Table 4: Average NGD of 10 Topics

| Topic | SeaNMF | NMF | LDA |
|---|---|---|---|
| 1 | 0.8905 | 0.6897 | 0.5529 |
| 2 | 0.8332 | 0.6397 | 0.3504 |
| 3 | 0.6167 | 0.4242 | 0.5014 |
| 4 | 0.5226 | 0.7585 | 0.5163 |
| 5 | 0.8643 | 0.5032 | 0.6776 |
| 6 | 0.6194 | 0.5933 | 0.8502 |
| 7 | 0.7359 | 0.4223 | 0.7262 |
| 8 | 0.6577 | 0.6186 | 0.7947 |
| 9 | 0.5372 | 0.7205 | 0.3189 |
| 10 | 0.5105 | 0.4663 | 0.6434 |
| **Average** | **0.6788** | **0.58363** | **0.5932** |

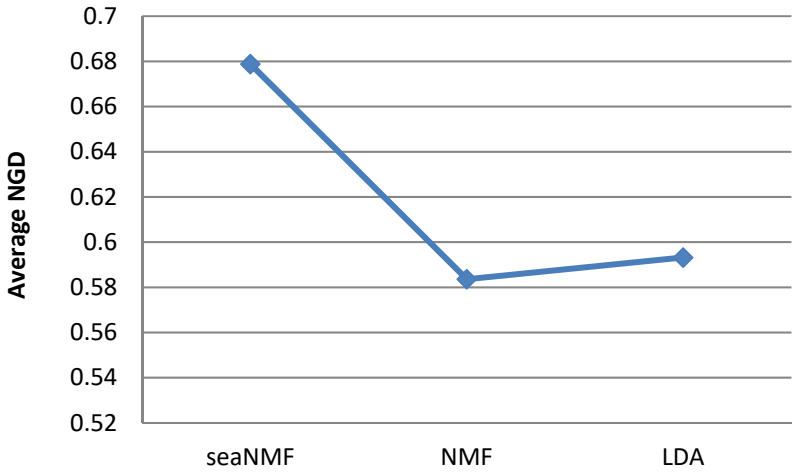


Fig 3: Average NGD for 10 Topics

The degree of correlativity of keywords computed is 67.88%, 58.6% ,59.32% for SeaNMF, NMF and LDA respectively. Fig. 3 shows that words clustered under each topic by SeaNMF are highly correlated.

## 4. Conclusions

Learning meaningful topics from short text is considered to be a challenge due to limited contextual information in it. This paper includes empirical study of three state-of- the-art methods of topic modeling. LDA is good for normal length text but not so for short text as it does not consider the relationships among keywords during topic discovery. NMF is a dimension reduction technique which yields clustering results based on the words in same region using term-document matrix whereas SeaNMF gives grouping of words using word-context semantic correlation matrix and skip-gram view of corpus that reveals word semantic association.SeaNMF outperforms NMF and LDA as it discovers more relevant topics from short text.

## References

[1] Alghamdi Rubayyi, Khalid Alfalqi, (2015), "A Survey of Topic Modeling in Text Mining", International Journal of Advanced Computer Science and Applications, vol. 6, no. 1, pp 147-153
[2] Alguliev Rasim, Aliguliyev Ramiz, Makrufa S Hajirahimova, Chingiz A Mehdiyev,(2011) "MCMR: Maximum Coverage and Minimum redundant text summarization model", Article in Expert Systems with Applications, Elsevier.
[3] Alhawarat M., Hegazi M.(2018), "Revisiting K-means and topic modeling, a comparison study to Cluster Arabic Documents", IEEE Access.
[4] Blei D, A. Ng, M. Jordan,(2003),"Latent Dirichlet Allocation", Journal of Machine Learning Research, 3: 993-1022
[5] Chhatbar Cirag Dilip, (2010), "Improving Statistical Topic Models by Using Ontological Concepts", COMP8740 Project Report, Dept., Computer Science, Australian National University.
[6] Choo Jaegul, Changhyun Lee, Chandan K. Reddy, Haesun Park,(2013), "Utopian:User-driven Topic Modeling based on Interactive Non-negative Matrix Factorization", IEEE transactions on Visualization and Computer Graphics, vol.19
[7] Choo Jaegul, Changhyun Lee, Chandan K. Reddy, Haesun Park, (2015) "Weakly supervised nonnegative matrix factorization for user-driven clustering", Data Mining and Knowledge Discovery, 1598–1621.
[8] Cilibrasi Rudi, Vitanyi Paul,(2001) "Automatic Meaning Discovery Using Google", BSIK/BRICKS project, Netherland
[9] Cohen Andrew R. and Paul M. B. Vitanyi,(2013), "Normalized Google Distance of Multisets with Applications", CoRR, abs|1308.3177
[10] Kim Jingu, Yunlong He,and Haesun park,(2014), "Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework", Journal of Global Optimization 58, 2(2014), 285-319
[11] Kuang Da, Jaegul Choo, and Haesun Park,(2015) "Nonnegative matrix factorization for interactive topic modeling and document clustering", In Partitional Clustering Algorithms, Springer, 215 - B
[12] Kulkarni Rohit,(2017), A Million News Headlines [CSV Data file], doi:10.7910/DVN/SYBGZL,Retrieved from: https://www.kaggle.com/therohk/million-headlines
[13] Levy Omer and Goldberg Yoav,(2014), "Neural Word Embedding as Implicit Matrix Factorization". In Advances in Neural Information Processing Systems 27. Curran Associates, Inc., 2177–2185.
[14] Likhitha S. , B. S. Harish, H. M. Keerthi Kumar,(2019), "A Detailed Survey on Topic Modeling for Document and Short Text Data", International Journal of Computer Applications, vol. 178, no. 39
[15] Milolov Tomas, Chen Kai, Corrado Greg, Dean Jeffrey,(2013), "Efficient Estimation of word representation in vector space", arXiv preprint arXiv:1301.3781
[16] Ramirez Eduardo H., Brena Ramon,(2011), "Topic Model Validation", Elsevier.
[17] Roder Michael, Both Andreas, (2015),s "Exploring the space of topic coherence Measures", ACM
[18] Shi Tian, Kyeongpil Kang , Jaegul Choo, Chandan Reddy, (2018),"Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations", ACM ISBN 978-1-4503-5639-8.
[19] Stevens Keith, Kegelmeyer Philip, (2012), "Exploring Topic Coherence over many models and many topics", proceeding of 2012 joint conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 952-961
[20] Dr. Vijayarani S., J. Ilamathi, Ms. Nithya , (2016)," Preprocessing Techniques for Text Mining-An Overview", International Journal of Computer Science & Communication Networks, vol 5.
[21] Yan Xiaohui, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, Yanfeng Wang,(2013), "Learning Topics in short Text by Non-negative Matrix Factorization on Term Correlation Matrix" , In Proceedings of the SIAM international Conference on Data Mining.
[22] Zuo Yuan, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, Hui Xiong, (2016),"Topic Modeling for Short Text: A Pseudo-Document View", KDD'16, ACM, August 13-17