KEMENTERIAN RISET, TEKNOLOGI, DAN PENDIDIKAN TINGGI
UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN" YOGYAKARTA
Jl. SWK 104 (Lingkar Utara) Condongcatur, Yogyakarta 55283 Telp.(0274)486188, 486733, Fax. 486400
Jl. Babarsari 2, Tambakbayan, Yogyakarta 55281 Telp.(0274)486911
Email : info@upnyk.ac.id. Homepage : http://www.upnyk.ac.id

UJIAN MATA KULIAH PRAKTIKUM SEMESTER GANJIL TA. 2020/2021

Mata Kuliah	:	Praktikum Data Science
Hari/Tanggal	:	Jumat, 22 Januari 2021
Asisten	:	Alfian Febriana Yusuf
		Ahmad Dzakiyyul Fuad
Kelas	:	B
Waktu	:	120 menit
Sifat Ujian	:	Online, Open Book

Kerjakan soal di bawah ini pada template Rmarkdown yang ada di SPADA!

 # BAGIAN 1

1. Tampilkan TOP 10 Aplikasi berdasarkan peringkat PENILAIAN/RATING yang diberikan user! **point 10**

2. Tampilkan rata-rata RATING yang dihitung menggunakan fungsi buatan untuk setiap kategori aplikasi! **point 15**

3. Berdasarkan soal nomor 2, buat plot untuk memvisualisasikan hasilnya! (Bentuk plot bebas) **point 15**

Info untuk 2 soal 4-5: Terdapat dua dataset yang digunakan. Satu dataset untuk info aplikasi dan satu dataset lagi untuk kumpulan reviewnya.

4. Dari kedua dataset tersebut, buat satu variable data baru yang isinya NAMA APLIKASI, RATING, dan JUMLAH REVIEW Positif dan/atau Negatif dan/atau Neutral (boleh semua, boleh pilih salah satu)lalu tampilkan isi data tabel tersebut! **point 20**

5. Dalam dunia data scientist, sebelum melakukan pemodelan ada baiknya data dilakukan preprocessing terlebih dahulu. Dengan dataset review yang sudah dimasukkan oleh user, lakukan sebuah preprocessing data SEDERHANA yang menurut kalian dapat dilakukan untuk dataset tersebut agar dataset bisa siap untuk dimodelkan (simpan hasil preprocessing dalam variabel baru dan tampilkan table hasilnya)!
   Clue : Clean, Tidy, no redundacy, no dupe, no null , dan lain-lain **point 40**

BAGIAN 2

1. Import library tidymodels, vroom, here, tidytext dan dua dataset ke dalam objek R **nilai 10**

2. Joining dua dataset menggunakan inner join **nilai 10**

3. Tahap pre-processing data. Ketika ingin melakukan analisis sentimen beberapa hal harus dilakukan sebelum data dapat digunakan. Bersihkan dan rapikan data dengan membuang data yang "nan" di bagian Translated_review. Setelah itu, data juga harus dibersihkan dari kata-kata yang mengandung stop_word (seperti: a, a's, after, dll). Data yang siap diolah juga harus ditokenisasi yaitu proses membagi teks dari paragraf atau kalimat ke kata. Hasil dari tokenisasi adalah tiap baris data hanya mengandung 1 kata. **nilai 15**

4. Sentimen analisis dapat menggunakan beberapa jenis metode berdasarkan sentiment lexicon. Ada beberapa sentiment lexicon seperti bing, afinn, dan nrc. Gunakan sentiment lexicon nrc untuk mendapatkan jumlah kata untuk 10 kategori nrc (positive, negative, fear, surprise, dll). **nilai 15**

5. Kita dapat mengetahui banyaknya kata tiap kategori nrc untuk tiap aplikasi. Cobalah untuk mencari banyak kata tiap kategori nrc yang dikelompokkan berdasarkan nama aplikasi. **nilai 15**

6. Setelah mendapatkan jumlah kata tiap kategori tiap aplikasi, kita dapat mngetahui aplikasi mana yang memiliki kata dengan kategori 'surprise' terbanyak untuk tiap aplikasi. Kita akan memvisualisasikan dengan grafik batang 10 aplikasi dengan jumlah kata kategori 'surprise' terbanyak. **nilai 20**

7. Selain menggunakan sentiment lexicon 'nrc', sentimen analisis juga dapat menggunakan sentiment lexicon 'bing'. Bing hanya akan memberikan label untuk tiap kata positif atau negatif saja. Carilah kata positif yang paling umum dan kata negatif yang paling sering digunakan saat memberikan review pada aplikasi! **nilai 15**

8. Pembacaan data akan lebih mudah jika ditampilkan dalam bentuk grafik. Tampilkan grafik 10 kata positif dan negatif terbanyak! **nilai 20**

9. Penganalisis data membutuhkan jumlah kata tiap kategori yang belum digabung dengan sentiment lexicon untuk menghitung rasio positif, ratio negatif dan net sentiment. Bantulah penganalisis tersebut untuk mendapatkan jumlah kata tiap kategori dari data yang sudah dirapikan! **nilai 15**

10. Selanjutnya penganalisis data ingin mendapatkan jumlah kata positif, jumlah kata negatif, rasio positif (jumlah kata positif/jumlah keseluruhan kata), rasio negatif (jumlah kata negatif/jumlah keseluruhan kata), dan net sentiment (jumlah kata positif - jumlah kata negatif) dengan menggunakan sentiment lexicon bing untuk tiap kategorinya. Tabel yang diinginkan oleh analisis adalah seperti berikut **nilai 40**