Algoritma pemisah tanda baca :

Input berupa array list kata. Dilakukan pengecekan untuk tiap kata, jika kata tersebut mengandung karakter non alfa numerik maka tiap karakter pada kata akan dicek satu-persatu. Dilakukan pemisahan kata menjadi beberapa bagian berdasarkan karakter non alfa numerik tersebut. Output berupa array list di mana kata dan karakter alfanumerik sudah terpisah.

Algoritma tokenisasi :

Input berupa array list yang dapat berupa kata atau karakter non alfanumerik. Tiap elemen pada array list diubah ke huruf kecil terlebih dahulu. Dilakukan pengecekan untuk setiap elemen dan elemen berikutnya pada array list. Jika elemen saat ini dan elemen berikutnya berada di daftar kata majemuk, maka kedua kata tersebut dianggap sebagai satu token dan digabungkan. Daftar kata majemuk disimpan di file teks frasa, yang dipanggil berdasarkan huruf awal pada elemen. Output berupa array list yang tiap itemnya dapat berupa kata dasar atau kata majemuk/frase.