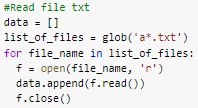
1. **Deskripsi Masalah**

Terdapat 20 file artikel berbahasa indonesia dengan topik ‘Cyber Crime Whatsapp’ akan dibuat sebuah model bahasa bigram. Kemudian akan diimplementasikan laplace smoothing dan akan dihitung nilai probabilitas bigram dan nilai perplexity. Terdapat 6 kalimat uji, 3 kalimat sesuai dengan topik, 3 kalimat tidak sesuai topik. Dari 6 kalimat uji, akan dibandingkan nilai perplexitynya, lebih bagus kalimat yang sesuai topik atau kalimat yang tidak sesuai topik.

1. **Perancangan Sistem**

* Membaca 20 file artikel .txt

****

* Case Folding

Dilakukannya case folding untuk mengubah semua huruf menjadi lowercase.

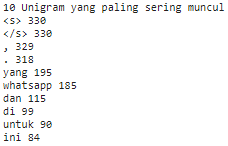
* Tokenisasi

Tokenisasi bertujuan untuk menambah tag ‘<s> pada awal kalimat, dan tag </s> pada akhir kalimat. Kemudian membagi kalimat menjadi per-kata.

* Frekuensi Unigram

Menghitung frekuensi unigram dengan menambahkan nilai 1 setiap menemukan kata yang sama.

* 10 Unigram yang palung sering muncul

****

* Probabilitas Unigram

Nilai probabilitas unigram didapat dengan cara frekuensi kata dibagi dengan jumlah semua kata

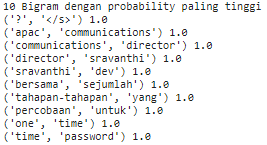
* Frekuensi Bigram

Menghitung frekuensi bigram dengan menambahkan nilai 1 setiap menemukan kombinasi kata yang sama

* Probabilitas Bigram

Nilai probabilitas bigram didapat dengan cara frekuensi kombinasi 2 kata dibagi dengan frekuensi 1 kata diawal kombinasi 2 kata.

* 10 Probabilitas Bigram tertinggi

****

* Cek Bigram untuk kalimat uji

Untuk cek probabilitas bigram pada kalimat uji, akan dilakukan case folding dan tokenisasi. Setelah itu dihitung nilai probabilitas bigram dengan mencari kombinasi 2 kata yang sama dengan probabilitas bigram dari kalimat latih. Jika sama, nilai probabilitas akan dikalikan dengan probabilitas bigram. Jika tidak sama, nilai probabilitas akan dikalikan dengan 0.

* Laplace Smoothing

Supaya nilai probabilitas bigram kalimat uji tidak 0, diperlukannya laplace smoothing. Dengan menambahkan nilai 1 di setiap frekuensi bigramnya. Kemudian untuk mendapatkan nilai probabilitas bigram kalimat uji adalah dengan frekuensi bigram (Ci) ditambah 1 kemudian dibagi dengan frekuensi unigram (N) kata sebelumnya ditambah dengan jumlah kata dari kalimat uji (V).

****

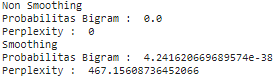
* Perplexity

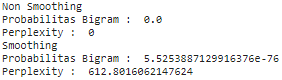
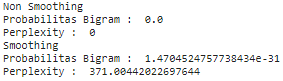
Nilai perplexity merupakan invers dari nilai probabilitas bigram. Semakin kecil nilai perplexity, maka semakin bagus kualitas language modelnya.

1. **Analisis**

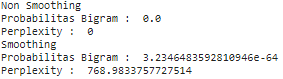
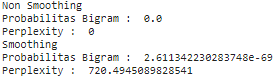
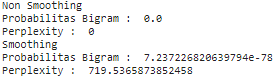
Terdapat 6 kalimat uji, 3 kalimat sesuai dengan topik, 3 kalimat tidak sesuai topik. Akan dibandingkan nilai perplexitynya

* Kalimat yang sesuai dengan topik





* Kalimat yang tidak sesuai dengan topik



1. **Kesimpulan**

Dari ketiga jenis kalimat yang berbeda dapat disimpulkan bahwa kalimat yang sesuai dengan topik, nilai probabilitas bigram yang sudah di smoothing lebih besar dan nilai perplexitynya lebih kecil daripada kalimat uji tidak sesuai topik. Hal ini dapat terjadi karena ada beberapa kata dalam kalimat sesuai topik yang sudah terdapat dalam corpus. Sehingga probabilitasnya akan semakin besar, dan nilai perplexity akan lebih kecil.