

# Automatic Metadata Extraction

## The High Energy Physics Use Case

Joseph Boyd

École Polytechnique Fédérale de Lausanne

*joseph.boyd@epfl.ch*

August 29, 2015

# Motivation

- ▶ INSPIRE-HEP digital library at CERN contains over 1 Million documents
- ▶ Manual curation of high energy physics (HEP) papers may be automated with machine learning techniques
- ▶ Custom datasets and specialised features required to model HEP paper characteristics

# Aims

Take existing state-of-the-art system for metadata extraction to:

- ▶ demonstrate a qualitative difference between HEP and general papers;
- ▶ propose new features to enhance models;
- ▶ run experiments to test these new features, and;
- ▶ draw conclusions about what characterises good feature engineering.

# Outline

Introduction

**Automatic Metadata Extraction**

Data and Features

Key Results

Conclusions

# Metadata Extraction

- ▶ *Metadata* refers to content useful to the bibliographic identification of the document
- ▶ *Extraction* refers to the classification of metadata within the document text
- ▶ Several automatic approaches exist: stylistic analysis, knowledge-base, machine learning (CRFs, HMMs, SVMs)...

# Metadata Extraction (Illustration)

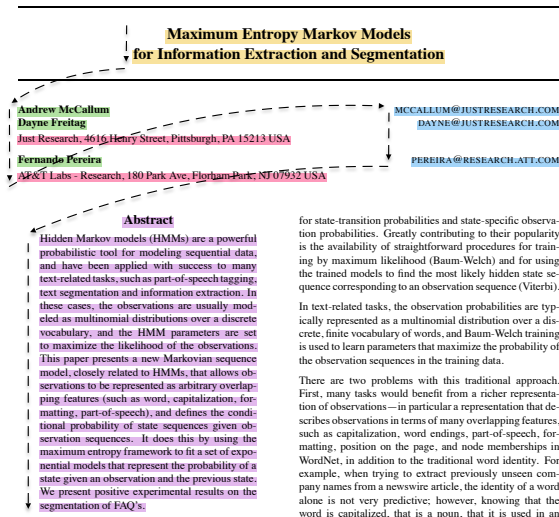


Figure: Tagging of a document header section.

# Why Conditional Random Fields?

- ▶ Transition interdependencies implies graphical structure  $\rightarrow$  model as a structured sequence
- ▶ Modelling conditional distribution,  $p(\mathbf{y}|\mathbf{x})$ , sufficient for classification
- ▶ Exploit rich information about observations,  $\mathbf{x}$ , without explicitly modelling the underlying probability distribution
- ▶ Classifying metadata may greatly benefit from modelling rich text features (punctuation, font size, layout ...)

# GROBID

- ▶ Selected according to performance in study comparing AME systems [1]
- ▶ Open source Java-based tool developed at INRIA, France
- ▶ Manages *cascade* of CRF models for annotating papers in progressively finer detail
- ▶ Uses C++ library *Wapiti* for back-end calculations (model training, prediction)



# GROBID - CRF Cascade

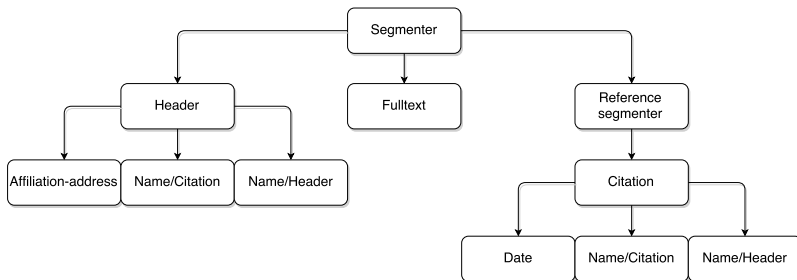


Figure: Cascade of models used by GROBID

# Outline

Introduction

Automatic Metadata Extraction

**Data and Features**

Key Results

Conclusions

## Identification of beauty and charm quark jets at LHCb

The LHCb collaboration<sup>†</sup>

### Abstract

Identification of jets originating from beauty and charm quarks is important for measuring Standard Model processes and for searching for new physics. The performance of algorithms developed to select  $b$ - and  $c$ -quark jets is measured using data recorded by LHCb from proton-proton collisions at  $\sqrt{s} = 7$  TeV in 2011 and at  $\sqrt{s} = 8$  TeV in 2012. The efficiency for identifying a  $b(c)$  jet is about 65%(25%) with a probability for misidentifying a light-parton jet of 0.3% for jets with transverse momentum  $p_T > 20$  GeV and pseudorapidity  $2.2 < \eta < 4.2$ . The dependence of the performance on the  $p_T$  and  $\eta$  of the jet is also measured.

Submitted to JINST

© CERN on behalf of the LHCb collaboration, license CC-BY-4.0.

Figure: Collaboration field in header section.

# HEP Paper Characteristics (ii)

## LHCb collaboration

R. Aaij<sup>38</sup>, B. Adeva<sup>37</sup>, M. Adinolfi<sup>46</sup>, A. Affolder<sup>52</sup>, Z. Ajaltouni<sup>5</sup>, S. Akar<sup>6</sup>, J. Albrecht<sup>9</sup>, F. Alessio<sup>38</sup>, M. Alexander<sup>51</sup>, S. Ali<sup>41</sup>, G. Alkhazov<sup>30</sup>, P. Alvarez Cartelle<sup>53</sup>, A.A. Alves Jr<sup>57</sup>, S. Amato<sup>2</sup>, S. Amerio<sup>22</sup>, Y. Amhis<sup>7</sup>, L. An<sup>3</sup>, L. Anderlini<sup>17,g</sup>, J. Anderson<sup>40</sup>, M. Andreotti<sup>16,f</sup>, J.E. Andrews<sup>58</sup>, R.B. Appleby<sup>54</sup>, O. Aquines Gutierrez<sup>10</sup>, F. Archilli<sup>38</sup>, P. d'Argent<sup>11</sup>, A. Artamonov<sup>35</sup>, M. Artuso<sup>59</sup>, E. Aslanides<sup>6</sup>, G. Auremma<sup>25,n</sup>, M. Baalouch<sup>5</sup>, S. Bachmann<sup>11</sup>, J.J. Back<sup>48</sup>, A. Badalov<sup>36</sup>, C. Baesso<sup>60</sup>, W. Baldini<sup>16,38</sup>, R.J. Barlow<sup>54</sup>, C. Barschel<sup>38</sup>, S. Barsuk<sup>7</sup>, W. Barter<sup>38</sup>, V. Batozskaya<sup>28</sup>, V. Battista<sup>39</sup>, A. Bay<sup>39</sup>, L. Beaucourt<sup>4</sup>, J. Beddow<sup>51</sup>, F. Bedeschi<sup>23</sup>, I. Bediaga<sup>1</sup>, L.J. Bel<sup>41</sup>, I. Belyaev<sup>31</sup>, E. Ben-Haim<sup>8</sup>, G. Bencivenni<sup>18</sup>, S. Benson<sup>38</sup>, J. Benton<sup>46</sup>, A. Berezhnov<sup>32</sup>, R. Bernet<sup>40</sup>, A. Bertolin<sup>22</sup>, M.-O. Bettler<sup>38</sup>, M. van Beuzekom<sup>41</sup>, A. Bien<sup>11</sup>, S. Bifani<sup>45</sup>, T. Bird<sup>54</sup>, A. Birnkraut<sup>9</sup>, A. Bizzeti<sup>17,i</sup>, T. Blake<sup>48</sup>, F. Blanc<sup>39</sup>, J. Blouw<sup>10</sup>, S. Blusk<sup>59</sup>, V. Bocci<sup>25</sup>, A. Bondar<sup>34</sup>, N. Bondar<sup>30,38</sup>, W. Bonivento<sup>15</sup>, S. Borghi<sup>54</sup>, M. Borsato<sup>7</sup>, T.J.V. Bowcock<sup>52</sup>, E. Bowen<sup>40</sup>, C. Bozzi<sup>16</sup>, S. Braun<sup>11</sup>, D. Brett<sup>54</sup>, M. Britsch<sup>10</sup>, T. Britton<sup>59</sup>, J. Brodzicka<sup>54</sup>, N.H. Brook<sup>46</sup>, A. Bursche<sup>40</sup>, J. Buytaert<sup>38</sup>, S. Cadeddu<sup>15</sup>, R. Calabrese<sup>16,f</sup>, M. Calvi<sup>20,k</sup>, M. Calvo Gomez<sup>36,p</sup>, P. Campana<sup>18</sup>, D. Campora Perez<sup>38</sup>, L. Capriotti<sup>54</sup>, A. Carbone<sup>14,d</sup>, G. Carboni<sup>24,l</sup>, R. Cardinale<sup>19,j</sup>, A. Cardini<sup>15</sup>, P. Carniti<sup>20</sup>, L. Carson<sup>50</sup>, K. Carvalho Akiba<sup>2,38</sup>, R. Casanova Mohr<sup>36</sup>, G. Casse<sup>52</sup>, L. Cassina<sup>20,k</sup>, L. Castillo Garcia<sup>38</sup>, M. Cattaneo<sup>38</sup>, Ch. Cauet<sup>9</sup>, G. Cavallero<sup>19</sup>, R. Cenci<sup>23,l</sup>, M. Charles<sup>8</sup>, Ph. Charpentier<sup>38</sup>, M. Chefdeville<sup>4</sup>, S. Chen<sup>54</sup>, S.-F. Cheung<sup>55</sup>, N. Chiapolini<sup>40</sup>, M. Chrzaszcz<sup>40</sup>, X. Cid Vidal<sup>38</sup>, G. Ciezarek<sup>41</sup>, P.E.L. Clarke<sup>50</sup>, M. Clemencic<sup>38</sup>, H.V. Cliff<sup>47</sup>, J. Closier<sup>38</sup>, V. Coco<sup>38</sup>, J. Cogan<sup>6</sup>, E. Cogneras<sup>5</sup>, V. Cogoni<sup>15,e</sup>, L. Cojocariu<sup>29</sup>, G. Collazuol<sup>22</sup>, P. Collins<sup>38</sup>, A. Comerma-Montells<sup>11</sup>, A. Contu<sup>15,38</sup>, A. Cook<sup>46</sup>, M. Coombes<sup>46</sup>, S. Coquereau<sup>8</sup>, G. Corti<sup>38</sup>, M. Corvo<sup>16,f</sup>, I. Counts<sup>56</sup>, B. Couturier<sup>38</sup>, G.A. Cowan<sup>50</sup>, D.C. Craik<sup>48</sup>, A. Crocombe<sup>48</sup>, M. Cruz Torres<sup>60</sup>, S. Cunliffe<sup>53</sup>, R. Currie<sup>53</sup>, C. D'Ambrosio<sup>38</sup>, J. Dalseno<sup>46</sup>, P.N.Y. David<sup>41</sup>, A. Davis<sup>57</sup>, K. De Bruyn<sup>41</sup>, S. De Capua<sup>54</sup>, M. De Cian<sup>11</sup>, J.M. De Miranda<sup>1</sup>, L. De Paula<sup>2</sup>,

Figure: Collaboration author list.

# HEP Paper Characteristics (iii)

- <sup>18</sup> *Laboratori Nazionali dell'INFN di Frascati, Frascati, Italy*
- <sup>19</sup> *Sezione INFN di Genova, Genova, Italy*
- <sup>20</sup> *Sezione INFN di Milano Bicocca, Milano, Italy*
- <sup>21</sup> *Sezione INFN di Milano, Milano, Italy*
- <sup>22</sup> *Sezione INFN di Padova, Padova, Italy*
- <sup>23</sup> *Sezione INFN di Pisa, Pisa, Italy*
- <sup>24</sup> *Sezione INFN di Roma Tor Vergata, Roma, Italy*
- <sup>25</sup> *Sezione INFN di Roma La Sapienza, Roma, Italy*
- <sup>26</sup> *Henryk Niewodniczanski Institute of Nuclear Physics Polish Academy of Sciences, Kraków, Poland*
- <sup>27</sup> *AGH - University of Science and Technology, Faculty of Physics and Applied Computer Science, Kraków, Poland*
- <sup>28</sup> *National Center for Nuclear Research (NCBJ), Warsaw, Poland*
- <sup>29</sup> *Horia Hulubei National Institute of Physics and Nuclear Engineering, Bucharest-Magurele, Romania*
- <sup>30</sup> *Petersburg Nuclear Physics Institute (PNPI), Gatchina, Russia*
- <sup>31</sup> *Institute of Theoretical and Experimental Physics (ITEP), Moscow, Russia*
- <sup>32</sup> *Institute of Nuclear Physics, Moscow State University (SINP MSU), Moscow, Russia*
- <sup>33</sup> *Institute for Nuclear Research of the Russian Academy of Sciences (INR RAN), Moscow, Russia*
- <sup>34</sup> *Budker Institute of Nuclear Physics (SB RAS) and Novosibirsk State University, Novosibirsk, Russia*
- <sup>35</sup> *Institute for High Energy Physics (IHEP), Protvino, Russia*
- <sup>36</sup> *Universitat de Barcelona, Barcelona, Spain*
- <sup>37</sup> *Universidad de Santiago de Compostela, Santiago de Compostela, Spain*
- <sup>38</sup> *European Organization for Nuclear Research (CERN), Geneva, Switzerland*
- <sup>39</sup> *Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*
- <sup>40</sup> *Physik-Institut, Universität Zürich, Zürich, Switzerland*
- <sup>41</sup> *Nikhef National Institute for Subatomic Physics, Amsterdam, The Netherlands*
- <sup>42</sup> *Nikhef National Institute for Subatomic Physics and VU University Amsterdam, Amsterdam, The Netherlands*
- <sup>43</sup> *NSC Kharkiv Institute of Physics and Technology (NSC KIPT), Kharkiv, Ukraine*
- <sup>44</sup> *Institute for Nuclear Research of the National Academy of Sciences (KINR), Kyiv, Ukraine*

Figure: Collaboration affiliation list.

# HEP Paper Characteristics (iv)

encode different attribute dimensions of an input data space. A good glyph design can enable users to conduct visual search more efficiently during interactive visualization, and facilitate effective learning, memorizing and using the visual encoding scheme. A less effective visual design may suffer from various shortcomings such as being perceptually confusing, semantically ambiguous, difficult to learn and remember, or unable to accommodate low-resolution display devices.

- 
- *Eamonn Maguire is with Oxford e-Research Centre and Department of Computer Science, University of Oxford, UK. E-mail: eamonn.maguire@st-annes.ox.ac.uk.*
  - *Philippe Rocca-Serra, Susanna-Assunta Sansone and Min Chen are with Oxford e-Research Centre, University of Oxford, UK. E-mail: {philippe.rocca-serra,susanna-assunta.sansone,min.chen}@oerc.ox.ac.uk.*
  - *Jim Davies is with Department of Computer Science, University of Oxford, UK. E-mail: jim.davies@cs.ox.ac.uk.*

*Manuscript received 31 March 2012; accepted 1 August 2012; posted online 14 October 2012; mailed on 5 October 2012.*

*For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.*

**Figure:** Discontinuous header data.

# Training Data

Two models addressed:

- ▶ *header* model, which classifies *word* tokens as <title>, <author>, <abstract>, etc.
- ▶ *segmentation* model, which classifies *line* tokens as <header>, <body>, <references>, etc.
- ▶ Custom HEP training sets collected for each
- ▶ HEP dataset combined with existing CORA datasets during experimentation

Model	HEP	CORA
Header	157 papers	<b>2506 papers</b>
Segmentation	<b>169 papers</b>	125 papers

**Table:** Number of training instances for each model from each dataset.

# Feature Engineering

- ▶ Experiments run for different features designed to enhance the models *header* and *segmentation*
- ▶ 66 experiments run testing combinations of features, model and CV configuration.

Method	Model
Baseline	both
Block Size	header
Character Classes	segmentation
Dictionaries	header
Levenshtein Distance	segmentation
Regularisation	header
Token Extensions	segmentation

**Table:** Feature engineering experiments



## Dictionary Features (*header*)

Dictionaries were derived from the INSPIRE-HEP corpus:

- ▶ affiliations
- ▶ authors
- ▶ collaborations
- ▶ journals
- ▶ titles
- ▶ stop words\*

Dictionary-based features were then modelled as,

$$f_{\text{dict}_i}(x_t) = \mathbb{1}_{\{x_t \in \text{dict}_i\}},$$

for each dictionary,  $\text{dict}_i$ .

## Character Class Features (*segmentation*)

Feature functions defined to be,

$$f_{\text{class}_i}(x_t) = \frac{1}{|x_t|} \sum_{n=1}^{|x_t|} \mathbb{1}_{\{x_{ti} \in \text{class}_i\}},$$

for each character class,  $\text{class}_i$ , where  $x_t$  is a token (line), and  $x_{ti}$  is the  $i$ th character in the line.

Class	Regex
Spacing	$r'[\backslash s]'$
Lower case	$r'[a-z]'$
Upper case	$r'[A-Z]'$
Numeric	$r'[\backslash d]'$
Punctuation	$r'[.,?;:]'$
Special character	$r'^{[\backslash sa-zA-Z d.,?;:]}'$

**Table:** Character classes used as features.

## Levenshtein Distance Features (*segmentation*)

Define similarity function,

$$\text{similarity}(a, b) = 1 - \frac{\text{lev}_{a,b}(|a|, |b|)}{\max(|a|, |b|)}.$$

Then feature function,

$$f_{lev}(x_t) = \begin{cases} 0 & \text{if } 0 \leq \text{similarity}(x_t, x_{t-1}) \leq T_1 \\ 1 & \text{if } T_1 \leq \text{similarity}(x_t, x_{t-1}) \leq T_2 \\ \vdots & \vdots \\ N-1 & \text{if } T_{N-1} \leq \text{similarity}(x_t, x_{t-1}) \leq 1 \end{cases}$$

where  $T_1, T_2, \dots, T_{N-1}$  are thresholds selected to create the  $N$  categories.

# Outline

Introduction

Automatic Metadata Extraction

Data and Features

Key Results

Conclusions

# Header Model (Best Features)

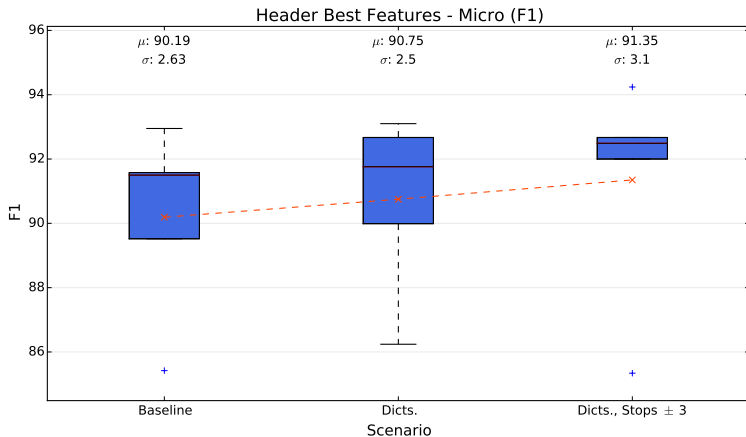


Figure: Best features for header model.

# Segmentation Model (Header Field)

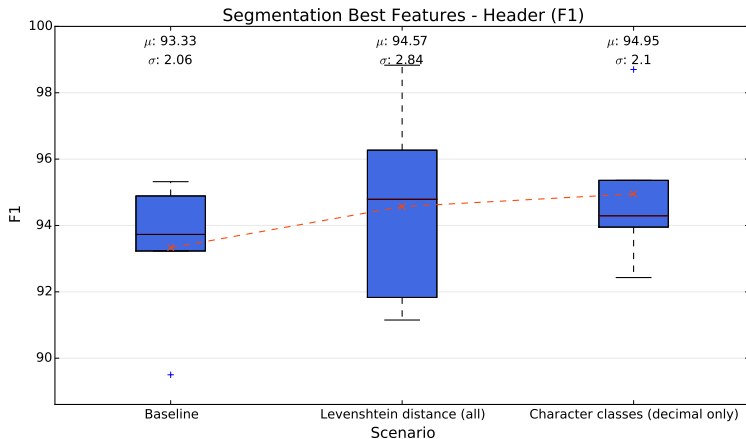


Figure: Best features for segmentation model <header> field.

# Segmentation Model (References Field)

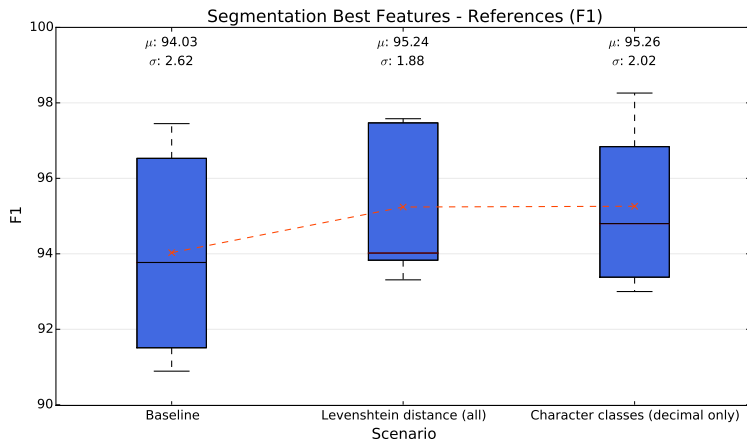


Figure: Best features for segmentation model <references> field.

# Segmentation Model (Baseline Confusion Matrix)

Confusion matrix - Segmentation (Baseline, HEP)

acknowledgement	544	5	258			10	1		
annex	13	1805	9078			67	1	3	46
body	142	2601	185466		162	303	45	41	590
cover				259	5	26	1		
footnote	6	8	495		1171	47	42	31	18
header		3	906	43	32	12211	11	6	36
headnote		18	372		56	76	1380	52	35
page	2	10	136		18	12	17	2257	12
references	115		392		11	124	18	5	10871
	acknowledgement	annex	body	cover	footnote	header	headnote	page	references

Figure: Baseline confusion matrix for segmentation model.



# Segmentation Model (Character Class Confusion Matrix)

Confusion matrix - Segmentation (Character Classes, HEP)

acknowledgement	616		185			17			
annex	16	1573	9293			3	3	1	124
body	51	2555	185735		87	530	52	28	312
cover				259	5	24	3		
footnote		19	472		1205	39	58	8	17
header			311	1	41	12863	8	3	21
headnote		6	354		55	59	1428	57	30
page		5	116		5	10	22	2297	9
references			433		14	96	21	6	10966
	acknowledgement	annex	body	cover	footnote	header	headnote	page	references

Figure: Confusion matrix for segmentation model with character classes.

# Outline

Introduction

Automatic Metadata Extraction

Data and Features

Key Results

Conclusions

# Conclusions

- ▶ Qualitative difference between HEP and general papers demonstrated (through inspection, subsampling).
- ▶ Valuable new datasets produced.
- ▶ Performance-enhancing features found for both HEP and general cases.
- ▶ Successful features offered a dimensionality reduction: dictionaries (12% error reduction), character classes (24% and 21% on <header> and <references> classifications).



M. Lipinski, K. Yao, C. Breiting, J. Beel, and B. Gipp,  
“Evaluation of header metadata extraction approaches and  
tools for scientific pdf documents,” in *Proceedings of the 13th  
ACM/IEEE-CS joint conference on Digital libraries*,  
pp. 385–386, ACM, 2013.