BLK_read Description

1 Goal

The BLK generator has been implemented to create synthetic datasets that can be subject to evaluation by Graph Spectral Analysis algorithms.

2 Parameters

It is driven by the following parameters:

- group_count (set to 4) – the number of groups/classes to which the generated documents belong
- ext (set to 2)
- noDocs (set to 2000) – the number of documents that will be generated
- overlap (set to 0.20) – the extent to which the vocabulary of distinct classes shall overlap
- minprob (set to 0.5)
- noiseprob (set to 10.01)

The name of the generated dataset consists of the components:
"BLK.",group_count,"_",overlap,"_",minprob.

3 Assumptions

It is assumed that the vocabulary used is twice as large as the number of documents. Furthermore, it is assumed that each group uses a separate basic vocabulary (of cardinality gnw plus an overlap with the preceding group), subject to potential noise and overlaps with other groups. Each document contains the same basic number of words dnw (1/30th of gnw times minprob). dnw samples from dnw normal distributions (one from each) are taken to point the position in the dictionary from which the word is to be taken. The standard deviations are the same (1/12th of the group dictionary size), while the means are separated by the group dictionary size divided by the group id plus ext. In this way a kind of different literary styles are simulated: each group has a different number of

words at which it is focusing (the first: ext, the second ext+1 etc.). The idea of different literary styles was drawn from observation that different groups of people discuss different number of topics.

4 Noise

After the basic process of generating documents noise is added. The number of noisy points equals noiseprob times number of documents (hence noiseprob is not really a probability, but rather a factor). A noisy point is added by picking two documents and a word from the entire vocabulary. Then with probability of minprob a word is inserted into each of them (the probability is applied separately to both, so that a word is inserted in both at the same time with probability $minprob^2$).

5 Group sizes

The generator tries to assign nearly the same number of documents to each group.

6 Availability

The generator code is available as BLK_read.R program, in the directory R of `https://github.com/ipipan-barstar/PLOS.EGSCoTD`

See also [1], Section A of Appendix.

References

[1] Piotr Borkowski, Mieczysław A. Kłopotek, Bartłomiej Starosta, Sławomir T. Wierzchoń, and Marcin Sydow. Eigenvalue based spectral classification. *PLOS ONE*, 18(4):1–35, 04 2023.