**CSCE604135 Kuliah Temu-Balik Informasi (*Web Search & Information Retrieval*)**

**Semester Genap 2024/2025**

**Alfan F. Wicaksono**

## Deskripsi

Kuliah *Information Retrieval* (Temu-Balik Informasi) menyampaikan ilmu seputar sains dan teknologi pengembangan Web *search engine*. Pada bagian pertama, peserta akan mempelajari struktur data dan algoritme fundamental untuk sebuah *text retrieval system*. Pada bagian kedua, peserta akan mendalami topik-topik terkait *data-driven retrieval model* dan juga bagaimana melakukan evaluasi terhadap efektifitas dari sebuah *search engine*.

## Prasyarat

Secara administratif, prasyarat dari kuliah ini adalah kuliah Kecerdasan Artifisial dan Sains Data Dasar. Namun, kuliah ini sangat dianjurkan bagi peserta yang sudah pernah mengambil kuliah Desain dan Analisis Algoritma serta Analisis Numerik. Akan lebih baik lagi jika perserta sudah pernah (atau sedang) mengambil kuliah Machine Learning atau Deep Learning.

## Buku Teks (BACA! JANGAN MALAS!)

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, **Introduction to Information Retrieval**, Cambridge University Press. 2008. Buku online tersedia secara gratis: https://nlp.stanford.edu/IR-book/information-retrieval-book.html
- Bhaskar Mitra & Nick Craswell, **An Introduction to Neural Information Retrieval**, https://www.microsoft.com/en-us/research/uploads/prod/2017/06/fntir2018-neuralir-mitra.pdf
- **Alice's Adventures in a Differentiable Wonderland**: A primer on designing neural networks -- Volume I, A Tour of the Land. https://arxiv.org/abs/2404.17625
- Beberapa paper penelitian yang dipublikasikan di top venue seperti SIGIR, WSDM, ACL, dan sebagainya.

## Asisten Dosen

- Muhammad Ilham Ghozali (IK 2020)
- Jaycent Gunawan Ongris (IK 2021)
- Lyzander Marciano Andrylie (IK 2021)

- Inaya Rahmanisa (IK 2021)
- Muhammad Falensi Azmi (IK 2021)

**Waktu Kuliah:**

- **Senin: 16:00 – 17:40 (waktu ketika kita semua mengantuk)**
- **Jumat: 08:00 – 09:40**

**Evaluasi**

1. UTS – 30%
2. UAS – 30%
3. Tugas Pemrograman – 4 kali dengan total bobot 25%
4. Tugas Kelompok – 1 kali dengan bobot 15%
5. Partisipasi Tutorial – 5% (tidak tentu)

**UTS dan UAS berupa essay**: analisis algoritme, pemrograman, dan kajian matematis

**Jadwal (Subject To Change)**

| | | |
|---|---|---|
| 1 | 3 Februari<br>(100 menit) | Introduction to IR (History & Big Picture; 70 minutes)<br>Inverted Index: Dictionary & Postings Lists (30 minutes) |
| | 7 Februari<br>(100 menit) | Inverted Index: Phrase Queries & Proximity Search<br><br>Index Construction: Tokenization, Normalization, Stemming, Stop Words; Byte-pair encoding tokenization, word-piece tokenization, unigram tokenization. |
| 2 | 10 Februari<br>(100 menit) | Index Construction: External Memory Indexing, Distributed Indexing, Dynamic Indexing, Logarithmic Indexing |
| | 14 Februari<br>(100 menit) | Index Compression<br>• Dictionary Compression; Optimal Binary Search Trees<br>• Compression Principles<br>• Postings Compression: Variable Byte Encoding, Elias-Gamma Encoding, OptPForDelta |
| 3 | 17 Februari<br>(100 menit) | **Tugas Pemrograman 1**<br>Ranked Retrieval<br>Term Frequency & Weighting: TF & IDF, TaaT scheme |
| | 21 Februari<br>(100 menit) | Vector Space Model for Scoring<br>Tolerant Retrieval: Spelling Correction (Isu Fondasi) |
| 4 | 24 Februari<br>(100 menit) | Tolerant Retrieval: Spelling Correction (Isu Fondasi) |

| | | |
|---|---|---|
| | 28 Februari<br>(100 menit) | Probabilistic IR: "Best Match 25" (BM25)<br>Efficient Query Processing: Inefficient Evaluation of BM25, Top-K Retrieval & WAND Algorithm |
| 5 | 3 Maret<br>(100 menit) | **Tugas Pemrograman 2**<br>Efficient Query Processing: Inefficient Evaluation of BM25, Top-K Retrieval & WAND Algorithm<br><br>Query Auto-Completion: Trie & RMQ-based Index |
| | 7 Maret<br>(100 menit) | IR Evaluation<br>• Online Evaluation: A/B Testing & Interleaving<br>• Offline Evaluation: Cranfield-Style Evaluation & Model-based metrics; Statistical Significance Test<br><br>Web Crawling |
| 6 | 10 Maret<br>(100 menit) | Relevance Feedback & Query Expansion |
| | 14 Maret<br>(100 menit) | Foundations of Machine Learning:<br>Gradients, Jacobians, Optimizations & Gradient Descent, Loss functions, Bayesian learning, Linear models, Fully-connected models, Automatic Differentation |
| 7 | 17 Maret<br>(100 menit) | **Tugas Pemrograman 3**<br>Foundations of Machine Learning:<br>Gradients, Jacobians, Optimizations & Gradient Descent, Loss functions, Bayesian learning, Linear models, Fully-connected models, Automatic Differentation |
| | 21 Maret<br>(100 menit) | Foundations of Machine Learning:<br>Gradients, Jacobians, Optimizations & Gradient Descent, Loss functions, Bayesian learning, Linear models, Fully-connected models, Automatic Differentation |
| 8 | 24 Maret<br>(100 menit) | Text Classification for IR |
| | 28 Maret<br>Cuti Bersama/ Libur | - |
| 9 | 31 Maret<br>Libur Idul Fitri | - |
| | 4 April<br>Cuti Bersama/ Libur | - |
| 10 | 7 April<br>Cuti Bersama/ Libur | - |
| | 11 April<br>(100 menit) | Distributed Word Representations for IR<br>• Konsep Distributional Semantics<br>• Latent Semantic Analysis & Singular Value Decomposition |

| | | |
|---|---|---|
| | | • Neural Embeddings |
| | 14 – 23 April | Ujian Tengah Semester |
| 11 | 25 April<br>(100 menit) | **Rilis Deskripsi Tugas KELOMPOK**<br>Learning to Rank<br>• Gradient & Newton Boosting Framework<br>• MART<br>• RankNet<br>• LambdaRank<br>• LambdaMART |
| 12 | 28 April<br>(100 menit) | Learning to Rank<br>• Gradient & Newton Boosting Framework<br>• MART<br>• RankNet<br>• LambdaRank<br>• LambdaMART<br><br>Link Analysis: PageRank |
| | 2 Mei<br>(100 menit) | Neural Language Models & Transformers (Encoders & Decoders) |
| 13 | 5 Mei<br>(100 menit) | **Tugas Pemrograman 4**<br>Neural Language Models & Transformers (Encoders & Decoders) |
| | 9 Mei<br>(100 menit) | Nearest Neighbor: KD-Trees<br>Approximate Nearest Neighbor:<br>Locality Sensitive Hashing (LSH)<br>Hirarchical Navigable Small World (HNSW) |
| 14 | 12 Mei<br>Libur | - |
| | 16 Mei<br>(100 Menit) | Nearest Neighbor: KD-Trees<br>Approximate Nearest Neighbor:<br>Locality Sensitive Hashing (LSH)<br>Hirarchical Navigable Small World (HNSW) |
| 15 | 19 Mei<br>(100 menit) | Recommender Systems |
| | 23 Mei<br>(100 menit) | Tutorial - Multimodal IR (Voice / Image Retrieval) |
| | 26 Mei – 6 Juni | UAS |