Text Mining

Ali Akbar Septiandri

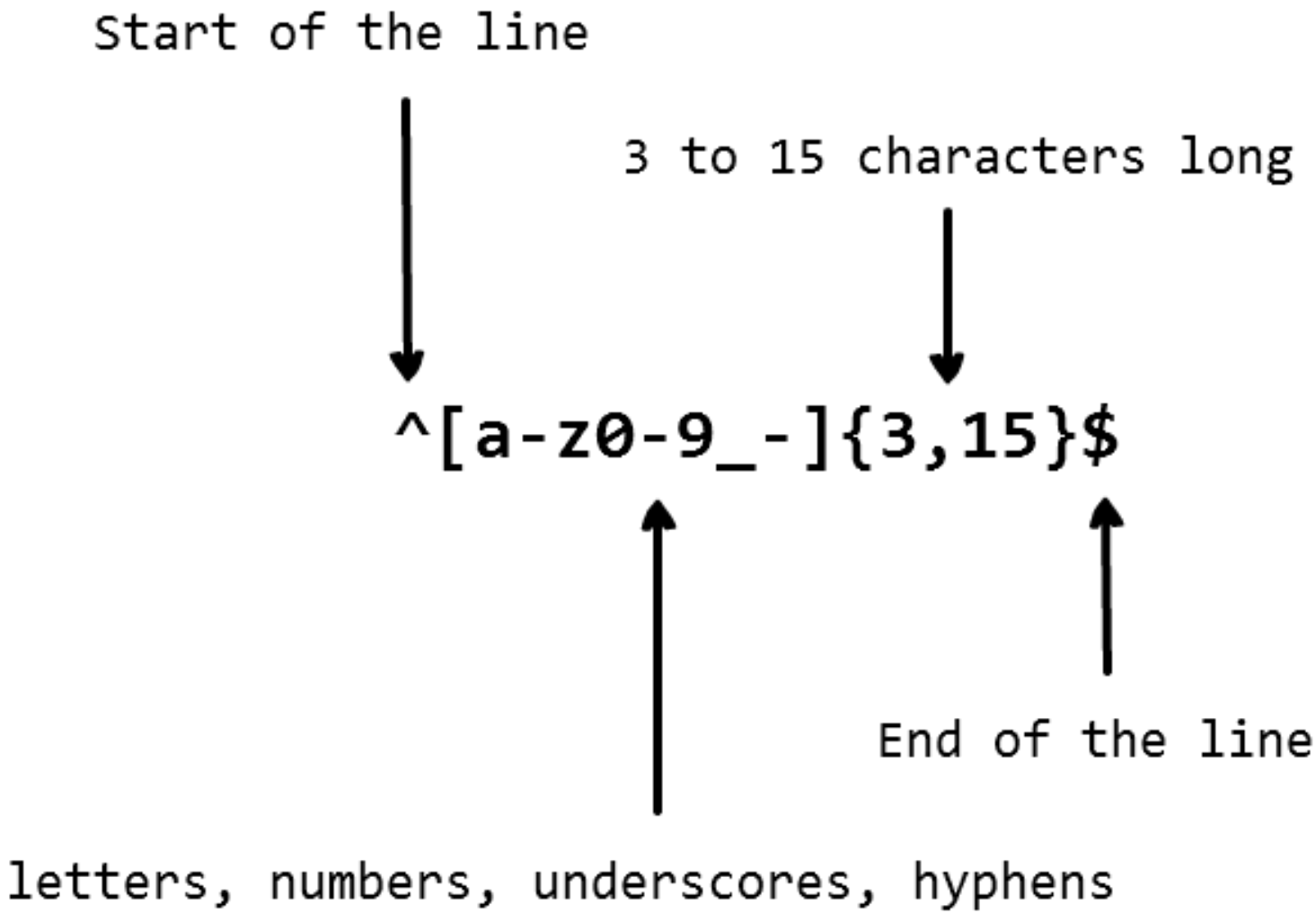
Universitas Al Azhar Indonesia

March 5, 2019

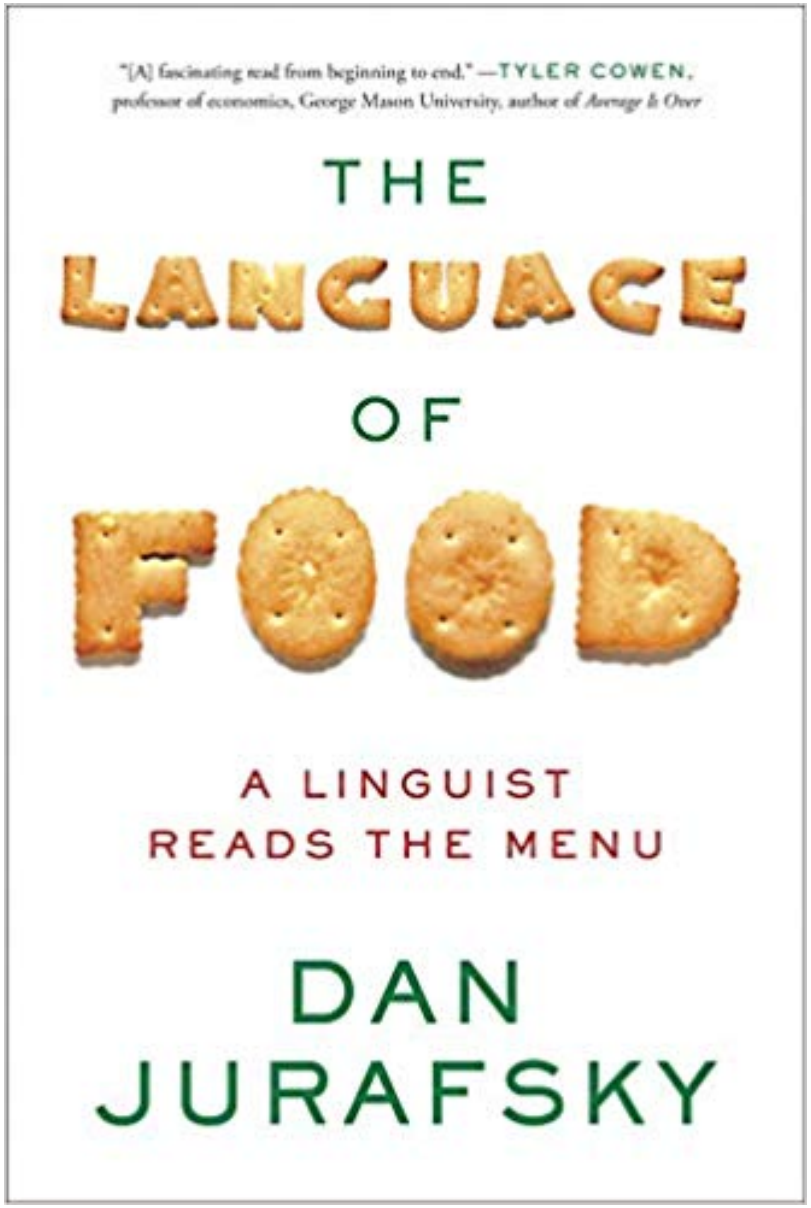
Pendahuluan

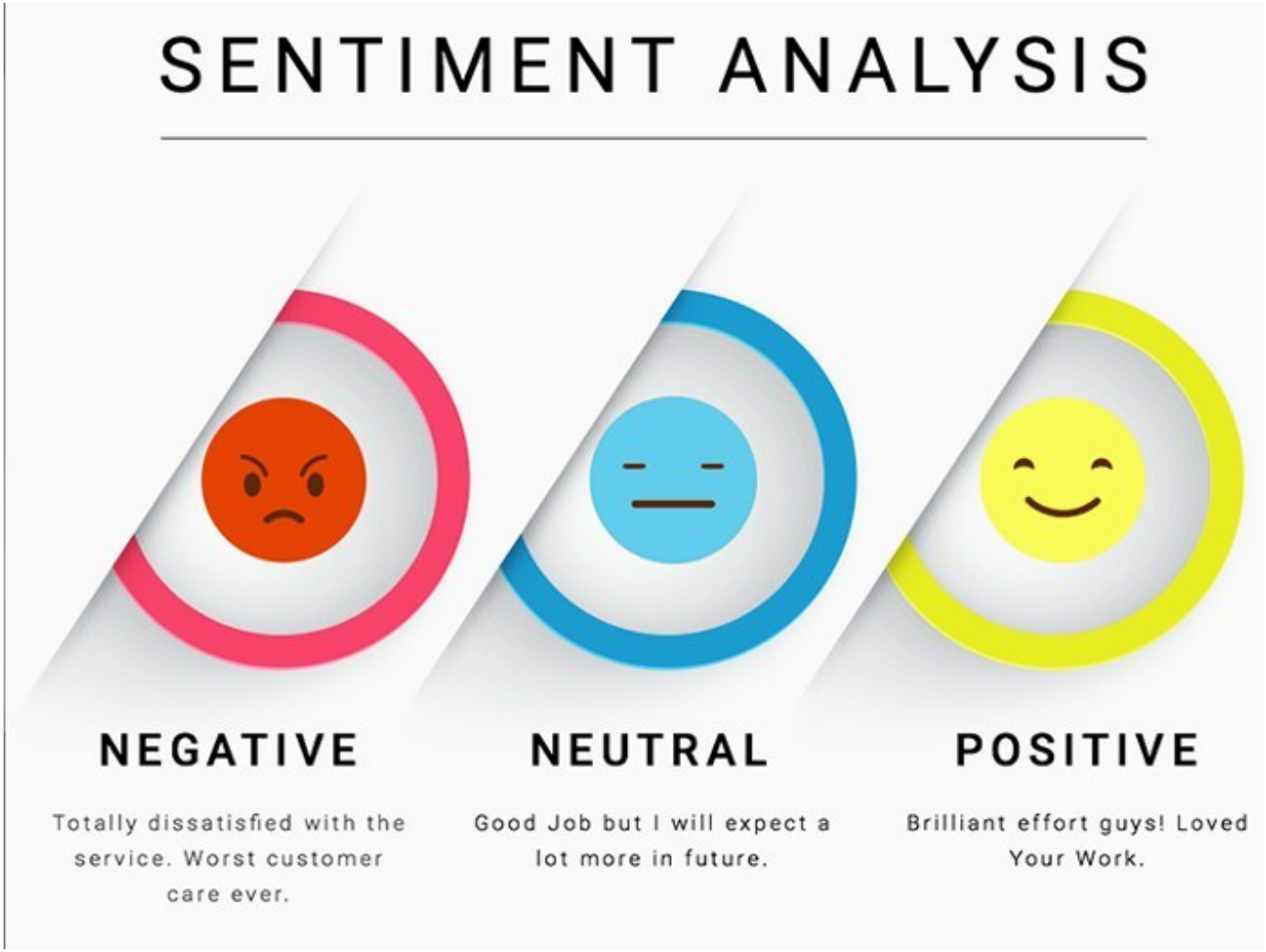
Kuliah apa ini?
Apa perbedaannya dengan Data Mining?

Data mining dengan data teks

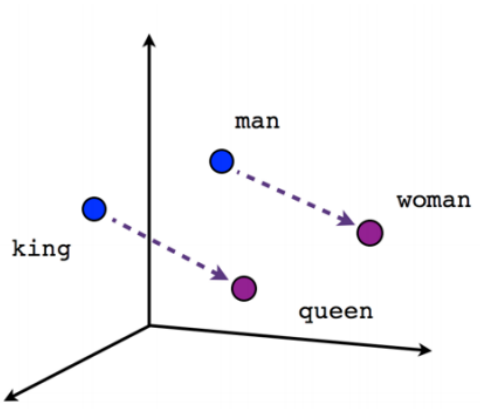


Gambar: Contoh regular expression. Sumber: tajawal

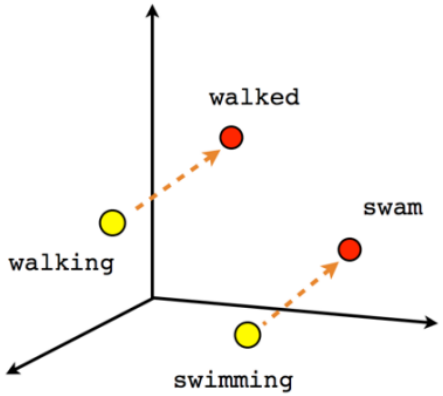

"[A] fascinating read from beginning to end." —TYLER COWEN,
professor of economics, George Mason University, author of Average Is Over
THE
LANGUAGE
OF
FOOD
A LINGUIST
READS THE MENU
DAN
JURAFSKY



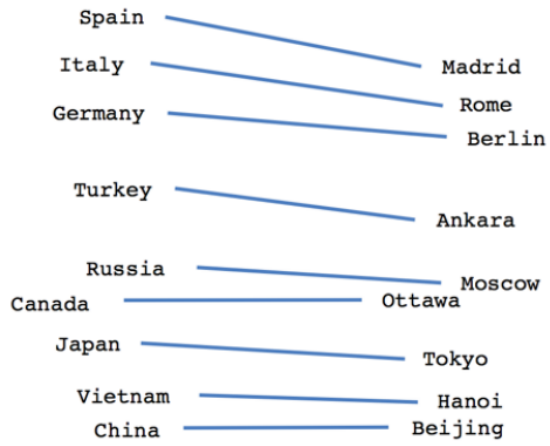
Gambar: "Apa yang menjadi sentimen dari ulasan ini?" Sumber: KDNuggets



Male-Female

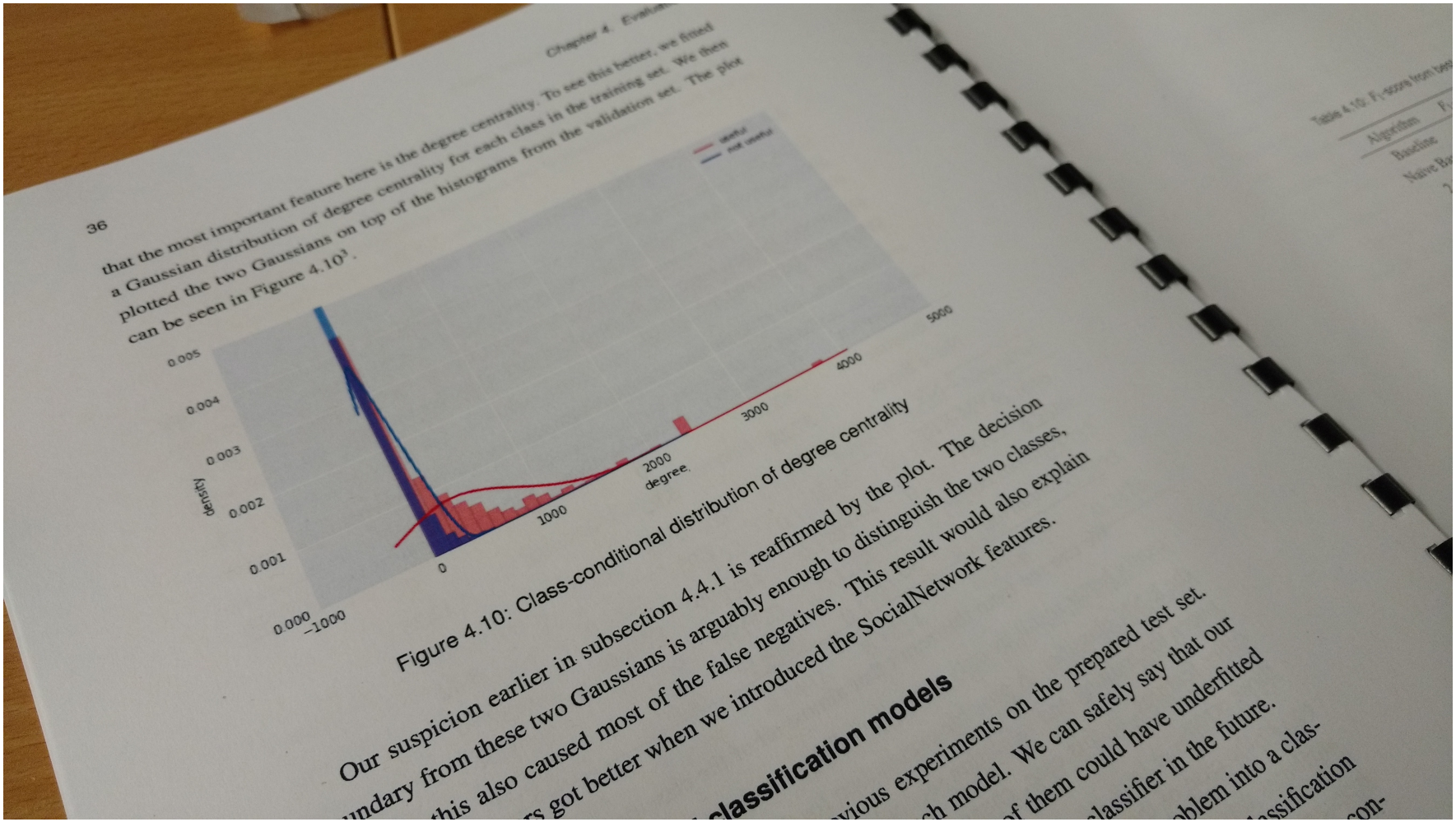


Verb tense

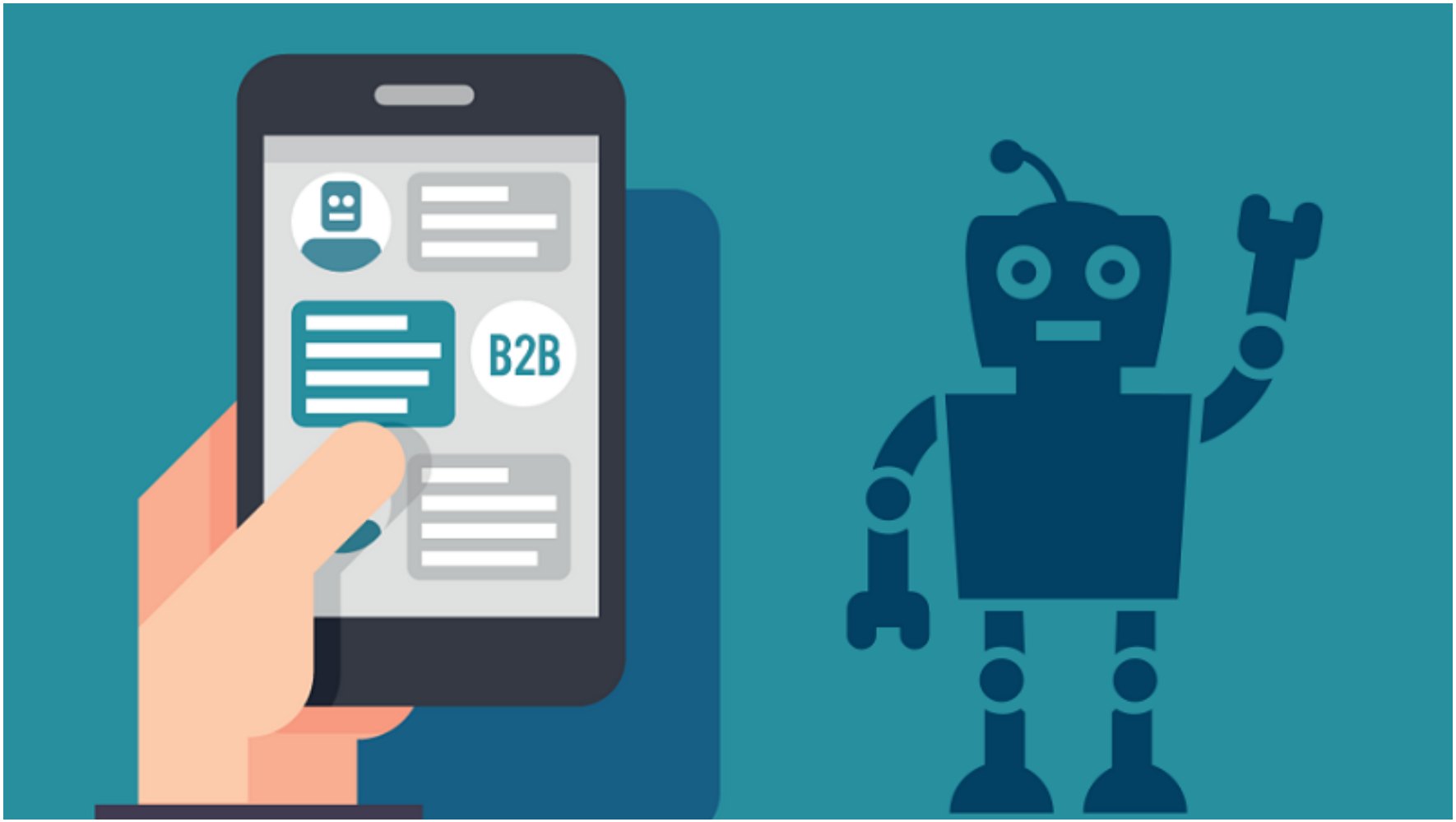


Country-Capital

Gambar: Representasi kata dalam vektor (Mikolov et al., 2013). Sumber: TensorFlow



Gambar: Deteksi plagiarisme dari makalah



Gambar: Penggunaan chatbot untuk bisnis. Sumber: Acquire

Topik dalam kuliah ini

Sebelum UTS

1. Intro
2. Regex, Text Normalization, Edit Distance
3. Language Modeling
4. Naïve Bayes & Sentiment Classification
5. Logistic Regression
6. Information Retrieval I
7. Information Retrieval II

Setelah UTS

8. Vector Semantics, Neural Embeddings, Word2Vec
9. Relation Extraction
10. Question Answering
11. Chatbots
12. Recommender Systems
13. Social Networks
14. Kuliah Tamu

Referensi

1. Stanford CS124: From Languages to Information
2. University of Edinburgh: Text Technologies for Data Science
3. Stanford CS276: Information Retrieval and Web Search (advanced)
4. Stanford CS224n: Natural Language Processing with Deep Learning (advanced)

Bahan Bacaan

1. Jurafsky & Martin. *Speech and Language Processing.*
2. Manning, Raghavan, and Schutze. 2008. *Introduction to Information Retrieval.*

Administrasi

Aturan perkuliahan

- Materi bisa dilihat di `http://uai.aliakbars.com/text-mining/`
- Kuliah setiap hari Senin, 07.00-09.30 (toleransi 15 menit)
- Teknologi: Python, Pylab, NLTK, SpaCy, gensim
- Terdapat 4 tugas
- Kuis
- Ujian Tengah Semester dan Ujian Akhir Semester (tidak ada perbaikan)
- Komponen nilai: 40% tugas, 30% UTS, 30% UAS

Aturan dalam tugas

- Secara *default*, setiap tugas bersifat individual
- Silakan berdiskusi, tapi **jangan menyalin kode atau tulisan teman**
- Keterlambatan pengumpulan akan berakibat pada pengurangan nilai
- Pengumpulan tugas dilakukan melalui situs e-learning

Aturan dalam tugas (lanjutan)

- Kode boleh diadaptasi dari internet, tapi selalu cantumkan sumbernya dengan benar
- Contoh:
  - Sumber: `google.com`, `stackoverflow.com` (salah)
  - Sumber: `https://github.com/keras-team/keras/blob/master/examples/mnist_mlp.py` (benar)
- Plagiarisme dapat berakibat pada **nilai E** untuk kuliah ini

Mulailah pengerjaan tugas segera setelah diberikan!

Jangan menghapalkan materinya!

Ujian bersifat buka buku

Terima kasih