Deskripsi Tugas NER

Proses identifikasi field-field yang penting untuk diekstrak sering kali membutuhkan informasi Named Entity, misal untuk field actor maka bagian teks yang sesuai untuk diekstrak adalah bagian yang diklasifikasikan sebagai Named Entity PERSON. Pada Tugas kali ini, kita fokuskan terlebih dahulu untuk dapat mengenali Named Entity (Named Entity Recognition/NER), dengan lingkup jenis Named Entity yang dikenali adalah: PERSON (PER), ORGANIZATION (ORG), dan LOCATION (LOC). Notasi yang digunakan adalah notasi B-I-O, dengan kode B menyatakan token yang merupakan awal (beginning) teks yang merupakan Named Entity, kode I menyatakan token adalah bagian tengah atau akhir Named Entity, dan kode O menyatakan token bukan merupakan Named Entity. Sebagai contoh, berikut adalah tagging NER dengan notasi BIO:

```
Presiden B-PER
Jokowi I-PER
akan O
meresmikan O
jembatan O
Holtekamp O
di O
Jayapura B-LOC
, I-LOC
Papua I-LOC
. O
```

Tugas Anda adalah membangun Named Entity Recognizer untuk ketiga tipe Named Entity tersebut, menggunakan 3 metode: pendefinisian aturan secara manual, klasifikasi, dan pendekatan berbasis model sekuens (dengan metode CRF/Conditional Random Field). Anda perlu melabeli dataset 100 kalimat yang kemarin sudah Anda kumpulkan dengan tiga tipe NE seperti contoh di atas, untuk dapat menganalisis hasil tagging dengan aturan dan melakukan pelatihan (metode klasifikasi dan model sekuens). Anda akan diberi file contoh kode untuk NER dengan ketiga pendekatan tersebut, namun aturan/fitur yang digunakan masih sangat sederhana, **tugas Anda untuk mengembangkan** (gunakan fitur lain yang lebih lengkap, dll)! File input yang digunakan sebagai contoh adalah file kalimat_POS_NE_train.txt. Sementara untuk Tugas yang Anda kerjakan, gunakan file kalimat_POSTag.txt yang berisi token-token kalimat yang sudah idberi

A. Metode berbasis aturan yang didefinisikan secara manual
   Seperti yang juga sudah dibahas saat pertemuan di kelas, informasi yang dapat digunakan untuk mengidentifikasi Named Entity PER, ORG, dan LOC adalah sebagai berikut:
   - POSTag token NNP (proper noun)
   - Token dituliskan dengan huruf awal kapital
   - Named Entity LOC biasanya diawali oleh token di atau ke
   - Named Entity ORG dituliskan dengan huruf kapital semua
B. Metode klasifikasi
   Fitur yang dapat digunakan untuk klasifikasi antara lain:

- Token tersebut sendiri
- POS token/kata
- Apakah token diawali dengan huruf kapital
- Prefix1: karakter pertama token
- Prefix2: 2 karakter pertama token
- Prefix3: 3 karakter pertama token
- Suffix1: karakter terakhir token
- Suffix2: 2 karakter terakhir token
- Suffix3: 3 karakter terakhir token

Kelas:

- B-PER
- I-PER
- B-ORG
- I-ORG
- B-LOC
- I-LOC
- O

C. Metode berbasis sekuens model
Fitur yang digunakan dapat berupa fitur yang juga digunakan pada metode klasifikasi, ditambah dengan kata di sekitarnya (token-1, token+1, postag-1, postag+1).