

Text Mining Wikipedia to extract historical facts

João Valente and João Gradim

Faculdade de Engenharia da Universidade do Porto,
Rua do Dr. Roberto Frias, s/n, Porto, Portugal

Abstract. The abstract should summarize the contents of the paper using at least 70 and at most 150 words. It will be set in 9-point font size and be inset 1.0 cm from the right and left margins. There will be two blank lines before and after the Abstract. ...

1 Natural Language Processing

1.1 Local extraction

```
...
(PP (IN in)
  (NP (NNP x)))

or

(PP (IN in)
  (NP (NNP x) (, ,) (NNP y)))
...
```

This kind of structure is present on most events parsed from Wikipedia and allows the retrieval of the physical location of the occurrence, mainly the country or a city in a determined country.

2 Main implementation problems

Wikipedia is an open knowledge base, relying mainly on user generated content. As such, it's difficult to ensure a proper and constant textual structure for information. Although the pages for the most recent centuries (aprox. 18th century) have a well defined and constant HTML structure that allows for reliable information retrieval, there are many years that don't follow this structure, leading to specific parsing cases.

This openness lead to another problem: an HTML structure not suited for easy parsing. A series of workarounds had to be implemented to successfully extract useful information.

References

- [1990] Santorini, B. : Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd Printing). (July 1990)
- [1995] Bies, A., Ferguson, M., Katz, K., MacIntyre, R. : Bracketing Guidelines for Treebank II Penn Treebank Project (June 1995)

Subject Index

- Absorption 327
- Absorption of radiation 289–292, 299, 300
- Actinides 244
- Aharonov-Bohm effect 142–146
- Angular momentum 101–112
 - algebraic treatment 391–396
- Angular momentum addition 185–193
- Angular momentum commutation relations 101
- Angular momentum quantization 9–10, 104–106
- Angular momentum states 107, 321, 391–396
- Antiquark 83
- α -rays 101–103
- Atomic theory 8–10, 219–249, 327
- Average value
 - (*see also* Expectation value) 15–16, 25, 34, 37, 357
- Baker-Hausdorff formula 23
- Balmer formula 8
- Balmer series 125
- Baryon 220, 224
- Basis 98
- Basis system 164, 376
- Bell inequality 379–381, 382
- Bessel functions 201, 313, 337
 - spherical 304–306, 309, 313–314, 322
- Bound state 73–74, 78–79, 116–118, 202, 267, 273, 306, 348, 351
- Boundary conditions 59, 70
- Bra 159
- Breit-Wigner formula 80, 84, 332
- Brillouin-Wigner perturbation theory 203
- Cathode rays 8
- Causality 357–359
- Center-of-mass frame 232, 274, 338
- Central potential 113–135, 303–314
- Centrifugal potential 115–116, 323
- Characteristic function 33
- Clebsch-Gordan coefficients 191–193
- Cold emission 88
- Combination principle, Ritz's 124
- Commutation relations 27, 44, 353, 391
- Commutator 21–22, 27, 44, 344
- Compatibility of measurements 99
- Complete orthonormal set 31, 40, 160, 360
- Complete orthonormal system, *see*
- Complete orthonormal set
- Complete set of observables, *see* Complete set of operators
- Eigenfunction 34, 46, 344–346
 - radial 321
 - calculation 322–324
- EPR argument 377–378
- Exchange term 228, 231, 237, 241, 268, 272
- f -sum rule 302
- Fermi energy 223
- H_2^+ molecule 26
- Half-life 65
- Holzwarth energies 68