Natural Language Processing

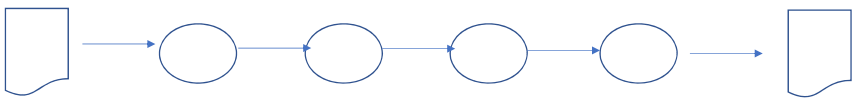
Regexes

Beautiful Soup

NLTK (Natural Language Toolkit)

Objective:
For some given webpage or blog or tweet:

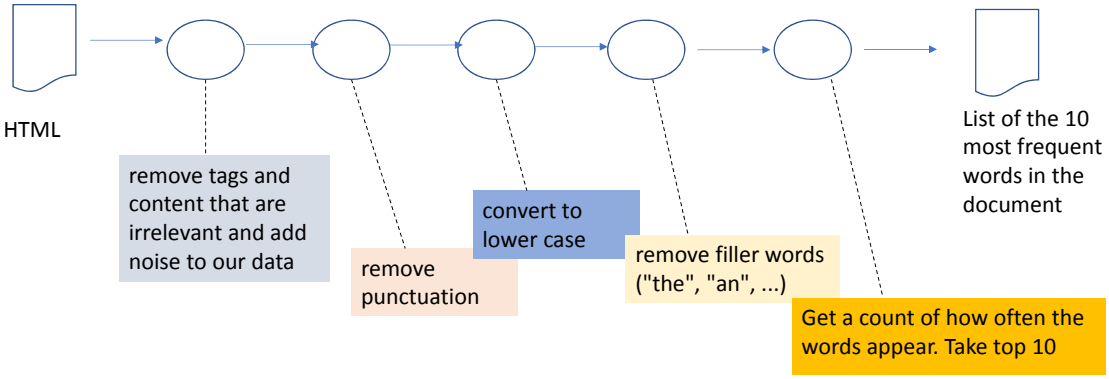
- write code to extract all the text
- analyze the text to determine what the page is about
- build JSON-LD based on the text
- do it using code



Data Flow : data flows through a pipeline of functions, each of which performs a single data transformation task.

Data Flow Thinking is Functional Thinking

Data Flow Pipeline : data flows through a pipeline of functions, each of which performs a single data transformation task.



The Challenge: Removing non-essential text from web pages

Where is the content, where is it not?

```
<!doctype html>
<html xmlns:og="http://opengraphprotocol.org/schema/" xmlns:fb="http://www.facebook.com/2008/fbml" xmlns:website="http://ogp.me/ns/website" lang="en-US"
  <head>
    <meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">

    <meta name="viewport" content="width=device-width, initial-scale=1">

    <!-- This is Squarespace. --><!-- artificial-intelligence-blog -->
<base href="">
<meta charset="utf-8" />
<title>2019 Artificial Intelligence News - AI News</title>
<link rel="shortcut icon" type="image/x-icon" href="https://static1.squarespace.com/static/585e8f70bebafb65339930b8/t/589cd24aff7c507ac47f8597/favicon.i
<link rel="canonical" href="https://www.artificial-intelligence.blog/news"/>
<meta property="og:site_name" content="Artificial Intelligence Blog + AI News"/>
<meta property="og:title" content="2019 Artificial Intelligence News - AI News"/>
<meta property="og:url" content="https://www.artificial-intelligence.blog/news"/>
<meta property="og:type" content="website"/>
<meta property="og:description" content="Daily fresh news from the field of Artificial Intelligence. There is so much happening in AI ... here's your on
<meta property="og:image" content="http://static1.squarespace.com/static/585e8f70bebafb65339930b8/t/5c61cac6652dea1549e1a8bb/1549912776477/artificial-in
<meta property="og:image:width" content="1500"/>
<meta property="og:image:height" content="750"/>
<meta itemprop="name" content="2019 Artificial Intelligence News - AI News"/>
<meta itemprop="url" content="https://www.artificial-intelligence.blog/news"/>
<meta itemprop="description" content="Daily fresh news from the field of Artificial Intelligence. There is so much happening in AI ... here's your one-s
<meta itemprop="thumbnailUrl" content="http://static1.squarespace.com/static/585e8f70bebafb65339930b8/t/5c61cac6652dea1549e1a8bb/1549912776477/artificia
<link rel="image_src" href="http://static1.squarespace.com/static/585e8f70bebafb65339930b8/t/5c61cac6652dea1549e1a8bb/1549912776477/artificial-intelligen
<meta itemprop="image" content="http://static1.squarespace.com/static/585e8f70bebafb65339930b8/t/5c61cac6652dea1549e1a8bb/1549912776477/artificial-intel
<meta name="twitter:title" content="2019 Artificial Intelligence News - AI News"/>
<meta name="twitter:image" content="http://static1.squarespace.com/static/585e8f70bebafb65339930b8/t/5c61cac6652dea1549e1a8bb/1549912776477/artificial-in
<meta name="twitter:url" content="https://www.artificial-intelligence.blog/news"/>
<meta name="twitter:card" content="summary"/>
<meta name="twitter:description" content="Daily fresh news from the field of Artificial Intelligence. There is so much happening in AI ... here's your on
<meta name="description" content="Daily fresh news from the field of Artificial Intelligence. There is so
much happening in AI ... here's your one-stop-shop for all artificial
intelligence news - AI news." />
<script type="text/javascript" src="//use.typekit.net/ik/VcFfX-zDKDf_tqsEUHt4HlBSiDrIMLmwPAFAhxMotLCfeCbffFHN4UJLFRbh52jhWD9tFA9tZRS3jcmKjAZKFR9DwDqowDg
<script type="text/javascript">try{Typekit.load();}catch(e){}</script>
<link rel="stylesheet" type="text/css" href="//fonts.googleapis.com/css?family=Lato:700,400,900"/>
<script type="text/javascript" src="//use.typekit.net/ik/tZg1cQd3xyTS-CN-9Wkzturd3l9ICoVllk_JWjIWleqfenGff4e6pUJ6wRMU5QwXFmvu52m85ejawe48jcBcjRjuwDwo5Qw
<script type="text/javascript">try{Typekit.load();}catch(e){}</script>
<script type="text/javascript">SQUARESPACE_ROLLUPS = {};</script>
<script>(function(rollups, name) { if (!rollups[name]) { rollups[name] = {}; } rollups[name].js = ["//static.squarespace.com/universal/scripts-compresse
```

Simpler page: Where is the content, where is it not?

```
<html>
<head>
<script> var x = 22; var y = 99; print(x,y) </script>
</head>
<body>
<h1 style="text-align: center;">The Best Artificial Intelligence Blogs</h1>
<h2>OpenAI</h2>
<p>The AI researchers at the non-profit AI research company OpenAI are working hard to help us all understand the power of AI as well as the issues that society must work through on this fascinating topic. An important nuance, they seek to enact the path to safe artificial general intelligence.</p>
<p>For more see: <a href="https://openai.com/" target="_blank" rel="nofollow">OpenAI.com</a></p>
<h2>The a16z AI Playbook</h2>
<p>This is not as frequently updated as a blog but it is such a tremendous resource it belongs high on our list. They bring insights on AI topics with a special focus on the creators who are building AI solutions.</p>
<p>For more see: <a href="http://aiplaybook.a16z.com/" target="_blank" rel="nofollow">The a16z AI Playbook</a></p>
<h2>Artificial Intelligence Blog</h2>
<p>They cover AI news, research, books, and thought leaders in the industry. Track for insights into companies and conferences as well.</p>
<p>For more see: <a href="https://www.artificial-intelligence.blog/news/" target="_blank" rel="nofollow">Artificial-Intelligence.blog</a></p>
<h2>Machine Learning Mastery</h2>
<p>Dr. Jason Browlee is a respected practitoner and master of machine learning and he writes for others seeking to really excel at machine learning. Since his focus is on the people who can really execute the blog gets technical, but it is still understandable for the non-technical person who needs to track the big issues.</p>
...
```

Goal:
Extract relevant text and create a Word Cloud

Beautiful Soup –
a Python Module for
extracting text from web pages

```
import bs4
from bs4 import BeautifulSoup
import numpy as np
import pandas as pd
from os import path

from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import requests
```

Extract text from simple web page about Rollo using BS

```
html = '''
<html>
<head>

<h1>Rollo WORLD</h1>
<p>Rollo ate lunch and then went to the pool to study AI.</p>
He knew AI was the way to take over the world.
Rollo dreamed of  AI as the solution to all his problems.
<p>Perhaps the world will someday
know about  Rollo and his ambitions. Yet now, the world does not know the Rollo.
</p>
</html>
'''

soup = BeautifulSoup(html, "lxml")
# we don't want to strip line breaks
htmlwords = soup.get_text(strip=True)
print (htmlwords)
```

```
RolloWORLDRollo ate lunch and then went to the pool to study AI.He knew AI was the way to take over the world.
Rollo dreamed of  AI as the solution to all his problems.Perhaps the world will someday
know about  Rollo and his ambitions. Yet now, the world does not know the Rollo.
```

when using strip="True", we remove extra space and linebreaks - but we may merge words ☹

```
html = '''
<html>
<head>

<h1>Rollo WORLD</h1>
<p>Rollo ate lunch and then went to the pool to study AI.</p>
He knew AI was the way to take over the world.
Rollo dreamed of  AI as the solution to all his problems.
<p>Perhaps the world will someday
know about  Rollo and his ambitions. Yet now, the world does not know the Rollo.
</p>
</html>
'''

soup = BeautifulSoup(html, "lxml")
# we don't want to strip line breaks
htmlwords = soup.get_text(strip=False)
print (htmlwords)
```

```
Rollo WORLD
Rollo ate lunch and then went to the pool to study AI.
He knew AI was the way to take over the world.
Rollo dreamed of  AI as the solution to all his problems.
Perhaps the world will someday
know about  Rollo and his ambitions. Yet now, the world does not know the Rollo.
```

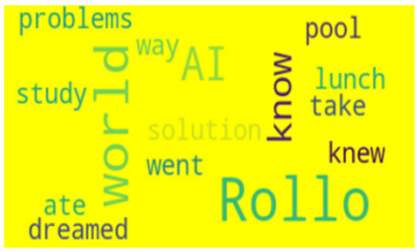
a string

We don't want to strip extra spaces and linebreaks. We get the single words!

WordCloud – from a string of words

```
wordcloud = WordCloud(max_font_size=50, max_words=15, background_color="yellow").generate(htmlwords)

# Display the generated image:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```



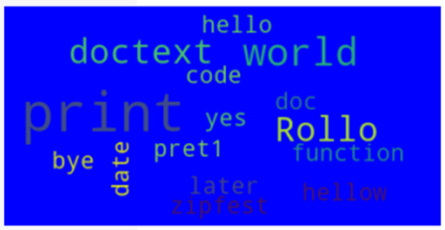
OK BUT what if our web page is more complex...

We get irrelevant words from the <script> element

```
html = '''
<html>
<head>
<script> print("hellow world")</script>
<script>
    (function(){
      print ("hello");  print ("bye"); print ("now");
      print ("later");  print ("doc"); print ("yes");
      print ("doctext");  print ("doctext"); print ("doctext");
    </script>
</head>

<h1>Rollo WORLD</h1>
<p class="foo1">pret1</p>
<p class="foo1">date is now</p>
<p class="foo1">zipfest</p>
<p>code</p>
<p>data</p>
<p class="foo1">aT DAWN Rollo <strong class='xyz'>ate
</strong>
that the lunch left for him. Oh that Rollo. He wants to take lunch and lunch over the that that  world.
The world does not know.
</p>
<html>
'''

soup = BeautifulSoup(html, "lxml")
# we don't want to strip line breaks
htmlwords = soup.get_text(strip=False)
print (htmlwords)
```




```
print("hellow world")

    (function(){
      print ("hello");  print ("bye"); print ("now");
      print ("later");  print ("doc"); print ("yes");
      print ("doctext");  print ("doctext"); print ("doctext");


Rollo WORLD
pret1
date is now
zipfest
code
data
aT DAWN Rollo ate

that the lunch left for him. Oh that Rollo. He wants to take lunch and lunch
over the that that  world.
The world does not know.
```

We get the TEXT inside every HTML element – even the script tag!

We need the power of **regex**

Quick Regex Overview....

regular expressions with the Python re module

Syntax

```
import re

result = re.sub(pattern, repl, string, count=0, flags=0);
```

Simple Examples

```
result = re.sub('abc',  '',    input)          # Delete pattern abc
result = re.sub('abc',  'def', input)          # Replace pattern abc -> def
result = re.sub(r'\s+', ' ',   input)          # Eliminate duplicate whitespaces
result = re.sub('abc(def)ghi', r'\1', input)   # Replace a string with a part of itself
```

Note: Take care to always prefix patterns containing \ escapes with raw strings (by adding an r in front of the string). Otherwise the \ is used as an escape sequence and the regex won't work.

Experiment with REGEXes at https://www.regexpal.com/

Basics

*	Match preceding character 0 or more times
+	Match preceding character 1 or more times
.	Match any single character
x|y	Match either 'x' or 'y'
\	Escape a special character
b	The character b
abc	The string abc

Character Classes

[abc]	Match any one of the characters in the set 'abc'
[^abc]	Match anything not in character set 'abc'

\d	Match a digit character
\D	Match a non-digit character
\s	Match a single white space character (space, tab, form feed, or line feed)
\S	Match a single character other than white space
\w	Match any alphanumeric character (including underscore)
\W	Match any non-word character

How to use **regex** to eliminate a tag (e.g.<script>) and all its content

The secret sauce

the '?' turns OFF greedy matching. note that the <script> element is gone in the text!

```
# remove the script element and its content
blogtext = re.sub(r'<script>.*?</script>', ' ', blogtext)
blogtext
```

```
'<html>\n<head>\n \n</head>\n<body>\n<h1 style="text-align: center;">The Best Artificial Intelligence Blogs</h1>\n<h2>OpenA
I</h2>\n<p>The AI researchers at the non-profit AI research company OpenAI are working hard to help us all understand the p
ower of AI as well as the issues that society must work through on this fascinating topic. An important nuance, they seek t
o enact the path to safe artificial general intelligence.</p>\n<p>For more see: <a href="https://openai.com/" target="_blan
k" rel="nofollow">OpenAI.com</a></p>\n<h2>The a16z AI Playbook</h2>\n<p>This is not as frequently updated as a blog but it
is such a tremendous resource it belongs high on our list. They bring insights on AI topics with a special focus on the cre
ators who are building AI solutions.</p>\n<p>For more see: <a href="http://aiplaybook.a16z.com/" target="_blank" rel="nofol
```

Challenge: Create WordCloud from a web page

The Best Artificial Intelligence Blogs

OpenAI

The AI researchers at the non-profit AI research company OpenAI are working hard to help us all understand the power of AI as well as the issues that society must work through on this fascinating topic. An important nuance, they seek to enact the path to safe artificial general intelligence.

For more see: OpenAI.com

The a16z AI Playbook

This is not as frequently updated as a blog but it is such a tremendous resource it belongs high on our list. They bring insights on AI topics with a special focus on the creators who are building AI solutions.

For more see: The a16z AI Playbook

Artificial Intelligence Blog

They cover AI news, research, books, and thought leaders in the industry. Track for insights into companies and conferences as well.

For more see: Artificial-Intelligence.blog

Machine Learning Mastery

Dr. Jason Browlee is a respected practitoner and master of machine learning and he writes for others seeking to really excel at machine learning. Since his focus is on the people who can really execute the blog gets technical, but it is still understandable for the non-technical person who needs to track the big issues.

For more see: Machine Learning Mastery


```
<html>
<head>
<script> var x = 22; var y = 99; print(x,y) </script>
</head>
<body>
<h1 style="text-align: center;">The Best Artificial Intelligence Blogs</h1>
<h2>OpenAI</h2>
<p>The AI researchers at the non-profit AI research company OpenAI are working hard to help us all understand the power of AI as well as the issues that society must work through on this fascinating topic. An important nuance, they seek to enact the path to safe artificial general intelligence.</p>
<p>For more see: <a href="https://openai.com/" target="_blank" rel="nofollow">OpenAI.com</a></p>
<h2>The a16z AI Playbook</h2>
<p>This is not as frequently updated as a blog but it is such a tremendous resource it belongs high on our list. They bring insights on AI topics with a special focus on the creators who are building AI solutions.</p>
<p>For more see: <a href="http://aiplaybook.a16z.com/" target="_blank" rel="nofollow">The a16z AI Playbook</a></p>
<h2>Artificial Intelligence Blog</h2>
<p>They cover AI news, research, books, and thought leaders in the industry. Track for insights into companies and conferences as well.</p>
<p>For more see: <a href="https://www.artificial-intelligence.blog/news/" target="_blank" rel="nofollow">Artificial-Intelligence.blog</a></p>
<h2>Machine Learning Mastery</h2>
<p>Dr. Jason Browlee is a respected practitoner and master of machine learning and he writes for others seeking to really excel at machine learning. Since his focus is on the people who can really execute the blog gets technical, but it is still understandable for the non-technical person who needs to track the big issues.</p>
<p>For more see: <a href="http://machinelearningmastery.com/blog/" rel="nofollow" target="_blank">Machine Learning Mastery</a></p>
<h2>The Algorithmia Blog</h2>
...
```

```
# read file and save content as text string
file = open("blogpage.html", "r")
blogtext = file.read()
blogtext
```

```
'<html>\n<head>\n<script> var x = 22; var y = 99; print(x,y) </script>\n</head>\n<body>\n<h1 style="text-align: center;">Th
e Best Artificial Intelligence Blogs</h1>\n<h2>OpenAI</h2>\n<p>The AI researchers at the non-profit AI research company Ope
nAI are working hard to help us all understand the power of AI as well as the issues that society must work through on this
fascinating topic. An important nuance, they seek to enact the path to safe artificial general intelligence.</p>\n<p>For mo
re see: <a href="https://openai.com/" target="_blank" rel="nofollow">OpenAI.com</a></p>\n<h2>The a16z AI Playbook</h2>\n<p>
This is not as frequently updated as a blog but it is such a tremendous resource it belongs high on our list. They bring in
sights on AI topics with a special focus on the creators who are building AI solutions.</p>\n<p>For more see: <a href="http
://aiplaybook.a16z.com/" target="_blank" rel="nofollow">The a16z AI Playbook</a></p>\n<h2>Artificial Intelligence Blog</h2>
\n<p>They cover AI news, research, books, and thought leaders in the industry. Track for insights into companies and confer
ences as well.</p>\n<p>For more see: <a href="https://www.artificial-intelligence.blog/news/" target="_blank" rel="nofollow
```

Tasks:

1) remove tags and content not relevant to the page (e.g. script and others?)
2) remove \n
3) remove tags
4) build Word Cloud for relevant text