Framework to Extract Context Vectors from Unstructured Data using Big Data Analytics

Department of Computer Engg. Jamia Millia Islamia, Delhi, India

Contributors: *Dr. Tanvir Ahmad, Rafeeq Ahmad, Farheen Nilofer, Sarah Masud*

Keywords: Big Data, Text Mining hadoop Analytics

QUICK LINKS:

- Paper: **http://ieeexplore.ieee.org/document/7880229/**
- Paper Code: **https://github.com/sara-02/hadoop_dump**
- Further Work In Progress: **https://github.com/sara-02/pylearn_spark**

PRESENTER: Sarah Masud

Alumni: **Jamia Millia Islamia**

Current Organization: **Red Hat**

Contact: **sarahmasud02@gmail.com**

MOTIVATION:

With an exponential increase in data, the process of information extraction becomes difficult. For text data this information is represented in form context vectors. The aim of this study is to examine and propose a framework for computing context vectors of large dimensions, trying to overcome the bottleneck of traditional systems.

Example:

Word: data

Context Vector:

<processing 0.6>,
<retrival 0.5>,
<warehouse 0.5>

PROPOSED SOLUTION:

Input: Text Corpus

Output: 2-d array of Concept to Concept Relation

Procedure:

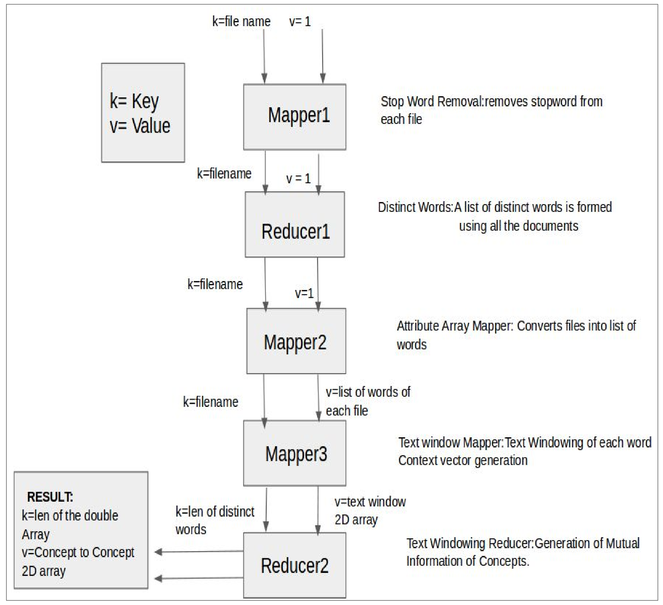
1. On a Mapper, for each document d ϵ D
  1.1 Select the term ti
  1.2 Remove if stop word

2. Reducer for Unique Word List from the corpus.
  2.1 Create unique attribute list.
  2.2 Create all attributes list.

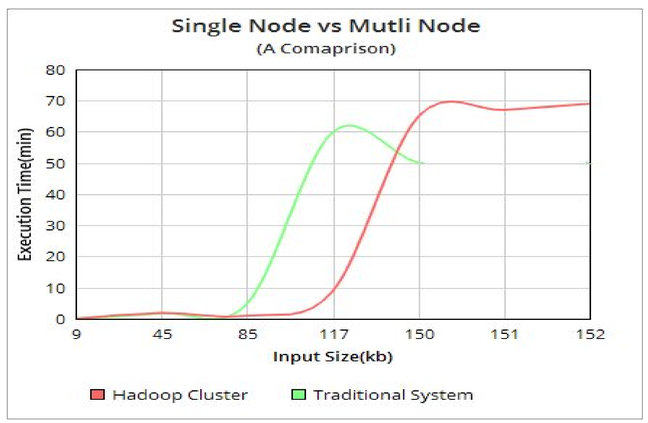
3. Chained Mapper for calculating the frequency of words that occur together in a window (Text windowing) stored in Attribute-Attribute matrix

4. Reducer for calculating mutual information from attribute-attribute matrix (Concept Extraction):
  4.1 $M.I = - \log_2(p(i\&j)) /(p(i) * p(j))$
  4.2 If $M.I$ > threshold then prob[i][j] = $M.I$

THE FRAMEWORK:



OBSERVATION:



FUTURE WORK :

- Experimenting with varying threshold.
- Benchmark performance with known algorithms.
- Extend this framework to leverage NLP and DL.

PRESENTED AT:

Grace Hopper Summit 2017
3rd-6th October