A Multilevel Clustering Model for Coherent Topic Discovery in Short Texts

Emmanuel Maithya[1]

[1]Affiliation not available

May 22, 2020

Abstract

In the world today, huge volumes of data are generated in short text form. Many topics modelling techniques have been developed. Among these techniques is the popular Latent Dirichlet Allocation (LDA), however the effectiveness of this and other techniques has been shown to increase linearly with the document size while being less effective when modelling topics from short texts. Furthermore, the generated topics exhibit poor semantic coherence. We present an n-gram based multi-level clustering model for discovering coherent topics from short texts. By taking advantage of the natural arrangement of words in the n-grams in the topic form stage, the model is able to discover semantically coherent topics that are easily interpreted. n-grams are discovered recursively starting with larger n-grams down to n-grams whose length is governed by a pre-defined lower limit. The discovered n-grams are then subjected to a multi-level clustering algorithm, with the lowest clustering level being constituted from the shortest n-grams that occur most frequently in the entire corpus, and the highest level from least common n-grams at level n. We evaluate the model against the standard LDA and Bi-Term models by measuring and presenting the comparative coherence scores achieved by the topics generated from two datasets.

Hosted file

`Thesis - Emmanuel Muthoka - May 2020.docx` available at https://authorea.com/users/325331/articles/453298-a-multilevel-clustering-model-for-coherent-topic-discovery-in-short-texts