ERUDITE: THE EDUCATIONAL RESOURCE DISCOVERY INDEX FOR DATA SCIENCE

Jose Luis Ambite, Kristina Lerman, Lily Fierro, Jonathan Gordon, Florian Geigl, Gully APC Burns, Priyambada Jain, Avnish Bhatraji, Jack Van Horn

University of Southern California
BD2K Training Coordination Center: Plan

• Identify high-quality training resources on the web
  • BD2K Centers training components
  • Massive Open Online Courses: Coursera, Udacity, edX, …
  • Tutorials, keynotes, talks: videolectures.net, youtube.com, …
  • Cross-listing agreement with Elixir

• Describe and organize resources into ERuDlte index
  • Automatic modeling plus human curation
    • Define common schema for training resources
    • Information extraction techniques to populate schema elements
    • Machine learning/Information retrieval techniques to categorize and relate resources

• Provide personalize training path for user learning goal
  • User modeling
  • Recommendation engine
Outline

• Automatic data extraction of learning resources
• ERuDIte: current database

• Modelling Learning Resources
  • Concept/Tag Taxonomy
    • Automatic Tag Discovery/Validation
  • Automatic Concept/Tag assignment
    • Machine Learning Methods
    • Information Retrieval Methods
  • Collaboration with Elixir/TESS

• Next Steps
Outline

• Automatic data extraction of learning resources
• ERuDIte: current database
• Modelling Learning Resources
  • Concept/Tag Taxonomy
    • Automatic Tag Discovery/Validation
  • Automatic Concept/Tag assignment
    • Machine Learning Methods
    • Information Retrieval Methods
  • Collaboration with Elixir/TESS
• Next Steps
Lesson 4: Outliers and Normal Distribution
Outliers, Quartiles; Binomial Distribution; Central Limit Theorem; Manipulating Normal Distribution

Lesson 5: Inference
Confidence intervals; Hypothesis Testing

Lesson 6: Regression
Linear regression; correlation

Lesson 7: Final Exam

Instructors & Partners

Sebastian Thrun
INSTRUCTOR
ERuDItte Automated Scraper System

- Programmed in Python: dryscrape, beautifulsoup
- Extensible, modular design: Module for each website
- Capable of handling Javascript pages, script interaction
- Framework portable as docker image
Outline

• Automatic data extraction of learning resources
• **ERuDIte: current database**
• Modelling Learning Resources
  • Concept/Tag Taxonomy
    • Automatic Tag Discovery/Validation
  • Automatic Concept/Tag assignment
    • Machine Learning Methods
    • Information Retrieval Methods
  • Collaboration with Elixir/TESS
• Next Steps
## Currently Indexed Learning Resources

<table>
<thead>
<tr>
<th>Provider</th>
<th>Types</th>
<th>Total</th>
<th>With Descriptions</th>
<th>With Transcripts</th>
<th>With Slides or Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD2K</td>
<td>Video / Written</td>
<td>173</td>
<td>155</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>edX</td>
<td>Course / Video</td>
<td>100</td>
<td>99</td>
<td>90</td>
<td>73</td>
</tr>
<tr>
<td>Coursera</td>
<td>Course / Video</td>
<td>84</td>
<td>84</td>
<td>64</td>
<td>62</td>
</tr>
<tr>
<td>Udacity</td>
<td>Course / Video</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Videolectures</td>
<td>Video</td>
<td>206</td>
<td>139</td>
<td>206</td>
<td>190</td>
</tr>
<tr>
<td>YouTube</td>
<td>Video</td>
<td>134</td>
<td>106</td>
<td>127</td>
<td>0</td>
</tr>
<tr>
<td>Elixir</td>
<td>Course / Written</td>
<td>240</td>
<td>48</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bioconductor</td>
<td>Course / Written</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cornell Virtual Workshop</td>
<td>Course / Written</td>
<td>38</td>
<td>19</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>997</td>
<td>669</td>
<td>350</td>
<td>190</td>
</tr>
</tbody>
</table>
Outline

• Automatic data extraction of learning resources
• ERuDlte: current database

• **Modelling Learning Resources**
  • **Concept/Tag Taxonomy**
    • Automatic Tag Discovery/Validation
  • Automatic Concept/Tag assignment
    • Machine Learning Methods
    • Information Retrieval Methods
  • Collaboration with Elixir/TESS

• Next Steps
Concept Map: Tagging Learning Resources with a Controlled Vocabulary

Dimension 1: Data Science Process

7 Hierarchical Dimensions
Outline

• Automatic data extraction of learning resources
• ERuDIte: current database
• Modelling Learning Resources
  • Concept/Tag Taxonomy
    • Automatic Tag Discovery/Validation
  • Automatic Concept/Tag assignment
    • Machine Learning Methods
    • Information Retrieval Methods
  • Collaboration with Elixir/TESS

• Next Steps
Vocabulary Discovery via Topic Models

• Automatically discover new concepts/tags using Topic Modeling (Non-negative matrix factorization)
• Representation of topics using Word Clouds
Topic Distribution of EruDite Resources exhibits Structure

[Dimensionality Reduction from topic distribution vectors to 2D using t-SNE ]
Introduction to Reinforcement Learning and Bayesian Learning

- reinforcement learning
- probability_statistics
- mathematics
- machine learning
Sequential Monte-Carlo Methods

- mathematics
- probability_statistics
- video
- in_depth

SMC methods gained popularity as powerful tools for solving intra of sequential data. Much effort was devoted to the development of estimating the filtering distribution algorithms, including auxiliary-variable techniques for control and planning. Various SMC algorithms were also studied experimental frameworks, thus allowing for the computation of eigen-pairs of large matrices. These methods were also studied experimental frameworks for building efficient high-dimensional proposal distributions for MCMC using SMC methods were proposed. These allow us to design effective MCMC algorithms in complex scenarios where standard strategies failed. Such methods have been demonstrated on a number of domains, including simulated folding and stochastic difference in static models.

Topic model suggests additional tags

Tags:
- mathematics
- probability_statistics
- video
- in_depth
Outline

• Automatic data extraction of learning resources
• ERuDIte: current database
• Modelling Learning Resources
  • Concept/Tag Taxonomy
    • Automatic Tag Discovery/Validation
  • **Automatic Concept/Tag assignment**
    • Machine Learning Methods
    • Information Retrieval Methods
  • Collaboration with Elixir/TESS
• Next Steps
Automated Learning Resource Tagging

• *Key to scalability: automated resource tagging*

• Approaches:
  • Machine Learning: Supervised learning
  • Information Retrieval: Resource similarity

• Two Experimental setups
  1. Consider only existing training tags
  2. Consider also parent tags up the hierarchy in training
Automated Tagging
Machine Learning

Extract Text
- Resource
- Resource
- ... Resource

Text Preprocessing
- Remove stopwords

Vectorize Text
- Tf-idf of words and bigrams

One vs. Rest Classifier
- Classifier for Tag1
- Classifier for Tag2
- ... Classifier for Tag2

Classifiers Tested:
- Random Forests
- Support Vector Machines
- Multinomial Naïve Bayes
- K-Nearest Neighbors
- Logistic Regression
Automated Tagging Results
Machine Learning

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Exact Tag (F1 score)</th>
<th>Exact + Parents Tags (F1 score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.66</td>
<td>0.75</td>
</tr>
<tr>
<td>Multinomial Naïve Bayes</td>
<td>0.70</td>
<td>0.80</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td><strong>0.73</strong></td>
<td><strong>0.83</strong></td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>0.73</td>
<td>0.81</td>
</tr>
<tr>
<td>K-Nearest Neighbors</td>
<td>0.70</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Techniques already useful for human curators (continue to work on improving results)
- Results of best hyperparameters optimized for F1 score
- F1 score= harmonic mean of Precision (Positive Predictive Value) and Recall (Sensitivity) = $2PR / (P + R)$
Automated Tagging Information Retrieval

For every tag \( t \):

- Training Resources assigned to \( t \):
  - TR1
  - TR2
  - TR3

Aggregate Cosine Similarities

Assign Tags to Resource \( R \):
- \( t1 \) 0.8 \( R \)
- \( t3 \) 0.7 \( R \)

\( R \) = Vectorized Representation of a Resource

Incoming Resource
- Compare with Cosine Similarity
## Automated Tagging Results

Information Retrieval

<table>
<thead>
<tr>
<th>Method</th>
<th>Title weight</th>
<th>Subtitle weight</th>
<th>Desc. weight</th>
<th>Max n-gram</th>
<th>Min DF</th>
<th>Max DF</th>
<th>Vector weight method</th>
<th>Similarity aggregate function</th>
<th>Rank</th>
<th>Dims</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0.6</td>
<td>Inc</td>
<td>max</td>
<td>10</td>
<td>5300-5700</td>
<td>0.784</td>
</tr>
<tr>
<td>NMF</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0.6</td>
<td>Inc</td>
<td>max</td>
<td>10</td>
<td>40</td>
<td>0.774</td>
</tr>
<tr>
<td>LSA</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>0.6</td>
<td>Inc</td>
<td>max</td>
<td>10</td>
<td>100</td>
<td>0.775</td>
</tr>
</tbody>
</table>

\[
\text{Inc} = \frac{1+\log(tf)}{\sqrt{\sum t_1^2 + t_2^2 + t_3^2 + t_4^2 + \ldots}}
\]
Outline

• Automatic data extraction of learning resources
• ERuDIte: current database
• Modelling Learning Resources
  • Concept/Tag Taxonomy
    • Automatic Tag Discovery/Validation
  • Automatic Concept/Tag assignment
    • Machine Learning Methods
    • Information Retrieval Methods
  • Collaboration with Elixir/TESS
• Next Steps
Schema Design & Collaboration with Elixir/BioSchemas

• Designed Schema for Learning Resources
  • Used existing standards to inform current model: Dublin Core, Learning Resource Metadata Initiative (LRMI), IEEE’s Learning Object Metadata (LOM), eXchanging Course Related Information (XCRI), XML Schema Definition (XSD), Metadata for Learning Opportunities (MLO), and Schema.org

• Reach a common schema for TCC and Elixir/Bioschemas
  • Shared and aligned TCC schema with Elixir/TESS/Bioschemas
  • Inform schema.org Course schema: http://pending.schema.org/Course
Outline

• Automatic data extraction of learning resources
• ERuDIte: current database
• Modelling Learning Resources
  • Concept/Tag Taxonomy
    • Automatic Tag Discovery/Validation
  • Automatic Concept/Tag assignment
    • Machine Learning Methods
    • Information Retrieval Methods
  • Collaboration with Elixir/TESS

• Next Steps
Next Steps

• Incorporate additional learning resources
  • Automatically identify training resources
    • Data science videos on youtube: On-going experiments point to 1000s of relevant videos
    • Scientific literature: books, tutorial/overview articles (building on TechKnAcq)
    • Educational resources from BD2K grantees, as they appear
  • Collect more content from each resource
    • Collect transcripts, slides from MOOCs
    • Speech recognition on video
  • Automatically identify learning resource pre-requisites/dependencies
    • Identify dependencies using cross-entropy (building on TechKnAcq)
    • Analyze concept ordering from syllabi, book chapters, data science curricula
    • Monitor user behaviour
• Personalize
  • Collect user data from database usage
  • Recommend related resources
THANK YOU
QUESTIONS?
Leveraging the Bibliome for BigDataU

• Written material provides the most extensive knowledge resource for study and learning
  • Textbooks, online tutorials, web-pages, blog-posts, research articles, lesson plans, online slide presentations, etc.

• Apply statistical NLP methods to learn conceptual structure of subjects, detect pedagogically valuable resources, organize and visualize recommended reading resources for students of BigDataU.
TechKnAcq
Technical Knowledge Acquisition – IARPA Seedling

Automated processing of a technical document corpus to:
(A) Expand the corpus to include pedagogically-relevant material (textbooks + chapters, online encyclopedia articles, tutorials)
(B) Construct a ‘Concept Graph’ with edges based on notion of dependencies/prerequisites.
(C) Provide an interface that creates a pedagogically derived reading list

Website: [http://techknacq.isi.edu](http://techknacq.isi.edu)
Demo: [http://techknacq-demo.isi.edu](http://techknacq-demo.isi.edu)
(user: test, password: ‘Iamthewalrus’)
Finding Topic Dependencies using Cross Entropy

Topic $t_1$ depends on Topic $t_2$ if:

1. If $t_2$ explains occurrences of $t_1$ more than the reverse
   \[ H(t_2; t_1) > H(t_1; t_2) \]

2. The joint occurrence between $t_1$ and $t_2$ is significant
   \[ H(t_1, t_2) < \text{threshold} \]

Cross Entropy
\[ H(p; q) = E_p[\log q] = H(p) + D_{KL}(p||q) \]

Kullback–Leibler divergence
\[ D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \]

Joint Entropy
\[ H(X,Y) = -\sum_x \sum_y P(x,y) \log_2[P(x,y)] \]
Topic Prerequisites Detection

- Approach based on Cross-Entropy $H$

- $t_1$ depends on $t_2$ if:
  - $H(t_2; t_1) > H(t_1; t_2)$