

## EDAMI TAXONOMY PROJECT

Task Ruben Romero:

Write a C++ program to apply a taxonomy to a specific data set.

Description:

1-Create taxonomies:

To create and organize the taxonomies the program uses a TDA N-AryTree. It TDA implements :

- Insert node
- Remove node:
- Remove sub tree:
- Search: Search is the most important method of this taxonomy for this program due to it will be use in each items of the data sets to know if we need to add any new items. So to get the max efficient the tree will be balance, and we will get the search in  $O(\log(n))$

2-Read, create and manipulate a data set.

To fulfill thi part, the program use a class DataSet that implement the next function:

- Read data set. Read from a file with the correct format
- Save data set. Save in a file with the correct format
- Apply taxonomy. Modified the data set adding the new items in base of the given taxonomy

3-Inputs.

The program will ask for a taxonomy and a data set with the next formats.

Taxonomy.

Level 0:\*root\*.

.  
.   
.   
.   
.

Level i:\*root\*->( \*child\_1\*, ..., \*child\_i\*,..... \*child\_n\*)

.

•  
•

Level n: \*child\_n1\*->>(\*child\_n1\_1\*) , \*child\_n2\*->>(\*child\_ni\_1\*)

Example:

		primero	
dos	tres	cuatro	cinco
seis	siete	ocho	nueve

Delimiter '-'

\$

-primero-

primero->( -dos-, -tres-, -cuatro-, -cinco- )

dos->( -seis- ), tres->( -cuatro- ), cuatro->( -ocho- ), cinco->( -nueve- )

Data set:

sequenceID eventID SIZE items  
x y s item1 ... itemi .... itemN

Example:

sequenceID eventID SIZE items  
2 1 5 design Webdesign inspiration icons images c++  
3 1 3 ubuntu linux util tools  
4 1 1 Apps Facebook  
5 1 1 board wordpress

4-Program:

1-Read Taxonomy.

2-Modified Taxonomy. Ask to the user for a modification of the taxonomy

3-Read DataSet.

#### 4-Modified DataSet with taxonomy.

##### 5-Efficient:

The main and heaviest function in the program is applyTaxonomy, this function read the item of a data set in memory and apply the taxonomy. Its deficient is the next:

n->size of the data set.

m->max size of a group of items in a row.

$O(\log(m)*n)$ : the cost of seach every items in a row times the number of rows.