Jovita Nurvania (1301174215)

LAPORAN TUGAS NLP PEKAN 5 – POSTAGGER

I. Deskripsi Masalah

POS Tagger adalah suatu cara untuk mengkategorikan kelas kata menjadi beberapa kategori seperti kata benda, kata kerja, kata sifat, dll. POS atau *part-of-speech* memberikan informasi tambahan pada sebuah kata dengan tetangga-tetangganya. Untuk POS Tagger sendiri yang dilakukan pada penelitian kali ini adalah dengan tiga pendekatan. Pertama, pendekatan *baseline*, yaitu tag tiap katanya dihitung untuk mencari tag paling sering muncul dengan itu akan mengetahui tag dari kata *unknown*, pendekatan ini memiliki nilai akurasi yang tinggi yaitu bisa melebihi 90%. Kedua, pendekatan *classification*, dengan mencari prefix dan suffix dari setiap kata yang ada. Ketiga, pendekatan HMM Viterbi, dengan memodelkan probabilitas sekuens variable secara acak, dimana dengan pendekatan ini dapat memprediksi apa yang akan terjadi selanjutnya.

II. Penyelesaian Masalah

Pada penelitian ini, dilakukan pelatihan pada sebuah file tsv yang berisi ribuan kata dengan tag nya masing-masing, dimana kata-kata tersebut jika disusun dapat menjadi beberapa kalimat. Untuk penelitian ini, yang diambil hanyalah 50 kalimat pertamanya untuk menjadi data latih. Setelah itu, frekuensi kata dengan tagnya dihitung frekuensinya masing-masing, dan ditampilkan juga tag apa yang paling sering muncul dari data latih tersebut dengan fungsi yang ada pada kode. Setelah itu baru dimasukkan data uji yang berupa tsv untuk dilakukan pengujian. Data uji sama dengan data latih berisi kata dengan tagnya yang benar. Tetapi, untuk menguji keakuratan metode ini, tag pada setiap kata dihilangkan terlebih dahulu.

Jika setiap kata sudah tidak memiliki tag lagi, yang dilakukan adalah pengujian pada data tersebut. Dari pengujian tersebut, maka akan muncul tag hasilnya sesuai dengan data latih yang ada. Jika kata tersebut tidak ada pada data latih, maka akan muncul keterangan bahwa kata tidak ada di data latih dan tag hasil tidak dapat ditampilkan. Tahap terakhir pada pendekatan ini adalah menghitung akurasi, yaitu dengan membandingkan tag hasil dari proses data uji dengan data latih dan tag asli dari data uji tersebut. Pendekatan *baseline* memiliki nilai akurasi 0,9481. Sesuai dengan deskripsi sebelumnya, bahwa pendekatan ini memiliki nilai akurasi diatas 90%.

Pendekatan kedua adalah *classification*, hal pertama yang dilakukan adalah sama seperti pendekatan *baseline.* Bedanya, pada pendekatan ini data latih tidak langsung ditrain melainkan dipisahkan dahulu dengan fungsi features antara prefix dan suffixnya. Kemudian setelah fungsi dibangun baru diaplikasikan pada data latih untuk dilatih dengan fungsi tersebut. Masuk pada data uji, data uji berupa tsv juga diinputkan ke program dengan cara yang sama dengan sebelumnya. Data uji pun juga melalui fungsi yang sama hanya datanya saja yang berbeda. Lalu setelah diuji maka dihitung nilai akurasinya, pendekatan ini memiliki nilai akurasi 0,8312. Dimana nilai ini lebih kecil dari pendekatan *baseline.*

III. Analisis Hasil

Hasil tagging dengan pendekatan *baseline :*

```
3] kata: menteri , tag: NNP
   kata: pertahanan , tag: NN
   kata: as , tag: NNP
   kata: dijadwalkan , tag: VB
   kata: mengunjungi , tag: VB
   kata: india , tag: NNP
   kata: . , tag: Z
   kata: tata , tag: NNP
   kata: power , tag: NNP
   kata: menyuplai , tag: VB
   kata: batu bara , tag: NN
   kata: pada , tag: IN
   kata: tahun , tag: NN
   kata: 2000 , tag: CD
   kata: . , tag: Z
   kata: pemerintah , tag: NN
   kata: hati-hati , tag: JJ
   kata: dalam , tag: IN
   kata: mengelola , tag: VB
   kata: bumn , tag: NN
   kata: . , tag: Z
```

```
kata: perusahaan , tag: NN
kata: baru , tag: JJ
kata: tersebut , tag: PR
kata: mencanangkan , tag: VB
kata: target , tag: NN
kata: perolehan , tag: NN
kata: laba bersih , tag: NN
kata: . , tag: Z
kata: menteri , tag: NNP
kata: pertahanan , tag: NN
kata: mengunjungi , tag: VB
kata: pangkalan , tag: NN
kata: udara , tag: NN
kata: . , tag: Z
kata: menurut , tag: IN
kata: laporan , tag: NN
kata: sekretaris , tag: NNP
kata: perusahaan , tag: NN
kata: , , tag: Z
kata: laba bersih , tag: NN
```

```
kata: transaksi , tag: NN
kata: penjualan , tag: NN
kata: barang mewah , tag: NN
kata: tahun , tag: NN
kata: 2007 , tag: CD
kata: turun , tag: VB
kata: . , tag: Z
kata: menkeu , tag: NNP
kata: memperkirakan , tag: VB
kata: inflasi , tag: NN
kata: akan , tag: MD
kata: meningkat , tag: VB
kata: dibanding , tag: VB
kata: tahun , tag: NN
kata: lalu , tag: CC
kata: . , tag: Z
kata: kenaikan , tag: NN
kata: tarif , tag: NN
kata: didorong , tag: VB
kata: oleh , tag: IN
kata: target , tag: NN
kata: laba bersih , tag: NN
kata: yang , tag: SC
kata: meningkat , tag: VB
kata: . , tag: Z
```

```
Nilai akurasi baseline :  0.9481
```

Hasil tagging dengan pendekatan *classification :*

```
Hasil tagging :  {'word': 'menteri', 'prefix-1': 'm', 'prefix-2': 'me', 'suffix-1': 'i', 'suffix-2': 'ri', 'prev_word': '', 'next_word': 'pertahanan'}
tag :  NNP
Hasil tagging :  {'word': 'pertahanan', 'prefix-1': 'p', 'prefix-2': 'pe', 'suffix-1': 'n', 'suffix-2': 'an', 'prev_word': 'menteri', 'next_word': 'as'}
tag :  NNP
Hasil tagging :  {'word': 'as', 'prefix-1': '', 'prefix-2': '', 'suffix-1': '', 'suffix-2': '', 'prev_word': 'pertahanan', 'next_word': 'dijadwalkan'}
tag :  NNP
Hasil tagging :  {'word': 'dijadwalkan', 'prefix-1': 'd', 'prefix-2': 'di', 'suffix-1': 'n', 'suffix-2': 'an', 'prev_word': 'as', 'next_word': 'mengunjungi'}
tag :  VB
Hasil tagging :  {'word': 'mengunjungi', 'prefix-1': 'm', 'prefix-2': 'me', 'suffix-1': 'i', 'suffix-2': 'gi', 'prev_word': 'dijadwalkan', 'next_word': 'india'}
tag :  VB
Hasil tagging :  {'word': 'india', 'prefix-1': 'i', 'prefix-2': 'in', 'suffix-1': 'a', 'suffix-2': 'ia', 'prev_word': 'mengunjungi', 'next_word': '.'}
tag :  NNP
Hasil tagging :  {'word': '.', 'prefix-1': '', 'prefix-2': '', 'suffix-1': '', 'suffix-2': '', 'prev_word': 'india', 'next_word': ''}
tag :  Z
Hasil tagging :  {'word': 'tata', 'prefix-1': 't', 'prefix-2': 'ta', 'suffix-1': 'a', 'suffix-2': 'ta', 'prev_word': '', 'next_word': 'power'}
tag :  NNP
Hasil tagging :  {'word': 'power', 'prefix-1': 'p', 'prefix-2': 'po', 'suffix-1': 'r', 'suffix-2': 'er', 'prev_word': 'tata', 'next_word': 'menyuplai'}
tag :  NNP
Hasil tagging :  {'word': 'menyuplai', 'prefix-1': 'm', 'prefix-2': 'me', 'suffix-1': 'i', 'suffix-2': 'ai', 'prev_word': 'power', 'next_word': 'batu bara'}
tag :  VB
Hasil tagging :  {'word': 'batu bara', 'prefix-1': 'b', 'prefix-2': 'ba', 'suffix-1': 'a', 'suffix-2': 'ra', 'prev_word': 'menyuplai', 'next_word': 'pada'}
tag :  NN
Hasil tagging :  {'word': 'pada', 'prefix-1': 'p', 'prefix-2': 'pa', 'suffix-1': 'a', 'suffix-2': 'da', 'prev_word': 'batu bara', 'next_word': 'tahun'}
tag :  IN
Hasil tagging :  {'word': 'tahun', 'prefix-1': 't', 'prefix-2': 'ta', 'suffix-1': 'n', 'suffix-2': 'un', 'prev_word': 'pada', 'next_word': '2000'}
tag :  NN
Hasil tagging :  {'word': '2000', 'prefix-1': '2', 'prefix-2': '20', 'suffix-1': '0', 'suffix-2': '00', 'prev_word': 'tahun', 'next_word': '.'}
tag :  CD
Hasil tagging :  {'word': '.', 'prefix-1': '', 'prefix-2': '', 'suffix-1': '', 'suffix-2': '', 'prev_word': '2000', 'next_word': ''}
tag :  Z
Hasil tagging :  {'word': 'pemerintah', 'prefix-1': 'p', 'prefix-2': 'pe', 'suffix-1': 'h', 'suffix-2': 'ah', 'prev_word': '', 'next_word': 'hati-hati'}
tag :  NN
Hasil tagging :  {'word': 'hati-hati', 'prefix-1': 'h', 'prefix-2': 'ha', 'suffix-1': 'i', 'suffix-2': 'ti', 'prev_word': 'pemerintah', 'next_word': 'dalam'}
tag :  JJ
Hasil tagging :  {'word': 'dalam', 'prefix-1': 'd', 'prefix-2': 'da', 'suffix-1': 'm', 'suffix-2': 'am', 'prev_word': 'hati-hati', 'next_word': 'mengelola'}
tag :  IN
Hasil tagging :  {'word': 'mengelola', 'prefix-1': 'm', 'prefix-2': 'me', 'suffix-1': 'a', 'suffix-2': 'la', 'prev_word': 'dalam', 'next_word': 'bumn'}
tag :  VB
Hasil tagging :  {'word': 'bumn', 'prefix-1': 'b', 'prefix-2': 'bu', 'suffix-1': 'n', 'suffix-2': 'mn', 'prev_word': 'mengelola', 'next_word': '.'}
tag :  NN
Hasil tagging :  {'word': '.', 'prefix-1': '', 'prefix-2': '', 'suffix-1': '', 'suffix-2': '', 'prev_word': 'bumn', 'next_word': ''}
tag :  Z
```

```
Nilai akurasi classification :  0.8312
```

Dari proses tagging yang dilakukan, mendapatkan nilai akurasi yang berbeda dimana akurasi pendekatan *baseline* lebih besar dibandingkan dengan pendekatan *classification.* Menurut analisis penulis, hal tersebut dikarenakan pada pendekatan *classification* setiap kata diklasifikasikan terlebih dahulu menurut prefix dan suffixnya oleh karena itu kata sangat detail dan rinci kategorinya sehingga pada saat diuji mencari kata yang ada pada data latih agar sama dengan data uji kemungkinannya lebih kecil dibanding dengan *baseline* yang tidak terlalu rinci diklasifikasikannya dan *baseline* juga sangat simple sehingga untuk menentukan tag hasil dari data uji lebih mudah oleh karena itu nilai akurasinya pun juga tinggi.

Untuk pendekatan HMM-Viterbi, dikarenakan kurangnya pemahaman penulis, penulis tidak melakukan penelitian menggunakan metode tersebut. Diharapkan untuk tugas kedepannya penulis lebih dapat memahami tugas-tugas yang diberikan agar semua tugas dapat dikerjakan dengan baik.