INNOVATIF
BUSINESS DIGITAL PLATFORMS

StripeCon Malta 2017

Which site search solution(s) do you provide?

The end of Google Site Search

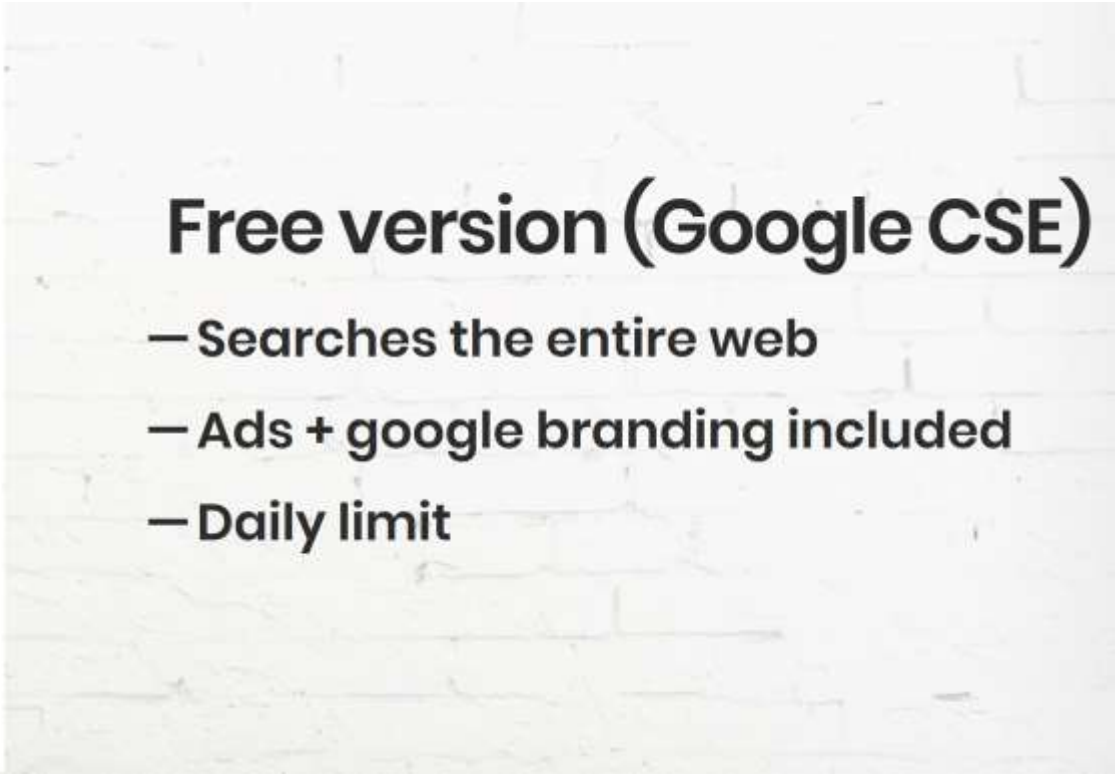
On April 1, 2017, Google will discontinue sales of the Google Site Search. All new purchases and renewals must take place before this date. The product will be completely shut down by April 1, 2018.

Google Site Search

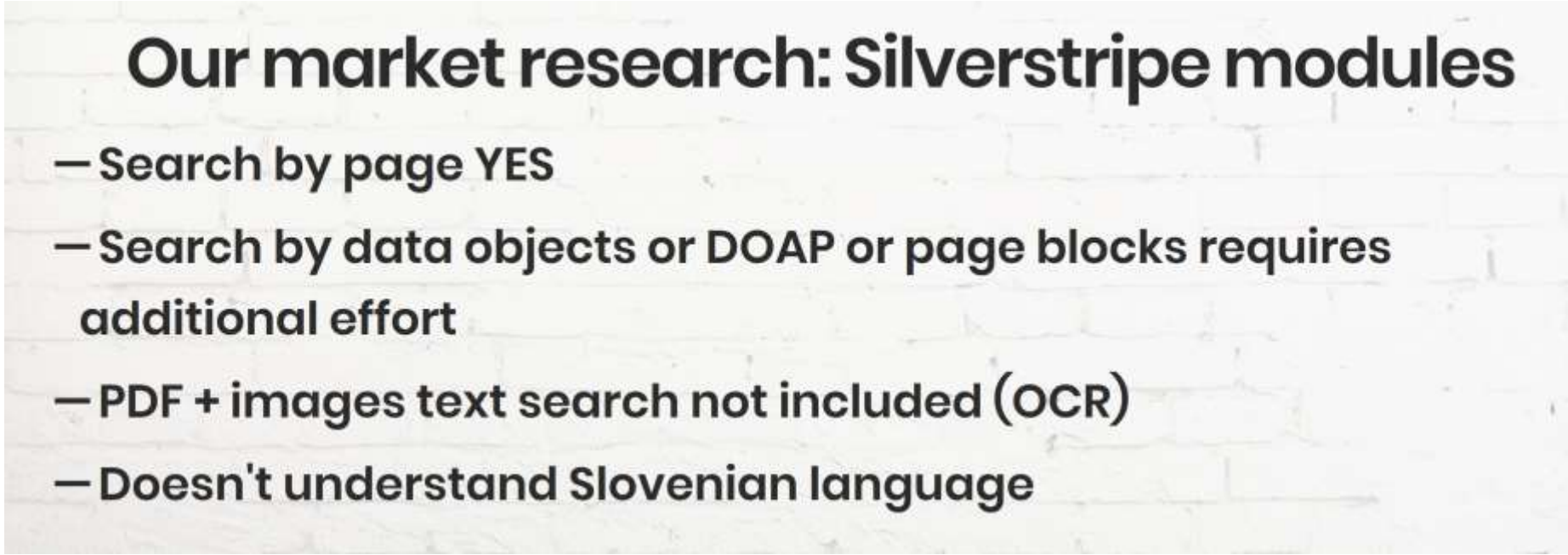
Google Site Search brings the same search technology that powers Google.com to your website, delivering relevant results with lightning speed.

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	✓	✗



Market research



Market research

SOLR + Silverstripe

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"
```

Lemmatisation concept

Lemmatisation

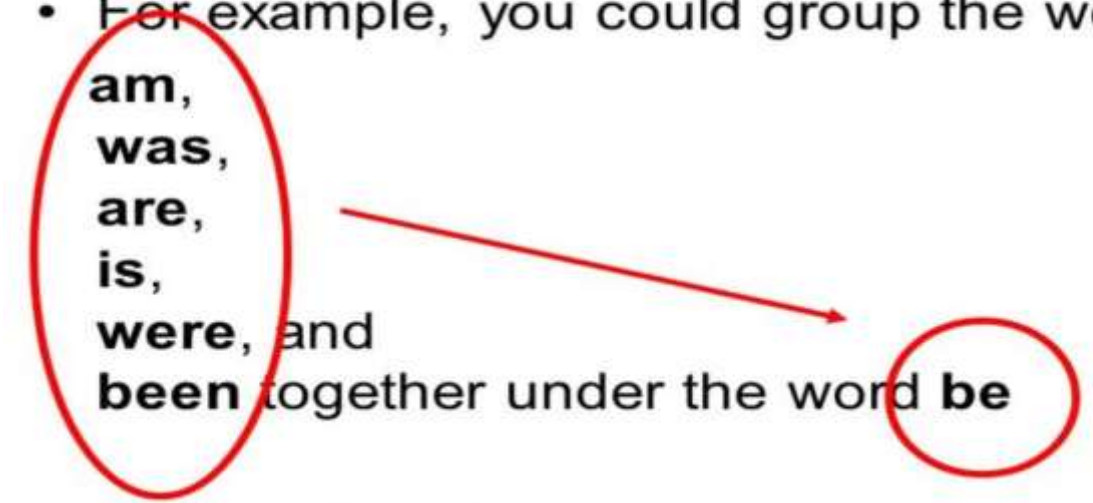
From Wikipedia, the free encyclopedia

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]

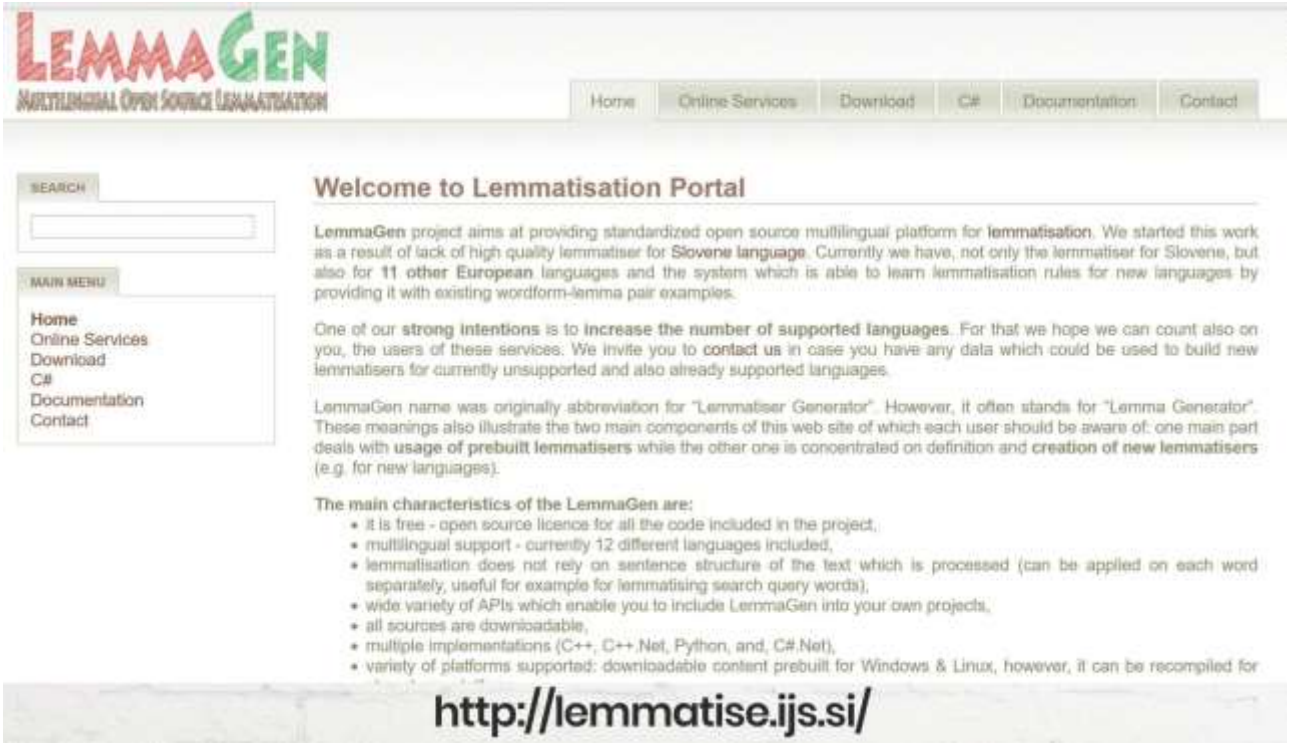
In computational linguistics, lemmatisation is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatisation depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context 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



LemmaGen


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

Crawler

Norconex Crawler

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 :)