TUGAS BESAR II
IF3170 - INTELIGENSI BUATAN

PENERAPAN PEMBELAJARAN MESIN UNTUK KATEGORISASI BERITA OTOMATIS

Jika terdapat perbedaan spesifikasi antara penjelasan saat presentasi dengan dokumen tugas besar ini, maka yang digunakan sebagai acuan resmi adalah dokumen ini.

I. Deskripsi Masalah

Banyaknya berita digital yang dihasilkan setiap hari dapat menyebabkan terjadinya *information overload* bagi para pembaca berita. Kategorisasi berita merupakan salah satu alternatif solusi, sehingga pengguna dapat lebih mudah mencari berita yang dibutuhkan. Berikut adalah contoh kategori berita yang disediakan portal berita.

http://www.republika.co.id/

http://www.tempo.co/

Task kategorisasi berita dapat dilakukan dengan menggunakan sebuah fungsi klasifikasi yang menerima teks berita dan menghasilkan satu kategori dari teks berita tersebut. Fungsi klasifikasi ini dapat diestimasi dengan model klasifikasi yang dibangun menggunakan pembelajaran mesin. Model berisi kumpulan pola teks setiap kategori berita yang ada.

Tugas Besar II kali ini akan membangun aplikasi kategorisasi berita yang mampu:

- membangun model klasifikasi dari dataset berita berlabel dengan menggunakan berbagai teknik pembelajaran mesin yang telah dipelajari sebelumnya di kuliah.
- menerima teks berita baru dan menentukan kategori dari berita tersebut, berdasarkan model klasifikasi yang dimiliki. Masukan berita dapat berupa teks berita (wajib), link berita berupa html (opsional), ataupun hasil crawling otomatis (opsional).
- Menerima kumpulan teks berita baru dan menghasilkan csv yang berisi id berita dan kategori setiap beritanya.
- Mengelola dataset berita berlabel. Jika terdapat berita yang telah salah ditentukan kategorinya (ditentukan oleh pengguna sebagai feedback), sistem dapat membangun model baru (diulang dari awal, bukan incremental).

II. Implementasi Program

Program yang diimplementasikan adalah program **berbasis web** yang memanfaatkan library WEKA untuk pemodelannya. Untuk pemrosesan teks, WEKA juga menyediakan filter StringToWordVector untuk mendapatkan fitur leksikal dari setiap artikel berita.

Program ini memiliki spesifikasi:

1. Masukan untuk pemodelan berisi dataset berita berlabel, dalam bentuk basisdata. Kategori yang terdefinisi sebagai berikut: Pendidikan, Politik, Hukum & Kriminal, Sosial Budaya, Olahraga, Teknologi & sains, Hiburan, Ekonomi dan Bisnis, Kesehatan, dan Bencana & Kecelakaan. Berikut adalah contoh berita dengan kategori Pendidikan. Keluaran dari masukan ini adalah model klasifikasi berita.

> **Tender Naskah UN Diduga Bermasalah Sejak Awal**
> TEMPO.CO, Jakarta - Forum Indonesia untuk Transparansi Anggaran (Fitra) telah menduga PT Ghalia Indonesia Printing tak akan berhasil menyelesaikan tender naskah ujian nasional. Koordinator Investigasi dan Advokasi Uchok Sky Khadafi menilai proses tender perusahaaan tersebut ganjil. "Dari awal saya sudah menduga ini bermasalah," katanya saat dihubungi, Ahad, 14 April 2013. Uchok menyebutkan, dalam lelang, Ghalia menawarkan harga yang lebih tinggi, Rp 22,8 miliar. Sedangkan perusahaan lainnya, PT Aneka Ilmu memberi menawarkan Rp 17 miliar, PT Jasuindo Tiga Perkasa Rp 21,1 miliar, dan PT Balebat Dedikasi Prima Rp 21,6 miliar. Namun kementerian Pendidikan dan Kebudayaan tetap memenangkan perusahaan tersebut. Selain itu, Ghalia ternyata tak hanya mengikuti satu paket lelang. Perusahaan itu juga ikut dalam lelang tiga paket lainnya. Menurut Uchok, ini merupakan bukti Ghalia tak mempertimbangkan kapasitas perusahannya. "Yang penting menang, dan akhirnya bermasalah," ujarnya. PT Ghalia Indonesia Printing adalah perusahaan yang mencetak naskah ujian untuk 11 provinsi. Provinsi tersebut yakni Kalimantan Selatan, Kalimantan Timur, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Gorontalo, dan Sulawesi Barat. Ghalia mengaku kesulitan memasukkan naskah ke boks per sekolah hingga membuat ujian nasional tingkat SMA, MA, dan SMK untuk ke-11 provinisi tersebut ditunda. "Kalau mencetak, kami sudah selesai, tapi ketika memasukan ke boks per sekolah, itu yang kami kesulitan," kata Direktur Ghalia, Hamzah Lukman. Ujian nasional awalnya akan diselenggarakan serentak Senin besok, 15 April 2013. Karena terlambat, jadwal untuk Senin, yakni ujian Bahasa Indonesia, dipindah pekan depan. Untuk Selasa, yakni Bahasa Inggris dan Fisika/Ekonomi, ditunda 23 April 2013. Sementara itu, untuk mata pelajaran Matematika yang seharusnya Rabu, 17 April, digeser ke hari Jumat, 19 April 2013.

2. Masukan untuk pengetesan model dapat berupa kumpulan artikel tidak berlabel . Keluaran dari masukan ini adalah csv kategori berita. Format masukan dan keluaran akan ditentukan oleh asisten.

3. Masukan untuk aplikasi kategorisasi dengan model yang dibangun dapat berupa teks berita (wajib), link berita berupa html (opsional), ataupun hasil crawling otomatis (opsional). Keluaran dari masukan ini adalah kategori berita.

III. Antarmuka

1. Antarmuka **harus** diimplementasi sebagai GUI.
2. Penggunaan library dalam pembuatan antarmuka program **diperbolehkan**.
3. Setiap kelompok harus membuat antarmuka program masing – masing.
4. Kreativitas mempengaruhi penilaian antarmuka.

IV. Deliverable

Aturan mengenai pengumpulan dan demo adalah:

1. Setiap kelompok wajib melakukan asistensi dan laporan progress minimal sebanyak **2 kali** masing-masing **setiap minggu** pada tanggal **16 - 29 November 2014**.
2. Batas pengumpulan Tugas Besar adalah tanggal **01 Desember 2014** pukul **16.28 WIB**. Terlambat mengumpulkan mengakibatkan pengurangan nilai akhir Tugas Besar.
2. Demo akan diadakan pada tanggal **02 – 05 Desember 2014**.
3. Pengisian jadwal demo sudah bisa dilakukan pada tanggal **01 - 02 Desember 2014** saat pengumpulan. Tempat pengisian jadwal akan diberitahukan kemudian. Terlambat mengisi jadwal demo mengakibatkan kehilangan kesempatan demo.
4. Pada saat demo akan dilakukan tes akhir secara individu terkait pembuatan Tugas Besar.
5. Tugas dikumpulkan dalam bentuk CD dengan nama CD:

   TB_II_<nim_terkecil_dalam_kelompok>
6. CD minimal mengandung:
   a. Source code (**source_code.zip**)
   b. Log activity setiap anggota kelompok (disatukan dalam 1 dokumen)
   c. Program yang siap dijalankan **(<nama_program>.zip** yang berisi **.jar** dan asset atau library lain yang diperlukan untuk menjalankan program)
   d. Panduan menggunakan program (**readme.txt**)

V. Kelompok

Kelompok Tugas Besar 1 IF3170 terdiri dari 4 – 5 orang peserta kuliah (**diperkenankan lintas kelas**). Untuk mempermudah proses pengaturan kelompok, peserta harus membuat sebuah dokumen berisi daftar seluruh kelompok dan anggota. Dokumen dikirim 1 kali saja kepada **seluruh** asisten IF3170 maksimal pada tanggal 14 Des 2014. Daftar dikirimkan melalui *e-mail* (kontak dapat dilihat pada poin IX) dalam file **kelompok.csv** dengan format:

1. Baris pertama diisi dengan daftar NIM peserta kuliah yang belum mendapatkan kelompok dipisahkan dengan koma (,).
2. Baris kedua dan seterusnya berisi daftar kelompok. Setiap baris berisi NIM anggota kelompok dipisahkan dengan koma.

Contoh:

```
13512700,13512800,13512900,13512701,13512801,13512901
13512702,13512802,13512902,13512713,13512704
13512712,13512822,13512942,13512703,13512774
13512722,13512812,13512932,13512733,13512794
…
13512732,13512832,13512922,13512743,13512734
13512742,13512842,13512912,13512723
```

catatan: asisten berhak melakukan modifikasi terhadap komposisi kelompok jika memang diperlukan.

VI. Penilaian

Bobot penilaian Tugas Besar 1 IF3054 adalah:

No	Komponen	Max
1	Pemodelan klasifikasi berita	20
2	Akurasi model	20
2	Kelengkapan fitur program	20
3	Keberhasilan ujicoba	20
4	Antarmuka	10
5	Tes akhir (per orang)	10
6	Akurasi terbaik (3 kelompok)	20
	Nilai maksimal	120

VII. Kontak Asisten

Daftar asisten IF3170 beserta *e-mail* yang dapat dihubungi:

1. Mahdan Ahmad Fauzi Al-Hasan (13510104@std.stei.itb.ac.id)
2. David Setyanugraha (13511003@std.stei.itb.ac.id)
3. Alifa Nurani Putri (13511074@std.stei.itb.ac.id)
4. Genta Indra Winata (13511094@std.stei.itb.ac.id)

VIII. Referensi

Mitchell, Tom M. "Machine learning." (1997).