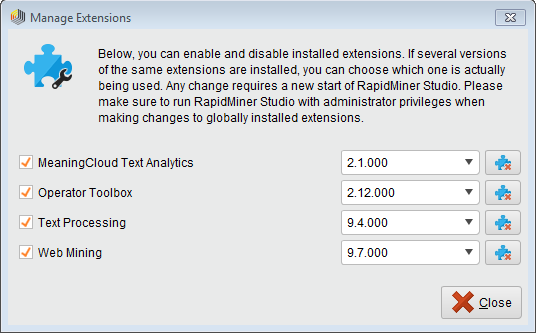
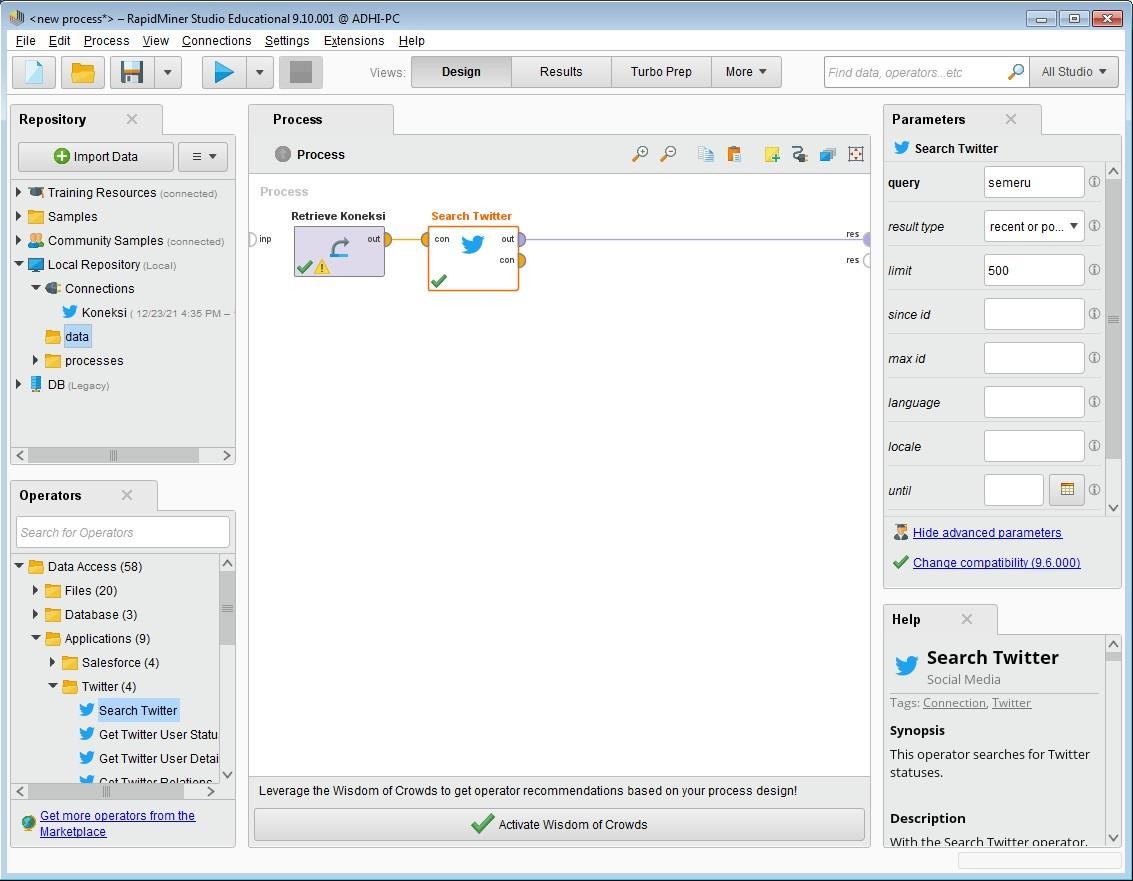
TUGAS 10 BIG DATA TF&IDF

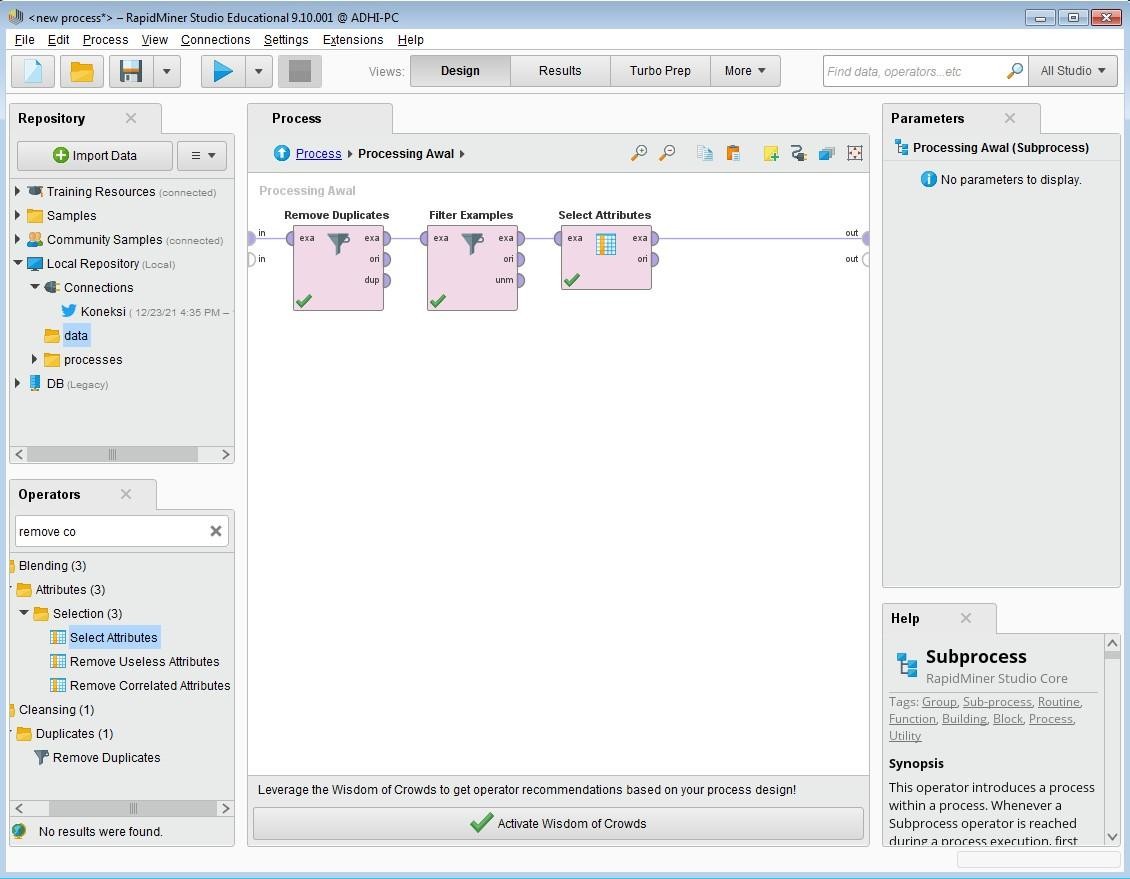
Nama Kelompok : 5200411128 Dava Adhi Yodya Pradipta

5200411195 Yesi Nurhana Dalimunthe 5200411398 Bilal Nurul Fauzi

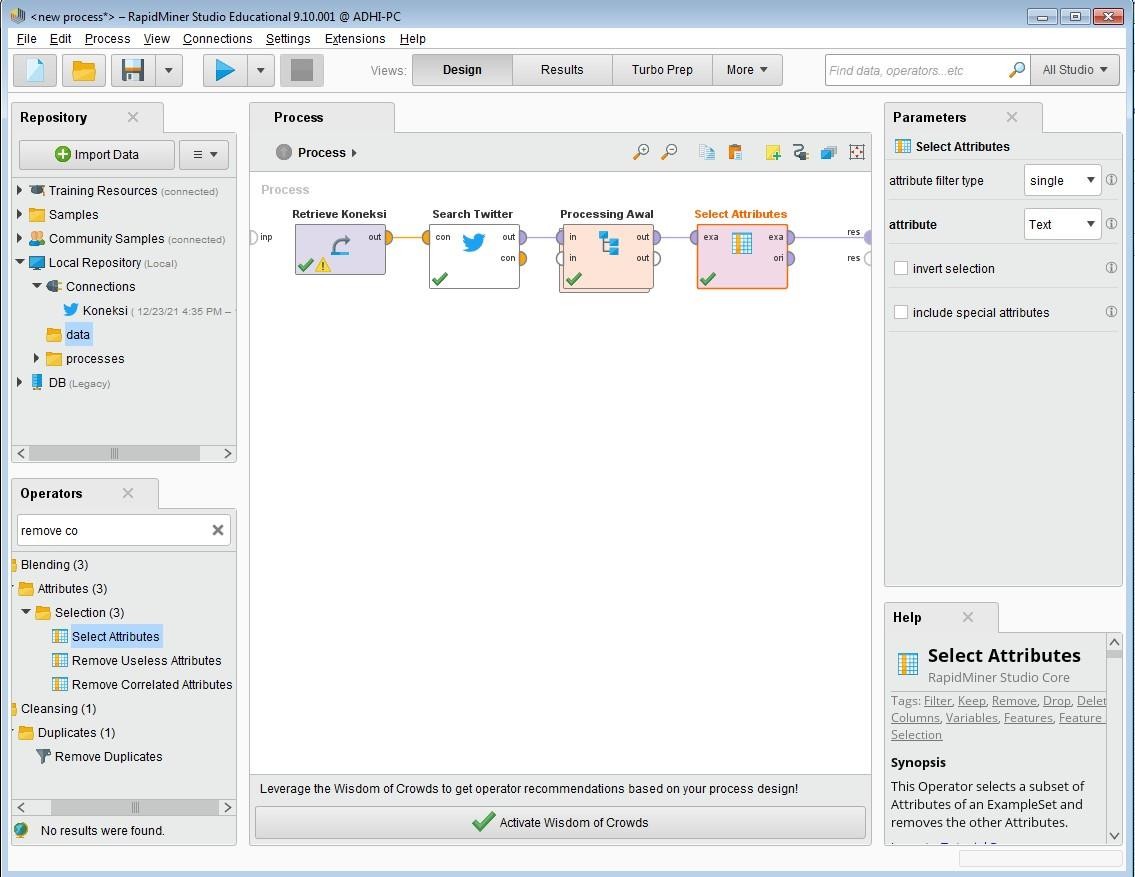
1. Download dan instalasi extensions untuk melakukan proses intallasi operator toolbox dan text processing di rapid miner yang akan digunakan
2. Langkah selanjutnya masukan data dari twitter dengan dragdown dan lakukan juga pada koneksi untuk operator search twitter, disini saya mengambil data 500 dan set query lokasi connection



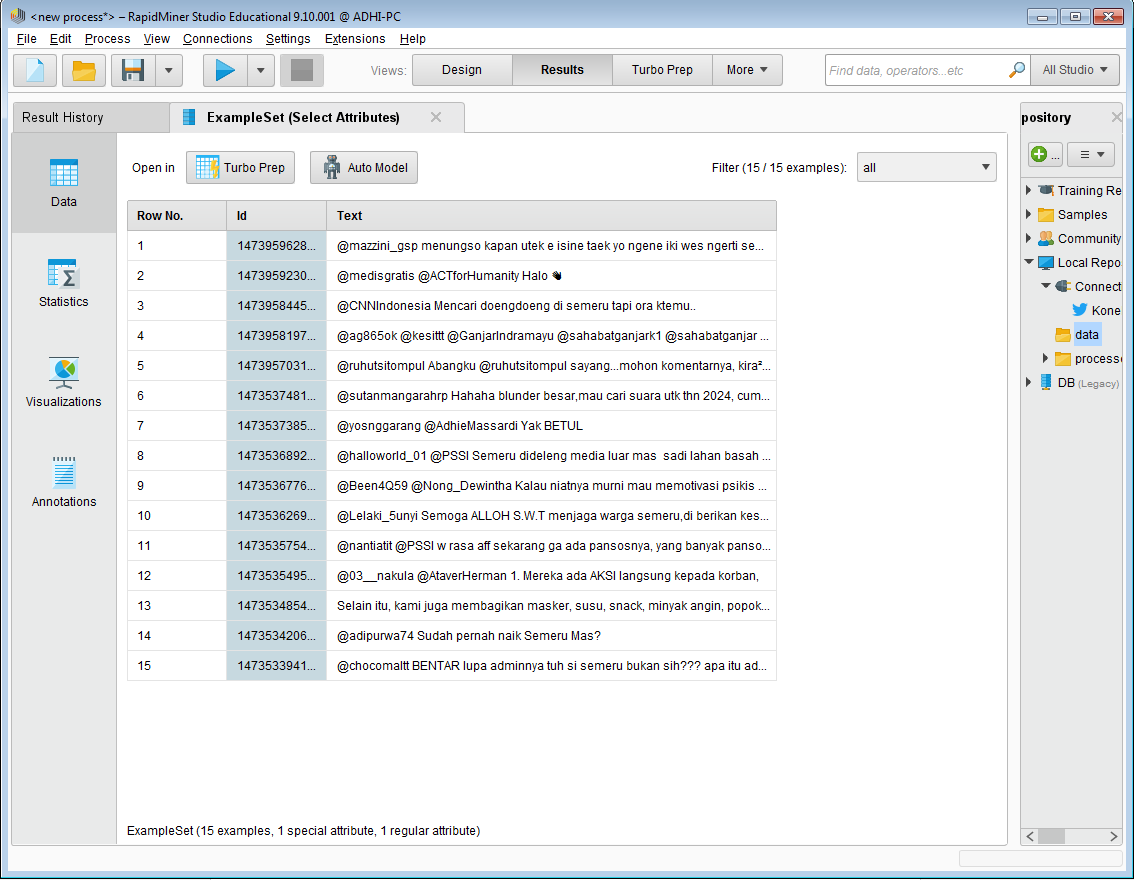
1. Lakukan processing awal dengan menambahkan operator
   * Remove duplicate (untuk menghapus data yang ganda)
   * Filter example (untuk menghilangkan data yang missing value)



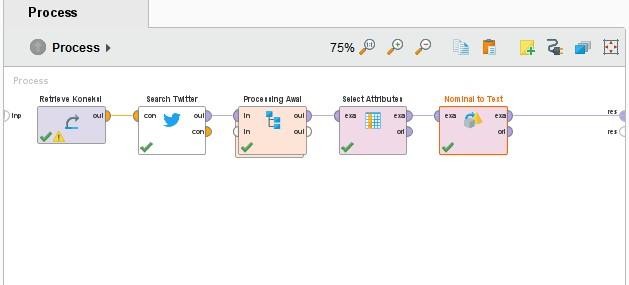
1. Setelah itu lakukan select atribut, disini operator select atribut digunakan untuk memilih atribut yang akan digunakan pada proses ini, yaitu type single dan atribut text.



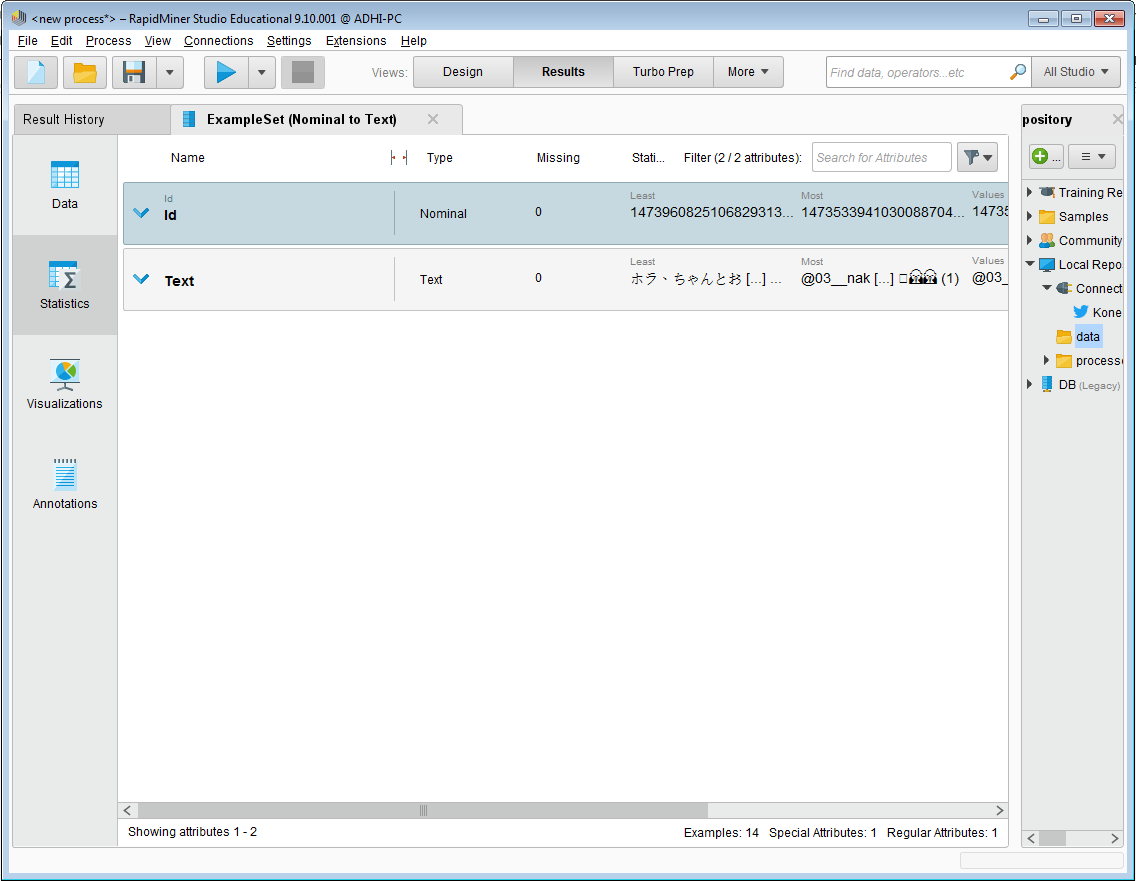
Ini tampilan data setelah melakukan proses dengan operator select attribute



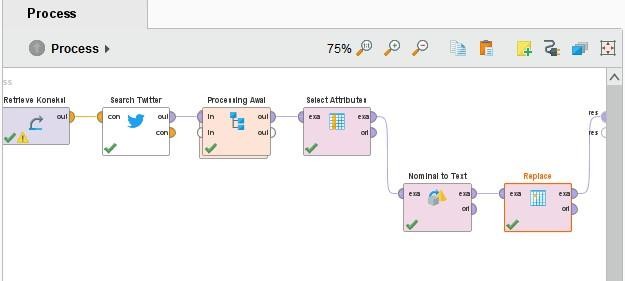
1. Langkah selanjutnya menentukan type data yang akan kita gunakan pada proses ini, disini kita type text sehingga harus merubah type pada data menjadi nominal to text, sehingga bisa melakukan text processing



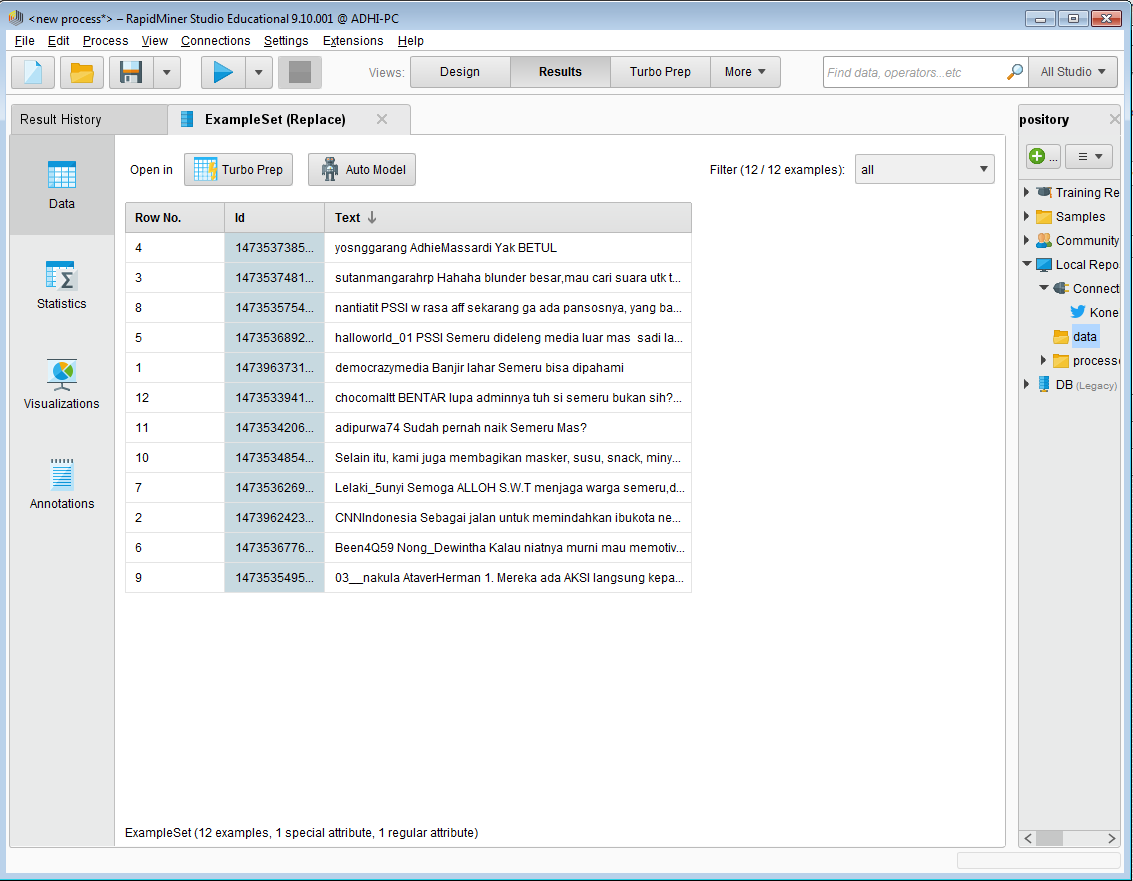
Ini tampilan setelah melakukan proses nominal to text

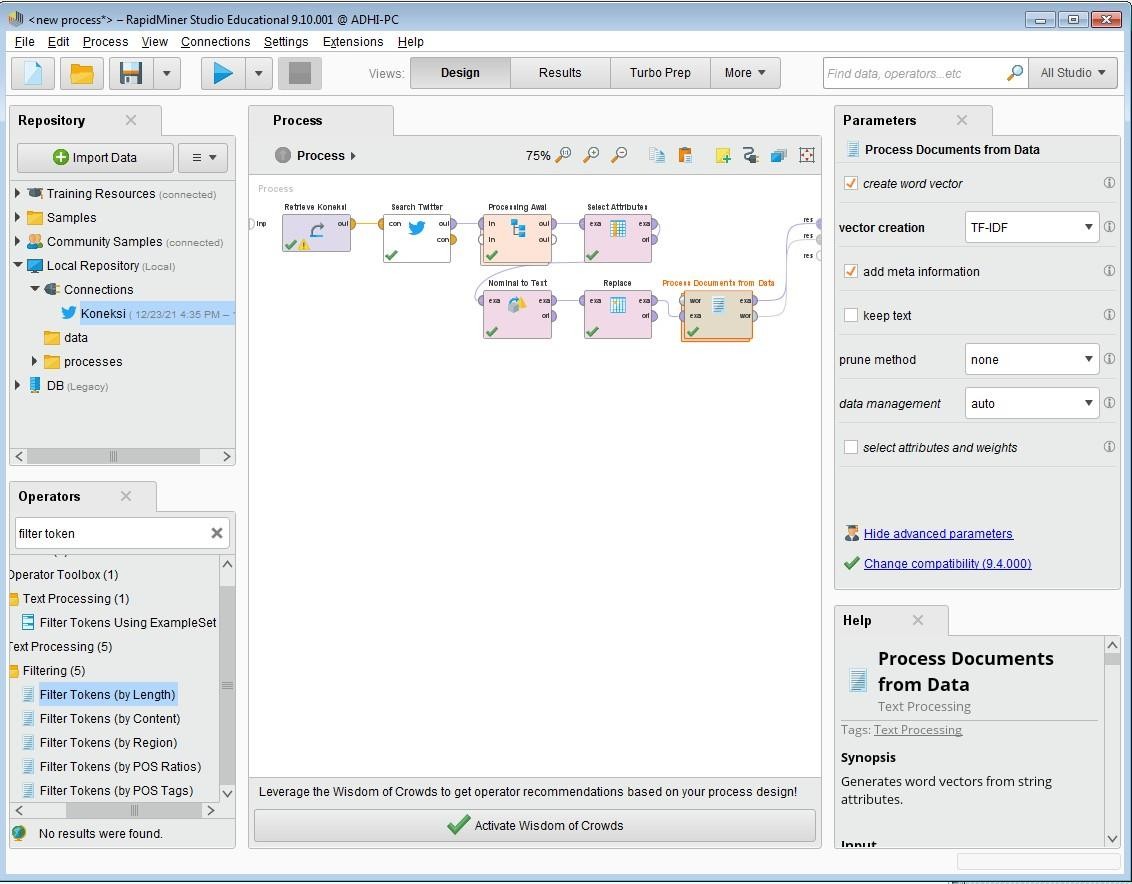


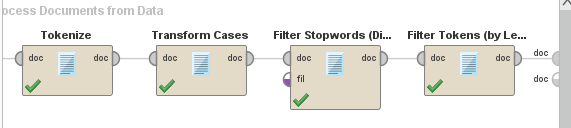
1. Kemudian kita lakukan proses dengan operator replace untuk menghilangkan kata yang tidak penting seperti tanda (@) dengan mengset parameter, type parameter all dan replace attribute @\*



Tampilan setalah melakuka proses replace



1. Langkah selanjutnya melakuka document from data yang didalamnya terdapat:
   * Tokenize yang berfungsi untuk membagi data menjadi beberapa kata
   * Operator selanjutnya adalah transform cases yang akan digunkan untuk merubah ke lower cases
   * Untuk memfilter data dengan menggunakan operator stopwords
   * Yang terakhir kita filter token by len untuk menghilangkan kata” yang tidak di perlukan



Ini tampilan akhir atau hasil dari text processing berdasarkan seberapa banyak kata tersebut terpakai

