

Simurg

An Extendable Multilingual Corpus for Abstractive Single Document Summarization

Pashutan Modaresi, Stefan Conrad (Heinrich-Heine University of Düsseldorf)

Automatic Text Summarization

Extractive Summarization (started nearly 60 years ago)

Abstractive Summarization (revolutionized with deep learning)

English Gigaword Corpus

Linguistic Data Consortium

Attentional Encoder-Decoder Neural Networks

4,111,240 Documents

Size

Accessibility

Licensing

Multilinguality


Change

Extendability

Neural Text Summarization

Politics

The Latest: Trump offers Housing secretary job to Ben Carson




President-elect Donald Trump gestures to people seated in a restaurant as he leaves the New York Times building following a meeting, Tuesday, Nov. 22, 2016, in New York. (Mark Lennihan/Associated Press)

By Associated Press November 22 at 6:29 PM

NEW YORK — The Latest on Donald Trump's transition to the presidency (all times local):

6:10 p.m.

President-elect Donald Trump has formally offered retired neurosurgeon Ben Carson the position of secretary of the Department of Housing and Urban Development.



HRS Das
Sch
ab
Kah
ab

Most Read

- 1 Trump Foundation admits violating ban on "self-dealing" to IRS shows
- 2 This is the single most disturbing thing Donald Trump said in New York Times interview
- 3 Trump has a plan for government workers. They're not going to like it
- 4 Trump will shut down Clinton investigations? That's not supposed to work.
- 5 The Latest: Former DC so-called chief not seeking a Trump job

Unlimited Access to The

HighSpeed Inter
bis zu 120 Mbit/s

Size

Accessibility

Licensing

Multilinguality

Change

Extendability

Simurg

Create the corpus on your own machine

<https://github.com/pasmod/simurg>

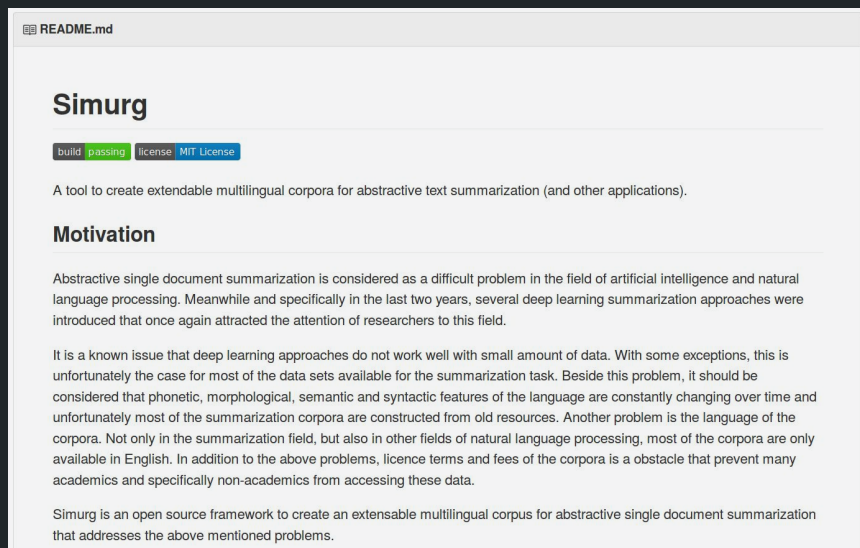
Supports multiple languages

No size limit

No licensing problems

Free

Open source



The screenshot shows the README.md file for the Simurg project. It features a title 'Simurg', a progress indicator for 'build passing' and 'license MIT License', and a description: 'A tool to create extendable multilingual corpora for abstractive text summarization (and other applications)'. The 'Motivation' section discusses the challenges of abstractive single document summarization, particularly with small data sets and non-English corpora. It concludes by stating that Simurg is an open source framework to address these issues.

README.md

Simurg

build passing license MIT License

A tool to create extendable multilingual corpora for abstractive text summarization (and other applications).

Motivation

Abstractive single document summarization is considered as a difficult problem in the field of artificial intelligence and natural language processing. Meanwhile and specifically in the last two years, several deep learning summarization approaches were introduced that once again attracted the attention of researchers to this field.

It is a known issue that deep learning approaches do not work well with small amount of data. With some exceptions, this is unfortunately the case for most of the data sets available for the summarization task. Beside this problem, it should be considered that phonetic, morphological, semantic and syntactic features of the language are constantly changing over time and unfortunately most of the summarization corpora are constructed from old resources. Another problem is the language of the corpora. Not only in the summarization field, but also in other fields of natural language processing, most of the corpora are only available in English. In addition to the above problems, licence terms and fees of the corpora is an obstacle that prevent many academics and specifically non-academics from accessing these data.

Simurg is an open source framework to create an extensible multilingual corpus for abstractive single document summarization that addresses the above mentioned problems.

Simurg

Template Corpus

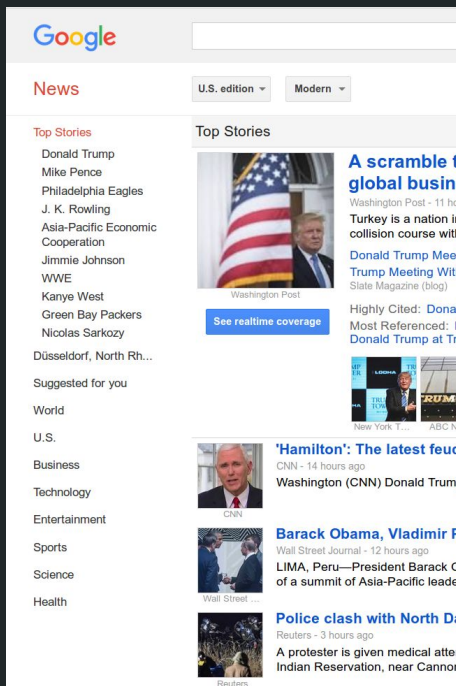
Corpus Population

Shareable

Top Stories

Headlines

URLs



Simurg

Template Corpus

Corpus Population

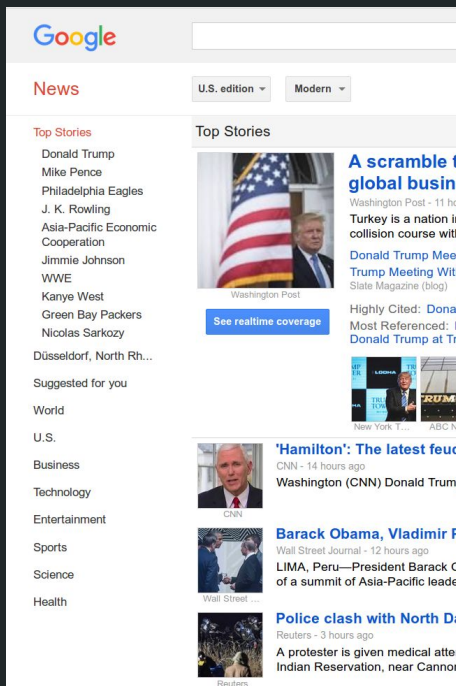
Shareable

Headlines

Top Stories

URLs

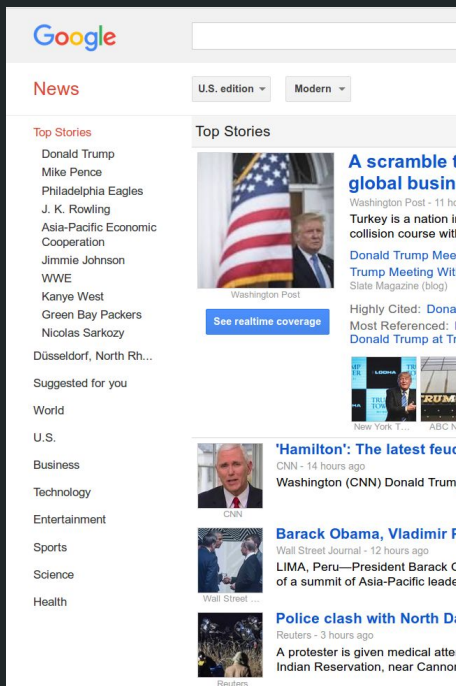
Contents behind URLs change



Simurg

Template Corpus

Corpus Population



```
{
  "archived_snapshots": {
    "closest": {
      "available": true,
      "url": "http://web.archive.org/web/20161121150052/http://example.com/",
      "timestamp": "20161121150052",
      "status": "200"
    }
  }
}
```

Top Stories

Headlines

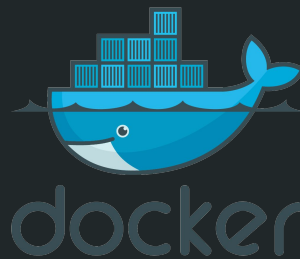
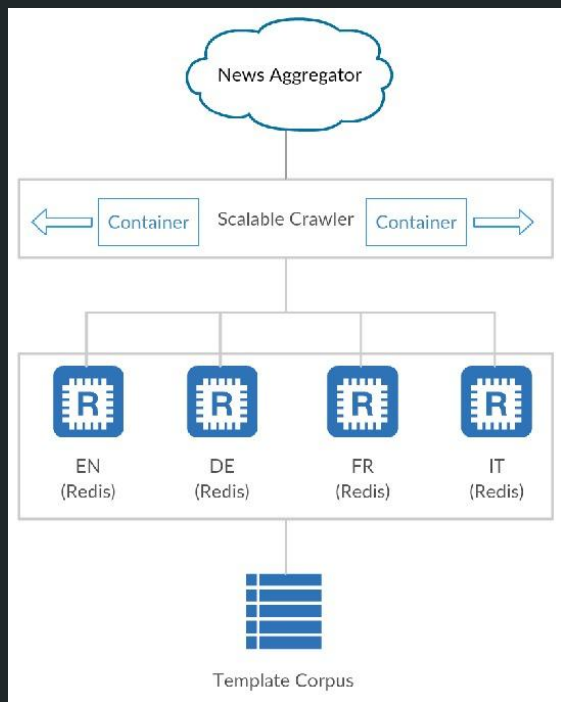
URLs

Contents behind URLs change

Simurg

Template Corpus

Corpus Population



key	http://www.sacbee.com/news/...
id	3409c881-8856-49fd-a1c0
url	http://www.sacbee.com/news/...
wayback_url	not found
headline_selector	h1.title
timestamp	2016-06-22T00:55:25

Simurg

Template Corpus

Corpus Population

How to get the content of the news?



Dragnet (Context Extraction)



Simurg

Template Corpus

Corpus Population



Trump denies any conflict of interest over business empire

1 hour ago | US & Canada



Donald Trump gave an interview to the New York Times on Tuesday

Billionaire US President-elect Donald Trump has said he is not obliged to cut ties to his business empire when he takes office on 20 January.

A Democratic senator is tabling a resolution calling on him to liquidate his assets to prove he does not intend to profit from the office of president.

There is no legal requirement to liquidate assets but past US presidents have set aside their business dealings.

Mr Trump also disowned far right activists who hailed his election win.

Trump elected

The people around Donald Trump

The rise of the alt-right

Trump's 'jail Clinton' U-turn backlash

Full coverage

Split HTML into visual blocks
<div><p><h1>

Train a logistic regression model

Model features:

Text density

Link density

Smoothed tag ratio

Semantic features:

Tokens in id and class attributes

Simurg

Template Corpus

Corpus Population

id	3409c881-8856-49fd-a1c0
timestamp	2016-07-11T12:09:11
lang	en
url	http://www.sacbee.com/news/...
wayback_url	None
headline	Natomas office park asks pastor who praised Orlando massacre to move out
body	The Natomas office park that houses Verity Baptist Church, whose pastor praised the recent massacre at a gay nightclub in Orlando, Fla., has asked the church to move out...

Setup Simurg

```
Make build
```

```
Make start_redis
```

Create Template Corpus

```
Make run  
python  
import simurg  
simurg.create_template_corpus(lang='fr')
```

Populating the Template Corpus

```
Make run  
python  
import simurg  
simurg.populate_template_corpus(lang='fr')
```

Future Work

- Automatic Headline Extraction
- Improving Content Extraction
 - Author Extraction
 - Date Extraction
 - Image Extraction
- Optimizing Parallel Execution

**Thank You
For
Your
Attention**