Web corpora

Serge Sharoff

Centre for Translation Studies
University of Leeds

August 14, 2024

UNIVERSITY OF LEEDS

Traditional corpora vs Web

- Brown or BNC: large infrastructural projects
- Texts of different kinds available on the Web
- deWac, ukWac, itWac, ruWac. . .
  2 million Web pages, 2 billion words each
- deTenTen, enTenTen, itTenTen, ruTenTen. . .
  8 million Web pages, 10 billion words each
- Common Crawl: 3 billion pages in all languages

Specialised domains

BNC `arts,medical,natsci,socsci,techeng`

Domains not well represented:

24 texts, 1.4 mln words for medicine
15 texts, 0.6 mln words for linguistics
4 texts, 0.1 mln words for chemistry

Example of corpus collection

- Keywords from Wikipedia's Renewable Energy category

fossil fuel	化石燃料	R{ископаемое топливо}
power station	发电厂	электростанция
hydroelectricity	水力发电	гидроэнергетика
photovoltaics	太阳能光伏	фотоэлектричество

- Queries to Yahoo: “photovoltaics “power station”
-

	En	Ru	Zh(CN)	Zh(TR)
URLs:	5762	5991	674	870
Words (MW):	6.5	5.8	1.9	1.68

Assessing the composition by keywords

7467	renewable energy	5629	источник энергия ‘energy source’
4352	wind turbine	4550	окружающий среда ‘environment’
3973	fossil fuel	2754	электрический энергия ‘electricity’
3127	greenhouse gas	2710	солнечный батарея ‘solar cell’
3049	natural gas	2274	солнечный энергия ‘solar energy’
2539	wind farm	2106	природный газ ‘natural gas’
2320	solar energy	1994	тепловой энергия ‘thermal energy’
2265	energy efficiency	1870	возобновлять источник ‘renewable energy’
1994	carbon dioxide	1561	производство электроэнергия ‘electricity generation’
1920	solar cell	1508	возобновлять источник энергия ‘renewable energy sourc
1782	wind energy	1439	изменение климат ‘climate change’
1722	generate electricity	1401	парниковый газ ‘greenhouse gas’
1559	solar patch	1315	альтернативный источник ‘alternative source’
1533	electricity generation	1289	энергия ветер ‘wind energy’

Crawling

- Parallel and near-parallel websites: `https://www.wipo.int/wipo_magazine_digital/en/`
- Simple crawling: `wget -m --no-parent`
- Focused crawling: selection of relevant pages
- Seed urls → links from those pages → urls
- Focus on what is relevant:
  1. Language: `glotlid, langdetect`
  2. Keywords: on-topic keywords as extracted from seed urls
- `https://www.bbcgoodfood.com/recipes/spatchcock-barbecue-chicken`

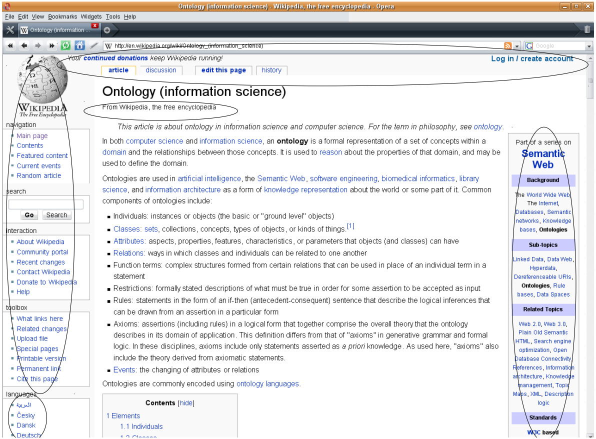
→ `https://www.bbc.com/zhongwen/simp`

→ `https://www.bbcgoodfood.com/review/best-air-fryers`

Parameters of cleaning

- Relevance: technical or not
- Duplicates and near-duplicates
- Saving to TXT from PDF, PPT, DOCX
  Encodings: not necessarily UTF8
- HTML boilerplate
- Trafilatura library in Python:
  `text=extract(fetch_url(url), output_format="xml")`

Webpage boilerplate


Ontology (information science) - Wikipedia, the free encyclopedia - Opera
File Edit View Bookmarks Widgets Tools Help
Ontology (information ...
http://en.wikipedia.org/wiki/Ontology_(information_science)
Google
Your continued donations keep Wikipedia running!
Log in / create account
article
discussion
edit this page
history
WIKIPEDIA
The Free Encyclopedia
navigation
Main page
Contents
Featured content
Current events
Random article
search
Go
Search
interaction
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Donate to Wikipedia
Help
toolbox
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link
Cite this page
languages
العربية
Česky
Dansk
Deutsch
Part of a series on
Semantic Web
Background
The World Wide Web, The Internet, Databases, Semantic networks, Knowledge bases, Ontologies
Sub-topics
Linked Data, Data Web, Hyperdata, Dereferenceable URIs, Ontologies, Rule bases, Data Spaces
Related Topics
Web 2.0, Web 3.0, Plain Old Semantic HTML, Search engine optimization, Open Database Connectivity, References, Information architecture, Knowledge management, Topic Maps, XML, Description logic
Standards
W3C based

Ontology (information science)

From Wikipedia, the free encyclopedia

This article is about ontology in information science and computer science. For the term in philosophy, see ontology.

In both computer science and information science, an **ontology** is a formal representation of a set of concepts within a domain and the relationships between those concepts. It is used to reason about the properties of that domain, and may be used to define the domain.

Ontologies are used in artificial intelligence, the Semantic Web, software engineering, biomedical informatics, library science, and information architecture as a form of knowledge representation about the world or some part of it. Common components of ontologies include:

- Individuals: instances or objects (the basic or "ground level" objects)
- Classes: sets, collections, concepts, types of objects, or kinds of things.[1]
- Attributes: aspects, properties, features, characteristics, or parameters that objects (and classes) can have
- Relations: ways in which classes and individuals can be related to one another
- Function terms: complex structures formed from certain relations that can be used in place of an individual term in a statement
- Restrictions: formally stated descriptions of what must be true in order for some assertion to be accepted as input
- Rules: statements in the form of an if-then (antecedent-consequent) sentence that describe the logical inferences that can be drawn from an assertion in a particular form
- Axioms: assertions (including rules) in a logical form that together comprise the overall theory that the ontology describes in its domain of application. This definition differs from that of "axioms" in generative grammar and formal logic. In these disciplines, axioms include only statements asserted as *a priori* knowledge. As used here, "axioms" also include the theory derived from axiomatic statements.
- Events: the changing of attributes or relations

Ontologies are commonly encoded using ontology languages.

Contents [hide]
1 Elements
1.1 Individuals
1.2 Classes

Basic points for preparing your corpora

- Web corpora are easy to create
- Comparable queries usually lead to comparable corpora
- Use specialised keywords to control their domains
- Clean your data: Trafilatura library