Final Gemastik
Penambangan Data

Task: Question Answering (Modul Answer Finder)

Deskripsi: Sebuah task NLP (Natural Language Processing) yang bertujuan untuk menghasilkan jawaban (*factoid*) berdasar dua input yang tersedia yaitu pertanyaan pengguna dan paragraf/kalimat yang merupakan sumber jawaban. Tipe *factoid* adalah jawaban yang berbentuk serangkaian kata yang merupakan bagian dari sebuah kalimat. Tipe *factoid* yang tersedia pada dataset ini adalah *Person*, *Organization*, *Location*, *Datetime*, dan *Quantity*.

Tipe Factoid	**Contoh Pertanyaan**
Person	Siapa nama penemu telpon?
Organization	Apa nama perusahaan minyak yang dimiliki Malaysia?
Location	Dimana Candi Mendut terletak?
Datetime	Kapan Alexander Graham Bell lahir?
Quantity	Berapa tinggi Gunung Tangkuban Perahu?

Contoh Isi Dataset
Input:
- Pertanyaan: Di mana Alexander Graham Bell meninggal?
- Paragraf: Alexander Graham Bell dilahirkan pada 3 Maret 1847 di Edinburgh, Skotlandia, Britania Raya dan meninggal pada 2 Agustus 1922 di Beinn Bhreagh, Nova Scotia, Kanada

Output:
- Jawaban: Beinn Bhreagh, Nova Scotia, Kanada

Proses:
Setiap peserta perlu membuat model yang dapat memberikan label pada setiap token / kata pada paragraf dan menandai token menggunakan salah satu dari tiga opsi yaitu: B (Begin), I (In), atau O (Out).

Pada contoh di atas, label yang diberikan pada setiap token adalah berikut:

Alexander (O) Graham (O) Bell (O) dilahirkan (O) pada (O) 3 (O) Maret (O) 1847 (O) di (O) Edinburgh (O) , (O) Skotlandia (O) , (O) Britania (O) Raya (O) dan (O) meninggal (O) pada (O) 2 (O) Agustus (O) 1922 (O) di (O) Beinn (B) Bhreagh (I) , (I) Nova (I) Scotia (I) , (I) Kanada (I)

Format file dataset akan mengikuti format csv sebagai berikut:

question,passage,seq_label
[“Di”,”mana”,”Alexander”,”Graham“,“Bell”,”meninggal”,”?”],
[“Alexander”,”Graham”,”Bell”,”dilahirkan”,”pada”,”3”,”Maret”,”1847”,”di”,”Edinburgh”,”,”,”Skotlandia”,”,”,”Britania”,”Raya”,”dan”,”meninggal”,”pada”,”2”,”Agustus”,”1922”,”di”,”Beinn”,”Bhreagh”,”,”,”Nova”,”Scotia”,”,”,”Kanada”],
[“O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”O”,”B”,”I”,”I”,”I”,”I”,”I”,”I”,”I”]

Token ditandai B untuk awal frasa yang menjadi jawaban, dan I untuk token berikutnya. Token ditandai O jika token tersebut bukan bagian dari jawaban.

Dataset:
Anda diberikan tiga set dataset untuk training, validation, dan test. Gunakan data training hanya untuk melatih model yang akan digunakan untuk memprediksi data di data validation dan test set. Data test set hanya akan diberikan oleh panitia 30 menit terakhir sebelum babak final berakhir dan tidak memiliki gold label. Hasil prediksi akan diumumkan setelah babak final berakhir.

Dataset	Jumlah data
train	2495
valid	311
test	311

Metric:
Metric yang digunakan untuk evaluasi adalah macro F1-score. Dan Anda akan diberikan script untuk mengevaluasi data validation.

Pengumpulan:
Anda diminta untuk mengumpulkan hasil prediksi di dalam format csv seperti berikut:
id_soal, prediksi **(tanpa header)**
1, [“O”,”O”,”B”,”O”]
2, [“O”,”O”,”B”,”I”,”O”]
…, …

Simpan file menggunakan format penamaan “<nama_tim>_final.csv”.