# Using Metadata for Query Refinement and Recommendation

## Automatic topic detection Example 1 - "Global Warming"

Query = "global warming"
Search Engine = http://nsdl.org/

Search result: 4,735
Experiment collection (duplicate results removed): 2,137
NSDL Collection size in search results: 86
Keywords size in search results: 6,186
Number of extracted topic: 10
Three example topic list

**Word (Stem), Entity Distribution (Beta Matrix):**

| ***Earth Science Topic*** | | ***Government Policy Topic*** | | ***Climate Change Topic*** | |
|---|---|---|---|---|---|
| scienc | 0.03678588 | energi | 0.0377121 | research | 0.032611 |
| earth | 0.03404345 | carbon | 0.0282291 | inform | 0.025435 |
| student | 0.01522045 | develop | 0.0174741 | nation | 0.018136 |
| includ | 0.01497114 | technolog | 0.0136578 | issu | 0.014974 |
| project | 0.01409855 | emiss | 0.0135421 | state | 0.01473 |
| resourc | 0.01409855 | fuel | 0.0131952 | center | 0.014487 |
| educ | 0.01247803 | power | 0.0130795 | news | 0.013027 |
| activ | 0.01210406 | dioxid | 0.0125013 | public | 0.012906 |
| learn | 0.00873836 | product | 0.0109979 | polici | 0.011933 |
| weather | 0.00848905 | cycl | 0.010651 | intern | 0.011689 |
| onlin | 0.00799043 | high | 0.0098414 | provid | 0.011568 |
| topic | 0.00774112 | CARBON_DIOXIDE | 0.0089163 | institut | 0.01023 |
| Earth_space_science | 0.0074918 | nuclear | 0.0089163 | Global_warming | 0.009622 |
| plan | 0.00736715 | GREENHOUSE_EFFECT | 0.0075285 | present | 0.009257 |
| Climate | 0.00711784 | process | 0.0063721 | articl | 0.00877 |

**Keyword Distribution:**

| *Earth Science Topic* | | *Government Policy Topic* | | *Climate Change Topic* | |
|---|---|---|---|---|---|
| earth_space_science | 0.16215084 | policy_economy | 0.2684802 | climatic_changes | 0.161443 |
| energy | 0.13454888 | energy_planning | 0.2669252 | air_pollution | 0.142076 |
| physics | 0.13390858 | carbon_dioxide | 0.2039896 | global_warming | 0.130123 |
| climate | 0.12670773 | greenhouse_effect | 0.1782857 | science_earth_science | 0.129516 |
| technology | 0.12003977 | environmental_impacts | 0.1625156 | climate_change | 0.114684 |
| space_science | 0.11821817 | greenhouse_gases | 0.1591277 | climate | 0.114438 |
| education_(general) | 0.11688358 | environmental_sciences | 0.143155 | health | 0.103269 |
| natural_hazards | 0.10845419 | climatic_change | 0.1347704 | greenhouse_gases | 0.099786 |
| atmospheric_science | 0.10832897 | progress_report | 0.1288264 | hydrology | 0.097104 |
| ecology | 0.1066971 | air_pollution | 0.1261824 | environment | 0.096516 |
| science | 0.10581011 | energy | 0.1190213 | environmental_science | 0.09647 |
| biology | 0.10481302 | health | 0.1088301 | science | 0.095201 |
| science_earth_science | 0.10404284 | environment | 0.0938561 | chemistry | 0.094571 |
| earth_science | 0.10383607 | general | 0.0937297 | geology | 0.094481 |
| climatology | 0.103464 | life_sciences | 0.0926709 | atmospheric_science | 0.094183 |

***NSDL Collection Distribution:***

| *Earth Science Topic* | | *Government Policy Topic* | | *Climate Change Topic* | |
|---|---|---|---|---|---|
| National Science Teachers Association (NSTA) | 0.19214482 | Office of Scientific and Technical Information (OSTI) | 0.1873345 | Infomine: Scholarly Internet Resource Collections | 0.178953 |
| On the Cutting Edge: Workshops for Geoscience Faculty | 0.17101018 | Dspace at MIT | 0.1339377 | NSDL Expert Voices Blogosphere | 0.128294 |
| Compadre: Resources for Physics and Astronomy Education | 0.14753963 | Directory of Open Access Journals: Technology and Engineering | 0.1247221 | Internet Scout Project | 0.126464 |

## Automatic topic detection Example 2 - "Pollution"

Query = "pollution"
Search Engine = http://nsdl.org/

Search result: 40,953
Experiment collection (Sampling, duplicate results removed): 3,099
NSDL Collection size in search results: 64
Keywords size in search results: 6,303
Number of extracted topic: 10
Three example topic list

***Word (Stem), Entity Distribution (Beta Matrix):***

| ***Polluted Waste Topic*** | | ***Auto & Fuel Topic*** | | ***Ecology Topic*** | |
|---|---|---|---|---|---|
| wast | 0.04168163 | energi | 0.0475954 | water | 0.052379 |
| prevent | 0.03809032 | fuel | 0.0210135 | pollut | 0.032116 |
| program | 0.0265682 | CONSUMPTION | 0.0177289 | area | 0.02253 |
| assess | 0.02529628 | technolog | 0.0171178 | studi | 0.017854 |
| pollut | 0.02305171 | 32_ENERGY_CONSERVATION | 0.0158957 | river | 0.016452 |
| generat | 0.01564464 | reduc | 0.0155901 | soil | 0.01614 |
| manag | 0.01564464 | product | 0.0148263 | metal | 0.016062 |
| project | 0.01437272 | research | 0.014368 | sampl | 0.012477 |
| POLLUTION_ABATEMENT | 0.01377417 | impact | 0.0141388 | investig | 0.012087 |
| WASTE_MANAGEMENT | 0.01317562 | UTILIZATION | 0.0125348 | general | 0.012009 |
| depart | 0.01265188 | industri | 0.0114654 | heavi | 0.00936 |
| design | 0.01250224 | effici | 0.0107015 | collect | 0.008814 |
| minim | 0.01085623 | engin | 0.0097085 | plant | 0.008658 |
| activ | 0.01003322 | vehicl | 0.0092502 | pollution | 0.008113 |
| oper | 0.00973394 | emiss | 0.0091738 | concentr | 0.008035 |

Keyword Distribution:

| ***Polluted Waste Topic*** | | ***Auto & Fuel Topic*** | | ***Ecology Topic*** | |
|---|---|---|---|---|---|
| minimization | 0.264938 | automobiles | 0.2284392 | life_sciences | 0.247618 |
| recycling | 0.23423475 | advanced_propulsion_systems | 0.2177552 | general | 0.225366 |
| radioactive_waste_management | 0.23285864 | electricity | 0.2175499 | ecology | 0.166066 |
| radioactive_wastes_from_nuclear_facilities | 0.23011078 | greenhouse_gases | 0.2167958 | environment | 0.151791 |
| pollution_abatement | 0.20816396 | energy_efficiency | 0.2068996 | chemistry | 0.126643 |
| non-radioactive_wastes_from_nuclear_facilities | 0.20767741 | production | 0.2068005 | water_quality | 0.124101 |
| nuclear_fuels | 0.20643665 | biomass_fuels | 0.1966765 | water_pollution | 0.110012 |
| management_of_radioactive_wastes | 0.20594331 | petroleum | 0.1923731 | science | 0.104265 |
| implementation | 0.19812427 | petroleum | 0.1909391 | astronomy | 0.103035 |
| waste_management | 0.19142779 | carbon_dioxide | 0.1893784 | pollution | 0.101148 |
| hanford_reservation | 0.18973429 | energy_conservation | 0.1859763 | waste_water | 0.099401 |
| waste_water | 0.18421317 | utilization | 0.1839778 | water | 0.097775 |
| recommendations | 0.17882325 | natural_gas | 0.1834659 | basic_biological_sciences | 0.095579 |
| waste_processing | 0.17299953 | efficiency | 0.1822442 | mercury | 0.093462 |
| planning | 0.17085338 | biomass | 0.1789522 | toxicity | 0.09332 |

***NSDL Collection Distribution:***

| ***Polluted Waste Topic*** | | ***Auto & Fuel Topic*** | | ***Ecology Topic*** | |
|---|---|---|---|---|---|
| Office of Scientific and Technical Information (OSTI) | 0.11000971 | Dspace at MIT | 0.1138752 | Directory of Open Access Journals: Biology and Life Sciences | 0.207128 |
| Virginia Tech Electronic Thesis and Dissertation Collection (VT-ETD) | 0.10617365 | Amser: Applied Math and Science Education Repository | 0.1101955 | Springerlink Online Journals | 0.191345 |
| Electronic Environmental Resources Library | 0.10155764 | Office of Scientific and Technical Information (OSTI) | 0.1068361 | Directory of Open Access Journals: Agriculture and Food Sciences | 0.164288 |

**Automatic topic detection Example 3 - "Ontology"**

Query = "ontology"
Search Engine = http://nsdl.org/

Search result: 5,497
Experiment collection (Sampling, duplicate results removed): 4,239
NSDL Collection size in search results: 45
Keywords size in search results: 2,355
Number of extracted topic: 10
Three example topic list

***Word (Stem), Entity Distribution (Beta Matrix):***

| *System Topic* | | *Gene Science Topic* | | *Semantic Topic* | |
|---|---|---|---|---|---|
| system | 0.04078 | gene | 0.1515065 | ontolog | 0.062433833 |
| base | 0.033126 | express | 0.0541178 | inform | 0.062296344 |
| inform | 0.030079 | analysi | 0.0498155 | knowledg | 0.041260501 |
| semant | 0.026486 | logic | 0.0177438 | develop | 0.022699462 |
| research | 0.020941 | method | 0.0168312 | access | 0.01981219 |
| knowledg | 0.020941 | experi | 0.0130503 | technolog | 0.016099982 |
| approach | 0.018285 | identifi | 0.01292 | domain | 0.014450112 |
| comput | 0.018207 | librari | 0.0127896 | increas | 0.013900155 |
| develop | 0.013442 | univers | 0.0107036 | open | 0.012112796 |
| requir | 0.012896 | common | 0.0107036 | standard | 0.012112796 |
| applic | 0.012818 | similar | 0.0107036 | need | 0.011837818 |
| servic | 0.012739 | interpret | 0.0105733 | engin | 0.011562839 |
| technolog | 0.012349 | cluster | 0.0101821 | scientif | 0.01142535 |
| Computer_Science | 0.01149 | current | 0.009791 | construct | 0.010737904 |
| design | 0.01024 | profil | 0.0096606 | retriev | 0.010462926 |

**Keyword Distribution:**

| *System Topic* | | *Gene Science Topic* | | *Semantic Topic* | |
|---|---|---|---|---|---|
| special_purpose_application-based_systems | 0.27748278 | animal_genetics_genomics | 0.2748929 | internet | 0.206239 |
| production/logistics | 0.26100579 | microbial_genetics_genomics | 0.2551853 | including | 0.206239 |
| multimedia_information_systems | 0.26090437 | plant_genetics_&_genomics | 0.2464685 | theoretical_..._libraries_information | 0.191741 |
| computer-aided_engineering | 0.26044964 | biochemistry | 0.214204 | content_analysis | 0.188862 |
| design | 0.26044964 | plant_pathology | 0.2041908 | class | 0.188862 |
| computer_communication_networks | 0.25784183 | methods_online | 0.2013155 | index_languages | 0.18723 |
| geographical_information_systems/cartography | 0.25731158 | cell_biology | 0.188085 | processes_schemes | 0.18723 |
| manufacturing | 0.25559207 | bioinformatics | 0.186472 | data_metadata_structures | 0.176082 |
| machines | 0.25559207 | plant_sciences | 0.1779689 | knowledge_representation | 0.171189 |
| computer_systems..._networks | 0.25361987 | life_sciences | 0.1725519 | artificial_intelligence | 0.15207 |
| electronic_computer_engineering | 0.25341618 | biotechnology | 0.159085 | ontologies | 0.151068 |
| operating_systems | 0.24763776 | zoology | 0.1581738 | database | 0.145331 |
| management_computing_information_systems | 0.24678207 | anatomy | 0.1576158 | classification | 0.143417 |
| business_information_systems | 0.24060285 | gene_ontology | 0.139415 | metadata | 0.140044 |

***NSDL Collection Distribution:***

| *System Topic* | | *Gene Science Topic* | | *Semantic Topic* | |
|---|---|---|---|---|---|
| Springer | 0.10513363 | BioMed | 0.1691316 | DOAJ | 0.091184 |
| PubMed Central | 0.08945866 | DOAJ | 0.1270941 | BioMed | 0.08878 |
| DOAJ | 0.08873607 | PubMed Central | 0.1068956 | PubMed Central | 0.083872 |