PERLOMBONGAN DATA TEKS

STQD6414 PERLOMBONGAN DATA

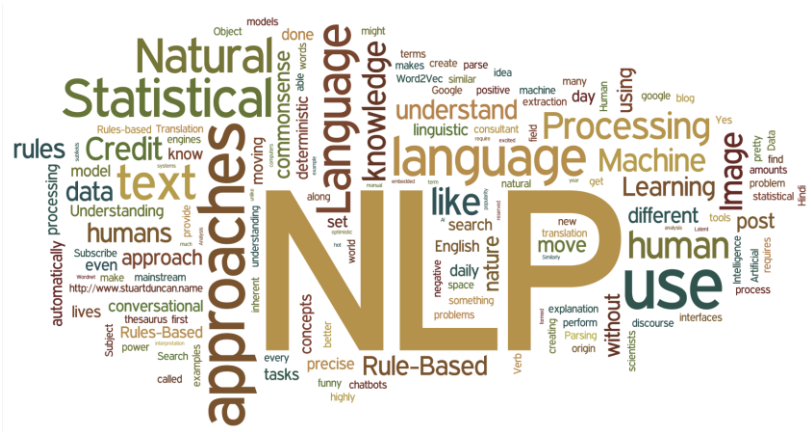
Prof. Madya Dr. Nurulkamal Masseran

Jabatan Sains Matematik

Universiti Kebangsaan Malaysia

PENGENALAN:

- Data teks merupakan data jenis tak berstruktur.
- Komputer perlu mentafsir data tak berstruktur untuk memahami bahasa manusia.
- Ini penting supaya maklumat data jenis teks boleh dianalisis oleh komputer.
- Prosedur ini dikenali sebagai pemprosesan bahasa semula jadi (*natural language processing (NLP)*).
- Perlombongan teks menggunakan teknik NLP untuk menjelmakan data tak berstruktur kepada bentuk berstruktur bagi tujuan mengenalpasti corak yang bermakna dan dapat mengekstrak maklumat data.



PENGENALAN:

- Data jenis teks boleh dijana oleh pelbagai sumber platform seperti: e-mel, ulasan produk, media sosial, surat khabar, maklum balas pelanggan, dokumen, fail, dan lain-lain.
- Walau bagaimanapun, teknik analisis data biasa tidak boleh digunakan untuk berurusan dengan data jenis teks.
- Oleh itu, teknik perlombongan teks memainkan peranan penting.

- Khususnya, teknik perlombongan teks berguna dalam:

i) mengenal pasti trend, topik popular dan tema yang berkaitan dengan isu-isu tertentu.

ii) mengekstrak sentimen dan emosi terhadap sesuatu isu tertentu.

- Contoh: Dalam perniagaan, data maklum balas daripada pelanggan membantu syarikat mendapatkan maklumat berkaitan persepsi dan pendapat terhadap produk atau perkhidmatan mereka.

KORPUS TEKS:

- Data teks biasanya dijelmakan kepada format korpus (*corpus*) sebelum analisis perlombongan data boleh dilakukan.
- Korpus teks merupakan koleksi teks bertulis.
- Berdasarkan format korpus, analisis perlombongan data, ujian hipotesis, penyemakan keberlakuan, pengesahan hukum linguistik dan lain-lain analisis boleh dijalankan terhadap data teks.

TEXT	CORPUS
Read whole	Read fragmented
Read horizontally	Read vertically
Read for content	Read for formal patterning
Read as a unique event	Read for repeated events
Read as an individual act of will	Read as a sample of social practice
Coherent communicative event	Not a coherent communicative event

PEMBERSIHAN DATA TEKS:

- Umumnya, data tidak berstruktur tidak disusun dengan teratur.
- Oleh itu, adalah sukar untuk menganalisis data tidak berstruktur secara terus daripada bentuk asalnya.
- Pembersihan data teks adalah prosedur yang sangat penting sebagai teknik pra-pemprosesan sebelum analisis data teks boleh dijalankan.
- Beberapa langkah penting dalam pembersihan data teks:

i) Keluarkan aksara khas daripada teks, iaitu simbol-simbol; /, @, | akan digantikan dengan ruang kosong.

ii) Tukar teks huruf besar kepada huruf kecil.

iii) Keluarkan nombor-nombor.

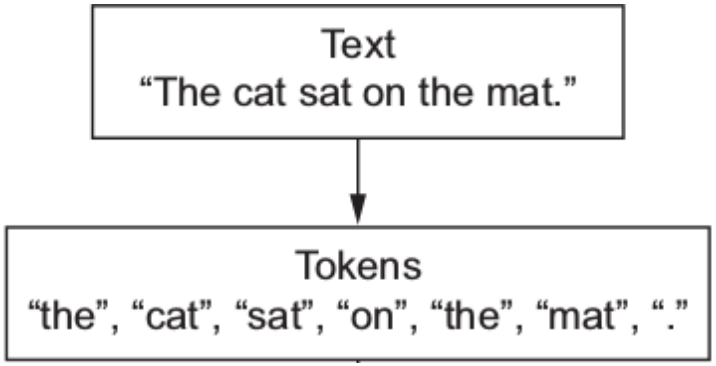
iv) Keluarkan kata henti (*stopwords*). Contoh kata henti dalam bahasa Inggeris “the, is, at, on”. Tiada senarai semesta (*universal*) kata henti yang digunakan dalam NLP.

v) Keluaran tanda baca (*punctuation*).

vi) Buang semua ruang tambahan yang tidak perlu dalam teks.

TOKENISASI PERKATAAN:

- Tokenisasi (*tokenization*) ialah teknik yang digunakan untuk mewakili perkataan kepada format angka yang kemudiannya boleh digunakan dalam perlombongan teks.
- Tokenisasi perkataan (word tokenization) melibatkan proses pembahagian teks kepada perkataan tunggal dengan setiap perkataan unik diberikan nombor yang unik.
- Token boleh mewakili perkataan, nombor atau tanda baca.
- Dalam tokenisasi, unit yang lebih kecil dibina dengan mendapatkan sempadan perkataan (*word boundaries*).
- Sempadan perkataan ialah titik akhir sesuatu perkataan dengan permulaan perkataan seterusnya.
- Tokenisasi ialah langkah pertama untuk proses pembendungan teks (*text stemming*).



TOKENISASI PERKATAAN :

Contoh:

- Pertimbangkan ayat, “I Love my cat”.
- Dalam tokenisasi, nilai integer unik akan diberikan kepada setiap perkataan unik dalam ayat tersebut, sedemikian hingga, 1 kepada “I”, 2 kepada “Love”, 3 kepada “my” dan 4 kepada “cat”.
- Jika kita mempunyai ayat lain: “I Love my car”, maka, perkataan “I Love my”, sudah mempunyai nombor unik 001 002 003.
- Oleh itu, hanya perkataan baharu akan diberikan nilai integer unik baru. Iaitu, integer 005 ditetapkan untuk perkataan “car”.
- Token-token bagi dua ayat tersebut ialah: 001 002 003 004, 001 002 003 005
- Berdasarkan tokenisasi, ciri-ciri kesamaan antara ayat boleh dianalisis.
- Secara intrinsik, komputer tidak memahami teks atau bahasa dalam erti kata manusia.
- Dengan tokenisasi, komputer mampu mengubah teks daripada bentuk yang difahami manusia kepada corak statistik yang boleh dipetakan menerusi teknik perlombongan data.

PEMBENDUNGAN TEKS:

- Pembendungan teks (*text stemming*) ialah proses menurunkan perkataan kepada bentuk akarnya (*root form*).
- Disamping itu, pembendungan juga berguna dalam mengurangkan dimensi data teks
- Oleh itu, pembendungan adalah teknik penting yang digunakan dalam NLP untuk mengurangkan kerumitan algorithma perlombongan teks.
- Contoh: proses pembendungan menurunkan perkataan: “fishing”, “fished” dan “fisher” kepada punca/akar perkataannya iaitu “fish”.
- Pembendungan meringkaskan beberapa perkataan berbeza kepada bentuk asalnya yang membawa makna yang sama dari segi konteks.

	original_word	stemmed_words
0	connect	connect
1	connected	connect
2	connection	connect
3	connections	connect
4	connects	connect

MATRIKS SEBUTAN-DOKUMEN:

- Matriks sebutan-dokumen (*document-term matrix*) mewakili hubungan antara sebutan dan dokumen.
- Baris matriks mewakili dokumen atau ayat tertentu dalam data teks.
- Kolum matriks mewakili perkataan yang unik dalam data teks.
- Pemasukan dalam matriks ini mewakili bilangan istilah dalam dokumen atau ayat tertentu.
- Matriks sebutan-dokumen juga dirujuk sebagai jadual maklumat kekerapan perkataan.

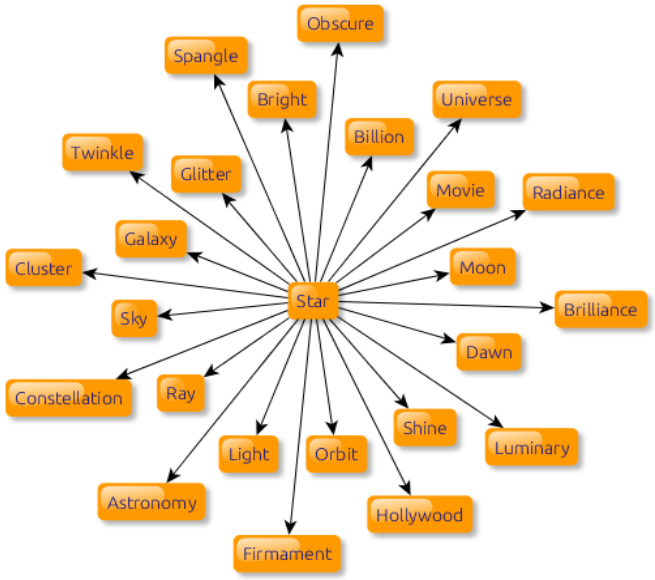
	text	mining	is	to	find	useful	information	from	text	mined	dark	came
D1	1	1	1	1	1	1	1	1	1	0	0	0
D2	0	0	1	0	0	1	1	1	1	1	0	0
D3	0	0	0	0	0	0	0	0	0	0	1	1

AWAN PERKATAAN:

- Awan perkataan (*word cloud*) ialah perwakilan visual bagi kekerapan perkataan dalam data teks.
- Awan perkataan menggunakan maklumat yang ditunjukkan oleh jadual kekerapan perkataan daripada matriks sebutan-dokumen.
- Awan perkataan berguna untuk menyerlahkan perkataan dan frasa yang popular berdasarkan kekerapan dan perkaitannya dalam data teks.
- Berdasarkan rajah awan perkataan, analisis yang lebih mendalam terhadap data teks boleh dijalankan.
- Rajah awan perkataan mengekstrak kata kunci yang terdapat dalam data teks.
- Saiz setiap perkataan menunjukkan magnitud kekerapannya.

PERKAITAN PERKATAAN:

- Perkaitan perkataan (*word association*) ialah teknik menganalisis kandungan data teks dengan mengenalpasti hubungan yang signifikan antara istilah/sebutan.
- Teknik ini menghitung kesamaan antara setiap perkataan dalam dokumen, selepas mengumpulnya melalui beg perkataan (*bag of words*)
- Ukuran korelasi seringkali digunakan untuk menentukan magnitud kekuatan pasangan perkataan yang berkaitan.
- Aturan perkataan (*word association*) boleh digambarkan menggunakan carta kesamaan perkataan dan carta graf korelasi perkataan.



ANALISIS SENTIMEN:

- Analisis sentimen ialah proses mengekstrak pendapat-pendapat yang mempunyai markah yang berbeza yang merujuk kepada sentimen positif, negatif atau neutral.
- Berdasarkan analisis sentimen, anda boleh mengetahui sifat dan emosi suatu pendapat atau ayat dalam teks.
- Analisis sentimen ialah teknik klasifikasi data teks.
- Data boleh diklasifikasikan ke dalam kelas yang berbeza seperti positif, negatif atau gembira, sedih, marah, dan sebagainya.
- Analisis sentimen digunakan untuk banyak aplikasi, terutamanya dalam risikan perniagaan (*business intelligence*).
- Contoh:

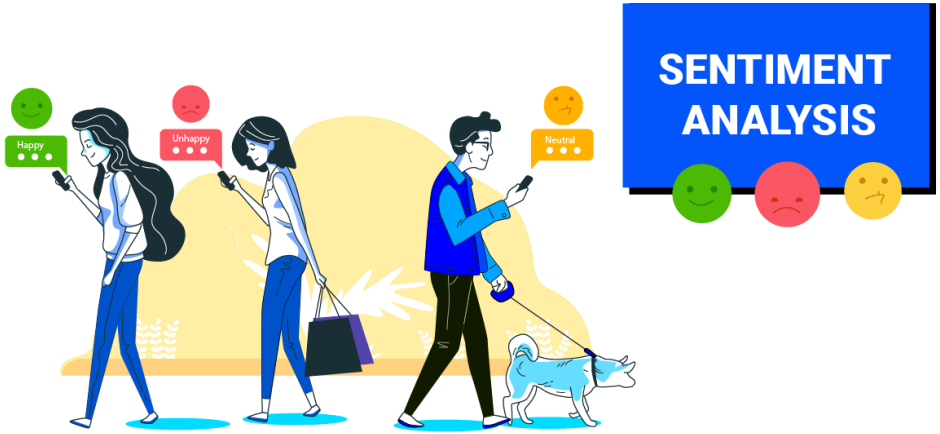
i) Menganalisis perbincangan media sosial berkaitan topik tertentu.

ii) Menganalisis respons suatu kaji selidik.

iii) Menentukan sama ada ulasan produk adalah positif atau negatif.

ANALISIS SENTIMEN:

- Walau bagaimanapun, analisis sentimen tidak dapat mencungkil maklumat mengapa sesetengah orang merasakan emosi tertentu.
- Walaupun begitu, analisis sentimen memberikan maklumat tentang perkataan yang dikaitkan dengan sentimen positif atau negatif yang kuat.
- Maklumat ini diperoleh dengan menghitung bilangan perkataan positif dan negatif dalam teks.
- Kemudian, analisis akan dijalankan untuk mencirikan gabungan perkataan positif dan negatif ini.



ANALISIS SENTIMEN:

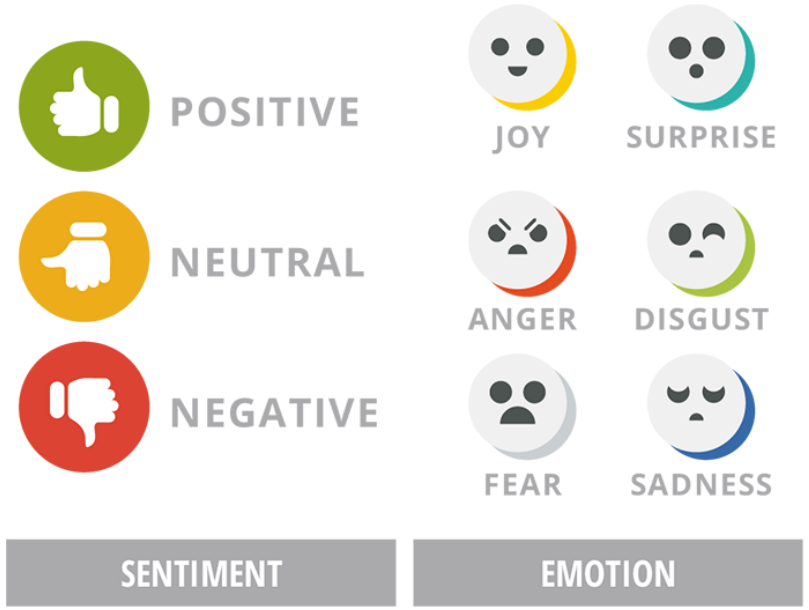
- Langkah pertama dalam analisis sentimen ialah mencipta teks leksikon (*lexicon text*).
- Teks leksikon merujuk kepada senarai perkataan.
- Umumnya, beberapa teks leksikon sudah ditakrif dalam perisian perlombongan teks.
- Namun, jika teks anda mempunyai topik tertentu yang khas, leksikon anda perlu ditambah atau diubah suai sebelum analisis sentimen dijalankan.

Lexicon	Positive Words	Negative Words
Simplest (SM)	good	bad
Simple List (SL)	good, awesome, great, fantastic, wonderful	bad, terrible, worst, sucks, awful, dumb
Simple List Plus (SL+)	good, awesome, great, fantastic, wonderful, best, love, excellent	bad, terrible, worst, sucks, awful, dumb, waist, boring, worse
Past and Future (PF)	will, has, must, is	was, would, had, were
Past and Future Plus (PF+)	will, has, must, is, good, awesome, great, fantastic, wonderful, best, love, excellent	was, would, had, were, bad, terrible, worst, sucks, awful, dumb, waist, boring, worse
Bing Liu	2006 words	4783 words
AFINN-96	516 words	965 words
AFINN-111	878 words	1599 words
enchantedlearning.com	266 words	225 words
MPAA	2721 words	4915 words
NRC Emotion	2312 words	3324 words

KLASIFIKASI EMOSI:

- Berbanding hanya dua sentimen (negatif dan positif), klasifikasi emosi juga memberikan maklumat berguna dalam analisis sentimen.
- Klasifikasi emosi yang dibina merujuk kepada NRC Word-Emotion Association Lexicon.
- NRC Emotion Lexicon merupakan senarai lapan emosi asas (dalam English) :

i) Anger
ii) Fear
iii) Anticipation
iv) Trust
v) Surprise
vi) Sadness
vii) Joy
viii) Disgust



RUJUKAN:

- Aggarwal, C. C., Zhai, C. (2012). *Mining Text Data*. Springer.
- Kwartler, T. (2017). *Text Mining in Practice with R*. Wiley.
- Lamba, M., Madhusudhan, M. (2022). *Text Mining for Information Professionals: An Uncharted Territory*. Springer.
- Silge, J., Robinson, D. (2017). *Text Mining with R: : A Tidy Approach*. O'Reilly Media, Inc.
- Zhai, C., Massung, S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. ACM Books.
- Žižka, J., Darena, F., Svoboda, A. (2021). *Text Mining with Machine Learning*. CRC Press.
- Zong, C., Xia, R., Zhang, J. (2021). Text Data Mining. Springer.

TOPIK SETERUSNYA:

Perlombongan Data Graf