Part Of Speech (POS) Tagging

Nama: Chlaudiah Julinar

NIM : 1301150434

Kelas : ICM-39-GAB

Deskripsi Tugas

Membuat POS-Tagger Bahasa Indonesia yang dapat menerima masukan berupa kalimat dan akan menghasilkan label POS-Tag untuk tiap kata dalam kalimat tersebut. Dalam membangun sebuah POS-Tag, dapat menerapkan beberapa pendekatan dimana seluruh pendekatan yang dipakai dapat menggunakan data train dan data test yang telah dianotasi dari berbagai penelitian maupun data train dan data test yang dianotasi manual oleh diri sendiri, serta dapat juga menggunakan kamus-kamus yang sudah ada dalam melakukan anotasi pada data train dan data test.

Pada tugas ini, dataset latih (data train dan data test) menggunakan corpus yang telah dibangun dan dianotasi pada suatu penelitian, yang dapat di akses pada http://bahasa.cs.ui.ac.id/postag/corpus . Tetapi, data train yang digunakan hanyalah **1000 kalimat** pertama dari korpus tersebut dan **20 kalimat** selanjutnya sebagai data test. Selanjutnya, pembangunan model dari data train dan pengujian data test akan menggunakan tiga pendekatan dibawah ini:

1. Baseline: membangun model data train berdasarkan frekuensi jumlah kemunculan suatu kata dengan tag tertentu, serta melakukan pengujian terhadap model melalui

data test dengan memberi tag pada kata didalam data test sesuai dengan jumlah kemunculan terbanyak suatu kata diberikan tag tertentu pada model yang dibangun.
2. Statistika: membangun model data train menggunakan feature yang dideskripsikan sendiri, lalu pengujian terhadap model dianggap sebagai persoalan klasifikasi sehingga proses pemberian tag pada kata-kata menggunakan metode klasifikasi berdasarkan model feature yang dibangun.
3. Model Sekuen (HMM-Viterbi): membangun model data train berdasarkan probabilitas kemunculan suatu kata diberi tag tertentu dan probabilitas kemunculan suatu tag setelah tag yang lain, lalu pengujian terhadap model akan dilakukan dengan memberikan tag pada setiap kata dalam datatest dengan mempertimbangkan probabilitas sekuens tag kata terbaik diantara probabilitas sekuens tag kata dalam suatu kalimat lainnya.

Skenario Pembangunan Model

1. Model dengan Metode Baseline

Pada metode ini, model dibangun dalam dua dictionary utama, yaitu:

- tag_count : berisi kumpulan tag yang terdapat dalam data train sebagai keys, dengan jumlah kemunculan setiap tag sebagai values
- jumlah_emission : berisi kumpulan kata diberi tag tertentu dalam data train sebagai keys, dengan jumlah kemunculan kata diberi tag tertentu sebagai values

2. Model dengan Metode Statistika

Pada metode ini, ada dua hal utama yang harus diperhatikan yaitu feature dan metode klasifikasi yang digunakan.

- Feature : Jumlah feature yang digunakan ada sebanyak 25 features, dimana feature tersebut dapat dikelompokkan menjadi tiga kategori: (a) feature berdasarkan tempat

kemunculan suatu kata, yaitu: kata diawal dan kata diakhir kalimat; (b) feature berdasarkan 10 jenis kata dari suatu kamus (lihat: 1-1 tag.txt) dimana seluruh kata dalam kamus tersebut jenisnya tidak akan pernah berubah; (c) feature berdasarkan karakter-karakter (huruf) dalam pada kata yang terletak diawal kata maupun diakhir kata.

- Metode Klasifikasi: Metode klasifikasi yang digunakan yaitu klasifikasi menggunakan Naive Bayes (Multinomial Bayes), dimana metode klasifikasi tersebut dalam suatu artikel dikatakan bahwa memiliki metode klasifikasi yang menghasilkan akurasi cukup tinggi tetapi tidak overfitting terhadap data train, sehingga pada data test dapat memprediksi tag kata dengan baik.

Pada saat membangun model, data train dibagi dua menjadi data train dan data validasi dimana pembagiannya adalah: 70% data train dan 30% data validasi. Berdasarkan hasil klasifikasi dengan naive bayes, score atau akurasi yang diperoleh adalah **0.79** atau **79%**

3. Model dengan Metode HMM-Viterbi

Pada metode ini, terdapat tiga dictionary utama dalam membangun model, yaitu:

- tag_count: berisi kumpulan tag yang terdapat dalam data train sebagai keys, dengan jumlah kemunculan setiap tag dalam data train sebagai values
- probability_emission : berisi kumpulan kata diberi tag tertentu dalam data train sebagai keys, dengan probabilitas kemunculan kata diberi tag tertentu dalam data train sebagai values
- probability_trans : beirisi kumpulan tag muncul setelah tag tertentu dalam data train sebagai keys, dengan probabilitas kemunculan tag tersebut setelah tag tertentu dalam data train sebagai values

Skenario Pengujian Model

A. Baseline

1. Pada baseline, data test akan dibaca dan dipisahkan antara tag dan kata, lalu dimasukkan kedalam masing-masing array yang menampung kata dan tag (array kumpulan_kata dan array kumpulan_tag).
2. Seluruh kata yang telah dipisahkan tersebut akan dicari satu persatu didalam dictionary jumlah_emission. Apabila kata tersebut ada didalam dict, maka akan dimasukkan kedalam array word_tag yang menampung kata itu sendiri, pasangan tag kata tersebut yang ada di dalam dictionary jumlah_emission, dan jumlah kemunculan kata dengan pasangan tag dari dictionary jumlah_emission
3. Jika ada kata yang tidak terdapat dalam dictionary, artinya kata tersebut tidak pernah muncul dalam model. Maka kata tersebut akan ditambahkan kedalam dict jumlah_emission dan diberi tag 'NN'. Diberi tag 'NN' karena tag 'NN' merupakan tag yang paling sering muncul setelah tag 'Z' didalam model. Lalu, dimasukkan kedalam array word_tag dengan jumlah kemunculan kata tersebut dan pasangan tagnya adalah 1.
4. Setelah seluruh kata dalam data test sudah memiliki pasangan tag, apabila terdapat kata yang memiliki pasangan tag lebih dari satu, maka akan dieliminasi sehingga hanya ada satu tag untuk kata tersebut. Proses eliminasi dilakukan berdasarkan jumlah kemunculan kata dengan pasangan tagnya, dimana yang dipilih adalah jumlah kemunculan yang terbanyak.

Seluruh kata dalam data test sudah memiliki pasangan tagnya masing-masing, maka dilakukan perhitungan akurasi dengan menjumlahkan hasil tag yang diperoleh dari proses pengujian yang sesuai dengan tag asli dari kata tersebut dalam array kumpulan_tag dan dibagi dengan jumlah keseluruhan tag, dikali dengan 100%.

B. Statistika

1. Sama seperti baseline, data test dibaca dan dipisahkan antara tag dan katanya.
2. Lalu, setiap kata dalam data test akan diprediksi tag-nya berdasarkan feature yang dibangun pada saat pembangunan model dengan menggunakan metode klasifikasi naive bayes

3. Perbedaan dengan baseline adalah setelah proses prediksi, seluruh kata hanya memiliki pasangan tag satu saja sehingga tidak perlu dieliminasi.
4. Setelah itu, dilakukan perhitungan akurasi sama seperti pada baseline

C. HMM-Viterbi

1. Data test dibaca dan dipisahkan antara tag dan katanya sama seperti pada kedua metode sebelumnya
2. Setiap kalimat didalam data test, masing-masing akan dibuatkan tabel emission dan transition dari setiap kalimat didalam data test. Hal ini untuk mempermudah pada saat perhitungan probabilitas sekuens tag untuk setiap kalimat
3. Pada saat membuat tabel emission dan tabel transition-nya, apabila ada kata dalam kalimat data test yang tidak terdapat didalam model data train yang dibangun, maka probabilitas kemunculan kata tersebut akan ditambahkan kedalam dictionary model data train untuk dict probabilitas_emission dengan tag yang diperoleh adalah seluruh tag yang pernah muncul didalam kalimat tersebut. Tetapi, apabila kata tersebut merupakan kata pertama dalam kalimat, maka pada dict model data train untuk dict jumlah_emission, kata tersebut kata akan ditambahkan dengan pasangan tagnya yaitu seluruh tag yang ada didalam model data train pada dict tag_count.keys()
4. Apabila, tabel emission dan tabel transition telah dibuat. Maka, proses pemberian tag untuk setiap kata dalam kalimat data test dilakukan sesuai dengan algoritma viterbi itu sendiri.
5. Terakhir yaitu dilakukan perhitungan akurasi, masih dengan cara yang sama dengan dua metode sebelumnya.

Hasil dan Analisis Hasil Pengujian Model

A. Baseline

- Akurasi Data Test: 88.97795591182364

- Jumlah kata yang memiliki tag yang berbeda: 55 Kata
- Analisis: Akurasi yang dihasilkan dapat dibilang sangat tinggi, hal ini dikarenakan rata-rata kata yang tidak pernah muncul didalam data train adalah kata dengan tag asli Noun (NN) dan tag asli CD. Untuk tag NN memang memiliki cakupan kata yang sangat luas dan tag ini bertipe open class karena jumlah kata dalam tag ini dapat terus bertambah dan sangat beragam. Sedangkan. untuk kata dengan tag CD merupakan kelas tag yang terdiri atas kombinasi-kombinasi angka mulai dari angka integer hingga angka decimal, maupun angka yang ditulis dengan huruf, sehingga dengan kombinasi yang sangat banyak tersebut maka banyak kombinasi angka yang ada didalam data test tetapi tidak pernah muncul didalam model. Kedua hal tersebut yang sangat mempengaruhi dalam hasil tag data test dengan metode baseline.

Dibawah ini, kata-kata yang memiliki tag berbeda dengan tag asli.

1. Kata: MPPA Tag Test: NN Tag Asli: NNP
2. Kata: 15,18 Tag Test: NN Tag Asli: CD
3. Kata: 24,43 Tag Test: NN Tag Asli: CD
4. Kata: 12,20 Tag Test: NN Tag Asli: CD
5. Kata: 1,68 Tag Test: NN Tag Asli: CD
6. Kata: 2,02 Tag Test: NN Tag Asli: CD
7. Kata: menyusut Tag Test: NN Tag Asli: VB
8. Kata: 58,27 Tag Test: NN Tag Asli: CD
9. Kata: 41,39 Tag Test: NN Tag Asli: CD
10. Kata: memangkas Tag Test: NN Tag Asli: VB
11. Kata: 41,98 Tag Test: NN Tag Asli: CD
12. Kata: 34,11 Tag Test: NN Tag Asli: CD
13. Kata: mencetak Tag Test: NN Tag Asli: VB
14. Kata: 160,50 Tag Test: NN Tag Asli: CD
15. Kata: 30,5 Tag Test: NN Tag Asli: CD
16. Kata: diantaranya Tag Test: NN Tag Asli: X
17. Kata: dibagikan Tag Test: NN Tag Asli: VB
18. Kata: 49 Tag Test: NN Tag Asli: CD
19. Kata: - Tag Test: Z Tag Asli: NNP

20. Kata: 8 Tag Test: CD Tag Asli: NNP
21. Kata: 16,9 Tag Test: NN Tag Asli: CD
22. Kata: sekitar Tag Test: RB Tag Asli: IN
23. Kata: 14,18 Tag Test: NN Tag Asli: CD
24. Kata: 14,8 Tag Test: NN Tag Asli: CD
25. Kata: Hidajat Tag Test: NN Tag Asli: NNP
26. Kata: Thandradjaja Tag Test: NN Tag Asli: NNP
27. Kata: kotor Tag Test: NN Tag Asli: JJ
28. Kata: 78,9 Tag Test: NN Tag Asli: CD
29. Kata: 229 Tag Test: NN Tag Asli: CD
30. Kata: 128 Tag Test: NN Tag Asli: CD
31. Kata: 93,3 Tag Test: NN Tag Asli: CD
32. Kata: 95,1 Tag Test: NN Tag Asli: CD
33. Kata: 183,9 Tag Test: NN Tag Asli: CD
34. Kata: Hidajat Tag Test: NN Tag Asli: NNP
35. Kata: 2,020 Tag Test: NN Tag Asli: CD
36. Kata: terganggu Tag Test: NN Tag Asli: VB
37. Kata: usahakan Tag Test: NN Tag Asli: VB
38. Kata: jaga Tag Test: NN Tag Asli: VB
39. Kata: kalaupun Tag Test: NN Tag Asli: SC
40. Kata: volatility Tag Test: NN Tag Asli: FW
41. Kata: terukur Tag Test: NN Tag Asli: VB
42. Kata: Jangan Tag Test: NN Tag Asli: NEG
43. Kata: sampai Tag Test: IN Tag Asli: SC
44. Kata: menahan Tag Test: NN Tag Asli: VB
45. Kata: ditahan Tag Test: NN Tag Asli: VB
46. Kata: berapa Tag Test: SC Tag Asli: WH
47. Kata: kali Tag Test: NND Tag Asli: NN
48. Kata: kedua Tag Test: CD Tag Asli: OD
49. Kata: berapa Tag Test: SC Tag Asli: WH
50. Kata: levelling Tag Test: NN Tag Asli: FW
51. Kata: off Tag Test: NN Tag Asli: FW
52. Kata: di mana Tag Test: SC Tag Asli: WH

53. Kata: mulai Tag Test: MD Tag Asli: VB
54. Kata: mulai Tag Test: MD Tag Asli: VB
55. Kata: inginkan Tag Test: NN Tag Asli: VB

B. Statistika

- Akurasi Data Validasi : 0.7956760877852203
- Akurasi Data Test : 81.76352705410822
- Jumlah kata yang memiliki tag salah: 91 Kata
- Analisis: Dapat dilihat bahwa, meskipun hasil data validasi yang diperoleh tidak terlalu tinggi tetapi hal tersebut memberikan efek yang baik terhadap data test, dimana akurasi data test cukup tinggi. Feature-feature yang didefinisikan diawal sangatlah berpengaruh, terutama feature berdasarkan karakter dalam kata. Apabila feature-feature yang didefinisikan lebih banyak lagi, dan jenis kata yang didefinisikan juga beragam, maka akurasi yang dihasilkan akan lebih tinggi daripada hasil yang sekarang. Dan bukti bahwa metode klasifikasi naive bayes tidak overfitting terhadap data validasi dapat dinyatakan benar dalam pengujian model yang dilakukan.

Dibawah ini, kata-kata yang memiliki tag berbeda dengan tag asli.

1. Kata: pertama Tag Test: NN Tag Asli: OD
2. Kata: lalu Tag Test: JJ Tag Asli: CC
3. Kata: bersih Tag Test: VB Tag Asli: JJ
4. Kata: pertama Tag Test: NN Tag Asli: OD
5. Kata: pertama Tag Test: NN Tag Asli: OD
6. Kata: namun Tag Test: NN Tag Asli: CC
7. Kata: beban Tag Test: VB Tag Asli: NN
8. Kata: bersih Tag Test: VB Tag Asli: JJ
9. Kata: pertama Tag Test: NN Tag Asli: OD
10. Kata: lalu Tag Test: JJ Tag Asli: CC
11. Kata: pertama Tag Test: NN Tag Asli: OD
12. Kata: sehingga Tag Test: NN Tag Asli: SC
13. Kata: tetap Tag Test: VB Tag Asli: RB

14. Kata: diantaranya Tag Test: PRP Tag Asli: X
15. Kata: untuk Tag Test: IN Tag Asli: SC
16. Kata: per Tag Test: NN Tag Asli: IN
17. Kata: lembar Tag Test: NN Tag Asli: NND
18. Kata: - Tag Test: Z Tag Asli: NNP
19. Kata: 8 Tag Test: CD Tag Asli: NNP
20. Kata: pertama Tag Test: NN Tag Asli: OD
21. Kata: sekitar Tag Test: JJ Tag Asli: IN
22. Kata: lalu Tag Test: JJ Tag Asli: CC
23. Kata: kotor Tag Test: NN Tag Asli: JJ
24. Kata: pertama Tag Test: NN Tag Asli: OD
25. Kata: pertama Tag Test: NN Tag Asli: OD
26. Kata: pertama Tag Test: NN Tag Asli: OD
27. Kata: karena Tag Test: IN Tag Asli: SC
28. Kata: hingga Tag Test: NN Tag Asli: IN
29. Kata: pertama Tag Test: NN Tag Asli: OD
30. Kata: sebanyak Tag Test: NN Tag Asli: CD
31. Kata: satu Tag Test: NN Tag Asli: CD
32. Kata: stabil Tag Test: NN Tag Asli: JJ
33. Kata: per Tag Test: NN Tag Asli: IN
34. Kata: agar Tag Test: NN Tag Asli: SC
35. Kata: riil Tag Test: NN Tag Asli: JJ
36. Kata: tidak Tag Test: NN Tag Asli: NEG
37. Kata: terganggu Tag Test: JJ Tag Asli: VB
38. Kata: Kita Tag Test: NNP Tag Asli: PRP
39. Kata: tetap Tag Test: VB Tag Asli: RB
40. Kata: stabil Tag Test: NN Tag Asli: JJ
41. Kata: kita Tag Test: NN Tag Asli: PRP
42. Kata: jaga Tag Test: NN Tag Asli: VB
43. Kata: kalaupun Tag Test: VB Tag Asli: SC
44. Kata: ia Tag Test: NN Tag Asli: PRP
45. Kata: tentu Tag Test: NN Tag Asli: RB
46. Kata: saja Tag Test: NN Tag Asli: RB

47. Kata: volatility Tag Test: NN Tag Asli: FW
48. Kata: sangat Tag Test: NN Tag Asli: RB
49. Kata: terukur Tag Test: NN Tag Asli: VB
50. Kata: Ia Tag Test: NNP Tag Asli: PRP
51. Kata: agar Tag Test: NN Tag Asli: SC
52. Kata: tidak Tag Test: NN Tag Asli: NEG
53. Kata: Jangan Tag Test: NNP Tag Asli: NEG
54. Kata: sampai Tag Test: IN Tag Asli: SC
55. Kata: sangat Tag Test: NN Tag Asli: RB
56. Kata: dekat Tag Test: NN Tag Asli: JJ
57. Kata: karena Tag Test: IN Tag Asli: SC
58. Kata: tidak Tag Test: NN Tag Asli: NEG
59. Kata: melalui Tag Test: VB Tag Asli: IN
60. Kata: Yang Tag Test: NNP Tag Asli: DT
61. Kata: bukan Tag Test: NN Tag Asli: NEG
62. Kata: kita Tag Test: NN Tag Asli: PRP
63. Kata: tidak Tag Test: NN Tag Asli: NEG
64. Kata: berapa Tag Test: VB Tag Asli: WH
65. Kata: Sementara Tag Test: NNP Tag Asli: SC
66. Kata: menurut Tag Test: VB Tag Asli: IN
67. Kata: Kalau Tag Test: NNP Tag Asli: SC
68. Kata: dua Tag Test: NN Tag Asli: CD
69. Kata: kali Tag Test: VB Tag Asli: NN
70. Kata: lelang Tag Test: SC Tag Asli: NN
71. Kata: pertama Tag Test: NN Tag Asli: OD
72. Kata: kedua Tag Test: NN Tag Asli: OD
73. Kata: berapa Tag Test: VB Tag Asli: WH
74. Kata: trend Tag Test: NN Tag Asli: FW
75. Kata: kalau Tag Test: NN Tag Asli: SC
76. Kata: tidak Tag Test: NN Tag Asli: NEG
77. Kata: levelling Tag Test: SC Tag Asli: FW
78. Kata: off Tag Test: NN Tag Asli: FW
79. Kata: trend Tag Test: NN Tag Asli: FW

80. Kata: Hal Tag Test: NNP Tag Asli: NN
81. Kata: menurut Tag Test: VB Tag Asli: IN
82. Kata: baik Tag Test: VB Tag Asli: JJ
83. Kata: di mana Tag Test: NN Tag Asli: WH
84. Kata: bergulir Tag Test: JJ Tag Asli: VB
85. Kata: bidang Tag Test: SC Tag Asli: NN
86. Kata: sudah Tag Test: NN Tag Asli: MD
87. Kata: memang Tag Test: VB Tag Asli: RB
88. Kata: belum Tag Test: NN Tag Asli: NEG
89. Kata: seperti Tag Test: NN Tag Asli: IN
90. Kata: kita Tag Test: NN Tag Asli: PRP
91. Kata: inginkan Tag Test: NN Tag Asli: VB

C. HMM-Viterbi

- Akurasi Data Test: 79.55911823647295
- Jumlah kata dengan tag salah : 102 kata
- Analisis: Dengan metode ini, akurasi data test yang dihasilkan lebih kecil daripada metode yang lain. Kemungkinan yang menyebabkan terjadinya hal ini adalah pada saat memasukkan kata-kata dalam data test yang tidak pernah dibaca dalam model, karena kata-kata tersebut hanya dimasukkan kedalam kamus probabilitas dengan pasangan tag yaitu seluruh tag yang pernah ada hanya didalam kalimat itu saja, bukan dengan seluruh tag yang ada didalam model. Karena, apabila hal tersebut dilakukan, maka durasi pemrosesan HMM-Viterbi akan berlangsung sangat lama. Sehingga, dapat dikatakan bahwa untuk menggunakan model HMM-Viterbi, alangkah lebih baik jika corpus yang digunakan memiliki kata-kata yang beragam dan sangat banyak, agar dapat menghasilkan akurasi data test yang sempurna.

Dibawah ini, kata-kata yang memiliki tag berbeda dengan tag asli.

1. Kata: Perusahaan Tag Test: NNP Tag Asli: NN
2. Kata: ritel Tag Test: NNP Tag Asli: NN
3. Kata: 2007 Tag Test: NNP Tag Asli: CD

4. Kata: laba bersih Tag Test: IN Tag Asli: NN
5. Kata: 24,43 Tag Test: SYM Tag Asli: CD
6. Kata: lalu Tag Test: NN Tag Asli: CC
7. Kata: dari Tag Test: VB Tag Asli: IN
8. Kata: menyusut Tag Test: CC Tag Asli: VB
9. Kata: semula Tag Test: VB Tag Asli: NN
10. Kata: lain Tag Test: PR Tag Asli: JJ
11. Kata: manajemen Tag Test: JJ Tag Asli: NN
12. Kata: memangkas Tag Test: IN Tag Asli: VB
13. Kata: dari Tag Test: VB Tag Asli: IN
14. Kata: 2006 Tag Test: NNP Tag Asli: CD
15. Kata: mencetak Tag Test: NN Tag Asli: VB
16. Kata: laba bersih Tag Test: SC Tag Asli: NN
17. Kata: sebesar Tag Test: CC Tag Asli: JJ
18. Kata: 30,5 Tag Test: SYM Tag Asli: CD
19. Kata: diantaranya Tag Test: CD Tag Asli: X
20. Kata: untuk Tag Test: IN Tag Asli: SC
21. Kata: sebesar Tag Test: CC Tag Asli: JJ
22. Kata: lembar Tag Test: NN Tag Asli: NND
23. Kata: saham Tag Test: NNP Tag Asli: NN
24. Kata: . Tag Test: NNP Tag Asli: Z
25. Kata: Perusahaan Tag Test: NNP Tag Asli: NN
26. Kata: telepon seluler Tag Test: NNP Tag Asli: NN
27. Kata: laba bersih Tag Test: CC Tag Asli: NN
28. Kata: 14,18 Tag Test: SYM Tag Asli: CD
29. Kata: kotor Tag Test: IN Tag Asli: JJ
30. Kata: 78,9 Tag Test: IN Tag Asli: CD
31. Kata: menjadi Tag Test: CD Tag Asli: VB
32. Kata: 93,3 Tag Test: RB Tag Asli: CD
33. Kata: miliar Tag Test: IN Tag Asli: CD
34. Kata: karena Tag Test: IN Tag Asli: SC
35. Kata: jumlah Tag Test: CD Tag Asli: NN
36. Kata: Gubernur Tag Test: VB Tag Asli: NNP

37. Kata: yang Tag Test: DT Tag Asli: SC
38. Kata: terganggu Tag Test: JJ Tag Asli: VB
39. Kata: usahakan Tag Test: MD Tag Asli: VB
40. Kata: stabil Tag Test: RB Tag Asli: JJ
41. Kata: jaga Tag Test: NNP Tag Asli: VB
42. Kata: dalam Tag Test: NNP Tag Asli: IN
43. Kata: volatility Tag Test: NNP Tag Asli: FW
44. Kata: terukur Tag Test: NNP Tag Asli: VB
45. Kata: . Tag Test: NNP Tag Asli: Z
46. Kata: penjagaan Tag Test: PRP Tag Asli: NN
47. Kata: Jangan Tag Test: UH Tag Asli: NEG
48. Kata: dalam Tag Test: X Tag Asli: IN
49. Kata: tempo Tag Test: UH Tag Asli: NN
50. Kata: . Tag Test: NNP Tag Asli: Z
51. Kata: niat Tag Test: VB Tag Asli: NN
52. Kata: melalui Tag Test: FW Tag Asli: IN
53. Kata: ' Tag Test: NNP Tag Asli: Z
54. Kata: ' Tag Test: NNP Tag Asli: Z
55. Kata: . Tag Test: NNP Tag Asli: Z
56. Kata: ditahan Tag Test: DT Tag Asli: VB
57. Kata: itu Tag Test: DT Tag Asli: PR
58. Kata: volatilitas Tag Test: DT Tag Asli: NN
59. Kata: -nya Tag Test: DT Tag Asli: PRP
60. Kata: , Tag Test: DT Tag Asli: Z
61. Kata: bukan Tag Test: DT Tag Asli: NEG
62. Kata: ' Tag Test: DT Tag Asli: Z
63. Kata: rate Tag Test: DT Tag Asli: FW
64. Kata: -nya Tag Test: DT Tag Asli: PRP
65. Kata: ' Tag Test: DT Tag Asli: Z
66. Kata: kita Tag Test: DT Tag Asli: PRP
67. Kata: tidak Tag Test: DT Tag Asli: NEG
68. Kata: mengarah Tag Test: DT Tag Asli: VB
69. Kata: ' Tag Test: DT Tag Asli: Z

70. Kata: rate Tag Test: DT Tag Asli: FW
71. Kata: ' Tag Test: DT Tag Asli: Z
72. Kata: berapa Tag Test: DT Tag Asli: WH
73. Kata: , Tag Test: DT Tag Asli: Z
74. Kata: kata Tag Test: DT Tag Asli: VB
75. Kata: Burhanudin Tag Test: DT Tag Asli: NNP
76. Kata: . Tag Test: DT Tag Asli: Z
77. Kata: untuk Tag Test: IN Tag Asli: SC
78. Kata: kali Tag Test: NND Tag Asli: NN
79. Kata: terakhir Tag Test: OD Tag Asli: JJ
80. Kata: penurunan Tag Test: CD Tag Asli: NN
81. Kata: , Tag Test: CC Tag Asli: Z
82. Kata: yang Tag Test: DT Tag Asli: SC
83. Kata: Rp Tag Test: SC Tag Asli: SYM
84. Kata: 16 Tag Test: SC Tag Asli: CD
85. Kata: yang Tag Test: DT Tag Asli: SC
86. Kata: berapa Tag Test: SC Tag Asli: WH
87. Kata: tidak Tag Test: FW Tag Asli: NEG
88. Kata: , Tag Test: FW Tag Asli: Z
89. Kata: menurun Tag Test: FW Tag Asli: VB
90. Kata: -nya Tag Test: FW Tag Asli: PRP
91. Kata: , Tag Test: FW Tag Asli: Z
92. Kata: kata Tag Test: FW Tag Asli: VB
93. Kata: Burhanudin Tag Test: FW Tag Asli: NNP
94. Kata: . Tag Test: FW Tag Asli: Z
95. Kata: di mana Tag Test: SC Tag Asli: WH
96. Kata: perkreditan Tag Test: PR Tag Asli: NN
97. Kata: mulai Tag Test: MD Tag Asli: VB
98. Kata: Ekspansi Tag Test: PRP Tag Asli: NN
99. Kata: perkreditan Tag Test: PR Tag Asli: NN
100. Kata: mulai Tag Test: MD Tag Asli: VB
101. Kata: kecepatan Tag Test: JJ Tag Asli: NN
102. Kata: inginkan Tag Test: X Tag Asli: VB

Kesimpulan

Kesimpulan yang diperoleh dari ketiga metode yang dilakukan terhadap proses POS-Tagging adalah POS-Tagging sangat bergantung terhadap corpus yang digunakan, mulai dari keberagaman kata didalam corpus hingga jumlah kata beserta tagnya didalam corpus, yang dimana semakin banyak kata dan pasangan tagnya maka akan semakin bagus pula dalam proses pemberian tag pada data-test.

Dari ketiga metode ini, akurasi yang paling tinggi adalah baseline tetapi menurut saya hal ini tidak akan berlangsung lama, karena apabila jumlah data test diperbanyak dengan keberagaman data test semakin tinggi, maka baseline akan jauh tertinggal. Karena, dapat dilihat pada kata-kata yang mendapatkan tag salah adalah terdapat banyak kelas tag yang tidak terdefinisi dalam data test, salah satunya adalah kelas tag CD. Sehingga. apabila data test yang dibangun berisi banyak kata dengan tag CD maka baseline akan memiliki akurasi yang jauh tertinggal daripada metode yang lain.

Sedangkan, untuk kedua metode yang lain kelas tag yang tidak terdefinisi tidak ada yang terlalu dominan, sehingga permasalahan dari kedua metode tersebut adalah jumlah korpus dan keberagaman yang harus lebih diperbanyak dan jumlah feature yang didefinisikan untuk kelas statistik harus lebih banyak dan lebih spesifik.