BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

ISYS8036 - Business Intelligent and Analytics

Topic 11

MEREPRESENTASIKAN DAN MEMINING TEXT

LEARNING OUTCOMES

Setelah mempelajari materi ini peserta kuliah diharapkan mampu melakukan penalaran/ penarikan kesimpulan berdasar / melalui :

- Constructing mining-friendly data representations;
- Representation of text for data mining.
- Bag of words representation;
- TFIDF calculation;
- N-grams;
- Stemming;

OUTLINE MATERI :

1. Bekerja dengan text
2. Representasi Text
3. Bag of Words
4. Term Frequency
5. Menggabungkan TFIDF
6. Kesimpulan

PENDAHULUAN

Hingga saat ini kita mengabaikan atau mengesampingkan tahap penting dari proses penambangan data yaitu: mempersiapkan data. Dunia tidak selalu menyediakan kita data dalam representasi vektor fitur yang disukai kebanyakan metode penambangan data ambil sebagai input/ masukan. Data selalu direpresentasikan secara alami sesuai bidangnya. Jika kita ingin menerapkan banyak tool penambangan data yang kita miliki, kita harus merekayasa representasi data agar sesuai dengan tool tool itu, atau membuat tool baru yang cocok dengan data yang tersedia. Para ilmuwan data biasanya menggunakan kedua strategi ini. Secara umum akan lebih mudah untuk mempersiapkan data agar sesuai dengan tool yang ada.

Dalam Sesi ini, kita akan fokus pada satu jenis data yang telah menjadi sangat umum karena Internet telah menjadi saluran komunikasi di mana-mana: data teks. Memeriksa data teks memungkinkan kita untuk mengilustrasikan banyak kerumitan nyata rekayasa data, dan juga membantu kita untuk memahami lebih baik jenis data yang sangat penting.

Pada prinsipnya, teks hanyalah bentuk lain dari data, dan pemrosesan teks hanyalah kasus khusus rekayasa representasi. Berurusan dengan teks membutuhkan langkah-langkah pra-pemrosesan khusus dan kadang-kadang keahlian khusus tim ilmu data. Pertama, mari kita bahas mengapa bekerja dengan teks sangat penting dan sulit.

Bekerja Dengan Text

Teks ada di mana-mana. Banyak aplikasi menghasilkan atau merekam teks. Catatan medis, log keluhan konsumen, permintaan informasi produk, dan catatan perbaikan sebagian besar masih ditujukan sebagai komunikasi antara orang, bukan komputer,. Memanfaatkan sejumlah besar data ini dibutuhkan proses mengubahnya menjadi bentuk yang berarti.

Internet merupakan "media baru," tetapi secara substansi masih sama seperti media tradisional. Internet berisi sejumlah besar teks dalam bentuk halaman web pribadi, Twitter, email, status Facebook, deskripsi produk, komentar Reddit, posting blog — dst dst. Tengok mesin pencari (Google dan Bing) yang kita gunakan sehari-hari. Keduanya berisi sejumlah besar ilmu

pengetahuan yang berorientasi pada teks. Musik dan video mungkin mendominasi trafik internet, tetapi ketika orang berkomunikasi satu sama lain di Internet biasanya melalui teks. Web 2.0 memungkinkan pengguna untuk berinteraksi satu sama lain sebagai komunitas, dan menghasilkan banyak konten tambahan dari sebuah situs. Konten dan interaksi yang dibuat pengguna ini biasanya berbentuk teks.

Dalam dunia bisnis, memahami umpan balik pelanggan biasanya menggunakan teks. Ini tidak selalu terjadi; harus diakui, beberapa sikap konsumen yang penting diwakili secara eksplisit sebagai data atau dapat disimpulkan melalui perilaku, misalnya melalui peringkat bintang lima, pola klik-tayang, tingkat konversi, dan sebagainya. Kita juga dapat membayar agar data dikumpulkan dan dikuantifikasi melalui kelompok fokus dan survei online. Tetapi dalam banyak kasus jika kita ingin “mendengarkan pelanggan” kita harus benar-benar membaca apa yang ditulisnya — dalam ulasan produk, formulir umpan balik pelanggan, potongan opini, dan pesan email.

Bekerja dengan Text Sulit

Teks sering disebut sebagai data "tidak terstruktur". Ini mengacu pada fakta bahwa teks tidak memiliki struktur yang biasanya kita harapkan untuk data: berbentuk tabel (koleksi vektor fitur). Teks tentu saja memiliki banyak struktur, tetapi struktur linguistik - dimaksudkan untuk konsumsi manusia, bukan untuk komputer. Kata-kata dapat memiliki panjang yang bervariasi dan dalam suatu dokumen jumlah kata dapat sangat bervariasi. Terkadang urutan kata menjadi penting.

Sebagai data, teks termasuk sangat tak teratur. Orang menulis dengan tatabahasa yang salah , salah mengeja kata-kata, dan membuat singkatan seenaknya. Bahkan ketika teks diungkapkan dengan sempurna mungkin berisi sinonim (beberapa kata dengan arti yang sama) dan homograf (satu ejaan memiliki arti yang berbeda). Terminologi dan singkatan dalam satu domain mungkin tidak berarti di domain lain. Konteks menjadi sangat penting dalam komunikasi menggunakan text.

Dengan alasan ini, teks harus menjalani tahap pra-processing yang memadai sebelum dapat digunakan sebagai input ke algoritma penambangan data. Biasanya semakin kompleks, semakin banyak aspek dari masalah teks yang perlu dipersiapkan. Sesi ini hanya menjelaskan beberapa metode dasar dalam menyiapkan teks untuk penambangan data.

Representasi Text

Setelah membahas tentang betapa sulitnya mengolah teks, mari kita pelajari langkah-langkah dasar untuk mengubah kumpulan teks menjadi sekumpulan data yang dapat dimasukkan ke dalam algoritma penambangan data. Namun demikian, ide-ide ini adalah teknologi kunci yang mendasari banyak pencarian web, seperti Google dan Bing.

Pertama, akan diperkenalkan beberapa terminologi dasar. Dokumen adalah satu bagian dari teks, tidak peduli seberapa besar atau kecil. Dokumen bisa berupa satu kalimat atau 100 halaman laporan, atau apa pun di antaranya, seperti komentar YouTube atau postingan blog. Dokumen terdiri atas token atau term term. Untuk saat ini, pikirkan token atau term sebagai sebuah kata; Kumpulan dokumen disebut Corpus.

Bag of Words

Penting untuk diingat tujuan dari tugas representasi teks. Intinya, kita mengambil satu set dokumen — masing-masingnya adalah rangkaian kata-kata yang relatif bebas — dan mengubahnya menjadi bentuk fitur-vektor yang kita kenal. Setiap dokumen adalah satu instans tanpa fitur, karena kita tidak tahu sebelumnya fitur apa yang akan ada.

Pendekatan yang kita perkenalkan lebih dulu disebut “bag of words.” Pendekatan ini memperlakukan setiap dokumen hanya sebagai kumpulan kata-kata individual. Pendekatan ini mengabaikan tata bahasa, susunan kata, struktur kalimat, dan (biasanya) tanda baca. Ini memperlakukan setiap kata dalam dokumen sebagai kata kunci yang berpotensi penting dari dokumen.

Jika setiap kata adalah calon sebuah fitur, apa yang akan menjadi nilai fitur dalam dokumen yang diberikan? Ada beberapa pendekatan untuk ini. Dalam pendekatan paling dasar, setiap kata adalah token, dan setiap dokumen diwakili oleh satu (jika token ada dalam dokumen) atau nol (token tidak ada dalam dokumen). Pendekatan ini merepresentasikan dokumen ke kumpulan kata-kata yang terkandung di dalamnya.

Frekwensi Term

Representasi lain adalah menggunakan hitungan kata (frekuensi) dalam dokumen, bukan hanya nol atau satu. Ini memungkinkan kita untuk membedakan antara berapa kali sebuah kata digunakan; dalam beberapa aplikasi, pentingnya istilah dalam dokumen harus meningkat seiring dengan berapa kali istilah itu terjadi. Ini disebut representasi frekuensi kata.

Setiap kalimat dianggap sebagai dokumen terpisah. Pendekatan bag-of-words sederhana menggunakan frekuensi term akan menghasilkan tabel jumlah jangka yang ditunjukkan pada Tabel 10-1.

Tabel 10-1. Representasi frekuensi term.

	a	explain	hard	has	is	jazz	music	natural	rhythm	swing	to
d1	1	0	0	1	0	1	1	0	1	1	0
d2	0	1	1	0	1	0	0	0	0	1	1
d3	1	0	0	0	1	0	0	1	2	1	0

Biasanya beberapa pemrosesan awal dilakukan pada kata-kata sebelum menempatkannya ke dalam tabel.

- Pertama, normalisasi: setiap istilah diubah ke huruf kecil. Ini agar kata-kata seperti Skype dan SKYPE dihitung sebagai hal yang sama.
- Menghilangkan prefix dan hanya menggunakan kata dasar (stem).
- Akhirnya, stopwords telah dihapus. Kata yang sangat umum dalam bahasa Inggris (atau bahasa apa pun yang diurai). Kata-kata yang, dan, dari, dan pada dianggap stopwords dalam bahasa Inggris sehingga mereka biasanya dihapus.

Alih-alih melakukan perhitungan secara kasar, beberapa sistem melakukan langkah normalisasi frekuensi istilah sehubungan dengan panjang dokumen. Untuk menyesuaikan panjang dokumen, frekuensi jangka waktu baku dinormalisasi dengan cara tertentu, seperti dengan membagi masing-masing dengan jumlah total kata dalam dokumen.

Ukuran Kelaziman: Inverse Document Frequency

Term frekuensi mengukur seberapa lazim suatu istilah dalam satu dokumen. Kita juga mungkin peduli, ketika menentukan bobot istilah, seberapa umum itu di seluruh korpus yang kita fokuskan. Ada dua pertimbangan yang saling bertentangan.

Pertama, sebuah istilah seharusnya tidak terlalu langka. Misalnya, kata pengaya yang tidak biasa terjadi hanya dalam satu dokumen dalam korpus Anda. Apakah ini istilah yang penting? Ini mungkin tergantung pada aplikasi. Untuk pengambilan, istilah ini mungkin penting karena pengguna mungkin mencari kata yang tepat. Untuk pengelompokan, tidak ada gunanya menyimpan istilah yang hanya terjadi satu kali: itu tidak akan pernah menjadi dasar dari kelompok yang berarti. Karena alasan ini, sistem pemrosesan teks biasanya menetapkan batas bawah yang kecil (sewenang-wenang) pada jumlah dokumen di mana suatu istilah harus terjadi.

Pertimbangan lain, berlawanan adalah bahwa istilah seharusnya tidak terlalu umum. Istilah yang muncul di setiap dokumen tidak berguna untuk klasifikasi (tidak membedakan apa pun) dan tidak dapat berfungsi sebagai basis untuk kluster (seluruh korpus akan mengelompok bersama).

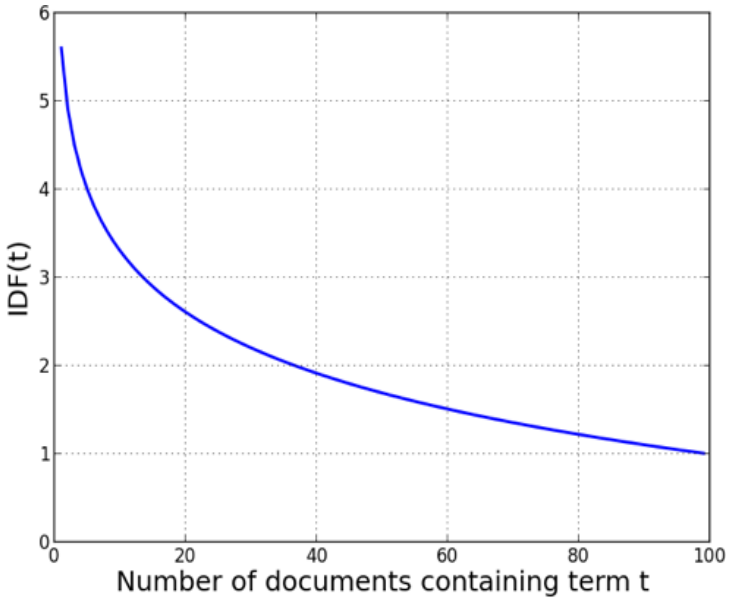
Istilah yang terlalu umum biasanya dihilangkan. Salah satu cara untuk melakukan ini adalah dengan memaksakan batas atas yang sewenang-wenang pada nomor (atau fraksi) dari dokumen di mana suatu kata dapat terjadi.

Selain menerapkan batas atas dan bawah pada frekuensi istilah, banyak sistem memperhitungkan distribusi istilah di atas korpus juga. Semakin sedikit dokumen di mana sebuah istilah terjadi, semakin signifikan kemungkinannya untuk berada di dokumen-dokumen. Terjadi istilah yang terbatas ini, biasanya diukur dengan persamaan yang disebut inverse document frequency (IDF), yang ditunjukkan dalam Persamaan 10-1.

$$\text{IDF} = 1 + \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing } t}\right)$$

Persamaan 10-1. Inverse Document Frequency (IDF) suatu istilah

IDF dapat dipandang sebagai ukuran kelangkaan. Gambar 10-1 menunjukkan grafik IDF (t) sebagai fungsi dari jumlah dokumen di mana t terjadi, dalam korpus 100 dokumen. Seperti yang Anda lihat, ketika sebuah istilah sangat langka (paling kiri) IDF cukup tinggi. Ini menurun dengan cepat karena t menjadi lebih umum dalam dokumen, dan asymptotes pada 1.0. Sebagian besar kata kunci, karena prevalensi mereka, akan memiliki IDF dekat satu.



Gambar 10-1. IDF dari istilah t dalam korpus 100 dokumen.

Menggabungkan TFIDF

Perhatikan bahwa nilai TFIDF spesifik untuk satu dokumen (d) sedangkan IDF bergantung pada seluruh korpus. Sistem yang menggunakan representasi sekotak kata biasanya melalui langkah-langkah menghentikan dan menghentikan kata-kata sebelum melakukan penghitungan waktu. Jumlah istilah dalam dokumen membentuk nilai-nilai TF untuk setiap istilah, dan jumlah dokumen di seluruh korpus membentuk nilai-nilai IDF.

Setiap dokumen dengan demikian menjadi vektor fitur, dan korpus adalah himpunan vektor fitur ini. Set ini kemudian dapat digunakan dalam algoritma penambangan data untuk klasifikasi, pengelompokan, atau pengambilan.

$$\mathrm{TFIDF}(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t)$$

Pendekatan representasi teks bag-of-words memperlakukan setiap kata dalam dokumen sebagai kata kunci potensial independen (fitur) dari dokumen, kemudian menetapkan nilai untuk setiap dokumen berdasarkan frekuensi dan kelangkaan. TFIDF adalah representasi nilai yang sangat umum untuk istilah, tetapi belum tentu optimal. Jika seseorang menggambarkan penambangan korpus teks menggunakan sekantong kata, itu artinya mereka memperlakukan setiap kata secara individual sebagai fitur. Nilai-nilai mereka bisa berupa biner, frekuensi istilah, atau TFIDF, dengan normalisasi atau tanpa. Para ilmuwan data mengembangkan intuisi tentang cara terbaik untuk menyerang masalah teks yang diberikan, tetapi mereka biasanya akan bereksperimen dengan berbagai representasi untuk melihat mana yang menghasilkan hasil terbaik.

Kesimpulan

Masalah kita tidak selalu menghadirkan kita data dalam representasi vektor fitur rapi yang sebagian besar metode penambangan data sebagai input. Masalah dunia nyata sering membutuhkan beberapa bentuk rekayasa representasi data agar mereka dapat menambang. Secara umum lebih mudah untuk terlebih dahulu mencoba merancang data agar sesuai dengan alat yang ada. Data dalam bentuk teks, gambar, suara, video, dan informasi spasial biasanya memerlukan preprocessing khusus —dan kadang-kadang pengetahuan khusus dari tim sains data.

Dalam sesi ini, kita membahas satu jenis data yang paling umum yang membutuhkan preprocessing: teks. Cara umum untuk mengubah teks menjadi vektor fitur adalah memecah setiap dokumen menjadi kata-kata individual (representasi "tas kata-kata"), dan menetapkan nilai untuk setiap istilah menggunakan rumus TFIDF. Pendekatan ini relatif sederhana, murah dan serbaguna, dan membutuhkan sedikit pengetahuan tentang domain, setidaknya pada awalnya. Terlepas dari kesederhanaannya, ia melakukan dengan sangat baik pada berbagai tugas.

DAFTAR PUSTAKA

1. Foster Provost & Tom Fawcett (2013) Data Science for Business: What you need to know about data mining and data analytic thinking, O'Reilly, ISBN: 978-1-449-36132-7.
2. Sharda, R., Delen, D., Turban, E., (2018). Business intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.