Nama **: Endang Prayoga Hidayatulloh**

NIM **: 2006189**

Mata Kuliah **: UTS Text Mining**

Keterangan hasil penelitian UTS Text Mining

- Semua bahan dan hasil file penelitian berada di dalam folder resource.
- Pada nomor satu ditugaskan untuk mentranskrip kan subtitle video ke dalam bentuk file dokumen. Video yang di transkrip adalah salah satu video dari youtube yang berjudul "ASTRAJINGGA SABDA GURU (FULL) - Ki Dalang Asep Sunandar Sunarya". Bagian video yang di transkripsi adalah video percakan pada menit ke 12 sampai menit ke 24.

  Hasil transkrip disimpan dengan nama file transkripsi_result.txt.
- Selanjutnya adalah tugas nomor dua yang diawali dengan melakukan tokenisasi text. Langkah pertama sebelum melakukan tokenisasi, teks transkrip dilakukan preprocessing untuk menghilangkan beberapa symbol dan huruf yang tidak dibutuhkan. Selanjutnya dilakukan tokenisasi dengan menggunakan salah satu module python yaitu nltk.
- Selanjutnya yaitu melakukan penghapusan data atau teks yang terjadi duplikasi. Selain itu pada tahap ini juga dilakukan sorting data menurut abjad.
- Lalu tugas selanjutnya yaitu melakukan pemisahan terhadap kata yang dianggap stopword. Pertama membuat variable list baru yang menampung kata stopword dalam Bahasa sunda, referensi kata didasarkan pada stopword Bahasa Indonesia yang terdapat pada dataset umum di internet. Selanjutnya adalah pemisahan teks stopword.
- Hasil dari nomor dua dari tokenisasi sampai pemisahan kata stopword dilakukan dengan tool jupyter notebook yang hasilnya terdapat pada file tokenisasi_preprocessing.ipynb
- Setelah itu melakukan export hasil tokenisasi dan pemisahan stopword kedalam bentuk csv yang selanjutnya dilakukan penambahan POS (Part Of Speech) Tagging dengan menggunakan Microsoft Excel secara manual karena tidak tersedianya POS Tag Bahasa sunda pada module python.
- Untuk menentukan tag pada teks, dilakukan pencarian kata dasar dari teks. Tool yang digunakan adalah sebuah website yang Bernama wictionary.
- Hasil dari POS Tag dalam excel diberi nama POSTag_BahasaSunda.xlsx.