Tugas 1 Data Mining

Teknik Informatika
Universitas Al-Azhar Indonesia

Tenggat: Jumat, 11 Oktober 2019 pukul 23.55

Mekanisme: Anda hanya diwajibkan untuk mengumpulkan kode Anda yang telah di-zip (tidak boleh menggunakan jenis kompresi yang lain!) ke pengunggah yang disediakan di `http://elearning.uai.ac.id`. Nama file yang Anda kumpulkan haruslah dalam format **tugas1_NIM.ipynb**. Penggunaan nama file selain nama tersebut dapat berakibat tugas Anda tidak diperiksa!

Keterlambatan: Pengumpulan tugas yang melebihi tenggat yang telah ditentukan tidak akan diterima. Keterlambatan akan berakibat pada nilai nol untuk tugas ini.

Kolaborasi: Anda diperbolehkan untuk berdiskusi dengan teman Anda, tetapi dilarang keras menyalin kode maupun tulisan dari teman Anda.

Kecurangan: Menyalin kode orang lain akan berakibat pada nilai nol untuk tugas ini.

1 Scraping [50 poin]

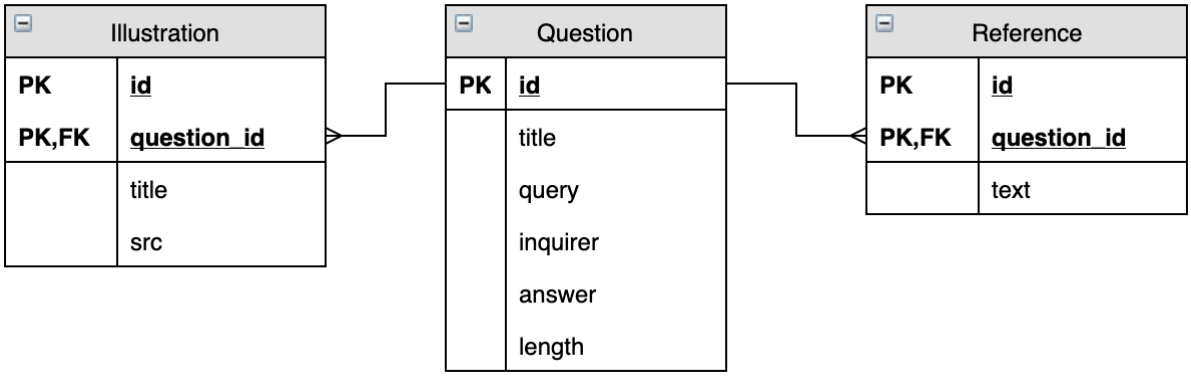
Dengan menggunakan `requests` dan `BeautifulSoup` atau `scrapy`, lakukan *scraping* untuk mengekstraksi:

1. judul,
2. pertanyaan,
3. penanya,
4. jawaban,
5. referensi, dan
6. ilustrasi yang disertakan dalam jawaban

dari situs what if?. Catatan: Anda tidak perlu menyimpan ilustrasinya. Anda hanya perlu mengambil URL ke ilustrasi tersebut.

2 Merapikan Hasil [30 poin]

Dari data yang telah Anda ambil dari situs tersebut, simpan datanya dalam 3 tabel dengan skema seperti pada gambar berikut: Pastikan bahwa:



1. Kolom `answer` hanya berisi teks saja (hapus setiap HTML tag yang ditemukan)
2. Buang semua formula yang ada dalam kolom `answer`
3. Referensi adalah semua informasi yang ada dalam tag `<span class="ref">`
4. Kolom `length` berisi panjang atau jumlah karakter dari kolom `answer`

Skema yang diberikan di atas adalah bentuk minimum yang diminta. Anda diperbolehkan menambahkan kolom atau tabel baru jika dirasa perlu. Namun, Anda tidak diperkenankan mengurangi kolom atau tabel yang ada.

3 Pertanyaan [20 poin]

Berdasarkan informasi yang sudah tersimpan seperti pada pertanyaan nomor 2, jawablah pertanyaan berikut:

1. Tunjukkan lima pertanyaan dengan jawaban terpanjang.
2. Berapa nilai rata-rata dan simpangan baku dari panjang jawaban?
3. Berapa nilai rata-rata dan simpangan baku dari jumlah referensi per pertanyaan?
4. Berapa nilai korelasi antara panjang pertanyaan dan panjang jawaban?
5. Berapa nilai korelasi antara panjang jawaban dan jumlah ilustrasi?

4 Bonus [5 poin]

Tunjukkan 10 kata yang paling sering muncul dalam jawaban, sertakan pula frekuensi untuk masing-masing kata tersebut.