Natural Language Processing for Nuclear Science Scholarship

W. Younes

Lawrence Berkeley Natl. Lab.

WANDA2021

Background

- Overview
  - Large body of existing nuclear science literature + steady daily additions
    - Significant challenge to archive/search/retrieve useful info
  - Goal: store/search by “meaning” rather than keywords
- Approach
  - Develop NLP framework to automatically categorize, summarize, and recommend nuclear science references
- Impact to Nuclear Data
  - Augments the nuclear data pipeline
  - Aids researchers in addressing current and future nuclear data needs

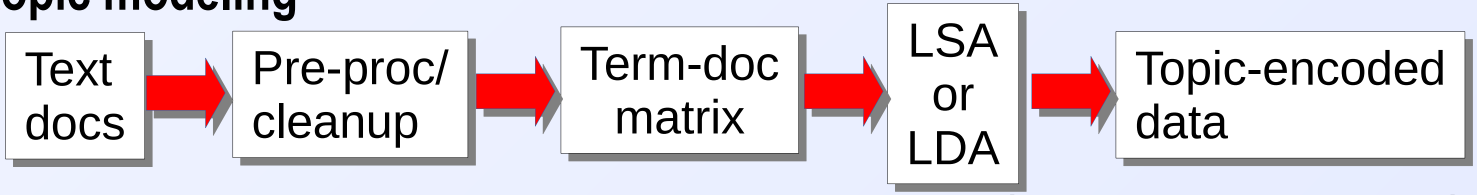
Computational Needs

- HPC resources
  - Development on multi-core desktop machines is ongoing
  - Scalable to HPC machines
- AI/ML resources
  - NLP algorithms to pre-process text (tokenization, stemming/lemmatizing, stop word removal)
  - Graph-based algorithm for keywording and summarization
  - Unsupervised ML algorithms for topic modeling
  - Planned extension to deep learning algorithms for summarization and article recommendation

Computational Techniques

- Techniques used
  - Topic modeling ⇒ documents as prob. distributions over topics, and topics as prob. distributions over words
- Algorithms/software
  - TextRank algorithm (similar to Google's PageRank) for keywording and summarization

  - Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) for topic modeling

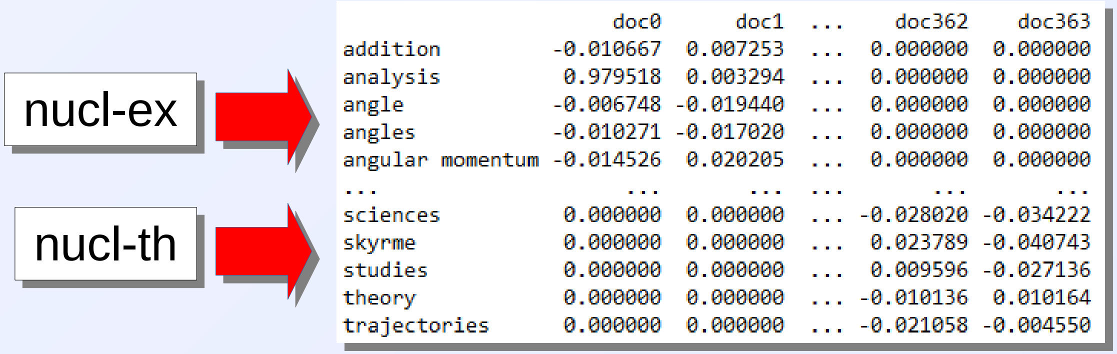


Text docs → Pre-proc/ cleanup → Term-doc matrix → LSA or LDA → Topic-encoded data

  - Homegrown implementations + python modules (nltk, sklearn, gensim)
- Hardware architecture
  - Present focus is on CPU-based computing

Results: LSA

- Papers retrieved from arXiv using two search strings:
  - “abs:fission AND cat:nucl-ex” → 382 papers in group #1
  - “abs:fission AND cat:nucl-th” → 346 papers in group #2
- Docs split at random into training and validation sets
- Pre-processing with TextRank to identify top keywords in each doc
- Training: term-doc matrices for both searches reduced by SVD and combined by forcing block-diagonal form:

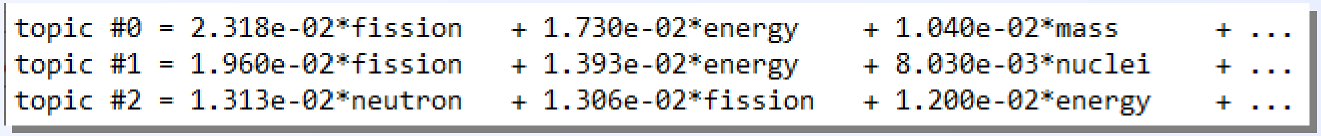

nucl-ex
nucl-th

```
                      doc0      doc1  ...    doc362    doc363
addition         -0.010667  0.007253  ...  0.000000  0.000000
analysis          0.979518  0.003294  ...  0.000000  0.000000
angle            -0.006748 -0.019440  ...  0.000000  0.000000
angles           -0.010271 -0.017020  ...  0.000000  0.000000
angular momentum -0.014526  0.020205  ...  0.000000  0.000000
...                    ...       ...  ...       ...       ...
sciences          0.000000  0.000000  ... -0.028020 -0.034222
skyrme            0.000000  0.000000  ...  0.023789 -0.040743
studies           0.000000  0.000000  ...  0.009596 -0.027136
theory            0.000000  0.000000  ... -0.010136  0.010164
trajectories      0.000000  0.000000  ... -0.021058 -0.004550
```

- Validation: similarity metric used to categorize new docs from validation set yields correct assignment in typically > 70% of cases

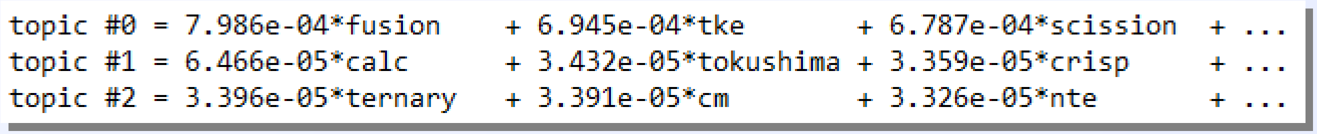
Results: LDA

- All docs from arXiv “fission” expt and theory searches taken together
- Pre-processing (including removal of stop words and stemming)
- Topics extracted with standard LDA:



```
topic #0 = 2.318e-02*fission   + 1.730e-02*energy   + 1.040e-02*mass     + ...
topic #1 = 1.960e-02*fission   + 1.393e-02*energy   + 8.030e-03*nuclei   + ...
topic #2 = 1.313e-02*neutron   + 1.306e-02*fission  + 1.200e-02*energy   + ...
```

- Topics extracted with weighted (TF-IDF) LDA:



```
topic #0 = 7.986e-04*fusion    + 6.945e-04*tke       + 6.787e-04*scission  + ...
topic #1 = 6.466e-05*calc      + 3.432e-05*tokushima + 3.359e-05*crisp     + ...
topic #2 = 3.396e-05*ternary   + 3.391e-05*cm        + 3.326e-05*nte       + ...
```

- Notes:
  - Weighted LDA gives more complementary topics
  - More work needed to filter out nuisance words (“tokushima”, “crisp”, …)
  - Work in progress on training and validation of LDA model

Outlook

- To do:
  - Build databases of stop words and meaningful words (vocabulary)
  - Expand testing of LSA/LDA to larger corpuses
  - Develop metrics for tailored article recommendations (e.g., by level/pedagogy)
  - Explore deep-learning applications for summarization and recommendation
- Benefits
  - Complements/augments capabilities of both archivists and users
  - Can be integrated with existing USNDP databases (NSR) and tools
  - Code development uses arXiv, but will eventually include original refereed papers (e.g. PRC)