

INNOVATIF BUSINESS DIGITAL PLATFORMS

StripeCon Malta 2017



Which site search solution(s) do you provide?



The end of Google Site Search

On April 1, 2017, Google will discontinue sales of the Google Site Search. All new purchases and renewals must take place before this date. The product will be completely shut down by April 1, 2018.

Google Site Search

Google Site Search brings the same search technology that powers Google.com to your website, delivering relevant results with lightning speed.

Google Site Search vs free CSE

	Google Site Search	Free CSE
Search options		
Search the entire web	✗	✓
Image-only search	✓	✗
Look and feel		
Option to remove ads	✓	✗
Access to XML API for results	✓	✗
JSON API	Unlimited	Daily limit
Make money with AdSense	✓	✓
Option to remove Google branding	✓	✗
Administration		
Transfer ownership	✓	✗
Share query quotas with a business group	✓	✗

Free version (Google CSE)

- Searches the entire web
- Ads + google branding included
- Daily limit

Market research

Our market research: Silverstripe modules

- Search by page YES
- Search by data objects or DOAP or page blocks requires additional effort
- PDF + images text search not included (OCR)
- Doesn't understand Slovenian language

Market research

Other open source options

- Elastic: <https://www.elastic.co/>
- Apache SOLR

SOLR + Silverstripe

APACHE SOLR™ 7.0.1

Solr is the popular, blazing-fast, open source enterprise search platform built on Apache Lucene™.

Apache SOLR

Why SOLR?

```
<fieldType name="text_en" class="solr.TextField" positionIncrementGap="100">
  <analyzer type="index">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.StopFilterFactory"
      ignoreCase="true"
      words="lang/stopwords_en.txt"
      enablePositionIncrements="true"
    />
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.EnglishPossessiveFilterFactory"/>
    <filter class="solr.KeywordMarkerFilterFactory" protected="protwords.txt"/>
    <filter class="solr.PorterStemFilterFactory"/>
  </analyzer>
  <analyzer type="query">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="solr.SynonymFilterFactory"
      synonyms="synonyms.txt"
      ignoreCase="true"
      expand="true"
    />
    <filter class="solr.StopFilterFactory"
      ignoreCase="true"
      words="lang/stopwords_en.txt"
      enablePositionIncrements="true"
    />
  </analyzer>
</fieldType>
```



Why SOLR?

- Advanced Full-Text Search Capabilities
- Optimized for High Volume Traffic
- Standards Based Open Interfaces
 - XML, JSON and HTTP
- Flexible and Adaptable with easy configuration
- Slovenian lemmatisation support

Lemmatisation concept

Lemmatisation

From Wikipedia, the free encyclopedia

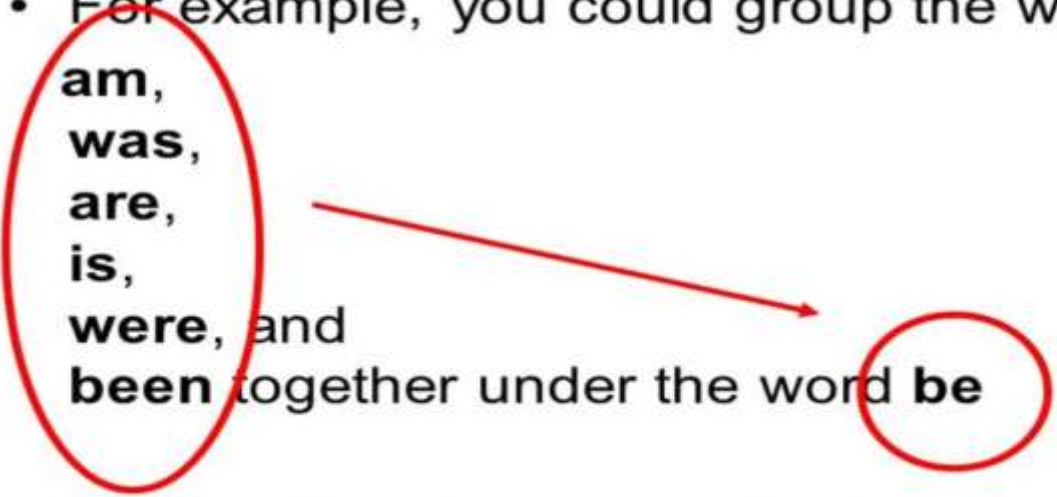
Lemmatisation (or **lemmatization**) in **linguistics** is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's **lemma**, or dictionary form.^[1]

In **computational linguistics**, lemmatisation is the algorithmic process of determining the **lemma** of a word based on its intended meaning. Unlike **stemming**, lemmatisation depends on correctly identifying the intended **part of speech** and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. As a result, developing efficient **lemmatisation** algorithms is an open area of research.^{[2][3]}

Lemma

- Lemmatising means
 - grouping morphological variants of words under a single headword
- For example, you could group the words

am,
was,
are,
is,
were, and
been together under the word **be**



LemmaGen

[Home](#)[Online Services](#)[Download](#)[C#](#)[Documentation](#)[Contact](#)[SEARCH](#)[MAIN MENU](#)[Home](#)[Online Services](#)[Download](#)[C#](#)[Documentation](#)[Contact](#)

Welcome to Lemmatisation Portal

LemmaGen project aims at providing standardized open source multilingual platform for **lemmatisation**. We started this work as a result of lack of high quality lemmatiser for **Slovene language**. Currently we have, not only the lemmatiser for Slovene, but also for **11 other European languages** and the system which is able to learn lemmatisation rules for new languages by providing it with existing wordform-lemma pair examples.

One of our **strong intentions** is to **increase the number of supported languages**. For that we hope we can count also on you, the users of these services. We invite you to **contact us** in case you have any data which could be used to build new lemmatisers for currently unsupported and also already supported languages.

LemmaGen name was originally abbreviation for "Lemmatiser Generator". However, it often stands for "Lemma Generator". These meanings also illustrate the two main components of this web site of which each user should be aware of: one main part deals with **usage of prebuilt lemmatisers** while the other one is concentrated on **definition and creation of new lemmatisers** (e.g. for new languages).

The main characteristics of the LemmaGen are:

- it is free - open source licence for all the code included in the project,
- multilingual support - currently 12 different languages included,
- lemmatisation does not rely on sentence structure of the text which is processed (can be applied on each word separately; useful for example for lemmatising search query words),
- wide variety of APIs which enable you to include LemmaGen into your own projects,
- all sources are downloadable,
- multiple implementations (C++, C++ .Net, Python, and, C#.Net),
- variety of platforms supported: downloadable content prebuilt for Windows & Linux, however, it can be recompiled for

<http://lemmatise.ijs.si/>

LemmaGen online test

Simple & Accurate Lemmatisation through Online Services

Instructions: Simply just copy/paste selected text into the big yellow text box below, select a language and press [Lemmatise].

Example text: part of wikipedia article about inflection:

Lexical items that do not respond to overt inflection are invariant or uninflected; for example, "must" is an invariant item: it never takes a suffix or changes form to signify a different grammatical category. Its category can only be determined by its context. Uninflected words do not need to be lemmatized in linguistic descriptions or in language computing. On the other hand, inflectional paradigms, or lists of inflected forms of typical words (such as sing, sang, sung, sings, singing, singer, singers, song, songs, songstress, songstresses in English) need to be analyzed according to criteria for uncovering the underlying lexical stem (here s+ng-); that is, the accompanying functional items (-i-, -a-, -u-, -s, -ing, -er, -o-, -stress, -es) and the functional categories of which they are markers need to be distinguished to adequately describe the language.

English

Lemmatise

Example text: part of wikipedia article about inflection:

Lexical ^{item} items ^{be} that do not respond to overt inflection ^{are} invariant or uninflected; for example, "must" ^{is} an invariant item: it never ^{take} takes a suffix or ^{change} changes form to signify a different grammatical category. Its category can only be ^{determine} determined by its context. Uninflected ^{word} words do not need to be lemmatized ^{description} in linguistic descriptions or in language ^{compute} computing. On the other hand, inflectional ^{paradigm} paradigms, or ^{list} lists of ^{inflect} inflected ^{form} forms of typical ^{word} words (such as ^{sing} sing, ^{sung} sung, ^{sings} sings, ^{singing} singing, ^{singer} singer, ^{singer} singer, ^{song} song, ^{songs} songs, ^{songstress} songstress, ^{songstresses} songstresses in English) need to be ^{analyze} analyzed according to ^{accord} criteria for ^{uncover} uncovering the underlying lexical stem (here s+ng-); that is, the accompanying functional items (-i-, -a-, -u-, -s, -ing, -er, -o-, -stress, -es) and the functional categories of which they are markers need to be distinguished to adequately describe the language.

Crawler



Selecting Crawler

- Stormcrawler
- Scrapy
- Nutch
- Norconex

Norconex Crawler

<http://www.norconex.com>

Language detection

Sitemap.xml support

SOLR support

Text extraction from documents

Detects modified and deleted documents

Metadata extraction

Customizable

Birth Of SearchINN product

- Slovenian language supported
- Silverstripe integration
- Simple setup
- Results weighted by relevancy (title, content, date of change, ...)
- Yearly licence fee (hosting, maintainance & further development)
- No daily limits - so far :)



Thank you :)