Chapter 1

The weScience Corpus

1.1 The Selection of Text

The weScience Corpus consists of Wikipedia articles. The intention has been that the articles should be related to linguistics, computational linguistics and natural language processing. We are in other words focusing on a limited domain rather than choosing random articles from the entire Wikipedia collection.

The corpus was collected from a downloaded collection of all available Wikipedia articles. These collections are regularly updated, and can be accessed at `http://download.wikimedia.org/enwiki/latest/`, the actual file is called `enwiki-latest-pages-articles.xml.bz2` and dates back to 2008-Mar-15 06:23:22.

From this collection, I set up an offline Wikipedia reader on my local computer. As Wikipedia articles are continuously changing, it was important to have this collection offline in order to have a data set that does not change overnight. The offline reader allows the user to browse and search for Wikipedia articles from one specific Wikipedia collection on a local Apache web server. A simple introduction on how to build an offline Wikipedia reader can be found on `http://users.softlab.ece.ntua.gr/~ttsiod/buildWikipediaOffline.html`

1.1.1 Link Collection

The corpus should consist of articles related to NLP and computational linguistics collected from Wikipedia. The total number of lines should be around 10 000. Choosing the articles myself would be less desirable than having an automatic algorithm that picked them. A handpicked corpus could arouse speculations regarding whether certain articles are chosen or left out due certain features with the articles that finally would boost the performance for the application that will make use of the corpus.

I decided to use a very simple way of clustering topics related to NLP based on some seed articles which all relate to the domain. This is a common staring point for more advanced unsupervised learning algorithms. I picked 11 seed articles, and harvested all hyperlinks from these articles to other Wikipedia articles. The seed articles are:

- Computational linguistics
- Context-free grammar
- Grammar
- Linguistics
- Machine learning
- Machine translation
- Natural language processing
- Parsing
- Part-of-speech tagging
- Statistical classification
- Syntax

I used `Linkchecker`[1] to collect hyperlinks to other Wikipedia articles from the seed articles. `Linkchecker` is a program which is used to check the validity of all links from one or more domains. For a given URL, it lists all links on the page. In a Wikipedia article, the vast majority of the links in an article refer to other Wikipedia articles.

I used all the seed articles as input for `Linkchecker`. The total number of (unique)links to other Wikipedia articles was 1158. I assumed that a link to an article that occurred only once was less likely to be relevant than a link that occurred twice or more. I therefore removed the links that only occurred once, and ended up with 153 articles. I based the corpus on these articles, without making any adjustments or modifications to the list that came out of this process. The entire list of articles can be inspected in Appendix A.

The list of articles were created using one single bash one-liner. This requires that a Wikipedia reader runs locally (in this case on 127.0.0.1:8000) and that `linkchecker` is installed. The python script `linkclean.py` removes a link if it only occurs once, and the grep commands removed external links, links to sub headlines and to templates. The actually command I used was:

```
linkchecker -oxml -r 1 -a -v -t 1
http://127.0.0.1:8000/article/Machine%20learning/
http://127.0.0.1:8000/article/Syntax/
http://127.0.0.1:8000/article/Natural%20language%20processing/
http://127.0.0.1:8000/article/Computational%20linguistics/
```

[1] http://linkchecker.sourceforge.net/

```
http://127.0.0.1:8000/article/Machine%20translation/
http://127.0.0.1:8000/article/Part-of-speech_tagging/
http://127.0.0.1:8000/article/Grammar/
http://127.0.0.1:8000/article/Parsing/
http://127.0.0.1:8000/article/Statistical%20classification/
http://127.0.0.1:8000/article/Context-free%20grammar/
http://127.0.0.1:8000/article/Linguistics/  | egrep "realurl" |
grep -v "#" | grep "127.0.0.1" | grep -v "Template" |
grep -v "searchbar" | sort |
python ~/workspace/linkclean/src/linkclean.py |
sort | uniq > allLinks.txt
```

When these articles were stripped for unwanted markup and chunks, the total number of lines were 12 938. The outcome of this process is the corpus that I have named the weScience Corpus. The process of cleansing the articles for unwanted markup will be discussed in Section 1.2

All the articles collected by this one-liner were stored in separate files. Some of the links refer to the same article (e.g. when the articles `Head-driven phrase structure grammar` and `Head-Driven Phrase Structure Grammar` both occur in the list, and the URL `http://127.0.0.1:8000/article/Head-Driven_Phrase_Structure_Grammar/` is redirected to `http://127.0.0.1:8000/article/Head-driven_phrase_structure_grammar/`). The script that stores the articles does however make sure that each article is stored only once. The articles are stored in the Wikipedia markup syntax, rather than HTML. This is done because some of the markup metadata may be of interest.

1.1.2 Alternative approach

The corpus could have been created with less seed articles and more clustering. One could argue that letting an automatic process be more in control of the selection process would have been a more balanced way of creating the corpus. I tried an alternative approach where I harvested all hyperlinks from the articles Natural Language Processing and Computational Linguistics. The total number of unique references to other articles were 178.[2]. I took each of the links that occurred twice in the two seed articles, and ended up with 15 articles. These articles were:

- Artificial intelligence
- Association for Computational Linguistics
- Computational linguistics
- Computer-assisted reviewing

[2]A number of these articles referred to Wikipedia articles in another language, thus the actual number of relevant articles were 129

- Corpus linguistics
- Grammar
- Information retrieval
- Lexical markup framework
- Machine learning
- Machine translation
- Natural language
- Natural language processing
- Semantics
- Speech recognition
- Technolangue/Easy

I then used the same approach for these articles as I did for the seed articles in 1.1.1. After having harvested all hyperlinks from these articles, and removed the links that only occurred once, 215 links remained. Most of these articles are on topic, but there are a few articles which clearly does not belong there (eg. an articles on Poker and an article about the year 2007). The list of links can be inspected in Appendix A.2.

After removing unwanted markup and chunks, the total number of lines for all the articles in this alternative corpus were 21672. This would have been too large for the corpus that I wanted to collect, but one could easily have sliced it in two and used half of it. 68 out the 153 articles in the WeScience corpus did also occur in the alternative corpus. The main reason for why I wanted to stick to the corpus generated from a higher number of seed articles is the quality. My subjective opinion is that the WeScience corpus is more accurate and on-topic, compared to the corpus gathered from only two seed articles

1.2 Wikipedia markup syntax

Wikipedia articles are edited in the wiki markup syntax system. When a Wikipedia article is accessed, the articles source syntax is converted into HTML, and displayed in a web browser. There are guidelines and conventions for how to edit or add content to Wikipedia, the intentions are to keep the architecture and design as consistent as possible. Here are some examples on Wikipedia markup, a more thorough manual can be found on `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style`

Linking conventions

(1) San Francisco also has public transportation.

a. `San Francisco also has [[public transport]]ation.`

b. `San Francisco also has [[public transport| public transportation]].`

In (1-a) the whole phrase “public transportation” is transformed to a hyperlink in the browser. In (1-b) the ’|’ separates the actual reference to another Wikipedia page (public transport) from the hyperlink that is shown in the browser (public transportation). When the page is converted to HTML, the references are converted to proper URLs, the actual URL in this case could be `http://en.wikipedia.org/wiki/Public_transport`.

Headings and sectioning

A top level heading is surrounded by two equal signs, and subheadings uses three equal signs, and so on.

(2) ==Top heading==

(3) ===Subheading==

It became however apparent that these conventions were not always followed, and both larger and minor deviations from the convention occurred frequently. One such example could be that the section ordering is wrong with respect to the Wikipedia conventions (e.g. that order in which the *References* and *Sources* section appear may vary from one article to anther). Such inconsistency is inevitable as there are millions of users who edit the articles, without any editors who read through the text before it is published. This poses a problem when trying to automatically transform the articles to a format which can serve as a good input for the Pet Parser (see Section **??**).

1.2.1 Desired format

The collected articles are fed into the Pet Parser. This parser assumes that each line contains exactly one sentence. Each Wikipedia source article should therefore before this stage be stripped for markup and parts of the article that do not contain relevant text, this may e.g. be a list of references or a snippet of source code used to show examples of a certain programming language.

Each sentences starts with an identifier. This is a eight-number digit where the first four refer to the article number, and the last four refer to the line/sentence number. Sentence number 16 in article 61 has thus the following identifier:

(4) [00610016] |This process is often called ”text normalization”, ”preprocessing”, or ”[[tokenization]]”.

Be aware that the first article and first sentence are enumerated 0000, thus sentence (4) is the actually the 17th sentence in the 62th article.

1.2.2 Cleansing of the articles

The articles are cleansed using the python script `Corpus Clean` which removes unwanted parts from the source text. This is entirely based on a large number of regular expressions. These python scripts remove amongst others content in curly brackets, images and HTML code, and entire sections like *See also*, *References*, *Bibliography* etc. Table 1.1 shows an example of only a few of the regular expressions that were used to strip the articles for unwanted content, markup and metadata.

```
regcurly1 = re.compile(r'{{?[^}^{]*?}?}',
re.MULTILINE | re.DOTALL)

regref = re.compile(r'<ref.*?/ref>',
re.MULTILINE | re.DOTALL)

regnotes = re.compile(r'==+\s?Notes\s?.*',
re.MULTILINE | re.DOTALL)

regreferences = re.compile(r'==+\s?References.*',
re.MULTILINE | re.DOTALL)

def regCleanFile(self,input):

        input = regex.regcurly1.sub('',input)
        input = regex.regref.sub('',input)
        (...)
        return input

def removeEnd(self,input):
        input = regex.regsealso.sub('',input)
        input = regex.regreferences.sub('',input)
        (...)
        return input
```

Table 1.1: Examples from the python script `Corpus Clean` used to remove unwanted content from the Wikipedia articles.

Note that the regular expressions `regnote` and `regreferences` remove the rest of the entire article. When either of these two sections appear in a

Wikipedia article that follows the conventions, we know that the rest of the article would contain mainly lists, references and other elements which will not serve as a good input sentences for our parser and it is therefore removed. The regular expressions `regcurley1` and `regref` will match anything that appears inside either curly brackets or <ref>...</ref>, and remove the entire content within this.

To convert the cleansed output to a one-sentence-per-line structure, all linebreaks were removed from the original source text. I than ran the sentence boundary detector `tokenizer v1.0` (Nagel, 2007), and used this to format the text according to the desired structure. After these stages, each separate article is written to a new file without any manual editing.

Sentence splitting

I tested a few different sentence boundary (SB) detectors before deciding which to use. A natural choice would have been to use the NTLK (Natural Language Toolkit) open source python modules (nlt) which includes a SB detector. I did however not find its performance sufficiently good, and decided to use `tokenizer v1.0` instead. This SB detector is used in a pipeline between stage one and stage two of the python scripts that remove and convert the articles from wiki markup to the desired format.

I manually read through the 11 first articles to inspect the performance of the SB detector, a total number of 1219 lines – roughly 10 % of the corpus. Based on these first articles, it seems apparent that the sentence segmentizer preforms very well on normal text. In the articles, I could only find one mistake in a sentence that did not contain any abnormalities with respect to punctuation or formatting.

(5) Example of incorrect sentence segmentation:

```
[00030052] |The weighted [[harmonic mean]] of precision and
recall, the traditional F-measure or balanced F-score is:
[00030053] |:<math>F = 2 \cdot (\mathrm \cdot \mathrm) /
(\mathrm + \mathrm).\,</math>
```

This sentence was found in article.0003, which is about Information Retrieval. Of the 13 mistakes that were found in the first eleven articles, five is found in this article. This article contains many formulas, and the segmentizer often interpreted the formula as a new sentence.

The formula (within the math-brackets) is the predicate of this sentence, and should of not be interpreted as a new sentence in a grammatical sense. It is whoever understandable that such mistakes occur, and the SB detector can hardly be expected to do a flawless segmentation in such complex texts. Table 1.2 provides the number of errors in the sentence segmentation in the first eleven articles.

Article #	Title	# of Lines	# of errors
article.0000	Russian language	183	1
article.0001	Lexical category	41	0
article.0002	Part-of-speech tagging	52	0
article.0003	Information retrieval	129	5
article.0004	Computer vision	153	0
article.0005	Linguistic typology	41	0
article.0006	Sentence (linguistics)	32	1
article.0007	Generative grammar	48	1
article.0008	Machine learning	68	0
article.0009	Noam Chomsky	261	2
article.0010	Panini	121	3

Table 1.2: Number of sentence segmenting mistakes in the first eleven articles

Most of the mistakes were made when the SB detector was mislead by punctuation. Examples of such mistakes could be that square brackets split the rightmost word in a sentence, thus the sentence boundary was often not detected:

```
[00120077] | [..] and a few [[grammatical aspect|aspect]]s.
Verbs are [..]
```

I found only very few mistakes where it was not obvious that the SB detector was mislead by the punctuation:

(6) Example of incorrect sentence segmentation:

```
[00100006] |Nothing definite is known about Panini's life, not
even the century he lived in. Scholarly mainstream favours a
[[4th century BC]] ''floruit'', corresponding to [[Achaemenid]]
[...]
```

For the latter mistake, I have no clear intuition regarding what triggered the SD detector to make this mistake. The overall conclusion is nevertheless that the sentence segmentizer performs acceptable. Of the 1129 sentences, I found 13 mistakes, approximately 1 %.

Bibliography

Natural language toolkit. URL `http://nltk.org`.

Sebastian Nagel. Tokenizer v1.0, 2007.

Appendix A

List of links

A.1 The links that were used to make the We-Science Corpus

http://127.0.0.1:8000/article/Adjective/
http://127.0.0.1:8000/article/Adverb/
http://127.0.0.1:8000/article/Algebraic_syntax/
http://127.0.0.1:8000/article/Algorithm/
http://127.0.0.1:8000/article/Ambiguous_grammar/
http://127.0.0.1:8000/article/Anthropological_linguistics/
http://127.0.0.1:8000/article/Applied_linguistics/
http://127.0.0.1:8000/article/Arabic_language/
http://127.0.0.1:8000/article/Artificial_intelligence/
http://127.0.0.1:8000/article/Association_for_Computational_Linguistics/
http://127.0.0.1:8000/article/Attribute_grammar/
http://127.0.0.1:8000/article/Bayesian_network/
http://127.0.0.1:8000/article/Bayesian_statistics/
http://127.0.0.1:8000/article/Bioinformatics/
http://127.0.0.1:8000/article/Biolinguistics/
http://127.0.0.1:8000/article/Boosting/
http://127.0.0.1:8000/article/Bottom-up_parsing/
http://127.0.0.1:8000/article/Brown_Corpus/
http://127.0.0.1:8000/article/Brown_University/
http://127.0.0.1:8000/article/Chinese_language/
http://127.0.0.1:8000/article/Cognitive_science/
http://127.0.0.1:8000/article/Comparative_linguistics/
http://127.0.0.1:8000/article/Compiler/
http://127.0.0.1:8000/article/Computational_linguistics/
http://127.0.0.1:8000/article/Computer-assisted_reviewing/
http://127.0.0.1:8000/article/Computer-assisted_translation/
http://127.0.0.1:8000/article/Computer_science/

http://127.0.0.1:8000/article/Computer_vision/
http://127.0.0.1:8000/article/Conditional_probability/
http://127.0.0.1:8000/article/Context-free_grammar/
http://127.0.0.1:8000/article/Context-sensitive_grammar/
http://127.0.0.1:8000/article/Controlled_language/
http://127.0.0.1:8000/article/Controlled_natural_language/
http://127.0.0.1:8000/article/Corpus_linguistics/
http://127.0.0.1:8000/article/Culture/
http://127.0.0.1:8000/article/CYK_algorithm/
http://127.0.0.1:8000/article/Data_mining/
http://127.0.0.1:8000/article/Decision_tree/
http://127.0.0.1:8000/article/Dependency_grammar/
http://127.0.0.1:8000/article/Descriptive_linguistics/
http://127.0.0.1:8000/article/Developmental_linguistics/
http://127.0.0.1:8000/article/Dionysius_Thrax/
http://127.0.0.1:8000/article/Dynamic_programming/
http://127.0.0.1:8000/article/English_language/
http://127.0.0.1:8000/article/Eo%3AKomputila_lingvistiko/
http://127.0.0.1:8000/article/Evolutionary_linguistics/
http://127.0.0.1:8000/article/Ferdinand_de_Saussure/
http://127.0.0.1:8000/article/Formal_grammar/
http://127.0.0.1:8000/article/Franz_Bopp/
http://127.0.0.1:8000/article/Functional_grammar/
http://127.0.0.1:8000/article/Generalised_phrase_structure_grammar/
http://127.0.0.1:8000/article/Generative_grammar/
http://127.0.0.1:8000/article/Gerald_Gazdar/
http://127.0.0.1:8000/article/Google/
http://127.0.0.1:8000/article/Google_tools/
http://127.0.0.1:8000/article/Grammar/
http://127.0.0.1:8000/article/Grammaticality/
http://127.0.0.1:8000/article/Grammatical_number/
http://127.0.0.1:8000/article/Grammatical_voice/
http://127.0.0.1:8000/article/Handwriting_recognition/
http://127.0.0.1:8000/article/Head-driven_phrase_structure_grammar/
http://127.0.0.1:8000/article/Head-Driven_Phrase_Structure_Grammar/
http://127.0.0.1:8000/article/Hidden_Markov_model/
http://127.0.0.1:8000/article/Historical-comparative_linguistics/
http://127.0.0.1:8000/article/Historical_linguistics/
http://127.0.0.1:8000/article/Idiom/
http://127.0.0.1:8000/article/Inflection/
http://127.0.0.1:8000/article/Information_extraction/
http://127.0.0.1:8000/article/Information_retrieval/
http://127.0.0.1:8000/article/Interlingual_machine_translation/
http://127.0.0.1:8000/article/Interlingua/

http://127.0.0.1:8000/article/Iron_Age_India/
http://127.0.0.1:8000/article/Language_acquisition/
http://127.0.0.1:8000/article/Language/
http://127.0.0.1:8000/article/Latin/
http://127.0.0.1:8000/article/Leonard_Bloomfield/
http://127.0.0.1:8000/article/Lexical_analysis/
http://127.0.0.1:8000/article/Lexical_functional_grammar/
http://127.0.0.1:8000/article/Lexical_markup_framework/
http://127.0.0.1:8000/article/Lexicon/
http://127.0.0.1:8000/article/Linguistics/
http://127.0.0.1:8000/article/Linguistic_typology/
http://127.0.0.1:8000/article/LL_parser/
http://127.0.0.1:8000/article/Lojban/
http://127.0.0.1:8000/article/LR_parser/
http://127.0.0.1:8000/article/Machine_learning/
http://127.0.0.1:8000/article/Machine_translation/
http://127.0.0.1:8000/article/Meaning_(linguistic)/
http://127.0.0.1:8000/article/Morphology_(linguistics)/
http://127.0.0.1:8000/article/Named_entity_recognition/
http://127.0.0.1:8000/article/Natural_language_processing/
http://127.0.0.1:8000/article/Natural_Language_Processing/
http://127.0.0.1:8000/article/Natural_language/
http://127.0.0.1:8000/article/Nearest_neighbor_(pattern_recognition)
/
http://127.0.0.1:8000/article/Neurolinguistics/
http://127.0.0.1:8000/article/Noam_Chomsky/
http://127.0.0.1:8000/article/Noun/
http://127.0.0.1:8000/article/Optical_character_recognition/
http://127.0.0.1:8000/article/Optimality_Theory/
http://127.0.0.1:8000/article/Orthography/
http://127.0.0.1:8000/article/Parse_tree/
http://127.0.0.1:8000/article/Parsing/
http://127.0.0.1:8000/article/Part_of_speech/
http://127.0.0.1:8000/article/Part-of-speech_tagging/
http://127.0.0.1:8000/article/Parts_of_speech/
http://127.0.0.1:8000/article/Pattern_recognition/
http://127.0.0.1:8000/article/P%C4%81%E1%B9%87ini/
http://127.0.0.1:8000/article/Perceptron/
http://127.0.0.1:8000/article/Phonetics/
http://127.0.0.1:8000/article/Phonology/
http://127.0.0.1:8000/article/Phrase/
http://127.0.0.1:8000/article/Pragmatics/
http://127.0.0.1:8000/article/Programming_language/
http://127.0.0.1:8000/article/Psycholinguistics/

http://127.0.0.1:8000/article/Quadratic_classifier/
http://127.0.0.1:8000/article/Reference/
http://127.0.0.1:8000/article/Reinforcement_learning/
http://127.0.0.1:8000/article/Russian_language/
http://127.0.0.1:8000/article/Search_engines/
http://127.0.0.1:8000/article/Semantics/
http://127.0.0.1:8000/article/Sentence_(linguistics)/
http://127.0.0.1:8000/article/Shallow_parsing/
http://127.0.0.1:8000/article/Sign_language/
http://127.0.0.1:8000/article/Sociolinguistics/
http://127.0.0.1:8000/article/Software/
http://127.0.0.1:8000/article/Source_text/
http://127.0.0.1:8000/article/Speech_recognition/
http://127.0.0.1:8000/article/Speech_synthesis/
http://127.0.0.1:8000/article/Statistical_classification/
http://127.0.0.1:8000/article/Statistics/
http://127.0.0.1:8000/article/Stochastic_context-free_grammar/
http://127.0.0.1:8000/article/Support_vector_machine/
http://127.0.0.1:8000/article/Syntactic_ambiguity/
http://127.0.0.1:8000/article/Syntax/
http://127.0.0.1:8000/article/Synthetic_language/
http://127.0.0.1:8000/article/Systemic_functional_grammar/
http://127.0.0.1:8000/article/SYSTRAN/
http://127.0.0.1:8000/article/Target_language/
http://127.0.0.1:8000/article/Technolangue/Easy/
http://127.0.0.1:8000/article/Text_corpus/
http://127.0.0.1:8000/article/Theoretical_computer_science/
http://127.0.0.1:8000/article/Theoretical_linguistics/
http://127.0.0.1:8000/article/Top-down_parsing/
http://127.0.0.1:8000/article/Transformational_grammar/
http://127.0.0.1:8000/article/Translation/
http://127.0.0.1:8000/article/Turing_machine/
http://127.0.0.1:8000/article/Universal_grammar/
http://127.0.0.1:8000/article/Universal_translator/
http://127.0.0.1:8000/article/Unsupervised_learning/
http://127.0.0.1:8000/article/Variable-order_Bayesian_network/
http://127.0.0.1:8000/article/Verb/
http://127.0.0.1:8000/article/Viterbi_algorithm/
http://127.0.0.1:8000/article/Word_sense_disambiguation/

A.2 Alternative links

These links were assembled using a clustering approached described in Section 1.1.2.

http://127.0.0.1:8000/article/2007/
http://127.0.0.1:8000/article/Acoustic_Model/
http://127.0.0.1:8000/article/Agent/
http://127.0.0.1:8000/article/AI-complete/
http://127.0.0.1:8000/article/AI_Winter/
http://127.0.0.1:8000/article/Alan_Newell/
http://127.0.0.1:8000/article/Applied_information_economics/
http://127.0.0.1:8000/article/Arimaa/
http://127.0.0.1:8000/article/Artificial_brain/
http://127.0.0.1:8000/article/Artificial_intelligence_in_fiction/
http://127.0.0.1:8000/article/Artificial_intelligence/
http://127.0.0.1:8000/article/Association_for_Computational_Linguistics/
http://127.0.0.1:8000/article/As_We_May_Think/
http://127.0.0.1:8000/article/Automated_planning_and_scheduling/
http://127.0.0.1:8000/article/Brain/
http://127.0.0.1:8000/article/Broca%27s_area/
http://127.0.0.1:8000/article/Brown_Corpus/
http://127.0.0.1:8000/article/Carnegie_Mellon_University/
http://127.0.0.1:8000/article/Chatterbot/
http://127.0.0.1:8000/article/Chinese_language/
http://127.0.0.1:8000/article/C._J._van_Rijsbergen/
http://127.0.0.1:8000/article/CMU/
http://127.0.0.1:8000/article/Cognitive_psychology/
http://127.0.0.1:8000/article/Cognitive/
http://127.0.0.1:8000/article/Cognitive_science/
http://127.0.0.1:8000/article/Collostructional_analysis/
http://127.0.0.1:8000/article/Combinatorial_explosion/
http://127.0.0.1:8000/article/Commonsense_knowledge/
http://127.0.0.1:8000/article/Computational_intelligence/
http://127.0.0.1:8000/article/Computational_lexicology/
http://127.0.0.1:8000/article/Computational_Linguistics_(journal)
/
http://127.0.0.1:8000/article/Computational_linguistics/
http://127.0.0.1:8000/article/Computational_semantics/
http://127.0.0.1:8000/article/Computation/
http://127.0.0.1:8000/article/Computer-assisted_reviewing/
http://127.0.0.1:8000/article/Computer-assisted_translation/
http://127.0.0.1:8000/article/Computer_science/
http://127.0.0.1:8000/article/Computer_vision/
http://127.0.0.1:8000/article/Configuration_space/

http://127.0.0.1:8000/article/Connectionism/
http://127.0.0.1:8000/article/Consciousness/
http://127.0.0.1:8000/article/Constructed_language/
http://127.0.0.1:8000/article/Controlled_language/
http://127.0.0.1:8000/article/Controlled_natural_language/
http://127.0.0.1:8000/article/Control_theory/
http://127.0.0.1:8000/article/Corpus_linguistics/
http://127.0.0.1:8000/article/C_(programming_language)/
http://127.0.0.1:8000/article/Cybernetic/
http://127.0.0.1:8000/article/Cybernetics/
http://127.0.0.1:8000/article/Cyc/
http://127.0.0.1:8000/article/DARPA/
http://127.0.0.1:8000/article/Dartmouth_Conferences/
http://127.0.0.1:8000/article/Data_mining/
http://127.0.0.1:8000/article/Decision_analysis/
http://127.0.0.1:8000/article/Decision_network/
http://127.0.0.1:8000/article/Decision_theory/
http://127.0.0.1:8000/article/Dialect/
http://127.0.0.1:8000/article/Dictionary/
http://127.0.0.1:8000/article/Direct_Voice_Input/
http://127.0.0.1:8000/article/Economic/
http://127.0.0.1:8000/article/Economics/
http://127.0.0.1:8000/article/E-mail_spam/
http://127.0.0.1:8000/article/Embodied_cognitive_science/
http://127.0.0.1:8000/article/English_language/
http://127.0.0.1:8000/article/Eo%3AKomputila_lingvistiko/
http://127.0.0.1:8000/article/Esperanto/
http://127.0.0.1:8000/article/European_Chapter_of_the_Association_
for_Computational_Linguistics/
http://127.0.0.1:8000/article/Evolutionary_computation/
http://127.0.0.1:8000/article/Extended_Boolean_model/
http://127.0.0.1:8000/article/Facial_recognition_system/
http://127.0.0.1:8000/article/French_language/
http://127.0.0.1:8000/article/Fuzzy_system/
http://127.0.0.1:8000/article/Game_theory/
http://127.0.0.1:8000/article/Generative_grammar/
http://127.0.0.1:8000/article/Gerard_Salton/
http://127.0.0.1:8000/article/Ghost_in_the_Shell/
http://127.0.0.1:8000/article/Google/
http://127.0.0.1:8000/article/Google_tools/
http://127.0.0.1:8000/article/GPL/
http://127.0.0.1:8000/article/Grammar/
http://127.0.0.1:8000/article/HAL_9000/
http://127.0.0.1:8000/article/Hans_Peter_Luhn/

http://127.0.0.1:8000/article/Herbert_Simon/
http://127.0.0.1:8000/article/Heuristic/
http://127.0.0.1:8000/article/Heuristics/
http://127.0.0.1:8000/article/Human-computer_interaction/
http://127.0.0.1:8000/article/Hybrid_intelligent_system/
http://127.0.0.1:8000/article/Hypertext/
http://127.0.0.1:8000/article/Idiom/
http://127.0.0.1:8000/article/IEEE/
http://127.0.0.1:8000/article/Inflection/
http://127.0.0.1:8000/article/Information_extraction/
http://127.0.0.1:8000/article/Information/
http://127.0.0.1:8000/article/Information_retrieval_applications/
http://127.0.0.1:8000/article/Information_retrieval/
http://127.0.0.1:8000/article/Information_science/
http://127.0.0.1:8000/article/Intelligent_agent/
http://127.0.0.1:8000/article/Intelligent_agents/
http://127.0.0.1:8000/article/Interdisciplinary/
http://127.0.0.1:8000/article/Interlingual_machine_translation/
http://127.0.0.1:8000/article/Interlingua/
http://127.0.0.1:8000/article/ISO_639-3/
http://127.0.0.1:8000/article/ISO/TC37/
http://127.0.0.1:8000/article/John_McCarthy_(computer_scientist)
/
http://127.0.0.1:8000/article/Judea_Pearl/
http://127.0.0.1:8000/article/Karen_Sp%C3%A4rck_Jones/
http://127.0.0.1:8000/article/Knowledge_representation/
http://127.0.0.1:8000/article/Language_model/
http://127.0.0.1:8000/article/Language/
http://127.0.0.1:8000/article/Latent_semantic_indexing/
http://127.0.0.1:8000/article/Latino_Sine_Flexione/
http://127.0.0.1:8000/article/Learning/
http://127.0.0.1:8000/article/Lexical_markup_framework/
http://127.0.0.1:8000/article/Lexical_resource/
http://127.0.0.1:8000/article/Lexicon/
http://127.0.0.1:8000/article/Linguistics/
http://127.0.0.1:8000/article/Linguistic_typology/
http://127.0.0.1:8000/article/Logic_programming/
http://127.0.0.1:8000/article/Logic/
http://127.0.0.1:8000/article/Logistics/
http://127.0.0.1:8000/article/Lojban/
http://127.0.0.1:8000/article/LREC/
http://127.0.0.1:8000/article/Machine_learning/
http://127.0.0.1:8000/article/Machine_perception/
http://127.0.0.1:8000/article/Machine-readable_dictionary/

http://127.0.0.1:8000/article/Machine_translation/
http://127.0.0.1:8000/article/Markov_model/
http://127.0.0.1:8000/article/Marvin_Minsky/
http://127.0.0.1:8000/article/Mathematical_logic/
http://127.0.0.1:8000/article/Mathematics/
http://127.0.0.1:8000/article/Meaning_(linguistic)/
http://127.0.0.1:8000/article/Medical_diagnosis/
http://127.0.0.1:8000/article/Mind/
http://127.0.0.1:8000/article/MIT/
http://127.0.0.1:8000/article/Morphology_(linguistics)/
http://127.0.0.1:8000/article/Named_entity_recognition/
http://127.0.0.1:8000/article/Native_speaker/
http://127.0.0.1:8000/article/Natural_language_processing/
http://127.0.0.1:8000/article/Natural_Language_Processing/
http://127.0.0.1:8000/article/Natural_language/
http://127.0.0.1:8000/article/Natural_Language_Toolkit/
http://127.0.0.1:8000/article/Natural_selection/
http://127.0.0.1:8000/article/Neats_vs._scruffies/
http://127.0.0.1:8000/article/Neural_network/
http://127.0.0.1:8000/article/Neural_networks/
http://127.0.0.1:8000/article/Neurology/
http://127.0.0.1:8000/article/Noam_Chomsky/
http://127.0.0.1:8000/article/Nyaya/
http://127.0.0.1:8000/article/OpenDocument/
http://127.0.0.1:8000/article/Open-source/
http://127.0.0.1:8000/article/Operations_research/
http://127.0.0.1:8000/article/Optical_character_recognition/
http://127.0.0.1:8000/article/Optimization_(mathematics)/
http://127.0.0.1:8000/article/Paradigm_pattern/
http://127.0.0.1:8000/article/Parser/
http://127.0.0.1:8000/article/Parsing/
http://127.0.0.1:8000/article/Pattern_matching/
http://127.0.0.1:8000/article/Pattern/
http://127.0.0.1:8000/article/Pattern_recognition/
http://127.0.0.1:8000/article/Philosophy/
http://127.0.0.1:8000/article/Phonetics/
http://127.0.0.1:8000/article/Phonology/
http://127.0.0.1:8000/article/Poker/
http://127.0.0.1:8000/article/Pragmatics/
http://127.0.0.1:8000/article/Probability/
http://127.0.0.1:8000/article/Programming_language/
http://127.0.0.1:8000/article/Prolog/
http://127.0.0.1:8000/article/Psychology/
http://127.0.0.1:8000/article/Qualification_problem/

http://127.0.0.1:8000/article/R2D2/
http://127.0.0.1:8000/article/Robert_Kowalski/
http://127.0.0.1:8000/article/Robert_R._Korfhage/
http://127.0.0.1:8000/article/Robotic_mapping/
http://127.0.0.1:8000/article/Robotics/
http://127.0.0.1:8000/article/Robot/
http://127.0.0.1:8000/article/Russian_language/
http://127.0.0.1:8000/article/Science_fiction/
http://127.0.0.1:8000/article/Search_engine/
http://127.0.0.1:8000/article/Semantics/
http://127.0.0.1:8000/article/Semantic_Web/
http://127.0.0.1:8000/article/Semiotics/
http://127.0.0.1:8000/article/Sentence_(linguistics)/
http://127.0.0.1:8000/article/Sign_language/
http://127.0.0.1:8000/article/Situated/
http://127.0.0.1:8000/article/Software/
http://127.0.0.1:8000/article/Source_text/
http://127.0.0.1:8000/article/Speech_recognition/
http://127.0.0.1:8000/article/Speech_synthesis/
http://127.0.0.1:8000/article/Stanford/
http://127.0.0.1:8000/article/Statistics/
http://127.0.0.1:8000/article/Stochastic/
http://127.0.0.1:8000/article/Stratego/
http://127.0.0.1:8000/article/Strong_AI/
http://127.0.0.1:8000/article/SVO_language/
http://127.0.0.1:8000/article/Swarm_intelligence/
http://127.0.0.1:8000/article/Syntax/
http://127.0.0.1:8000/article/Synthetic_language/
http://127.0.0.1:8000/article/SYSTRAN/
http://127.0.0.1:8000/article/Target_language/
http://127.0.0.1:8000/article/Taxonomic_classification/
http://127.0.0.1:8000/article/Technolangue/Easy/
http://127.0.0.1:8000/article/Text_corpus/
http://127.0.0.1:8000/article/Text_Retrieval_Conference/
http://127.0.0.1:8000/article/Transformational_grammar/
http://127.0.0.1:8000/article/Translation_memory/
http://127.0.0.1:8000/article/Translation/
http://127.0.0.1:8000/article/Unified_Modeling_Language/
http://127.0.0.1:8000/article/Universal_grammar/
http://127.0.0.1:8000/article/Universal_translator/
http://127.0.0.1:8000/article/Vannevar_Bush/
http://127.0.0.1:8000/article/WordNet/
http://127.0.0.1:8000/article/Word_sense_disambiguation/
http://127.0.0.1:8000/article/World_Wide_Web/

```
http://127.0.0.1:8000/article/Writing_system/
http://127.0.0.1:8000/article/XML/
```