Laporan Analisis POSTagging

1. Pendahuluan

Part of Speech (POS), umum disebut juga sebagai kelas kata, atau kategori sintaktik. Part of Speech memberikan informasi tambahan untuk sebuah kata dan tetangga-tetangganya.

2. Pembangunan Sistem

-Membaca File tsv sebagai data latih

```
colnames = ['Word','Tag']
tsv_read = pd.read_csv('train.01.tsv', sep='\t',names=colnames,header=None)
tsv_read = tsv_read.astype(str)
tsv_read.head()
```

-Membaca File tsv sebagai data uji

```
colnames = ['Word','Tag']
tsv_read_test = pd.read_csv("test_sentences.tsv",sep="\t", names=colnames, header=None)
tsv_read_test = tsv_read_test.astype(str)
tsv_read_test.head()
```

3. Hasil dan Analisis

1. -Hasil tagging kalimat uji berdasarkan metode baseline

```
['NNP', 'NN', 'NNP', 'VB', 'VB', 'NNP', 'Z', 'NNP', 'NNP', 'VB', 'NN', 'IN',
'NN', 'CD', 'Z', 'NN', 'JJ', 'IN', 'VB', 'NN', 'Z', 'NN', 'JJ', 'PR', 'VB',
'NN', 'NN', 'NN', 'Z', 'NNP', 'NN', 'VB', 'NN', 'NN', 'Z', 'IN', 'NN', 'NNP',
'NN', 'Z', 'NN', 'VB', 'Z', 'NN', 'NN', 'NN', 'NN', 'CD', 'VB', 'Z', 'NNP',
'VB', 'NN', 'MD', 'VB', 'VB', 'NN', 'CC', 'Z', 'NN', 'NN', 'VB', 'IN', 'NN',
'NN', 'SC', 'VB', 'Z', 'NN', 'IN', 'NN', 'NEG', 'MD', 'VB', 'NN', 'JJ', 'Z']
```

-Hasil tagging kalimat uji berdasarkan metode classification

```
['IN', 'NN', 'NNP', 'VB', 'VB', 'NNP', 'Z', 'CD', 'NNP', 'VB', 'NN', 'IN',
'NN', 'CD', 'Z', 'NN', 'JJ', 'IN', 'VB', 'NN', 'Z', 'NN', 'JJ', 'PR', 'VB',
'NN', 'NN', 'NN', 'Z', 'IN', 'NN', 'VB', 'NN', 'NN', 'Z', 'IN', 'NN', 'NNP',
'NN', 'Z', 'NN', 'VB', 'Z', 'NN', 'NN', 'NN', 'NN', 'CD', 'NNP', 'Z', 'IN',
'VB', 'NN', 'MD', 'VB', 'VB', 'NN', 'CC', 'Z', 'NN', 'NN', 'VB', 'IN', 'NN',
'NN', 'SC', 'VB', 'Z', 'NN', 'IN', 'NN', 'NEG', 'MD', 'VB', 'NN', 'JJ', 'Z']
```

-Hasil tagging kalimat uji berdasarkan metode HMM

```
['NNP', 'NNP', 'NNP', 'VB', 'VB', 'NNP', 'Z', 'NNP', 'NNP', 'VB', 'NN',
'<start>', 'NN', 'CD', 'Z', 'NN', 'JJ', 'NN', 'VB', 'NN', 'Z', 'NNP', 'JJ',
'PR', 'VB', 'NN', 'NN', 'NN', 'Z', 'NNP', 'NNP', 'VB', 'NN', 'NN', 'Z',
'<start>', 'NN', 'NNP', 'NNP', '<start>', 'NN', 'VB', 'Z', 'NN', 'NN', 'NN',
'NN', 'CD', 'VB', 'Z', 'NNP', 'VB', 'NN', '<start>', 'VB', 'VB', 'NN', 'CC',
'Z', 'NN', 'NN', 'VB', '<start>', 'NN', 'NN', '<start>', 'VB', 'Z', 'NN',
'<start>', 'NN', 'NEG', '<start>', 'VB', 'NN', 'JJ', 'Z']
```

---

2. - Akurasi berdasarkan metode baseline

```
Akurasi dengan metode Baseline:  0.948051948051948
```

- Akurasi berdasarkan metode classification

```
Akurasi dengan metode Classification: 0.8441558441558441
```

-Akurasi berdasarkan metode HMM

```
Akurasi dengan metode HMM Viterbi :  0.8311688311688312
```

Berdasarkan hasil yang didapat dari 3 metode yaitu baseline, classification dan HMM yaitu dengan metode baseline dengan akurasi tertinggi mencapai 94%, dengan metode classification sebesar 84% dan akurasi paling rendah menggunakan HMM sebesar 83% . Akurasi dengan metode baseline tertinggi karena pada metode baseline kalimat uji akan dicari dengan most-frequent tag pada kata tersebut dan tag *unknown words* dengan tag yang paling sering muncul, jika terdapat kata yang sama pada data latih, maka tag tersebut akan di assign pada most-frequent tag, jika menggunakan metode classification metode yang digunakan adalah decision tree terdapat kesalah pada aturan yang ambigu sehingga akurasi yang didapat masih dibawah metode baseline, sedangkan dengan menggunakan metode HMM akurasi yang paling rendah sebesar 83% dikarenakan saat proses decoding dan backtracking sehingga masih adanya nilai 0 menyebabkan kesalahan pada hasil tagging