**LAPORAN TUGAS PROGRAM 2**

**POSTagger Bahasa Indonesia**

****

**Disusun Oleh**

**1301174385**

**Adriansyah Dwi Rendragraha**

**ICM-40-GAB**

**PROGRAM STUDI S1 INFORMATIKA**

**FAKULTAS INFORMATIKA**

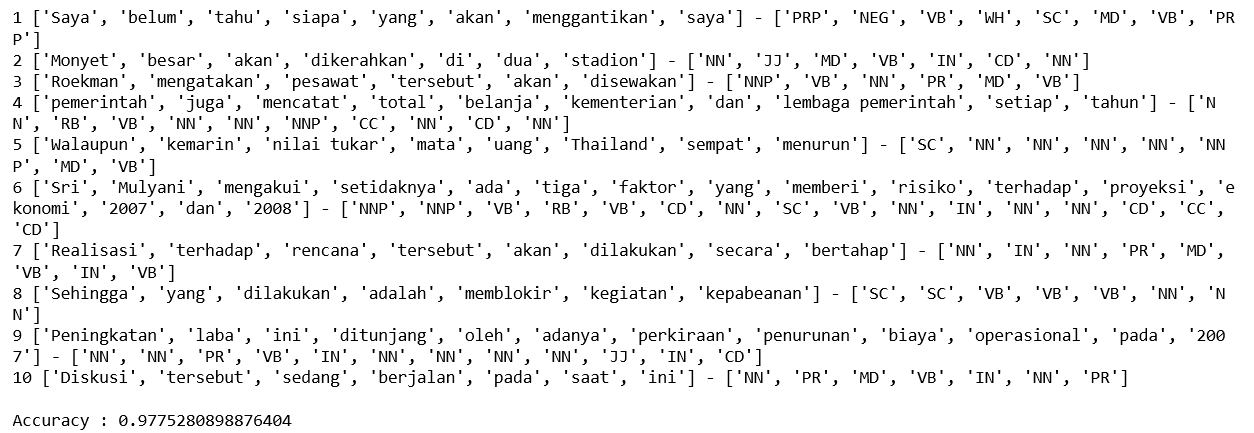
**UNIVERSITAS TELKOM**

**BANDUNG**

Dataset yang digunakan dalam tugas program POSTagger ini diambil dari website <http://bahasa.cs.ui.ac.id/postag/corpus>. Saya juga menyediakan 10 buah kalimat di file ‘Testing.txt’ dimana nantinya digunakan untuk menguji ke 3 model POSTagger yang sudah dibuat dan juga untuk mengukur nilai akurasinya. Dataset tersebut digunakan untuk men-*training* model yang sudah ditentukan, dimana model nya terdiri dari 3 model, yaitu:

1. **Metode Baseline**

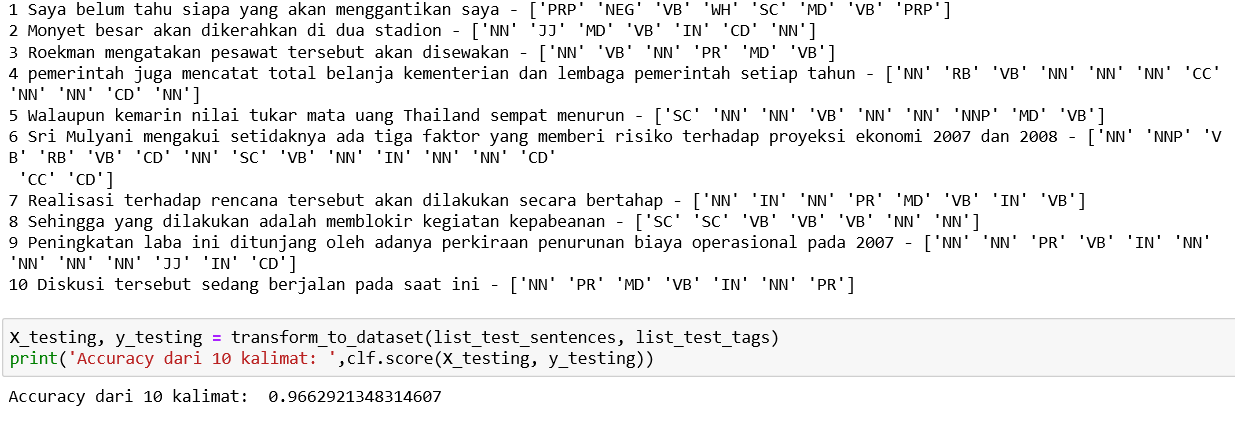
Pada metode ini, menggunakan frekuensi tag pada sebuah kata yang paling sering ditemui pada corpus yang telah di *training*. Jadi jika kata ‘saya’ memiliki frekuensi tag ‘PRP’ yang paling sering muncul, maka tag dari kata ‘saya’ adalah ‘PRP’. Hasil pengujiannya ke data testing dan akurasinya dapat dilihat sebagai berikut.



Pada hasil pengujiannya ini didapati akurasi sebesar 97.75% dimana ini cukup besar akurasinya dengan menggunakan metode baseline. Hal ini terjadi mungkin dikarenakan menggunakan corpus dengan data yang banyak. Sehingga frekuensi kebenaran POSTag nya menjadi lebih akurat.

1. **Metode Statistika**

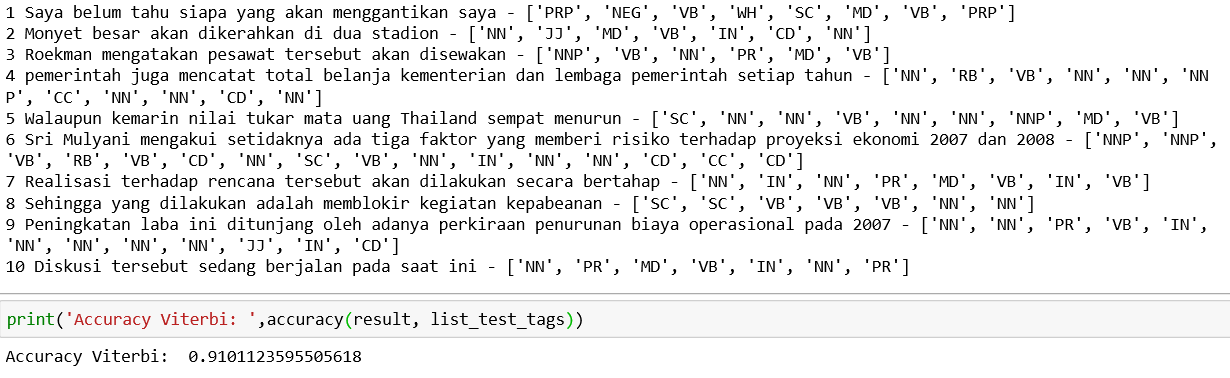
Pada metode ini, persoalan POSTagging diselesaikan sebagai suatu persoalan klasifikasi. Pada kasus ini saya menggunakan Decision Tree, dimana decision tree merupakan salah satu metode klasifikasi pada POSTag. Dalam metodenya, dapat dilakukan pengambilan fitur yang berfungsi untuk memastikan hasil nya. Dataset di metode ini dibagi menjadi dua, yaitu data train dan data test dengan presentasi (75%-25%). Dikarenakan untuk data train nya masih terlalu banyak, untuk training nya saya membuat menjadi 4 batch agar tidak terjadi out of memory pada saat melakuka training. Setelah selesai saya menyimpan modelnya supaya jika ingin menjalankan metode ini tinggal mengload model yang sudah disimpan. Hasil pengujiannya ke data testing dan akurasinya dapat dilihat sebagai berikut.



Pada hasil pengujiannya ini didapati akurasi sebesar 96.629%, terlihat hasilnya juga sudah baik. Hal yang berperan dalam pengujian adalah penggunaan fitur, pemilihan classifiern dan vectorizer yang digunakan dalam proses training dari data tersebut.

1. **Metode HMM-Viterbi**

Hidden Markov Model adalah sebuah pemodelan sekuens yang dimana statenya tidak dapat diamati secara langsung (tersemunyi), tetapi dapat diobesrvasi melalui suatu himpunan pengamanatan. Pada metode ini, saya diharuskan mencari trans\_probabilities dan emission\_probabilities untuk setiap kata dan POSTag nya. Kemudian mencari trans\_probabilities dan emission\_probabilities untuk kata dari 10 kalimat tersebut, yang nanti nya digunakan untuk pencarian POSTag seharusnya. Hasil pengujiannya ke data testing dan akurasinya dapat dilihat sebagai berikut.



Pada hasil pengujiannya ini didapati akurasi sebesar 91.01%. Terlihat dari kedua metode sebelumnya, akurasi di metode ini memiliki akurasi yang terkecil. Disini saya belum bisa memberikan kesimpulan pasti mengenai model HMM-Viterbi, karena saya masih belum begitu paham untuk masalah HMM-Viterbi ini, terutama dikasus *backtrace*-nya.

Kesimpulan dari tugas programming 2 POSTagger saya ini adalah metode baseline memiliki tingkat akurasi yang paling tinggi dari ketiga metodenya, banyak nya data latih yang digunakan dapat mempengaruhi hasil pengujian dari masing masing metode. Pada metode statistik, mungkin bisa lebih baik lagi jika mencoba dengan berbagai macam fiturnya, vectorizer, dan classifier. Untuk metode HMM-Viterbi, kemungkinan kesalahan nya terjadi saat kalkulasi nilai sehingga mempengaruhi POSTag yang didapat.