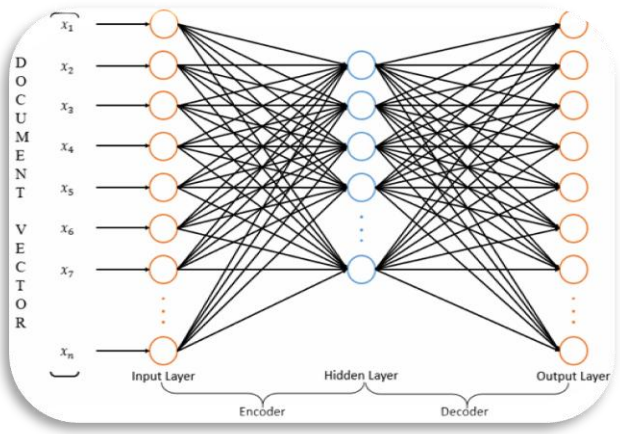
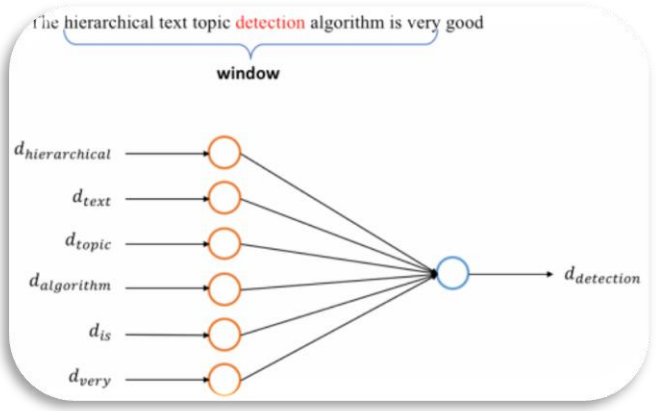
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
School of Computer Science and Engineering

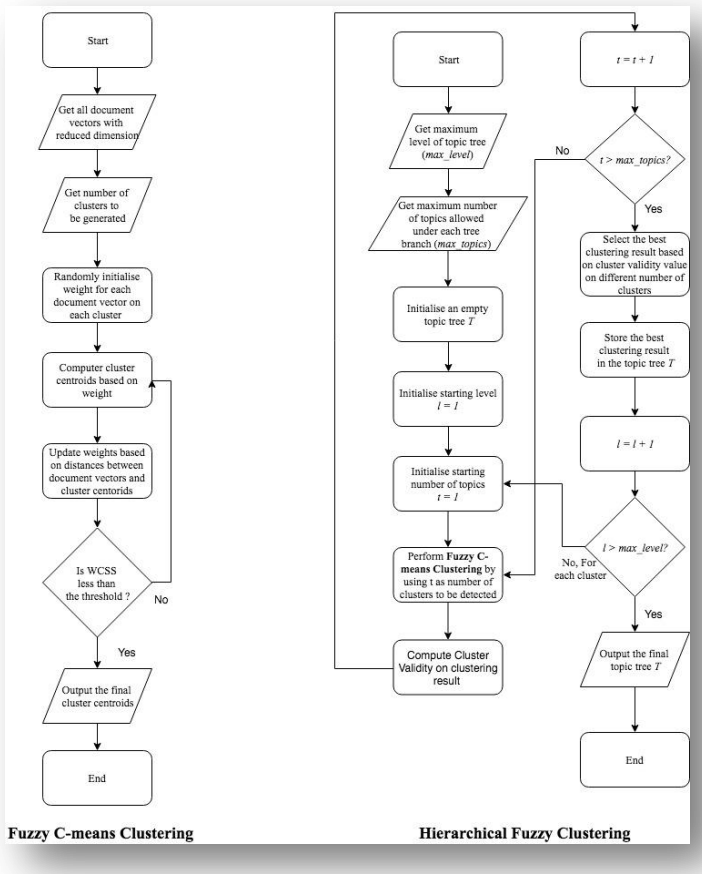
# Hierarchical Text Topic Detection From Text Data

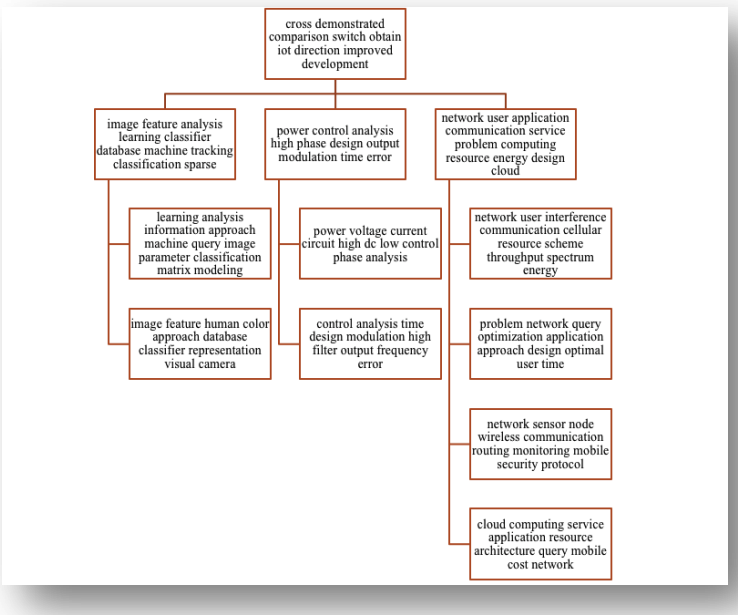



In an age of information boom, people get used to obtain information from online documents. However, useful information is not able to directly be extracted from most online documents due to the unstructured and unorganized nature of text. Therefore, to facilitate the searching and retrieval of information, documents are always labeled with tags, topics or categories. This paper presents with an efficient hierarchical topic detection algorithms which detect topics in a hierarchical manner from a large collection of documents. The algorithm starts with document processing which turns each document into a vector after a series of text preprocessing using some Natural Language Processing techniques. Since vectorized text data is normally very sparse, two dimensionality reduction techniques: Autoencoder and Word2Vec are applied respectively to reduce the data sparsity. The two figures shown on the left are simple illustration of how Autoencoder and Word2Vec work.

Hierarchical Fuzzy Clustering is then performed on the two sets of text vectors respectively to generate two hierarchical topic trees. After that, two topic trees are evaluated based on topic quality metrics such as association index and duplication index to select the best algorithm. The best algorithm is named as Hierarchical Fuzzy Topic Detection (HFTD). Finally, HFTD is tested on the text data of computer science papers from 1993 to 2017 on IEEE to obtain insights of computer science research topic changing trend. The flow chart on the right shows the basic flow of the hierarchical fuzzy clustering algorithm




At the end of HFC, a complete text topic tree is generated where each node on the tree is represented by a topic centroid. The centroid needs to be amplified back to the original dimension which is the dimension of the document vector right after tf-idf vectorization. Using the amplified topic vector, words with highest tf-idf value can be selected, and they can be used to represent this topic. Figure on the left shows an example on a 3-level topic tree generated by Autoencoder using IEEE 2017 paper data

SCE017-0153
Student: Chen Ziao
Email: B140007@e.ntu.edu.sg
Supervisor: Assoc Prof Adams Wai Kin Kong