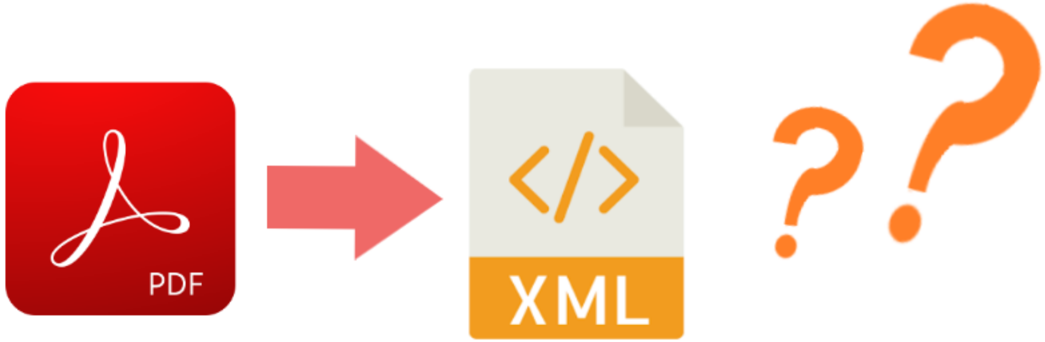
TECHNISCHE
UNIVERSITÄT
DARMSTADT

Index Extraction From Textbooks

KOM Lab Design Workshop – Group KM-2



Yujin Wang
guantao0815@gmail.com

Ruolin Huang
hrl96fz@foxmail.com

Fabian Rösch
fabianroesch@gmail.com

Tim Steuer

KOM – Multimedia Communications Lab
Technical University of Darmstadt
Prof. Dr.-Ing. Ralf Steinmetz (Director)
Dept. of Electrical Engineering and Information Technology
Dept. of Computer Science (adjunct Professor)
www.KOM.tu-darmstadt.de

19. Mai 2020

Our Project: Index Extraction From Textbooks

Motivation

- Index of a book can contain information about its contents that can be analyzed
- If Index can be extracted it can be used for keyword extraction or algorithmic paragraph headline creation

Problem

- Most Textbooks are PDF
- PDF is not well structured and thus can not easily be interpreted by a machine

Idea

- It’s difficult for a machine to read and contextualize the content of a PDF file
- XML is well structured with labels

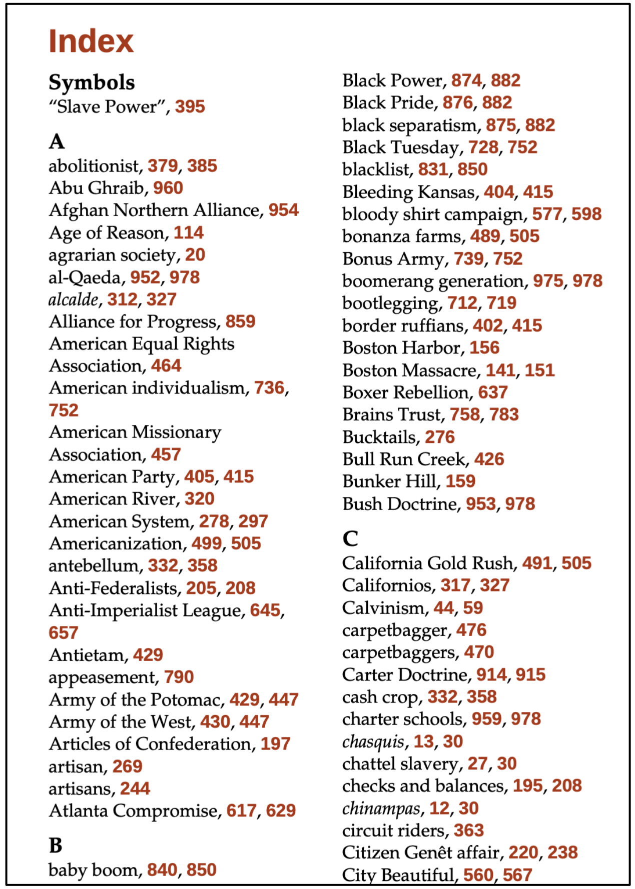
Solution

- Automate index extraction from PDF files to standardized XML files

Existing Related Projects and Services

Amazon Textract	• AWS cloud based service to extract data from scanned documents using machine learning • Made for scanned documents that do not contain actual text • Expensive and not open source
Apache Tika	• Java based toolkit to parse PDF Documents and it‘s metadata and extract into HTML format extraction • Powerful toolset
PDFBox	• Open Source Java Library • Powerful toolset • Well documented and integrated into Java

Implementation Idea: Intermediate Format



Index

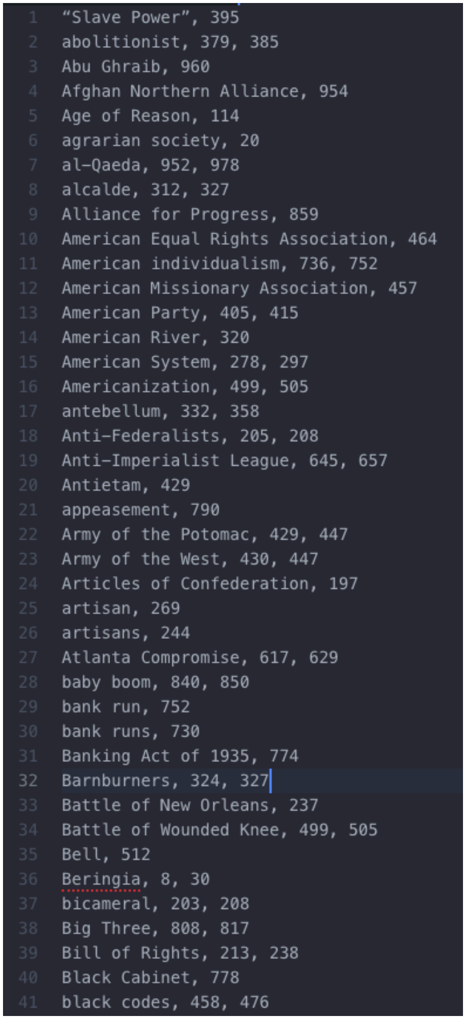
Symbols
"Slave Power", 395

A
abolitionist, 379, 385
Abu Ghraib, 960
Afghan Northern Alliance, 954
Age of Reason, 114
agrarian society, 20
al-Qaeda, 952, 978
alcalde, 312, 327
Alliance for Progress, 859
American Equal Rights Association, 464
American individualism, 736, 752
American Missionary Association, 457
American Party, 405, 415
American River, 320
American System, 278, 297
Americanization, 499, 505
antebellum, 332, 358
Anti-Federalists, 205, 208
Anti-Imperialist League, 645, 657
Antietam, 429
appeasement, 790
Army of the Potomac, 429, 447
Army of the West, 430, 447
Articles of Confederation, 197
artisan, 269
artisans, 244
Atlanta Compromise, 617, 629

B
baby boom, 840, 850
Black Power, 874, 882
Black Pride, 876, 882
black separatism, 875, 882
Black Tuesday, 728, 752
blacklist, 831, 850
Bleeding Kansas, 404, 415
bloody shirt campaign, 577, 598
bonanza farms, 489, 505
Bonus Army, 739, 752
boomerang generation, 975, 978
bootlegging, 712, 719
border ruffians, 402, 415
Boston Harbor, 156
Boston Massacre, 141, 151
Boxer Rebellion, 637
Brains Trust, 758, 783
Bucktails, 276
Bull Run Creek, 426
Bunker Hill, 159
Bush Doctrine, 953, 978

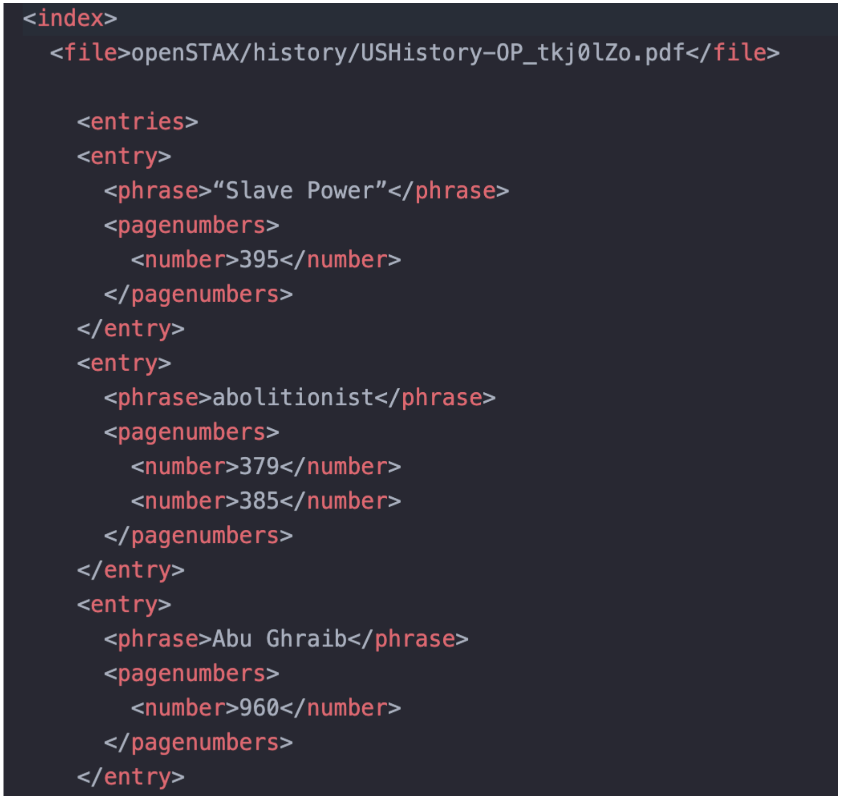
C
California Gold Rush, 491, 505
Californios, 317, 327
Calvinism, 44, 59
carpetbagger, 476
carpetbaggers, 470
Carter Doctrine, 914, 915
cash crop, 332, 358
charter schools, 959, 978
chasquis, 13, 30
chattel slavery, 27, 30
checks and balances, 195, 208
chinampas, 12, 30
circuit riders, 363
Citizen Genêt affair, 220, 238
City Beautiful, 560, 567

Step 1



```
"Slave Power", 395
abolitionist, 379, 385
Abu Ghraib, 960
Afghan Northern Alliance, 954
Age of Reason, 114
agrarian society, 20
al-Qaeda, 952, 978
alcalde, 312, 327
Alliance for Progress, 859
American Equal Rights Association, 464
American individualism, 736, 752
American Missionary Association, 457
American Party, 405, 415
American River, 320
American System, 278, 297
Americanization, 499, 505
antebellum, 332, 358
Anti-Federalists, 205, 208
Anti-Imperialist League, 645, 657
Antietam, 429
appeasement, 790
Army of the Potomac, 429, 447
Army of the West, 430, 447
Articles of Confederation, 197
artisan, 269
artisans, 244
Atlanta Compromise, 617, 629
baby boom, 840, 850
bank run, 752
bank runs, 730
Banking Act of 1935, 774
Barnburners, 324, 327
Battle of New Orleans, 237
Battle of Wounded Knee, 499, 505
Bell, 512
Beringia, 8, 30
bicameral, 203, 208
Big Three, 808, 817
Bill of Rights, 213, 238
Black Cabinet, 778
black codes, 458, 476
```

Step 2



```
<index>
  <file>openSTAX/history/USHistory-OP_tkj0lZo.pdf</file>

    <entries>
    <entry>
      <phrase>"Slave Power"</phrase>
      <pagenumbers>
        <number>395</number>
      </pagenumbers>
    </entry>
    <entry>
      <phrase>abolitionist</phrase>
      <pagenumbers>
        <number>379</number>
        <number>385</number>
      </pagenumbers>
    </entry>
    <entry>
      <phrase>Abu Ghraib</phrase>
      <pagenumbers>
        <number>960</number>
      </pagenumbers>
    </entry>
```

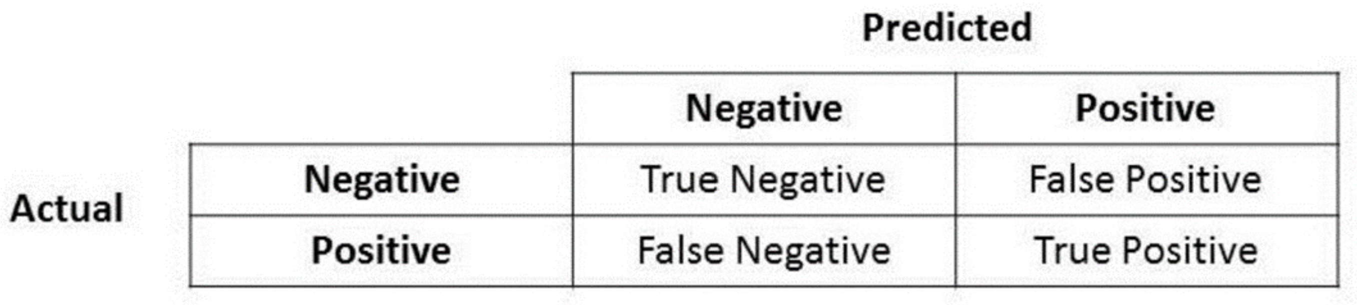
U. S. History [1]

Evaluation Method: Precision, Recall and F1 Metric

Precision = $$\frac{\text{Correctly Extracted Index Entries}}{\text{Correctly Extracted Index Entries} + \text{Falsely Extracted Index Entries}}$$

Recall = $$\frac{\text{Correctly Extracted Index Entries}}{\text{Correctly Extracted Index Entries} + \text{Falsely Considered as Other Text Part}}$$

F1 = $$2 * \frac{Precision * Recall}{Precision + Recall}$$



		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Accuracy, Precision, Recall or F1?[2]

Status Quo

Golden Standard

- Five manually created sample XML files for verification and testing

Programming language and library

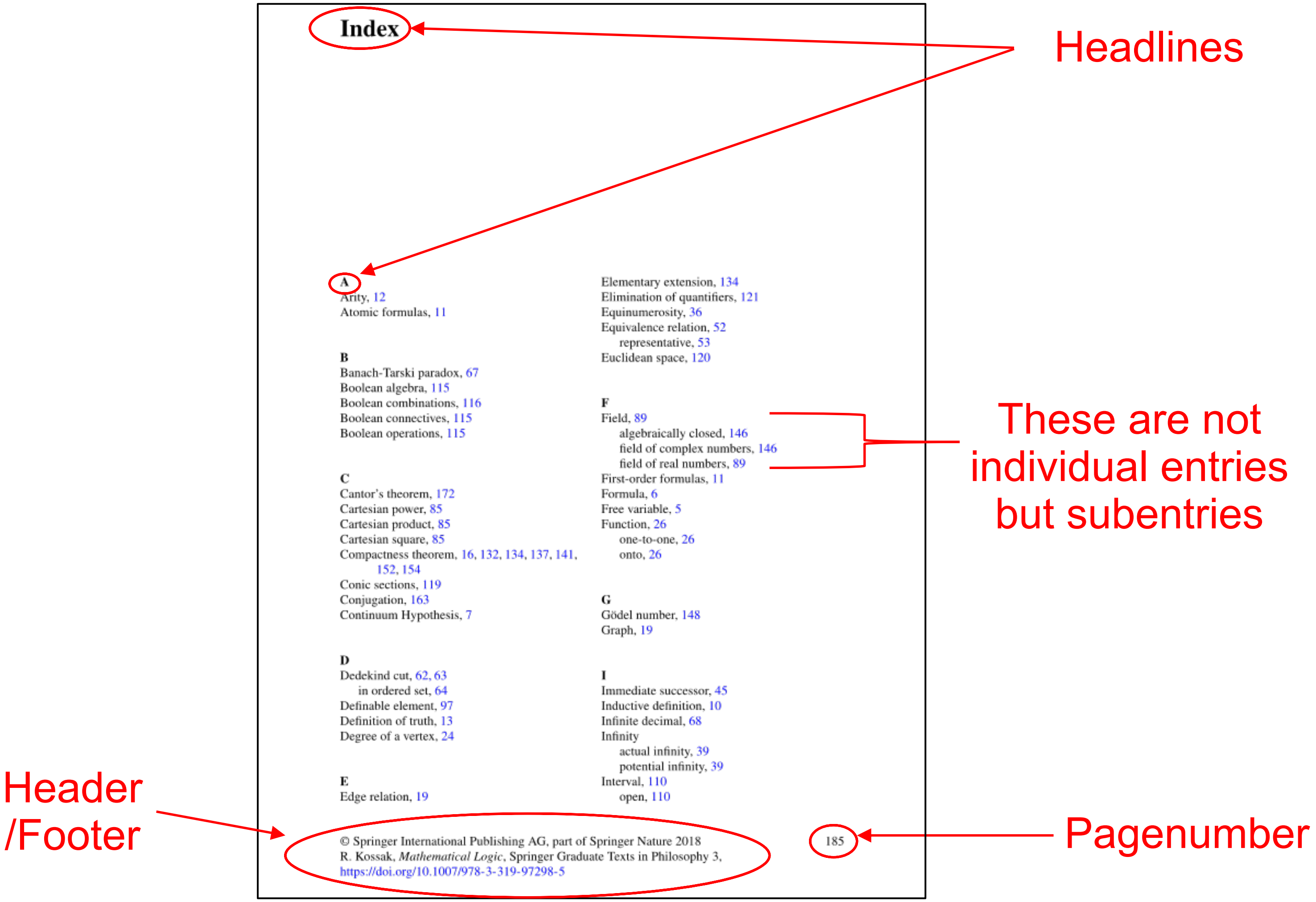
- Java
- PDFBox library

Raw text extraction

- Raw PDF to TXT extraction including all undesired content
- Basic filtering of undesired content and separating phrases from page numbers
- Performs well in one PDF but it still need to be optimized for general using

Main Obstacles Right Now

- Determining where the Index in any given PDF is
- Filtering and removal of non-relevant content (e.g. page number, header, etc)
- How to identify subphrases



Index

A

- Arity, 12
- Atomic formulas, 11

B

- Banach-Tarski paradox, 67
- Boolean algebra, 115
- Boolean combinations, 116
- Boolean connectives, 115
- Boolean operations, 115

C

- Cantor's theorem, 172
- Cartesian power, 85
- Cartesian product, 85
- Cartesian square, 85
- Compactness theorem, 16, 132, 134, 137, 141, 152, 154
- Conic sections, 119
- Conjugation, 163
- Continuum Hypothesis, 7

D

- Dedekind cut, 62, 63
  - in ordered set, 64
- Definable element, 97
- Definition of truth, 13
- Degree of a vertex, 24

E

- Edge relation, 19
- Elementary extension, 134
- Elimination of quantifiers, 121
- Equinumerosity, 36
- Equivalence relation, 52
  - representative, 53
- Euclidean space, 120

F

- Field, 89
  - algebraically closed, 146
  - field of complex numbers, 146
  - field of real numbers, 89
- First-order formulas, 11
- Formula, 6
- Free variable, 5
- Function, 26
  - one-to-one, 26
  - onto, 26

G

- Gödel number, 148
- Graph, 19

I

- Immediate successor, 45
- Inductive definition, 10
- Infinite decimal, 68
- Infinity
  - actual infinity, 39
  - potential infinity, 39
- Interval, 110
  - open, 110

© Springer International Publishing AG, part of Springer Nature 2018
R. Kossak, *Mathematical Logic*, Springer Graduate Texts in Philosophy 3,
https://doi.org/10.1007/978-3-319-97298-5

185

Headlines

These are not individual entries but subentries

Header /Footer

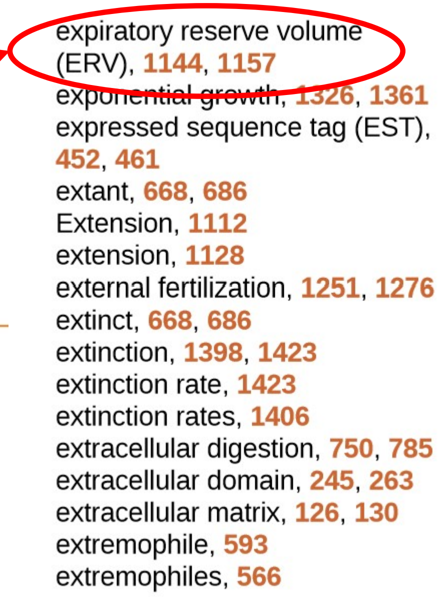
Pagenumber

Automata and Computability [3]

Further Problems: Special Cases

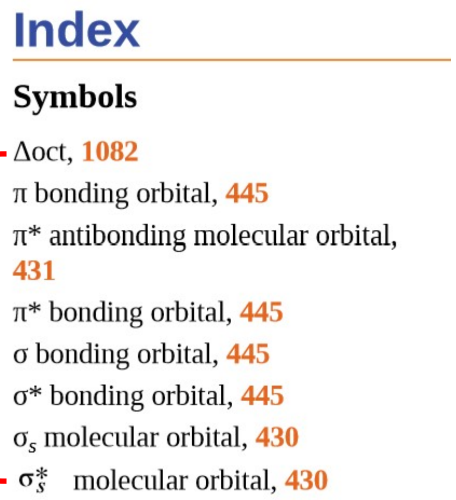
- Some special symbols
- The phrase is too long and split into two lines
- How to distinguish between the hyphen existing in the phrase itself and the hyphen due to line breaks

Line break due to long entrys



expiratory reserve volume (ERV), 1144, 1157
exponential growth, 1326, 1361
expressed sequence tag (EST), 452, 461
extant, 668, 686
Extension, 1112
extension, 1128
external fertilization, 1251, 1276
extinct, 668, 686
extinction, 1398, 1423
extinction rate, 1423
extinction rates, 1406
extracellular digestion, 750, 785
extracellular domain, 245, 263
extracellular matrix, 126, 130
extremophile, 593
extremophiles, 566

special symbols

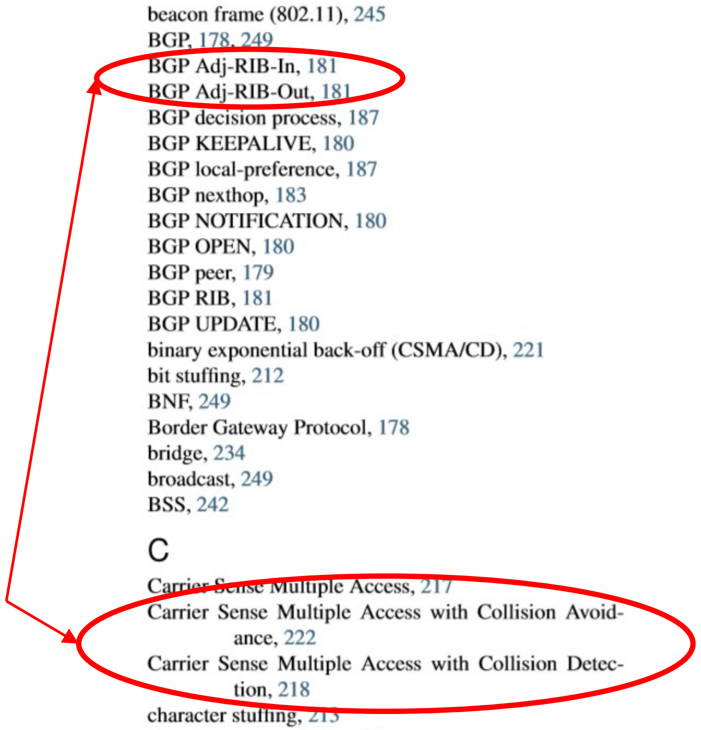


Index

Symbols

Δoct, 1082
π bonding orbital, 445
π* antibonding molecular orbital, 431
π* bonding orbital, 445
σ bonding orbital, 445
σ* bonding orbital, 445
σ_s molecular orbital, 430
σ_s^* molecular orbital, 430

hyphens due to line breaks



beacon frame (802.11), 245
BGP, 178, 249
BGP Adj-RIB-In, 181
BGP Adj-RIB-Out, 181
BGP decision process, 187
BGP KEEPALIVE, 180
BGP local-preference, 187
BGP nexthop, 183
BGP NOTIFICATION, 180
BGP OPEN, 180
BGP peer, 179
BGP RIB, 181
BGP UPDATE, 180
binary exponential back-off (CSMA/CD), 221
bit stuffing, 212
BNF, 249
Border Gateway Protocol, 178
bridge, 234
broadcast, 249
BSS, 242

C

Carrier Sense Multiple Access, 217
Carrier Sense Multiple Access with Collision Avoidance, 222
Carrier Sense Multiple Access with Collision Detection, 218
character stuffing, 213

Chemistry-OP [4]
Biology-OP[5]
Computer-Networking-Principles-Bonaventure-1-30-31-OTC1 [6]

Sources

[1] Corbett, P. S., Volker, J., Lund, J. M., Pfannestiel, T., Vickery, P. S., & Janssen, V. (2014). *U. S. History*. Amsterdam, Niederlande: Amsterdam University Press.

[2] Koo Ping Shung, (Mar 15, 2018) Accuracy, Precision, Recall or F1? https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

[3] Kozen, D. C. (2013). *Automata and Computability*. New York, Vereinigte Staaten: Springer Publishing.

[4] Flowers, P., Theopold, K., Langley, R., STEPHEN F., ROBINSON, W. R. (2015). Chemistry-OP. Rice University Press

[5] RYE, C., WISE, R., JURUKOVSKI, V., DESAIX, J., CHOI, J., AVISSAR, Y. (2013). Biology-OP. Rice University Press

[6] Olivier Bonaventure (October 30, 2011) *Computer-Networking-Principles-Bonaventure-1-30-31-OTC1.*