Nama : Rike Adelia
NIM : 1301154565

LAPORAN
TUGAS 1 PEMROSESAN BAHASA ALAMI

1. Dataset

Dataset yang digunakan untuk membangun model adalah kumpulan 100 artikel berbahasa Indonesia. Sumber artikel adalah kumpulan artikel yang dikumpulkan oleh Feryandi Nurdiantoro dari berbagai portal berita Indonesia, seperti Kompas.com, detik, sindonews, dan lainnya. Topik artikel yang dipilih adalah topik yang berkaitan dengan olahraga pada tanggal 20-21 Agustus 2018 karena pada topik olahraga memiliki sangat banyak kata asing seperti nama pemain dan nama klub. Selain itu, pada topik olahraga juga ada kata yang memiliki perbedaan makna jika dibandingkan dengan topik selain olahraga. Seperti contoh kata "partai" pada topik olahraga berarti babak atau fase pertandingan, sedangkan pada topik politik artinya "suatu organisasi yang menjalani ideologi tertentu".

2. Data uji

Untuk pengujian model dengan prediksi kata yang muncul, maka data yang digunakan adalah 10 kata berikut:

Kata	**Alasan**
asian	Jika model tepat, maka kalimat selanjutnya harus "games"
jakarta	Kata yang muncul harus berkaitan dengan kata "Jakarta" sesuai dengan korpus
babak	Kata yang muncul harus berkaitan dengan olahraga
kata	Kata yang muncul seharusnya merupakan tokoh olahragawan
melawan	Kata yang muncul seharusnya merupakan nama klub atau nama negara
kekalahan	Kata yang muncul harus berkaitan dengan olahraga
motogp	Kata yang muncul seharusnya berkaitan dengan MotoGP
manchester	Karena klub yang berawalan "Manchester" ada 2, United dan City, maka seharusnya yang muncul diantara dua itu

dua	Kata yang muncul harus berkaitan dengan olahraga
kampanye	Kata tersebut merupakan kata yang hanya muncul pada topik politik, akan diuji pada model yang dibangun dengan korpus topik olahraga saja

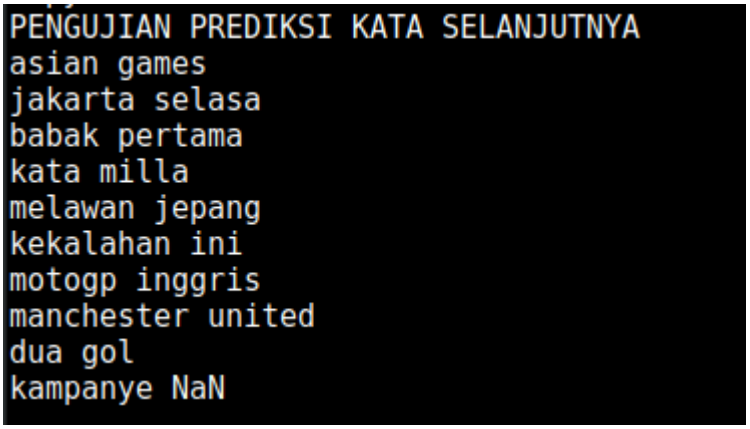
Untuk pengujian model menggunakan *perplexity*, maka data yang digunakan adalah 5 kalimat berikut:

Kalimat	**Alasan**
Perolehan medali di Asian Games 2018	Artikel memuat berita Asian Games
Ridwan Kamil yakin bisa memenangkan pilgub jawa barat 2018	Mengecek apakah model yang dibangun cocok untuk kalimat topik politik atau tidak, walaupun ada kata "memenangkan" yang juga bisa digunakan di topik olahraga
Manchester United mengalami kekalahan di pertandingan melawan Chelsea	Kalimat tersebut merupakan topik olahraga, namun tidak ada berita tentang klub Chelsea
Eko Yuli meraih medali emas	Kalimat tersebut merupakan topik olahraga
Timnas Indonesia percaya diri melaju ke babak selanjutnya	Kalimat tersebut merupakan topik olahraga

3. Analisis

a. Analisis hasil pengujian prediksi kemunculan kata

Hasil pengujian prediksi kemunculan kata selanjutnya adalah sebagai berikut:



Model yang dibangun sudah cukup bagus untuk memprediksi kata yang muncul setelah kata yang berkaitan dengan topik olahraga. Kata yang ambigu seperti “babak” dan “melawan” sudah diprediksi dengan baik karena kemunculan kata selanjutnya merupakan kata yang berkaitan dengan topik olahraga. Namun, model masih susah memprediksi kata selanjutnya setelah kata “jakarta”, yang muncul merupakan nama hari. Model ini juga sama sekali tidak bisa memprediksi kata yang hanya berkaitan dengan politik yaitu kata “kampanye”, karena model memang hanya dilatih dengan kata-kata yang berkaitan dengan topik olahraga.

b. Analisis hasil pengujian *perplexity*

Hasil pengujian dengan *perplexity* adalah sebagai berikut:

```
PENGUJIAN DENGAN PERPLEXITY
Perolehan medali di Asian Games 2018 | 4.539
Ridwan Kamil yakin bisa memenangkan pilgub jawa barat 2018 | 1021.153
Manchester United mengalami kekalahan di pertandingan melawan Chelsea | 57.043
Eko Yuli meraih medali emas | 16.095
Timnas Indonesia percaya diri melaju ke babak selanjutnya | 65.077
```

Pada pengujian *perplexity*, semakin kecil nilainya maka semakin baik model diterapkan pada kalimat tersebut. Pada kalimat-kalimat yang berkaitan dengan olahraga (kalimat 1, 3, 4, dan 5), nilai *perplexity* yang dihasilkan kurang dari 100. Kalimat 1 memiliki performa paling baik karena kalimat tersebut memiliki nilai 4,539; yang berarti model tersebut memiliki banyak korpus tentang asian games. Sedangkan performa model sangat buruk pada kalimat politik yaitu kalimat 2, karena memiliki nilai 1021,153; yang berarti model tersebut kurang cocok untuk topik selain olahraga.