_m	Q	k	F_M	F_m	Q	k	F_M	F_m	Q
Nips	268	0.274	**0.328**	**0.946**	129	**0.348**	0.222	0.874	289	0.292	0.075	0.783
Kos	117	0.386	**0.366**	**0.980**	115	**0.446**	0.300	0.854	247	0.392	0.295	0.903
R8	253	0.338	0.342	**0.996**	146	**0.542**	**0.408**	0.886	102	0.497	0.117	0.620
Classic4	177	0.241	**0.203**	**0.962**	108	0.289	0.130	0.841	126	**0.293**	0.143	0.85
Reuters	292	0.286	0.273	**0.987**	259	**0.332**	**0.275**	0.943	273	0.287	0.133	0.846
TDT2	234	0.230	**0.135**	**0.905**	227	**0.265**	0.123	0.858	293	0.203	0.085	0.882
TDT2all	254	0.153	**0.095**	**0.942**	219	**0.245**	0.084	0.839	296	0.239	0.084	0.845
20NewsGroups	294	**0.185**	0.098	0.913	267	0.178	**0.112**	**0.934**	259	0.160	0.092	0.932
Enron	224	**0.455**	**0.353**	0.898	298	0.315	0.346	**0.969**	271	0.255	0.292	0.963
NYTimes	125	**0.345**	**0.273**	**0.928**	173	0.319	0.216	0.897	196	0.292	0.199	0.907

Experimental Result on Nips Dataset

Table: Title of Nips papers belonging to one of the document clusters obtained by KSDC, *k*-means and Neural Gas algorithms respectively. KSDC and NG algorithms find a homogeneous cluster of 3 documents related to speech recognition. However, the *k*-means clustering method discovers a document cluster which is more heterogeneous as it also includes documents related to continuous speech recognition.

Algorithm	**Document Clusters**
KSDC	"Speech Recognition: Statistical and Neural Information Processing Approaches", "HMM Speech Recognition with Neural Net Discrimination", "Improved Hidden Markov Model Speech Recognition Using Radial Basis Function Networks".
k-means	"Speech Recognition: Statistical and Neural Information Processing Approaches", "A Continuous Speech Recognition System Embedding MLP into HMM", "Connectionist Approaches to the Use of Markov Models for Speech Recognition", "Continuous Speech Recognition by Linked Predictive Neural Networks", "Speech Recognition Using Connectionist Approaches", "Multi-State Time Delay Networks for Continuous Speech Recognition".
NG	"Speech Recognition: Statistical and Neural Information Processing Approaches", "Connectionist Approaches to the Use of Markov Models for Speech Recognition", "Connectionist Approaches to the Use of Markov Models for Speech Recognition".

Conclusion

- Utilized quality metrics based on Precision and Recall in unsupervised manner to generate new clustering quality metric Q.
- Q metric optimizes the model selection procedure of KSC method.
- Results in a Kernel Spectral Document Clustering technique.
- Showed the effectiveness of KSDC in comparison to k-means and NG method.
- KSDC method results in more balanced small-sized homogeneous clusters & prevent formation of degenerate large-sized heterogeneous clusters.
- Need to further analyze the results obtained from KSDC method.


D. Cutting, D. Karger, Pederson J., and J. Tukey.
Scatter-gather: A cluster-based approach to browsing large document collections.
In *Proc. of ACM SIGIR Conference*, 1992.


L. Baker and A. McCallum.
Distributional clustering of words for text classification.
In *Proc. of ACM SIGIR Conference*, 1998.


R. Bekkerman, R. El-Yaniv, Y. Winter, and N. Tishby.
On feature distributional clustering for text categorization.
In *Proc. of ACM SIGIR Conference*, 2001.


R. Xu and D. Wunsch.
Survey of clustering algorithms.
IEEE Transactions on Neural Networks and Learning Systems, 16(3):645–678, 2015.


A.Y. Ng, M.I. Jordan, and Y. Weiss.
On spectral clustering: analysis and an algorithm.
In *Proceedings of the Advances in Neural Information Processing Systems*, pages 849–856. MIT Press: Cambridge, MA, 2002.


J. Shi and J. Malik.
Normalized cuts and image segmentation.
IEEE Transactions on Pattern Analysis and Intelligence, 22(8):888–905, 2000.


U. von Luxburg.
A tutorial on spectral clustering.
Statistical Computing, 17:395–416, 2002.


W. Xu, X. Liu, and Y. Gong.
Document clustering based on non-negative matrix factorization.
In *Proc. of ACM SIGIR Conference*, 2003.


C. Alzate and J.A.K. Suykens.
Multiway spectral clustering with out-of-sample extensions through weighted kernel pca.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(2):335–347, 2010.


R. Mall, R. Langone, and Johan A.K. Suykens.
Kernel spectral clustering for big data networks.
Entropy (Special Issue: Big Data), 15(5):1567–1586, 2013.


R. Mall, R. Langone, and J.A.K. Suykens.
Multilevel hierarchical kernel spectral clustering for real-life large scale complex networks.
PLOS One, 9(6), 2014.


R. Kassab and J.C. Lamirel.
Feature based cluster validation for high dimensional data.
In *Proc. of IASTED International Conference on Artificial Intelligence and Applications (AIA)*, pages 97–103, 2008.


J.C. Lamirel, P. Cuxac, R. Mall, and G. Safi.
A new efficient and unbiased approach for clustering quality approach.
In *PAKDD 2011 Workshops*, pages 209–220, 2012.


R. Mall, R. Langone, and J.A.K. Suykens.
FURS: Fast and unique representative subset selection retaining large scale community structure.
Social Network Analysis and Mining, 3(4):1075–1095, 2013.


R. Mall, V. Jumutc, R. Langone, and J.A.K. Suykens.
Representative susbets for big data learning using *k*-nn graphs.
In *Proc. of IEEE Big Data*, pages 37–42, 2014.


R. Langone, R. Mall, and J.A.K. Suykens.
Clustering data over time using kernel spectral clustering with memory.
In *In Proc. of IEEE SSCI CIDM*, 2014.