Machine Learning Research Exploration

TOPIC MODEL

Model Selection:

Latent Dirichlet Model's Perplexity

■ Abstract ■ FullText



PROBLEM & APPROACH

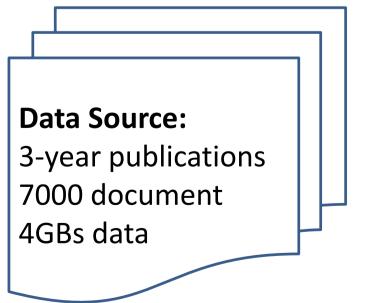
Research Questions:

- What are recent topics in machine learning research?
- Is there any interesting relationship among these topics?
- Is it possible to find research document based on relevancy instead of simple keywords?

Approaches:

- Analyze machine learning publications from major research journals
- Identify underlying topics of these document using topic modelling
- Identify interesting relationship among topics using network analysis
- Compute relevance score of document based on topic distribution similarity

DATA COLLECTION:





Web Scraping

PDF Processing

Abstract Content URL

MongoDB:

Authors

Springer Link



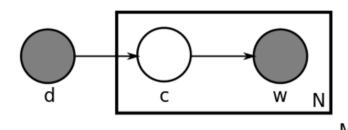
DATA ANALYSIS PIPELINE

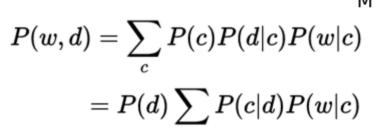
Data Pre-Processing: Word Lemmatization Stop Word Removal

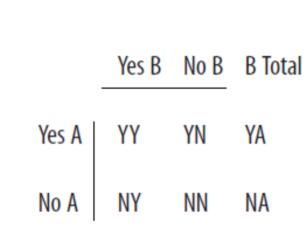


Exploratory Analysis: Word Cloud & Frequency

Feature Selection: Abstract vs Full Text TF vs TF-IDF







B Total YB NB



- Perplexity - Word Intrusion

Topic Model:

Latent Dirichlet Allocation



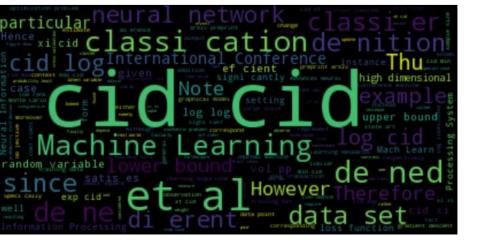
Network Analysis: Co-occurrence

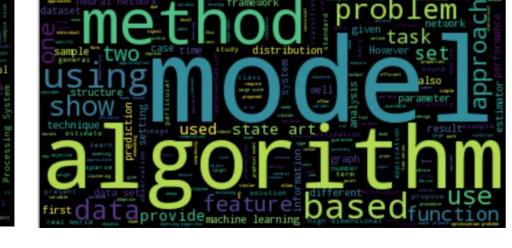
Independence

Data Product:

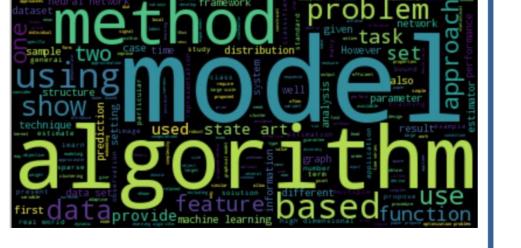
Document Search based on topic distribution similarity

DATA EXPLORATION:





"Full Text" Word Cloud "Abstract" Word Cloud



Number of latent topics (k) **Topics:** effect clinical Overall Topic Contribution Abnomaly Detection patient Biomedical Causual Inference Classification Computer Vision Topic: Monte Carlo Method Data Labelling Decomposition Deep Learning Dialog Dimensionality Reduction Distance Eigen Method Game chain Graph carlo Imputation Lasso Markov Decision Monte Carlo Method Multimodal Neurology Optimization Pairwise Comparison Probablistics Mode Recommender System Reinforecement Learning collaborative matrix Riemannian Method recommendation rating preference Shortest Path Signal Processing Social Network Text Mining Unknonw (finance) Unknown (Acoustic) Unknown (Energy) Jnknown (Localization) Jnknown (community) information retrieval inherit

Choose k = 45

Number of latent topics (k)

Topic: Causual Inference

Topic: Probablistics Model

Topic: Deep Learning

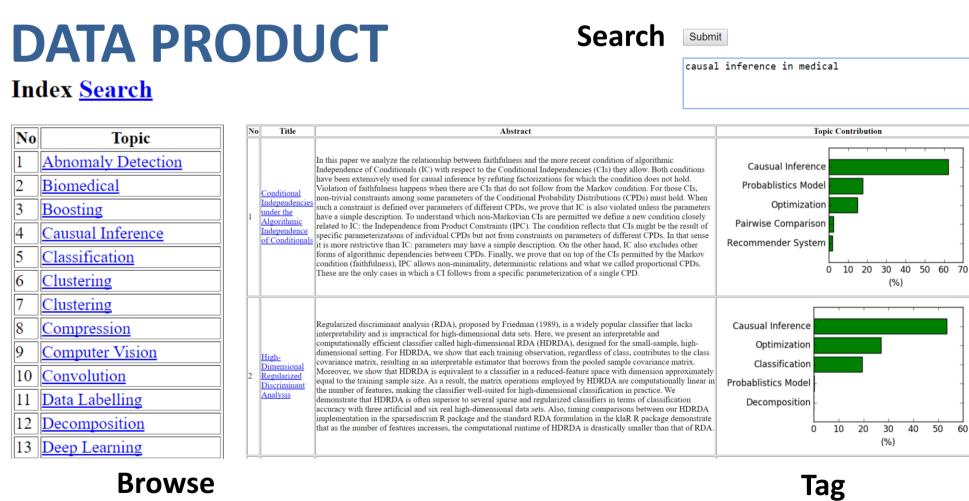
Topic: Unknown (Energy)

mind mind

energy gy

Unknown (community) Latent Dirichlet Model's Perplexity interpretable feasibilityside relation Node Size: Eigen Centrality Cadalas integral independency Index <u>Search</u> Causual Inference Classification Clustering Clustering variable Computer Vision 11 Data Labelling 2 Decomposition 13 Deep Learning ob Tie Crin learn network

NETWORK ANALYSIS Unknown coustic) Inknon (() finance) Monte Oarlo Metho Reinforecement Learning MarkovDecision



Edge: X² significant association

SUMMARY

- Major recent topics and relationships might be different from our impression
- The proposed topic distribution similarity-based approach can provide relevant search results and help auto-tagging.

Learning:

Methodology: Data Visualization – Model Interpretation – Project Planning

Technology: Scrapy, wordcloud, Gephi, LDA, GraphX, MongoDB, Django