Tebak Jenis Kelamin Berdasarkan Nama

Pengantar

Nama adalah sebuah identifier, termasuk nama seseorang. Dalam memberikan nama untuk anak yang baru lahir, orang tua sering kali mempertimbangkan jenis kelamin anak tersebut, walaupun, ada beberapa nama yang dianggap cocok digunakan untuk anak perempuan dan juga anak laki-laki.

Nah, akhir-akhir ini, identifikasi jenis kelamin berdasar nama seseorang (baik nama resmi, nama panggilan, ataupun nama akun-akun), jadi topik penelitian juga. Mengapa? Oleh karena data gender ini bisa sangat berguna untuk kepentingan-kepentingan lain, misal: melihat pola konsumsi produk, dll.

Bagaimana kita bisa mengidentifikasi gender seseorang berdasar namanya? Cara paling sederhana, dengan melihat nama depannya. Ada beberapa kata yang sifatnya universal, yang menjadi penanda jenis kelamin, misal: Peter, Amanda, Alyssa, Muhammad, dll. Penelitian yang fokus mengidentifikasi jenis kelamin berdasar nama depan antara lain oleh Liu et.al[1], dan juga oleh Ali Septiandri[2] dan Ridho Akbar[3] yang lebih khusus membahas nama orang Indonesia.

Penelitian-penelitian tadi sudah melibatkan penggunakan metode pembelajaran yang cukup rumit. Oleh karena kita baru belajar Pembelajaran Mesin, yuk kita coba versi yang jauh lebih sederhananya terlebih dahulu. Pertama, kita perlu menentukan fitur. Fitur adalah karakteristik dari data yang dianggap tepat untuk menyelesaikan persoalan. Kita ambil contoh dari pekerjaan Ridho Akbar saja dahulu, yaitu kita akan memakai fitur berupa jumlah kemunculan huruf alfabet (a-z) pada nama pertama. Lalu bagaimana menentukan jenis kelamin pemilik nama tersebut?

“letter 'i' occurs more often in female names than in males'. The same apply to 'a','u','e', 't', and 'l'. Otherwise, letter 'b', 'd' and 'o' appears more frequent on males' names.”

Tugas

Buatlah sebuah program dalam bahasa Python, yang akan memproses sebuah masukan berupa nama lengkap seseorang. Keluaran yang diharapkan adalah prediksi jenis kelamin orang tersebut: *laki-laki*, *perempuan*, atau *tidak-diketahui*.

Deskripsi tugas ini masih sangat bebas, jadi mestinya saya akan menjumpai cukup banyak variasi cara pengerjaan ^_^.

Komponen penilaian:

- 40 poin untuk kebenaran program dalam mengesktrak fitur (nilai penuh jika saya bisa memahami program yang kalian tulis, jadi jangan lupa sertakan komentar secukupnya 😊)

- 40 poin untuk kebenaran program dalam implementasi aturan identifikasi jenis kelamin (nilai penuh jika saya bisa memahami program yang kalian tulis, jadi jangan lupa sertakan komentar secukupnya 😊)
- 20 poin untuk komentar kalian, setelah dicobakan ke beberapa nama (termasuk nama sendiri), apa yang kalian amati?

[1] Liu, W., & Ruths, D. (2013, March). What's in a Name? Using First Names as Features for Gender Inference in Twitter. In *AAAI spring symposium: Analyzing microtext* (Vol. 13, p. 01).
[2] Septiandri, A. A. (2017). Predicting the Gender of Indonesian Names. *arXiv preprint arXiv:1707.07129*.
[3] Akbar, Ridho. (2016). Gender Classification of Indonesian Names Using Multinomial Naive Bayes and Random Forrest Classifiers.