Clustering in Natural Language Processing & Text Mining

CpE 6330 Project Presentation

Stuart Miller

Master's Student in Computer Engineering

Introduction

- IEEE Xplore provides a developer API
  - 176 Journals, Magazines, and Transactional Collections
  - Countless more conference and workshop proceedings

Questions/Goals

> How similar are papers published in journals?

> Can papers be neatly clustered by subject matter?

> Do Journals and Transactions vary more or less than Conference and Workshop Proceedings?

> Is this kind of clustering a useful tool for researchers?

Introduction

- Use Abstracts
  - Assume representative of paper in general
  - More manageable
  - About 200 words each

The Internet of Things (IoT) shall be able to incorporate transparently and seamlessly a large number of different and heterogeneous end systems, while providing open access to selected subsets of data for the development of a plethora of digital services. Building a general architecture for the IoT is hence a very complex task, mainly because of the extremely large variety of devices, link layer technologies, and services that may be involved in such a system. In this paper, we focus specifically to an urban IoT system that, while still being quite a broad category, are characterized by their specific application domain. Urban IoTs, in fact, are designed to support the Smart City vision, which aims at exploiting the most advanced communication technologies to support added-value services for the administration of the city and for the citizens. This paper hence provides a comprehensive survey of the enabling technologies, protocols, and architecture for an urban IoT. Furthermore, the paper will present and discuss the technical solutions and best-practice guidelines adopted in the Padova Smart City project, a proof-of-concept deployment of an IoT island in the city of Padova, Italy, performed in collaboration with the city municipality.

A. Zanella, N. Bui, A. Castellani, L. Vangelista and M. Zorzi, "Internet of Things for Smart Cities,"
in *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22-32, Feb. 2014.
doi: 10.1109/JIOT.2014.2306328

Outline

- > Introduction
- > **2018 Papers Analysis** **by LDA and LSA**
- > Model Applicability
- > TF-IDF Analysis
- > Perplexity & Insights by Publication
- > Topic Accuracy
- > Conclusions and Ending Points

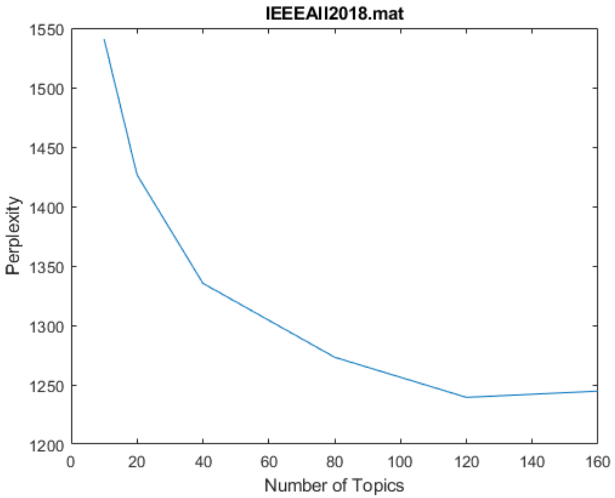
MISSOURI S&T

First Data Set

- Consider all available papers with 2018 publishing dates
  - 21,985 papers
  - 30 MB of abstract and associated bookkeeping information

Latent Dirichlet Allocation Approach

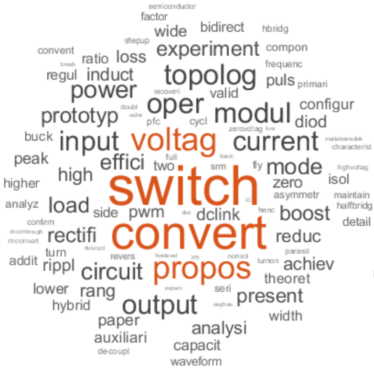
- Perplexity
  - $perplexity(t) = e^{-\sum_d p(t,d)\cdot\log(p(t,d))}$
  - Where $p(t,d)$ is the probability of topic t in document d, summed over all documents
  - Higher number mean even distribution (lower number means better fit)



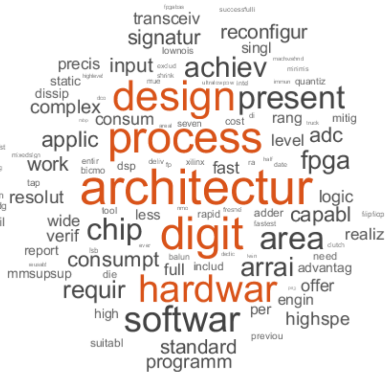
- Total summed perplexity for **IEEAll2018** data set
  - Choose 120 topics

IEEEAll2018 First 8 Topics

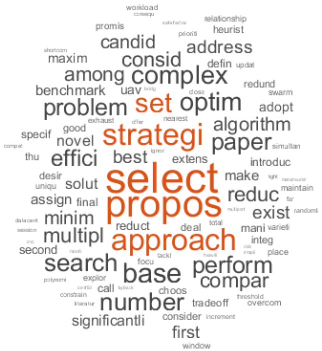
Topic: 1



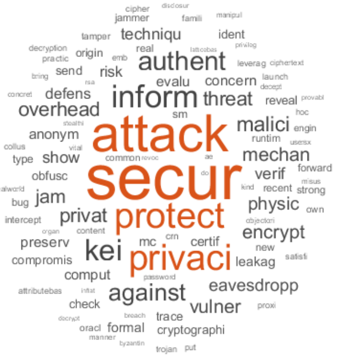
Topic: 2



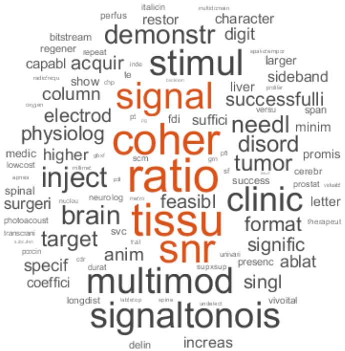
Topic: 3



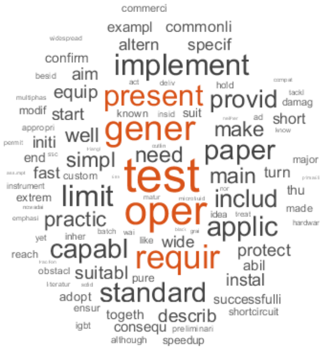
Topic: 4



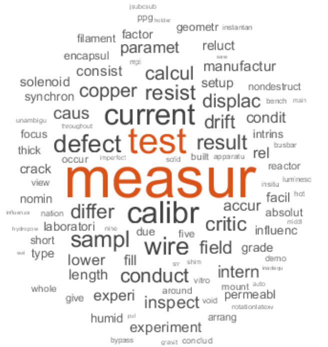
Topic: 5



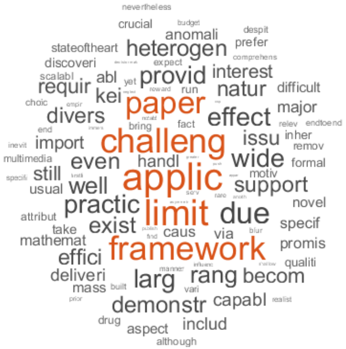
Topic: 6



Topic: 7

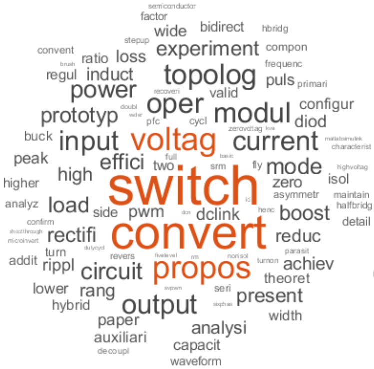


Topic: 8

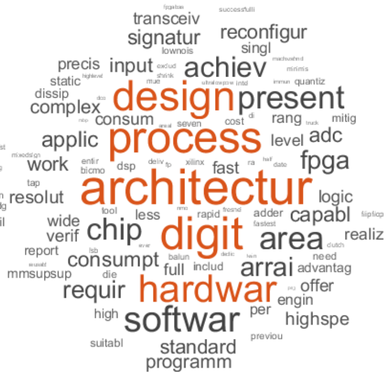


IEEEAll2018 First 8 Topics (of 120)

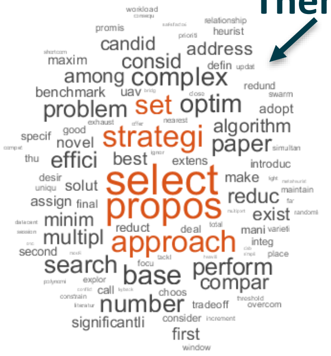
Topic: 1



Topic: 2

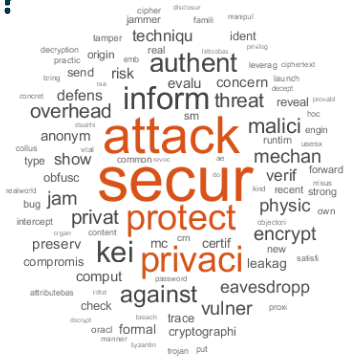


Topic: 3

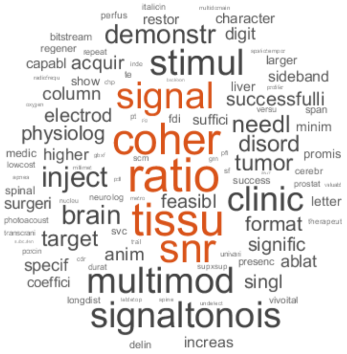


Theme?

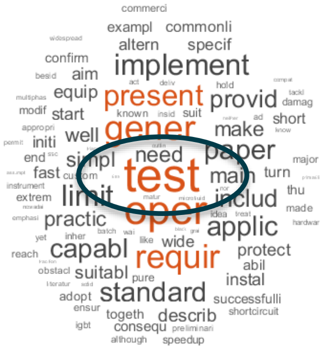
Topic: 4



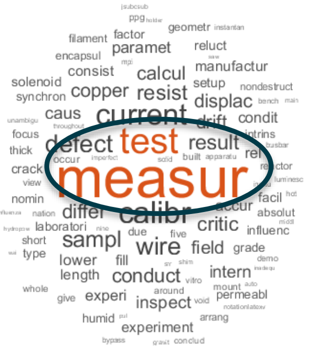
Topic: 5



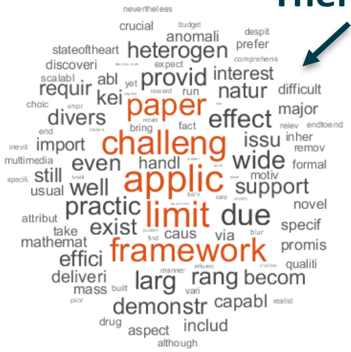
Topic: 6



Topic: 7

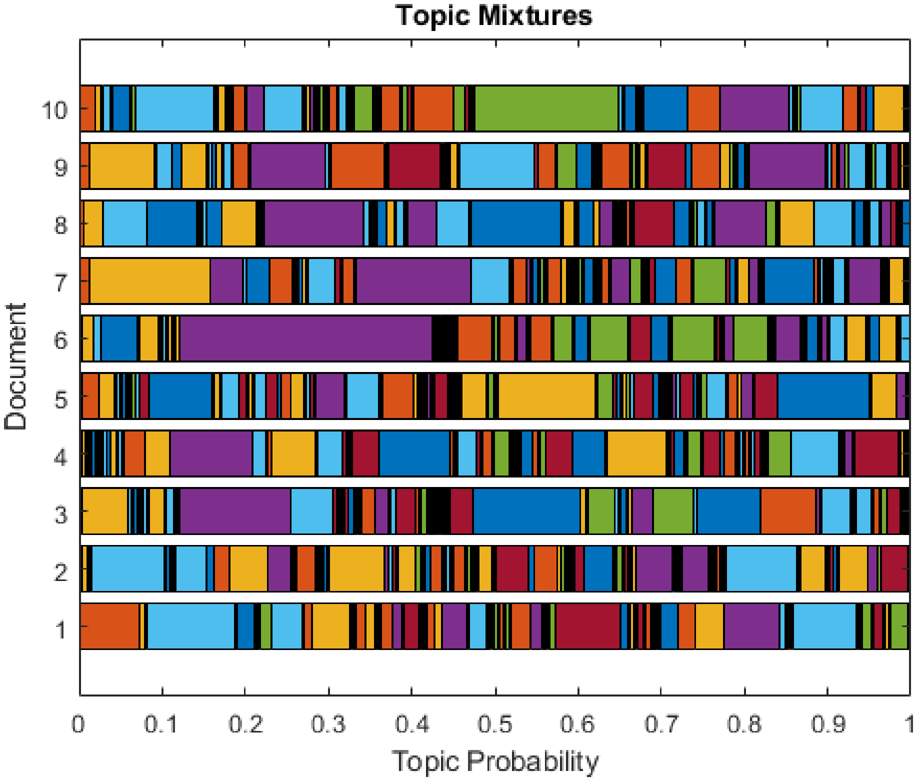


Topic: 8



Theme?

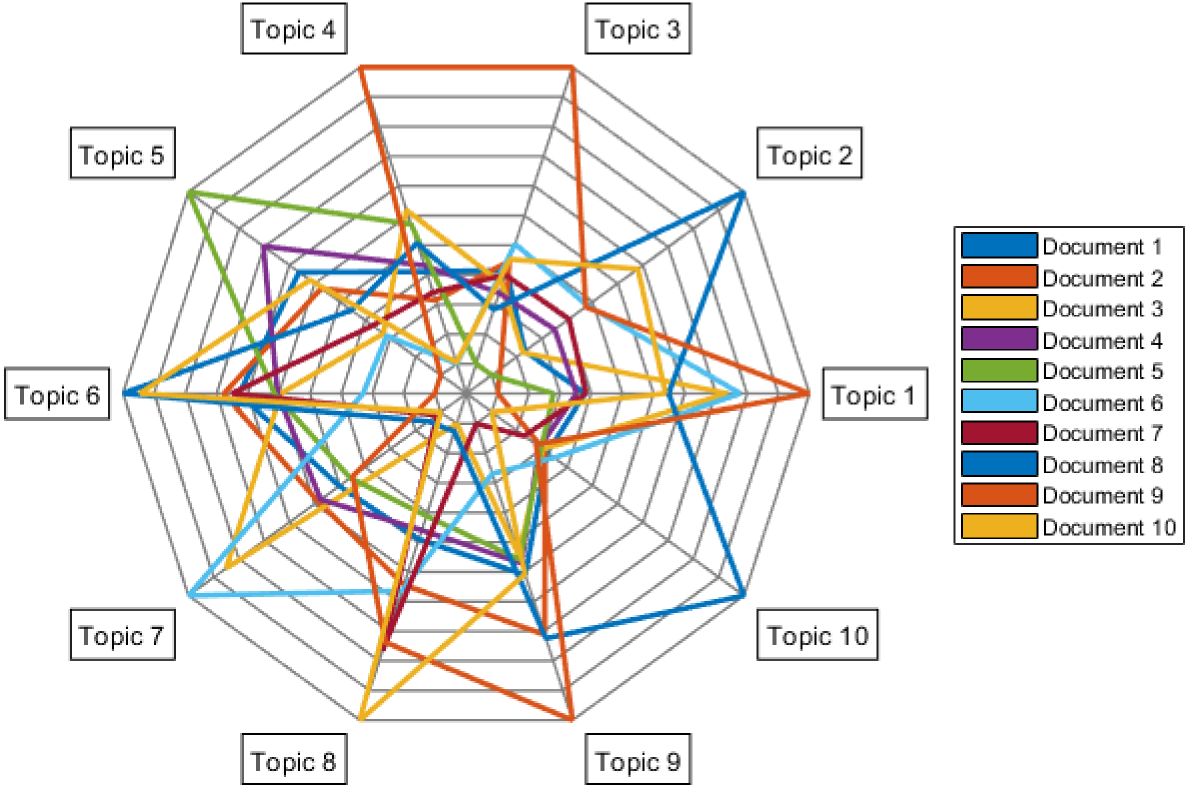
IEEEAll2018 First 10 Documents



Latent Semantic Analysis Approach

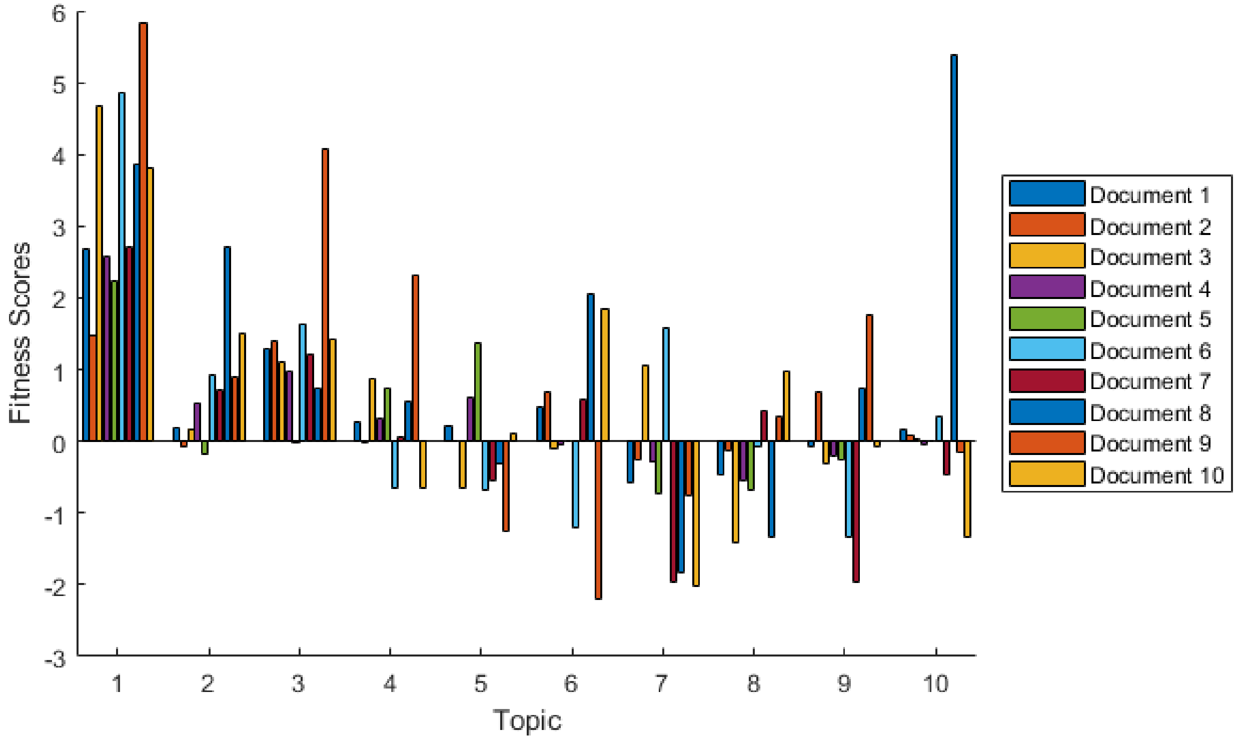
- Choose 120 topics once again
- Remember, fitness scores, not probability distributions

IEEEAll2018 First 10 Documents, First 10 Topics



MISSOURI S&T

IEEEAll2018 First 10 Documents, First 10 Topics

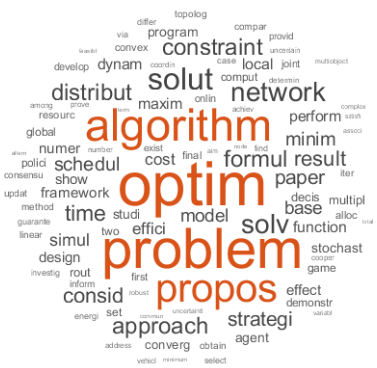


Overview

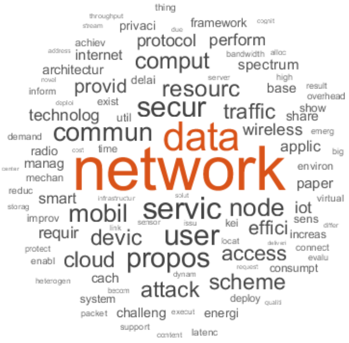
- Pretty bad classification
  - Unusable number of topics
  - Topic assignments are messy
- Re-run with fewer topics
  - 20 topics
  - Sticking with this assignment for the remainder of the presentation

IEEEAll2018 First 8 Topics (of 20)

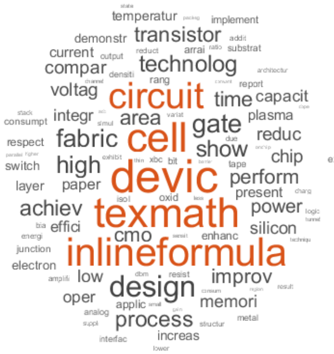
Topic: 1



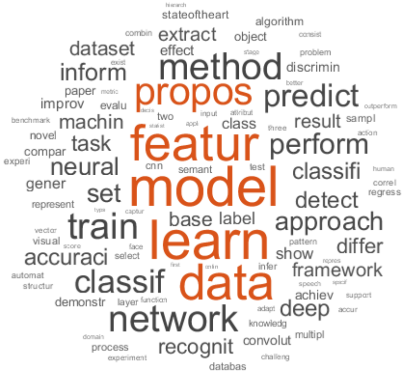
Topic: 2



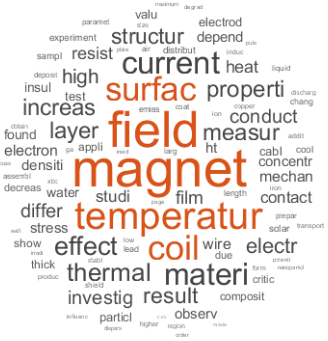
Topic: 3



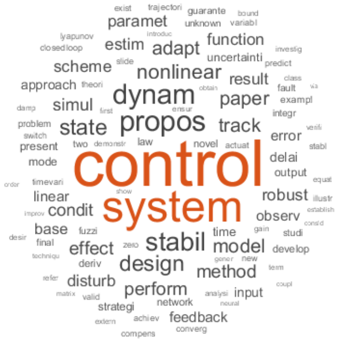
Topic: 4



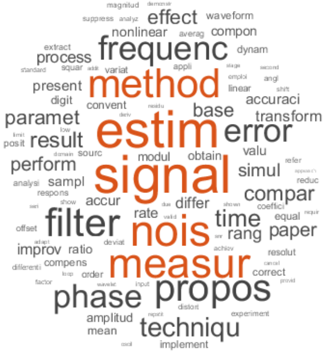
Topic: 5



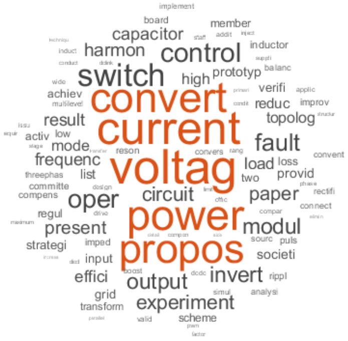
Topic: 6



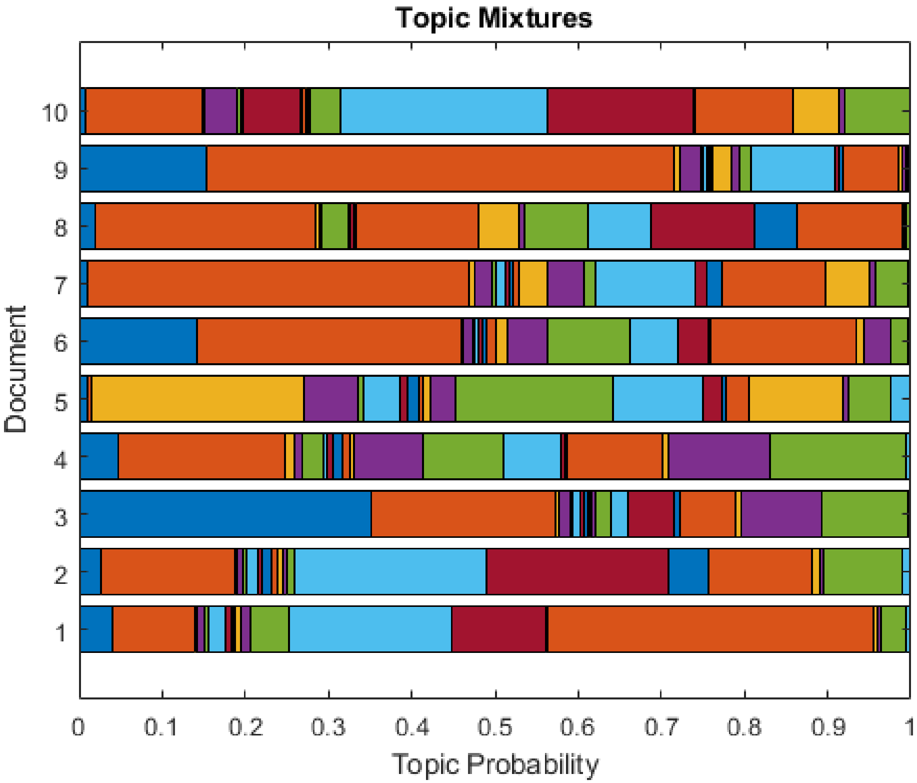
Topic: 7



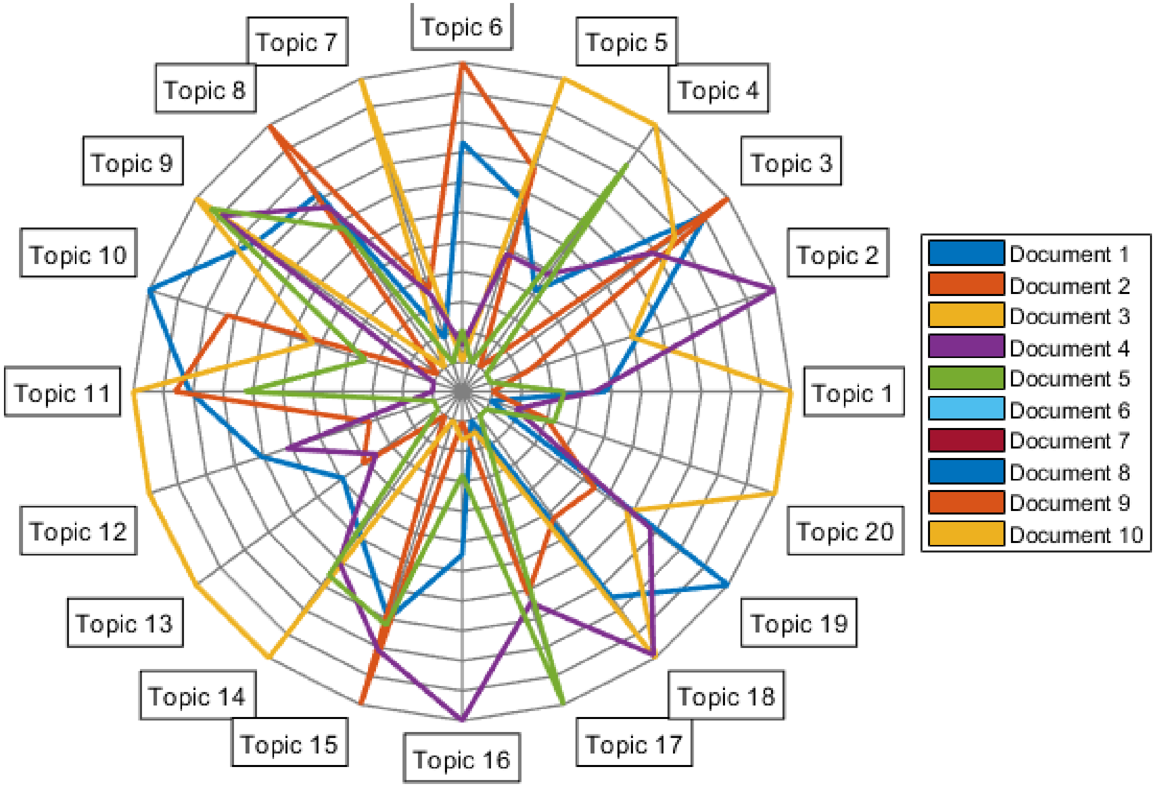
Topic: 8



IEEEAll2018 First 10 Documents (20 topics)



IEEEAll2018 First 10 Documents (20 of 20 topics)



Outline

- > Introduction
- > 2018 Papers Analysis by LDA and LSA
- > **Model Applicability**
- > TF-IDF Analysis
- > Perplexity & Insights by Publication
- > Topic Accuracy
- > Conclusions and Ending Points

MISSOURI S&T

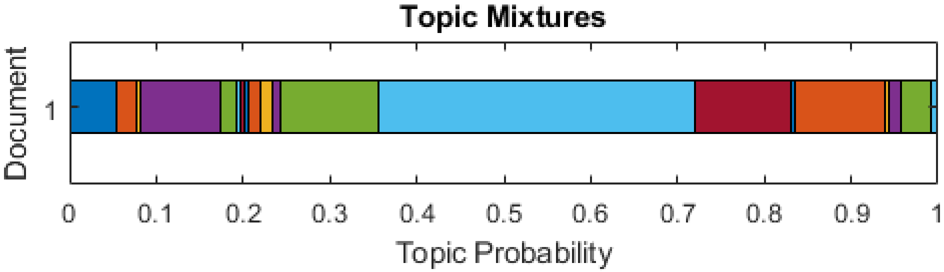
Model Applicability

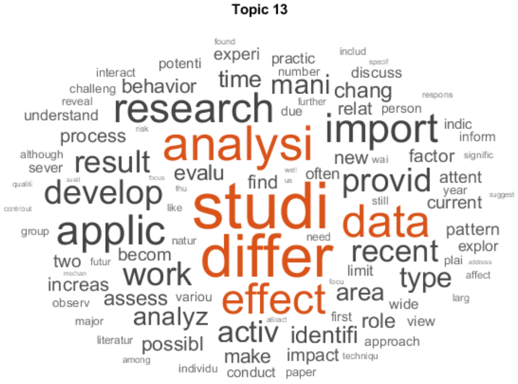
> Can we apply this model to **any** paper?

> Dr. Wunsch's **Survey of clustering algorithms**

Abstract
Data analysis plays an indispensable role for understanding various phenomena. Cluster analysis, primitive exploration with little or no prior knowledge, consists of research developed across a wide variety of communities. The diversity, on one hand, equips us with many tools. On the other hand, the profusion of options causes confusion. We survey clustering algorithms for data sets appearing in statistics, computer science, and machine learning, and illustrate their applications in some benchmark data sets, the traveling salesman problem, and bioinformatics, a new field attracting intensive efforts. Several tightly related topics, proximity measure, and cluster validation, are also discussed.

Rui Xu and D. Wunsch, "Survey of clustering algorithms," in IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, May 2005.



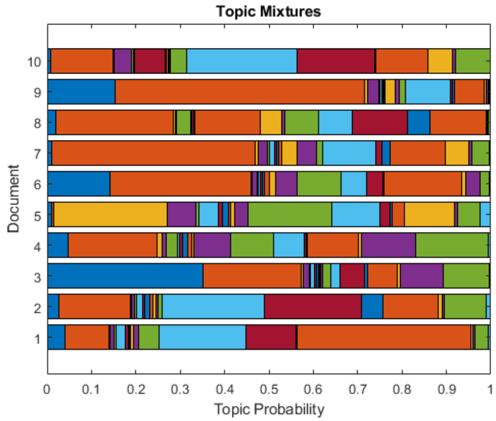


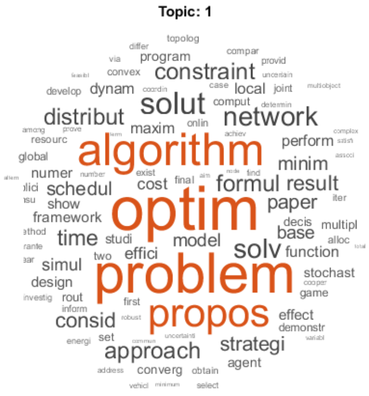
Question #1: How similar are journal papers?

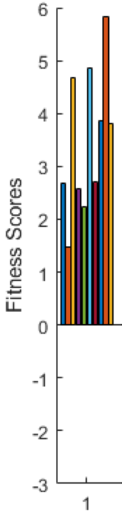
> Topic #1 - generic

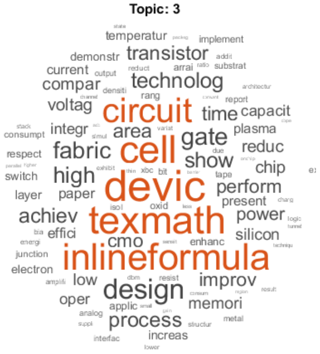
> Overall Topic Mixtures

> Topic #3 - TeX syntax









Outline

> Introduction

> 2018 Papers Analysis by LDA and LSA

> Model Applicability

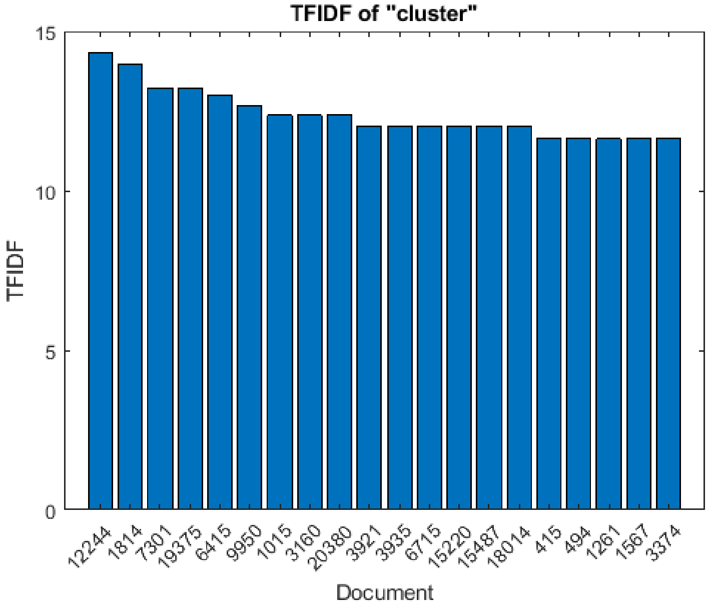
> **TF-IDF Analysis**

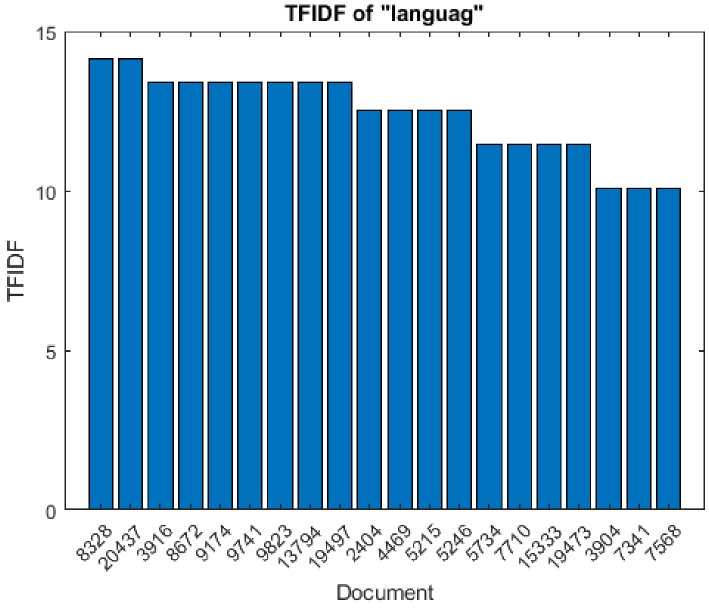
> Perplexity & Insights by Publication

> Topic Accuracy

> Conclusions and Ending Points

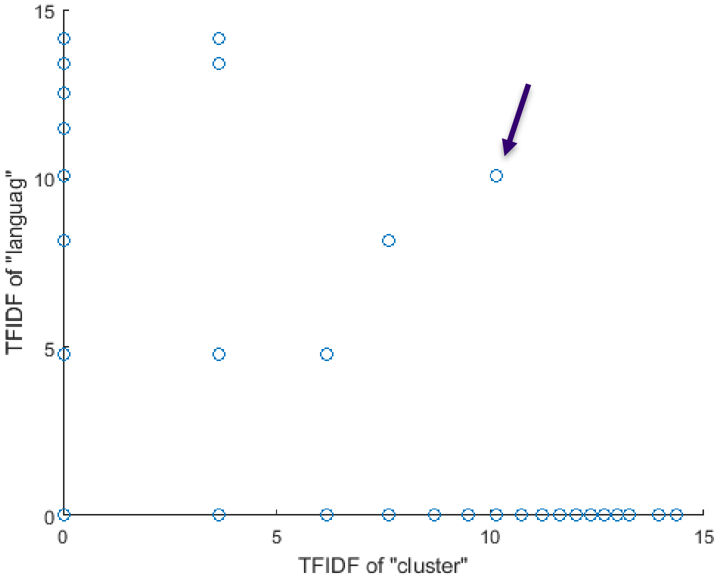
Term Frequency-Inverse Document Frequency Approach





TF-IDF Results

- Document 7568
  - On Pareto front
  - IEEE web search for “cluster language” returns this as well

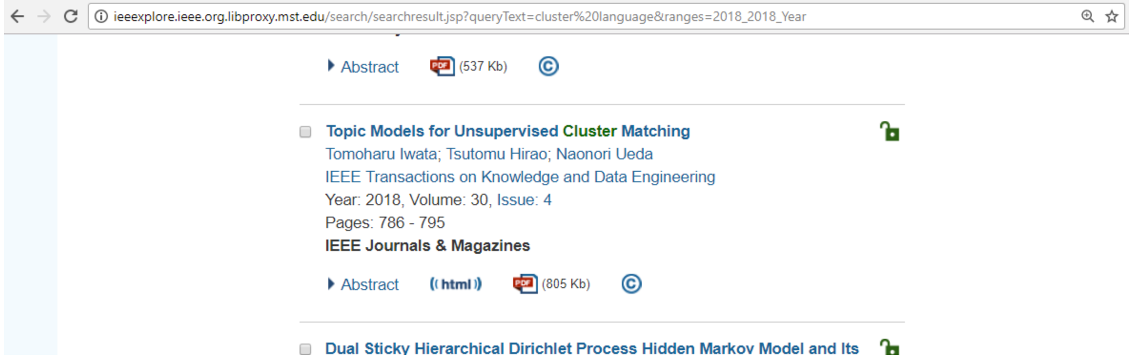


We propose topic models for unsupervised cluster matching, which is the task of finding matching between clusters in different domains without correspondence information. For example, the proposed model finds correspondence between document clusters in English and German without alignment information, such as dictionaries and parallel sentences/documents. The proposed model assumes that documents in all languages have a common latent topic structure, and there are potentially infinite number of topic proportion vectors in a latent topic space that is shared by all languages. Each document is generated using one of the topic proportion vectors and language-specific word distributions. By inferring a topic proportion vector used for each document, we can allocate documents in different languages into common clusters, where each cluster is associated with a topic proportion vector. Documents assigned into the same cluster are considered to be matched. We develop an efficient inference procedure for the proposed model based on collapsed Gibbs sampling. The effectiveness of the proposed model is demonstrated with real data sets including multilingual corpora of Wikipedia and product reviews.

T. Iwata, T. Hirao and N. Ueda, "Topic Models for Unsupervised Cluster Matching," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 4, pp. 786-795, April 1 2018.

TFIDF Conclusions

- Not useful
- Nothing more than a basic internet search query
- As per the algorithm’s design – meant to be fast and simple
- Will omit for rest of this project



ieeexplore.ieee.org.libproxy.mst.edu/search/searchresult.jsp?queryText=cluster%20language&ranges=2018_2018_Year

□ **Topic Models for Unsupervised Cluster Matching**
Tomoharu Iwata; Tsutomu Hirao; Naonori Ueda
IEEE Transactions on Knowledge and Data Engineering
Year: 2018, Volume: 30, Issue: 4
Pages: 786 - 795
IEEE Journals & Magazines

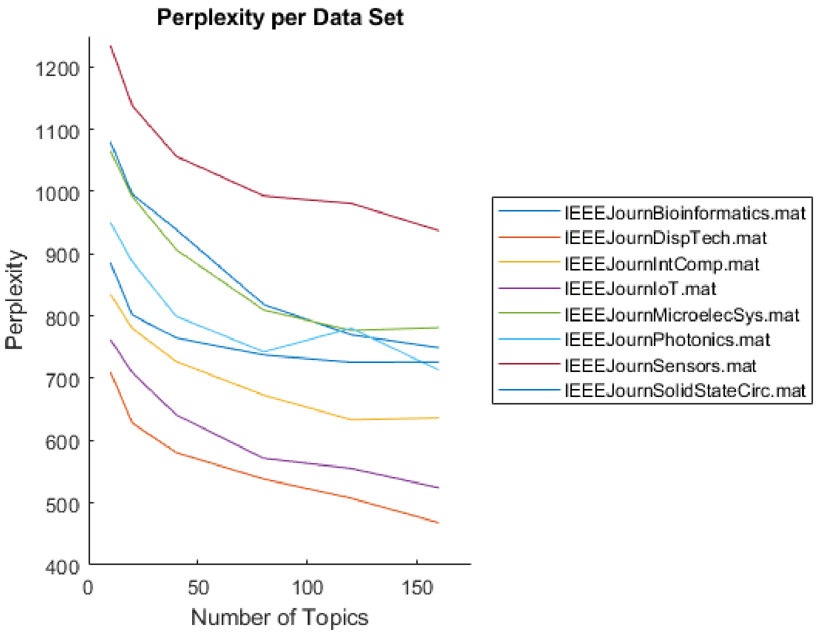
□ **Dual Sticky Hierarchical Dirichlet Process Hidden Markov Model and Its**

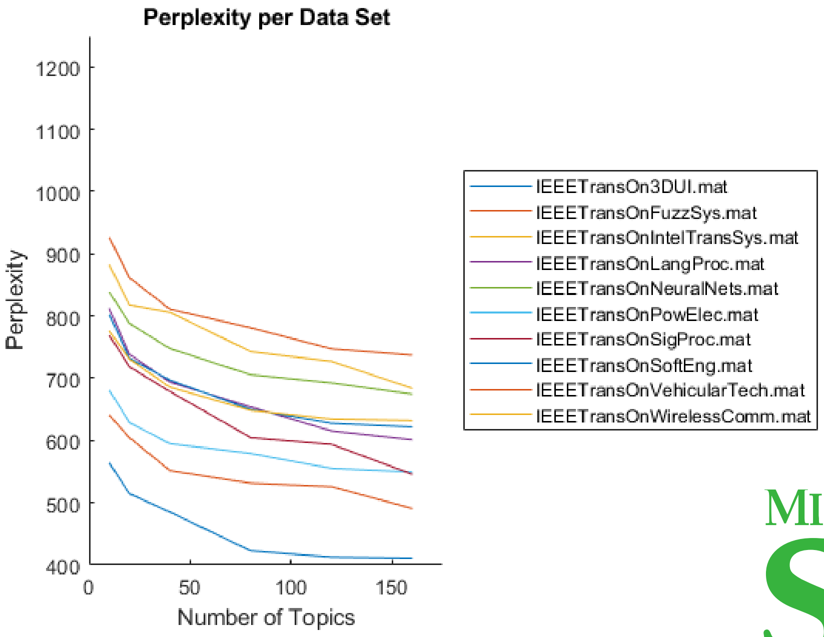
MISSOURI S&T

Outline

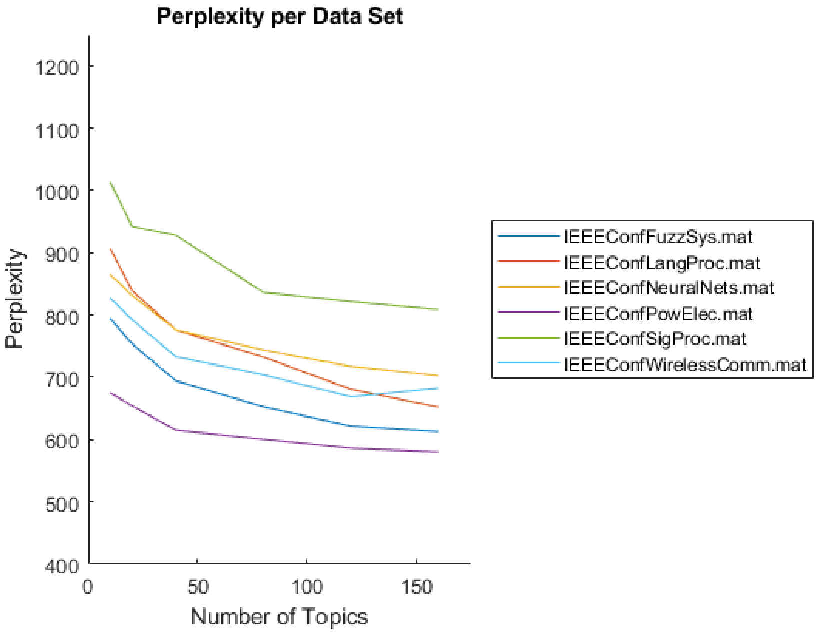
- Introduction
- 2018 Papers Analysis by LDA and LSA
- Model Applicability
- TF-IDF Analysis
- **Perplexity & Insights by Publication**
- Topic Accuracy
- Conclusions and Ending Points

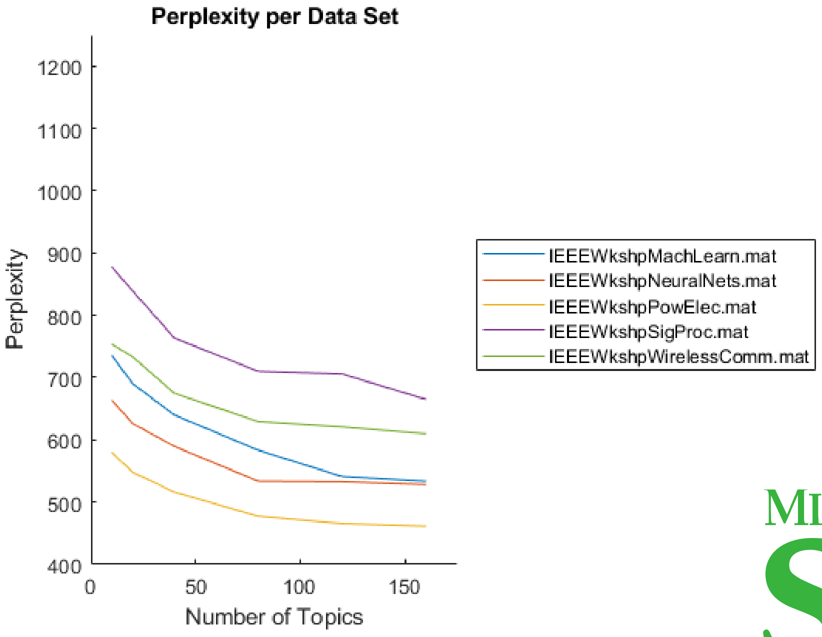
Perplexity for Journals and Transactions





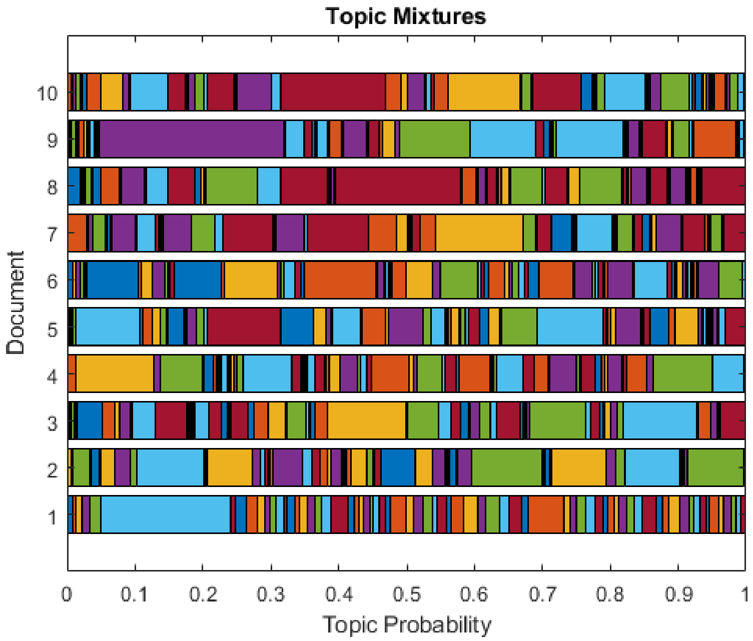
Perplexity for Journals and Transactions



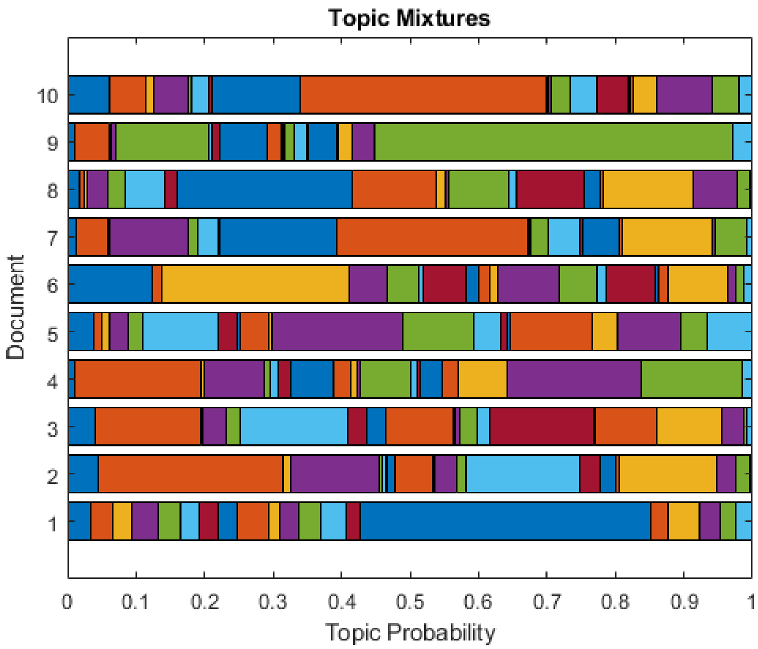


IEEE Sensors Journal Insights

Recommended 70 Topic LDA Model

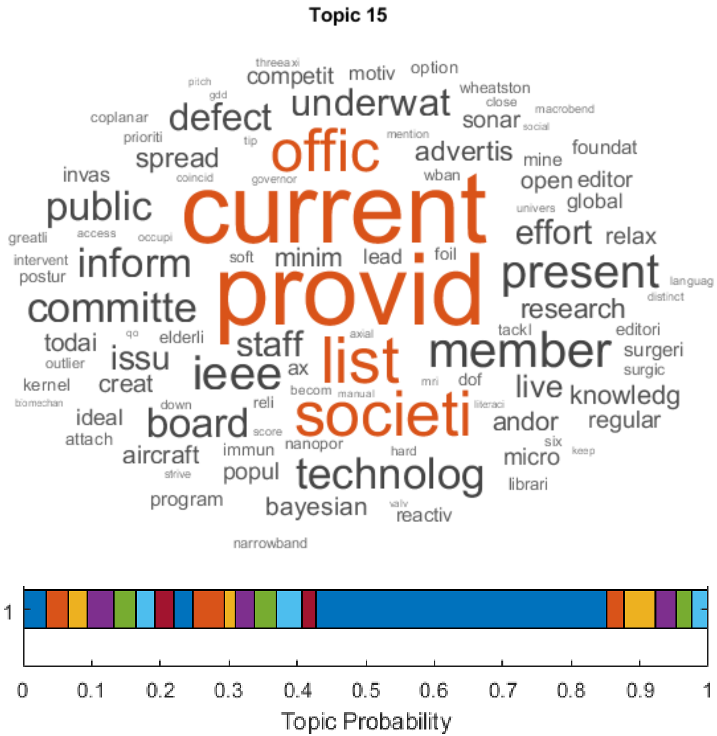


Simplified 20 Topic LDA Model



SOURI S&T

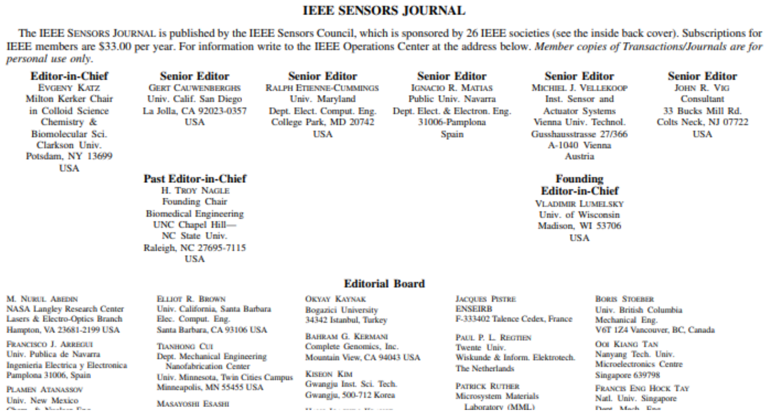
Topic 15 (Primary Topic of Docs 1, 14, 58, 90…)



Abstract

Provides a listing of current society officers.

"IEEE Sensors Journal publication information," in *IEEE Sensors Journal*, vol. 9, no. 11, pp. C2-C2, Nov. 2009.



IEEE SENSORS JOURNAL

The IEEE SENSORS JOURNAL is published by the IEEE Sensors Council, which is sponsored by 26 IEEE societies (see the inside back cover). Subscriptions for IEEE members are $33.00 per year. For information write to the IEEE Operations Center at the address below. *Member copies of Transactions/Journals are for personal use only.*

Editor-in-Chief
EVGENY KATZ
Milton Kerker Chair
in Colloid Science
Chemistry &
Biomolecular Sci.
Clarkson Univ.
Potsdam, NY 13699
USA

Senior Editor
GERT CAUWENBERGHS
Univ. Calif. San Diego
La Jolla, CA 92023-0357
USA

Senior Editor
RALPH ETIENNE-CUMMINGS
Univ. Maryland
Dept. Elect. Comput. Eng.
College Park, MD 20742
USA

Senior Editor
IGNACIO R. MATIAS
Public Univ. Navarra
Dept. Elect. & Electron. Eng.
31006-Pamplona
Spain

Senior Editor
MICHIEL J. VELLEKOOP
Inst. Sensor and
Actuator Systems
Vienna Univ. Technol.
Gusshausstrasse 27/366
A-1040 Vienna
Austria

Senior Editor
JOHN R. VIG
Consultant
33 Bucks Mill Rd.
Colts Neck, NJ 07722
USA

Past Editor-in-Chief
H. TROY NAGLE
Founding Chair
Biomedical Engineering
UNC Chapel Hill—
NC State Univ.
Raleigh, NC 27695-7115
USA

Founding Editor-in-Chief
VLADIMIR LUMELSKY
Univ. of Wisconsin
Madison, WI 53706
USA

Editorial Board

M. NURUL ABEDIN
NASA Langley Research Center
Lasers & Electro-Optics Branch
Hampton, VA 23681-2199 USA

FRANCISCO J. ARREGUI
Univ. Publica de Navarra
Ingenieria Electrica y Electronica
Pamplona 31006, Spain

PLAMEN ATANASSOV
Univ. New Mexico

ELLIOT R. BROWN
Univ. California, Santa Barbara
Elec. Comput. Eng.
Santa Barbara, CA 93106 USA

TIANHONG CUI
Dept. Mechanical Engineering
Nanofabrication Center
Univ. Minnesota, Twin Cities Campus
Minneapolis, MN 55455 USA

MASAYOSHI ESASHI

OKYAY KAYNAK
Bogazici University
34342 Istanbul, Turkey

BAHRAM G. KERMANI
Complete Genomics, Inc.
Mountain View, CA 94043 USA

KISEON KIM
Gwangju Inst. Sci. Tech.
Gwangju, 500-712 Korea

JACQUES PISTRE
ENSEIRB
F-333402 Talence Cedex, France

PAUL P. L. REGTIEN
Twente Univ.
Wiskunde & Inform. Elektrotech.
The Netherlands

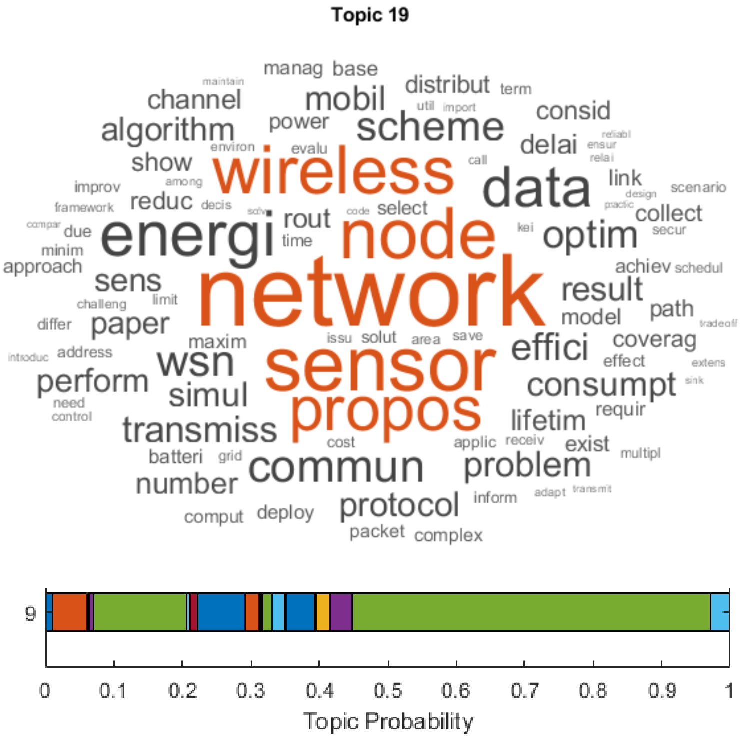
PATRICK RUTHER
Microsystem Materials
Laboratory (MML)

BORIS STOEBER
Univ. British Columbia
Mechanical Eng.
V6T 1Z4 Vancouver, BC, Canada

OOI KIANG TAN
Nanyang Tech. Univ.
Microelectronics Centre
Singapore 639798

FRANCIS ENG HOCK TAY
Natl. Univ. Singapore

Topic 19 (Primary Topic of Doc 9)

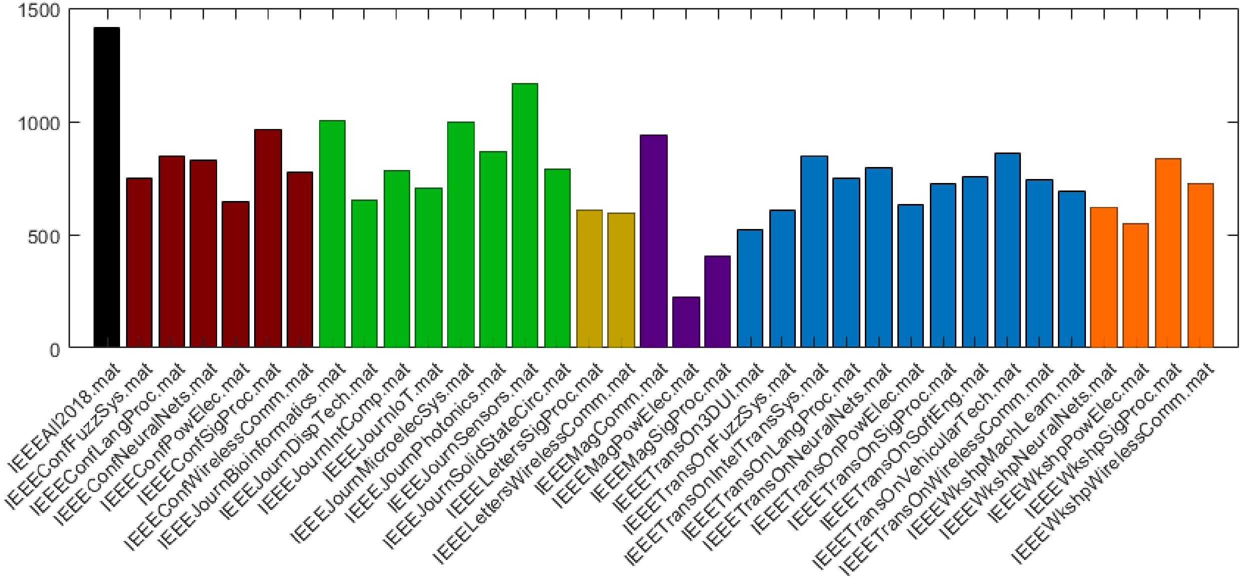


Abstract

As an extension of wireless sensor network in underwater environment, underwater acoustic sensor networks (UASNs) have caused widespread concern of academia. In UASNs, the efficiency and reliability of data transmission are very challenging due to the complex underwater environment in variety of ocean applications, such as monitoring abnormal submarine oil pipelines. Motivated by the importance of energy consumption in many deployments of UASNs, we therefore propose an energy-efficient data transmission scheme in this paper, called energy-efficiency grid routing based on 3D cubes (EGRCs) in UASNs, considering the complex properties of underwater medium, such as 3D changing topology, high propagation delay, node mobility and density, as well as rotation mechanism of cluster-head nodes. First, the whole network model is regarded as a 3D cube from the grid point of view, and this 3D cube is divided into many small cubes, where a cube is seen as a cluster. In the 3D cube, all the sensor nodes are duty-cycled in the media access control layer. Second, in order to make energy efficient and extend network lifetime, the EGRC shapes an energy consumption model considering residual energy and location of sensor nodes to select the optimal cluster-heads. Moreover, the EGRC utilizes residual energy, locations, and end-to-end delay for searching for the next-hop node to maintain the reliability of data transmission. Simulation validations of the proposed algorithm are carried out to show the effectiveness of EGRC, which performs better than the representative algorithms in terms of energy efficiency, reliability, and end-to-end delay.

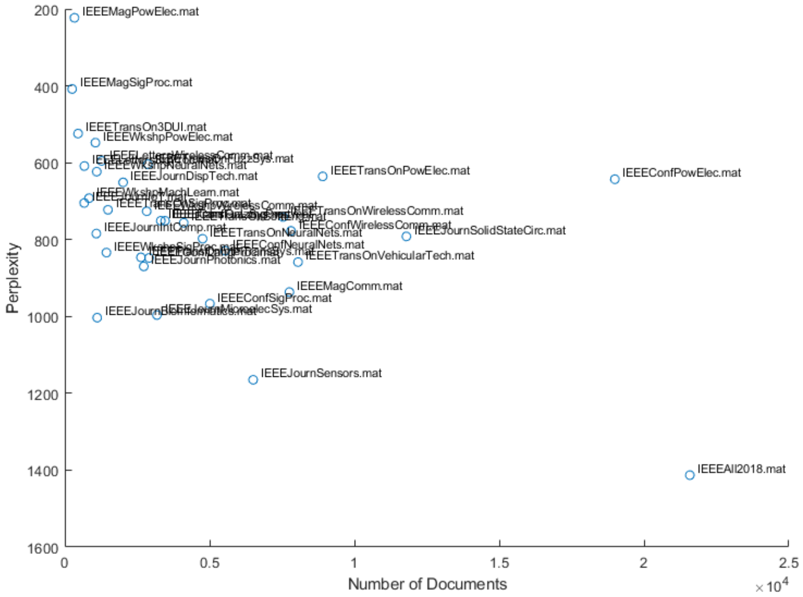
K. Wang, H. Gao, X. Xu, J. Jiang and D. Yue, "An Energy-Efficient Reliable Data Transmission Scheme for Complex Environmental Monitoring in Underwater Acoustic Sensor Networks," in IEEE Sensors Journal, vol. 16, no. 11, pp. 4051-4062, June1, 2016.

Perplexity by Journal



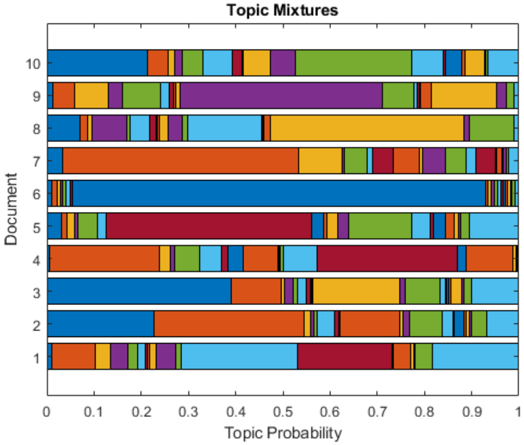
Data Set Size

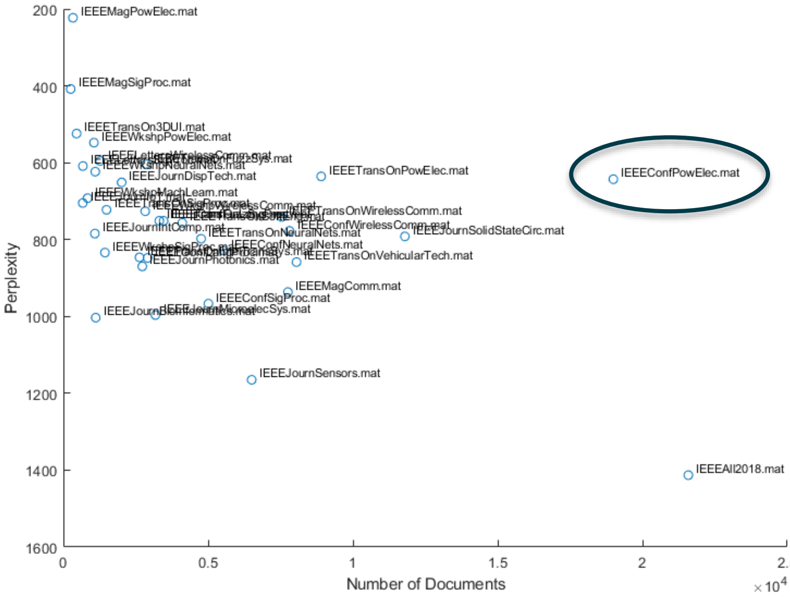
- Smaller data set ∝ lower perplexity
  - Magazines vs. All2018
- Plenty of exceptions though
  - IEEE Conferences on Power Electronics



Data Set Size

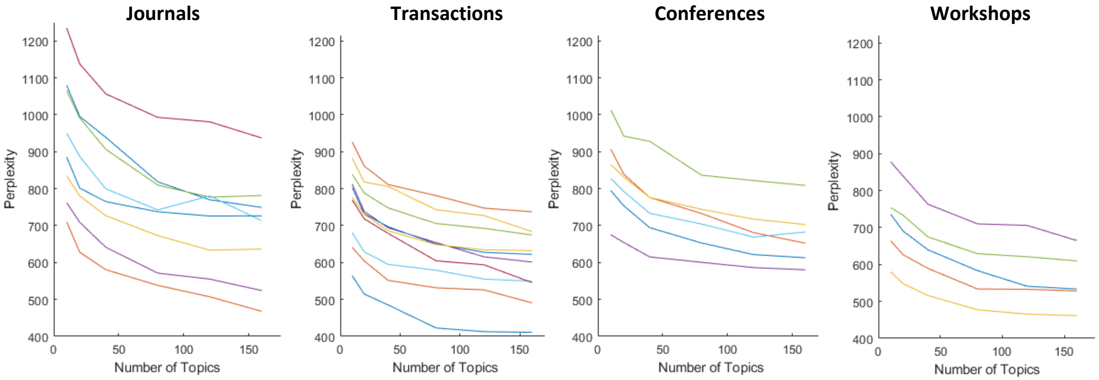
- Smaller data set ∝ lower perplexity
  - Magazines vs. All2018
- Plenty of exceptions though
  - IEEE Conferences on Power Electronics





Question #2: Do Journals and Transactions vary more or less than Conference and Workshop Proceedings?

- **Not really**
- Journals are more varied
- Subject matter is far more influential

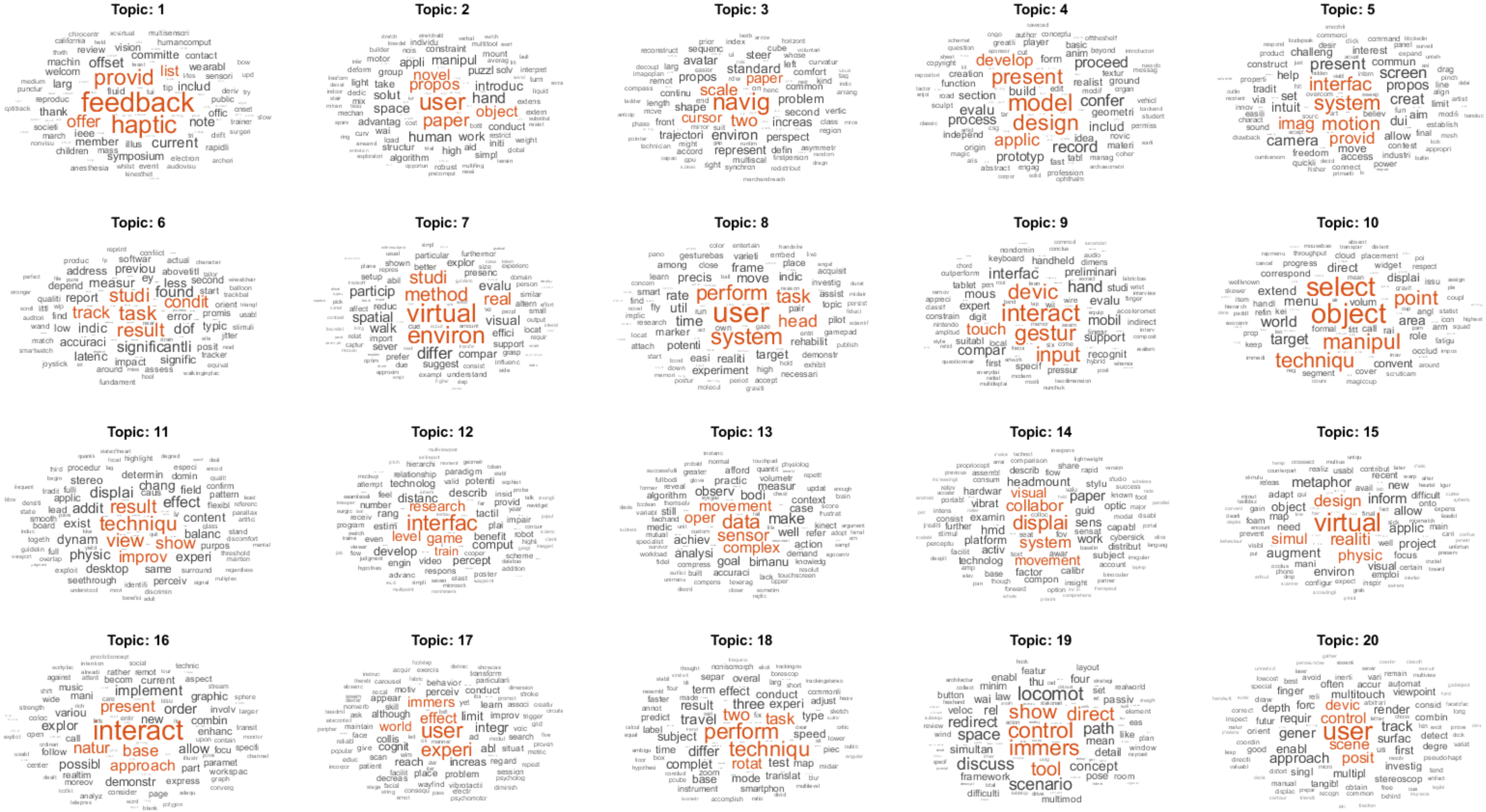


MISSOURI S&T

Outline

- Introduction
- 2018 Papers Analysis by LDA and LSA
- Model Applicability
- TF-IDF Analysis
- Perplexity & Insights by Publication
- **Topic Accuracy**
- Conclusions and Ending Points

Highly Specific Topic:
IEEE Transactions on 3D User Interfaces



Topic: 1

Topic: 2

Topic: 3

Topic: 4

Topic: 5

Topic: 6

Topic: 7

Topic: 8

Topic: 9

Topic: 10

Topic: 11

Topic: 12

Topic: 13

Topic: 14

Topic: 15

Topic: 16

Topic: 17

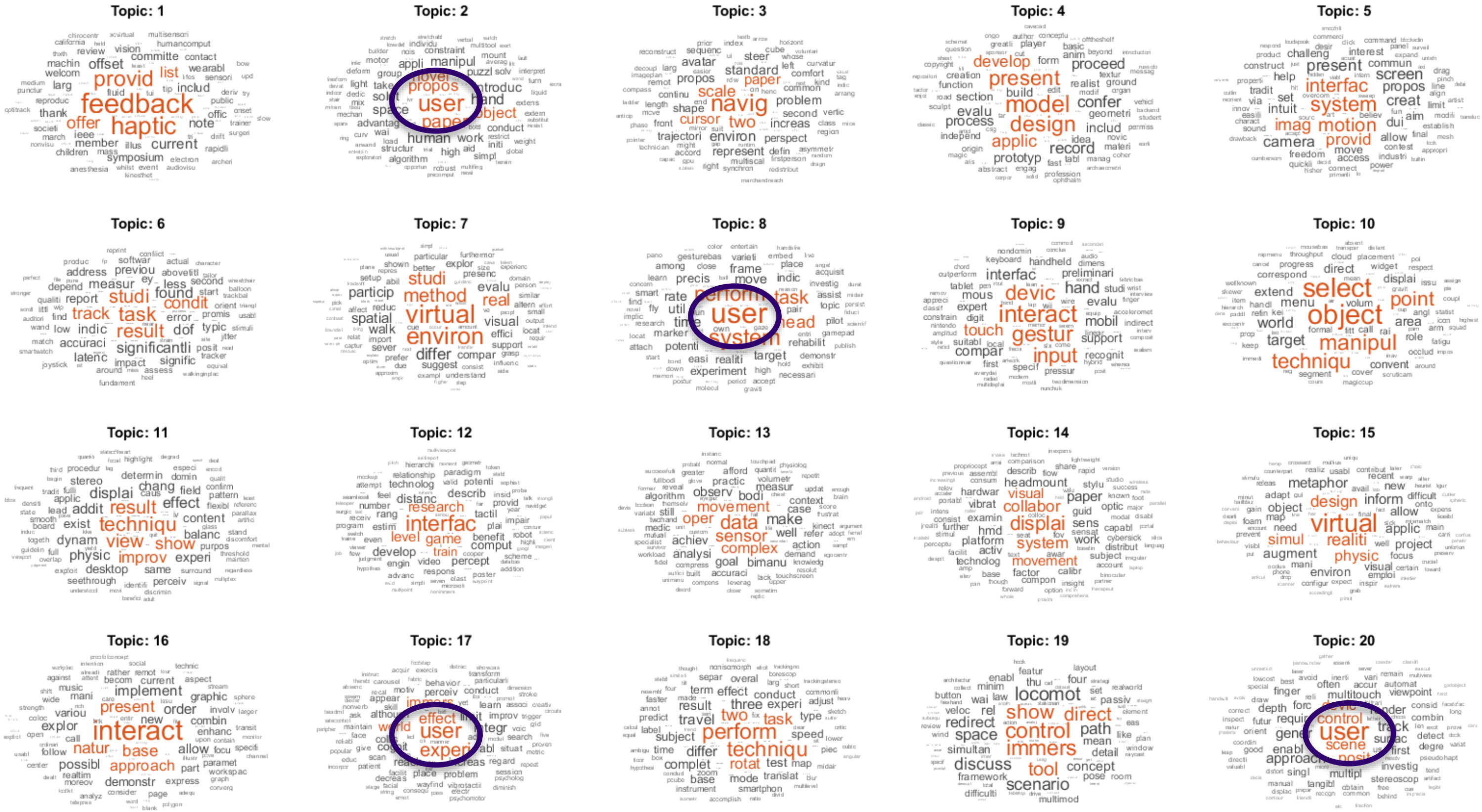
Topic: 18

Topic: 19

Topic: 20

MISSOURI S&T

Highly Specific Topic:
IEEE Transactions on 3D User Interfaces


Topic: 1

Topic: 2

Topic: 3

Topic: 4

Topic: 5

Topic: 6

Topic: 7

Topic: 8

Topic: 9

Topic: 10

Topic: 11

Topic: 12

Topic: 13

Topic: 14

Topic: 15

Topic: 16

Topic: 17

Topic: 18

Topic: 19

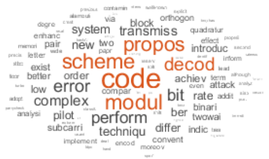
Topic: 20

MISSOURI
S&T

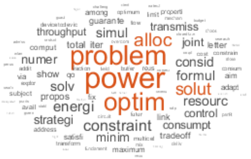
Broader Topic:

IEEE Transactions on Wireless Communications

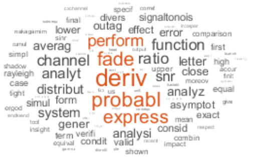
Topic: 1



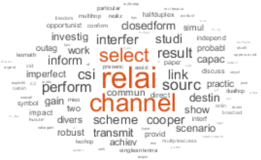
Topic: 2



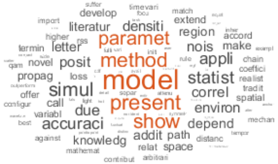
Topic: 3



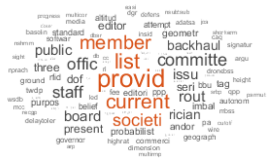
Topic: 4



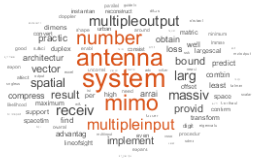
Topic: 5



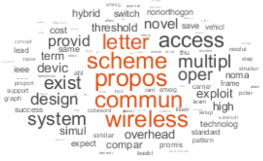
Topic: 6



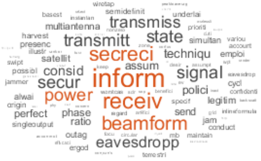
Topic: 7



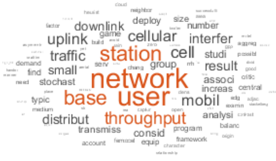
Topic: 8



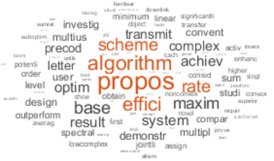
Topic: 9



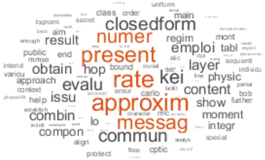
Topic: 10



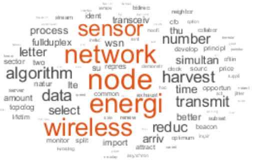
Topic: 11



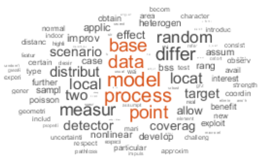
Topic: 12



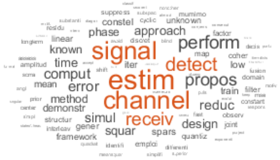
Topic: 13



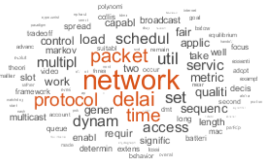
Topic: 14



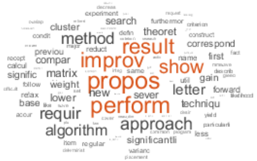
Topic: 15



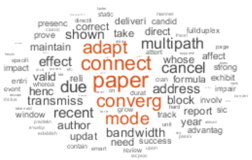
Topic: 16



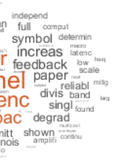
Topic: 17



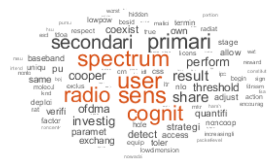
Topic: 18



Topic: 19

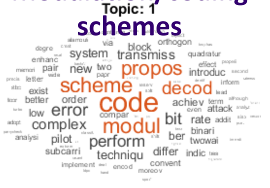


Topic: 20

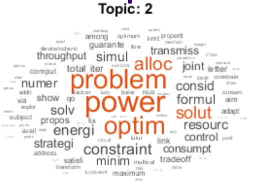


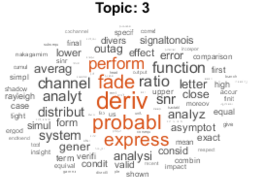
Topic Estimations

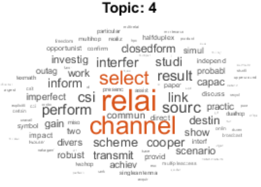
Modulation/coding schemes



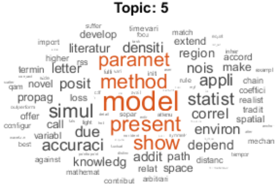
Power Optimization



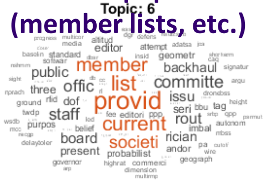




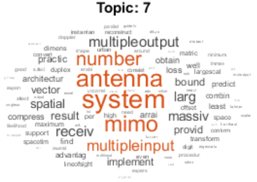
New model proposition

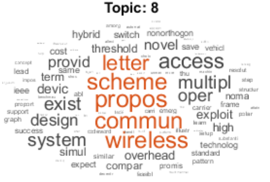


Meta papers (member lists, etc.)

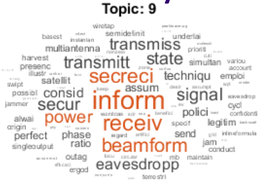


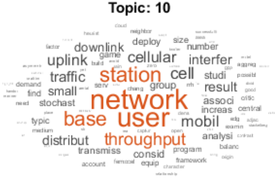
MIMO



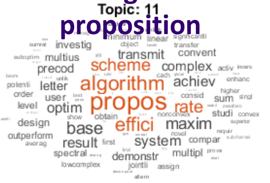


Security

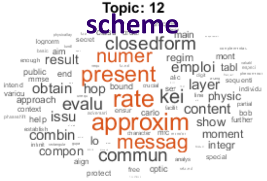




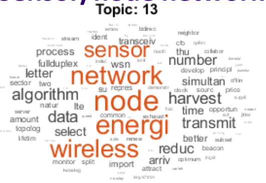
New algorithm proposition

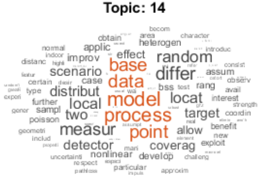


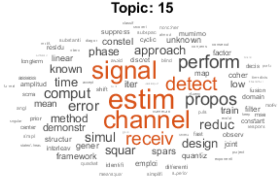
Rate estimation scheme



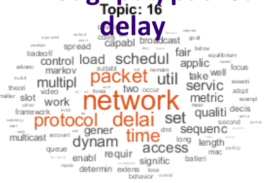
Sensor/node networks



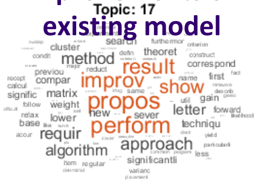




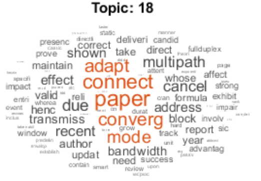
Throughput/packet delay



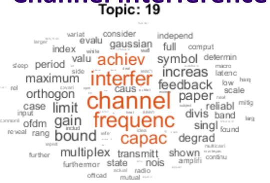
Improvements on existing model

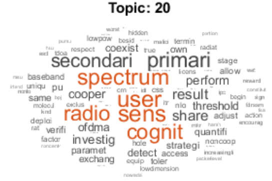


Survey Papers



Channel interference





Question #3: Can papers be neatly clustered by subject matter?

- Key aspects
  - Number of topics: smaller is better
    - Optimum perplexity often way too high
  - Subject matter is important: broad subject accommodate models better
  - Human insight is always needed, ambiguous topics

- Conclusion? **Yes**, fairly well

MISSOURI S&T

Outline

- Introduction
- 2018 Papers Analysis by LDA and LSA
- Model Applicability
- TF-IDF Analysis
- Perplexity & Insights by Publication
- Topic Accuracy
- **Conclusions and Ending Points**

Conclusions

- Is this a valuable research tool for you?
- Can reveal current research trends
- What is this valuable for?
  - Classifying upcoming papers
  - Scoring your paper according to current trends

Code

- Scripts and plots developed in MATLAB
- Code available on my Github
  - https://github.com/stewythe1st/Research-Mining
- IEEE Web API
  - https://developer.ieee.org/docs

Moving Forward

- Figure out a good way to determine optimal number of topics
- Try out with documents from other sources
- Perhaps perform analysis on paper content, not just abstract

Citations

Research Used

- K. S. Jones, “A Statistical Interpretation Of Term Specificity And Its Application In Retrieval”, Journal of Documentation, vol. 28, issue 1, pp. 11-21, 1972
- D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, Jan. 2003.
- S. C. Deerwester, S. T. Dumais, G. W. Furnace, R. A. Harshman, T. K. Landauer, K. E. Lochbaum, and L. A. Streeter, “Computer information retrieval using latent semantic structure ,” 13-Jun-1989.

Works Used in Examples

- A. Zanella, N. Bui, A. Castellani, L. Vangelista and M. Zorzi, "Internet of Things for Smart Cities," in IEEE Internet of Things Journal, vol. 1, no. 1, pp. 22-32, Feb. 2014.
- Rui Xu and D. Wunsch, "Survey of clustering algorithms," in IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, May 2005.
- T. Iwata, T. Hirao and N. Ueda, "Topic Models for Unsupervised Cluster Matching," in IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 4, pp. 786-795, April 1 2018.
- K. Wang, H. Gao, X. Xu, J. Jiang and D. Yue, "An Energy-Efficient Reliable Data Transmission Scheme for Complex Environmental Monitoring in Underwater Acoustic Sensor Networks," in IEEE Sensors Journal, vol. 16, no. 11, pp. 4051-4062, June1, 2016.
- "IEEE Sensors Journal publication information," in IEEE Sensors Journal, vol. 9, no. 11, pp. C2-C2, Nov. 2009.

Questions?

- Questions/Goals
  - How similar are papers published in journals? **More similar than you think!**
  - Can papers be neatly clustered by subject matter? **Yes!**
  - Do Journals and Transactions vary more or less than Conference and Workshop Proceedings? **Not really**
- Conclusions
  - Not much of a valuable research tool for most people
  - Does reveal current research trends
  - Is this valuable for…
    - Classifying upcoming papers
    - Scoring your paper according to current trends