Appendix: Experimental Setups of the Baseline Methods

We select three dimensionality reduction methods. After the dimensionality is reduced, ensembled K-means methods are used to obtain clusters. We remark that most manifold learning approaches, including those we've experimented, are transductive, scaling quadratically w.r.t. the sample size. This makes those approaches difficult to be applied in online or large-scale clustering settings. (3) *Spectral Clustering (Laplacian)*. We apply the Laplacian Eigenmaps (Belkin and Niyogi, 2001) with varying numbers of components to reduce the dimension of Tfidf vectors. Their cosine affinities are constructed based on nearest neighbors. (4) *Latent Semantic Indexing (LSI)* (Deerwester et al., 1990). We compute SVD of the Tfidf matrix, and choose varying numbers of components to form the dimension reduced vectors. (5) *Locality Preserving Projection (LPP)* (He and Niyogi, 2004; Cai et al., 2005). LPP is a popular document indexing method which produces low-dimensional representations. We follow the suggested setup in Cai *et al.* (Cai et al., 2005) and use the cosine affinity matrix.[2] We note that extra hyper-parameters are chosen from a pre-selected set empirically achieving the best performances.

We select two topic models. topic modeling techniques are originally developed to characterize documents with multiple topics, rather than cluster them into disjoint groups. Nevertheless, by assigning each document to its most significant topic, a clustering result can be obtained. We highlight that mixture-of-topics assumption that commonly utilized in topic modeling makes many of their approaches less sensitive to explore the homogeneity of clusters when increasing topics are estimated. (6) *Non-negative Matrix Factorization (NMF)* (Lee and Seung, 1999; Xu et al., 2003). We compute NMF of the Tfidf matrix by choosing K components, where K is the desired number of clusters. Documents (by row) are assigned to their largest column respectively in the factorization matrix to form clusters. (7) *Latent Dirichlet Allocation (LDA)* (Blei et al., 2003; Hoffman et al., 2010). Similar to NMF, we solve a LDA model of the word counting matrix by setting K topics. Because there are many hyper-parameters in a LDA model, our chosen set of hyper-parameters are set to achieve the low perplexity empirically. Adapting LDA to clustering, we naively group documents based on their most likely topics. Empirically, we find it works better than K-means of topic proportion vectors of documents.

Three methods based on word embeddings: we use four pre-trained word embeddings in our experiment to cross validate the effects of different pre-trained word embeddings. Three of them are trained on general large corpora, such as news articles and wikipedia pages. They are 300-dimensional SkipGram (Mikolov et al., 2013a) using negative sampling trained on GoogleNews, 300-dimensional Glove (Pennington et al., 2014) trained on Wikipedia corpus, standard 400-dimensional SkipGram trained on a 2010 Wikipedia dump[3] by our own (with window size of 10 and minimum count of 10). The first two can be downloaded publicly.[4] When compared with other non-embedding methods, best results of the three are reported.

We also use SkipGram to train domain-specific word embeddings using Ohsumed dataset (with window size of 20 and minimum count of 2), which is the fourth model. (8) *Average of word vectors (AvgDoc)* computes the average of the embeddings of distinctive words in a document. In practice, we find it outperforms one using weighted schema like Tfidf or Tf, especially for long texts. (9) *Paragraph Vectors (PV)* (Le and Mikolov, 2014). Two unsupervised methods, *i.e.* PVDM and PVDBOW, are proposed in Le and Mikolov (Le and Mikolov, 2014), in which pre-trained word vectors can be fine-tuned to obtain embeddings for documents. We have experimented with both methods PVDM and PVDBOW, and find PVDM performs significantly worse than PVDBOW on all datasets, thus only the results of PVDBOW are reported.

[2] http://www.cad.zju.edu.cn/home/dengcai/Data/DimensionReduction.html

[3] http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html

[4] https://code.google.com/p/word2vec/ http://nlp.stanford.edu/projects/glove/