Machine Learning Research Exploration

Simon Fraser University

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:

Data Source:
3-year publications
7000 document
4GBs data

Web Scraping
PDF Processing

MongoDB:
- Title
- Authors
- Abstract
- Content
- URL

Springer Link

JMLR NIPS arXiv IEEE

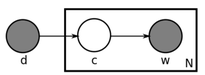
DATA ANALYSIS PIPELINE

Data Pre-Processing:
Word Lemmatization
Stop Word Removal

Exploratory Analysis:
Word Cloud & Frequency

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

Evaluation:
- Perplexity
- Word Intrusion



$$P(w,d) = \sum_c P(c)P(d|c)P(w|c)$$
$$= P(d)\sum_c P(c|d)P(w|c)$$

	Yes B	No B	B Total
Yes A	YY	YN	YA
No A	NY	NN	NA
B Total	YB	NB	T

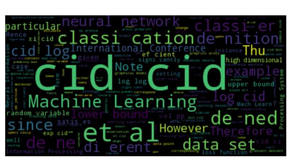
$$\chi^2 = T\frac{(YY*NN - YN*NY)^2}{YA*NA*YB*NB}$$

Topic Model:
Latent Dirichlet Allocation

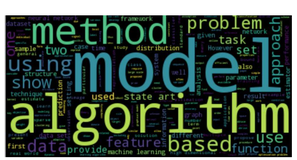
Network Analysis:
Co-occurrence
Independence

Data Product:
Document Search based on topic distribution similarity

DATA EXPLORATION:



"Full Text" Word Cloud

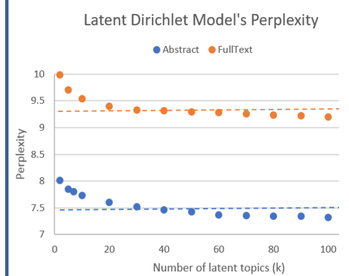


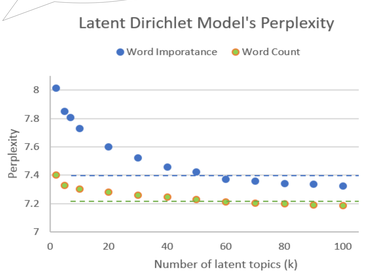
"Abstract" Word Cloud

TOPIC MODEL

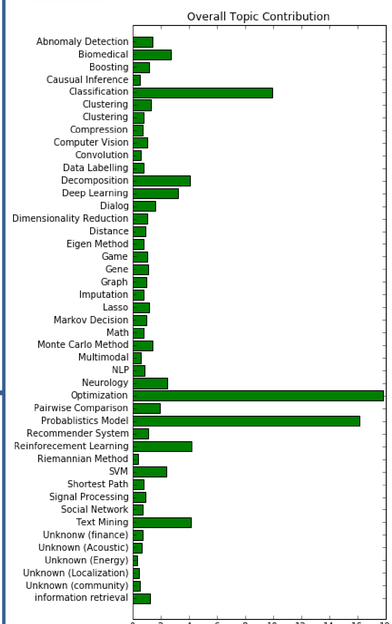
Model Selection:

Choose k = 45

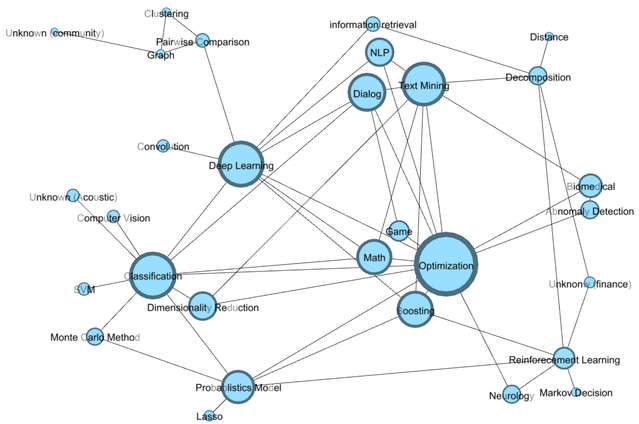




Topics:



NETWORK ANALYSIS



Node Size: Eigen Centrality

Edge: X^2 significant association

DATA PRODUCT

Index Search

Search

Browse

Tag

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