Report

**Design**

Terminology:

Counts:

* count(F): The total number of features.
* count(F = f): The number of occurrences of a specific feature f.
* count(L): The total number of lexemes.
* count(L = l): The number of occurrences of a specific lexeme l.
* count(F = f, L = l): The number of times the specific feature f appears with the specific lexeme l.

Intro

The system consists of four parts:

1. Step01, Step02 – Preprocessing: Filter the relevant lexemes and features.
2. Step1, Step2 – Corpus Statistics: Calculate count(F = f), count(L = l), and count(F = f, L = l).
3. Step3, Step4 – Algorithm Calculation: Measure association with context and compute vector similarity.
4. Step4 – Assessment: Evaluate the model's accuracy.

Steps

* Step 01: create a LexemeSet with the all lexemes in word-relatedness.txt.
* Step 02: create a DepLabelSet with the all dependencies label in the corpus.
* Step 1: calculates count(F=f) and count(L=l) at the corpus. Used for creating lexemeFeatureToCountMap.

Output: (Text feature/lexeme, LongWritable quantity).

* Step 2: for each lexeme presented in both the corpus and word-relatedness.txt, calculates a vector of counts(F=f,L=l). The step uses TreeMap to create a lexicographically ordered map, ensuring a consistent structure for all lexeme vectors.

Output: (Text lexeme, Text spaces\_separated\_counts(F=f, L=l))

* Step 3: measure association with the context and create four vectors, one for each association method.

Output: (Text lexeme, Text v5:v6:v7:v8, vi is space separated vector.

* Step 4: using fuzzy join, for each lexemes pair, create a 24-dimensional vector that measures vector similarity (distance) using six distance measure methods. Output: (Text lexeme, Text paces\_separated\_vector)
* Step 5: (Not part of the MapReduce pattern) Using Weka to assess the model's accuracy.

**Communication:**

* Map output records: This counter indicates the total number of key-value pairs emitted by the mappers. In your log, it shows:​

Step 01: Map output records=29094

Step02: Map output records=617726426

Step 1: Map output records= 85196780

Step 2: Map output records= 71781585

Step 3: Map output records=667

Step 4: Map output records=506920

* Map output bytes: This counter represents the total size (in bytes) of all key-value pairs emitted by the mappers before any compression. Your log shows:​

Step 01: Map output bytes=29094

Step02: Map output bytes=3841626296

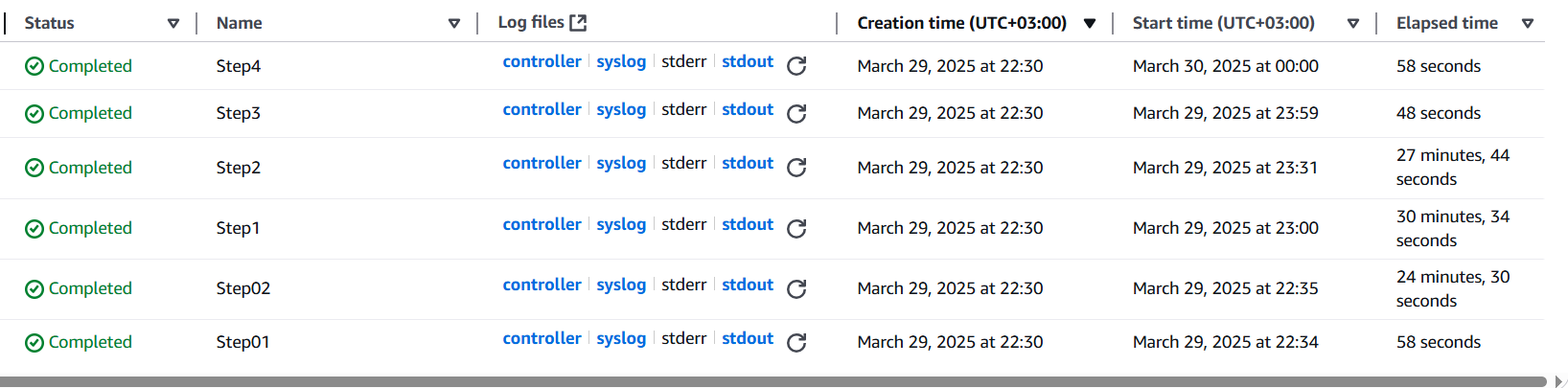
Step 1: Map output bytes= 1364364966

Step 2: Map output bytes= 1468179361

Step 3: Map output bytes= 352586

Step 4: Map output bytes=271650535

**Results:**



Evaluating:

10 NGRAM Files (Classifier - trees J48):

