

Chrome based Keyword Visualizer (under sparse text constraint)

COMPUTER SCIENCE 2013210085 SANGHO SUH
INDUSTRIAL ENGINEERING 2010170835 MOONSHIK KANG
COMPUTER SCIENCE 2010210012 HOONHEE CHO

12/7/2015

Team



Sangho Suh



Hunhee Cho



Moonshik Kang

INDEX

- ▶ Proposal Recap
- ▶ Implementation
- ▶ Evaluation
- ▶ Future Works

Proposal Recap



Keyword Visualizer (chrome plugin)

The screenshot shows a Chrome browser window displaying an article from The New York Times. The article is about Farhad Manjoo and Mike Isaac's weekly column in technology. The Keyword Visualizer plugin is active, highlighting specific words in yellow and displaying a large, semi-transparent word cloud on the right side of the page.

Each Saturday, Farhad Manjoo and Mike Isaac, **technology reporters at **The New York Times**, review **the week's news**, offering analysis and maybe a joke or two about the most important developments in the **tech** industry.**

Mike: Hello, **Farhad**! How are you? I just got back from a week of "new hire orientation" for **The Times**. I've been an employee here for 18 months. I guess we run on a slightly different timetable.

Farhad: To be fair, 18 months doesn't seem like nearly enough time to complete a thorough background check on you.

Mike: Fair. So, I was sort of checked out of the news flow, but came back to a deluge of **tech** craziness.

Google is doing this thing with a bunch of other **tech** companies to make web pages **load faster on our phones**, which is theoretically a good thing. Jet, the online shopping Amazon competitor, killed one of its **main business model decisions** — to charge for a membership — which seems theoretically like a bad thing (for Jet, at least).

Farhad: To be fair, 18 months doesn't seem like nearly enough time to complete a thorough background check on you.

Mike: Fair. So, I was sort of checked out of the news flow, but came back to a deluge of **tech** craziness.

Google is doing this thing with a bunch of other tech companies to make web pages **load faster on our phones**, which is theoretically a good thing. Jet, the online shopping Amazon competitor, killed one of its **main business model decisions** — to charge for a membership — which seems theoretically like a bad thing (for Jet, at least). And **Dell and EMC** are considering some sort of giant merger or takeover, which is still in its theoretical stage. Also, I have no idea how to speak intelligently on the cloud — the cloud! — so let's just tiptoe past that.

Farhad: You forgot about **the Microsoft event**. They made a laptop! With a screen that's actually a tablet! I thought it was pretty cool. Maybe PCs are cool again?

about news orientation guess
Farhad
newly shopping online phones new maybe
Times merges important Amazon
medium faster killed Oh industry event Each decisions open
Fair part let's Saturday hire Microsoft intelligently
months other made Magpie model pretty like
years screen ideas deluge pages
eyes flow job fell different least came back
seems screen load
PCs offer both one most
Dell actually Hello enough just tech reports bad complete Isaac
Hello forgot membership employee best
business craziness review Jet giant
analysis slightly background considering cloud Zzzzzzzz
sort cool theoretically
Mike

THE NAVITIMER 46 mm
BREITLING
INSTRUMENTS FOR PROFESSIONALS™

Implementation



Architecture

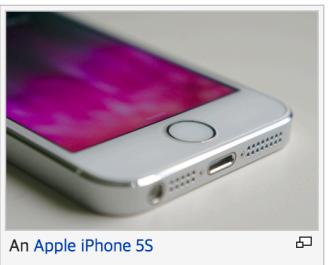
Smartphone

From Wikipedia, the free encyclopedia

"*Smartphones*" redirects here. For the song by Trey Songz, see *SmartPhones (song)*.

Not to be confused with *Feature phone*.

A **smartphone** or **smart phone** is a **mobile phone** with an advanced **mobile operating system** which combines features of a **personal computer** operating system with other features useful for mobile or handheld use.^{[1][2][3]} They typically combine the features of a cell phone with those of other popular **mobile devices**, such as **personal digital assistant** (PDA), **media player** and **GPS navigation unit**. Most smartphones can access the **Internet**, have a **touchscreen user interface**, can run third-party **apps**, **music players** and are **camera phones**. Most smartphones produced from 2012 onwards also have high-speed mobile broadband **4G LTE internet**, **motion sensors**, and **mobile payment mechanisms**.



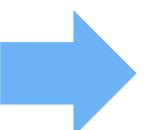
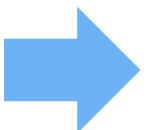
An Apple iPhone 5S

Keyword Visualizer

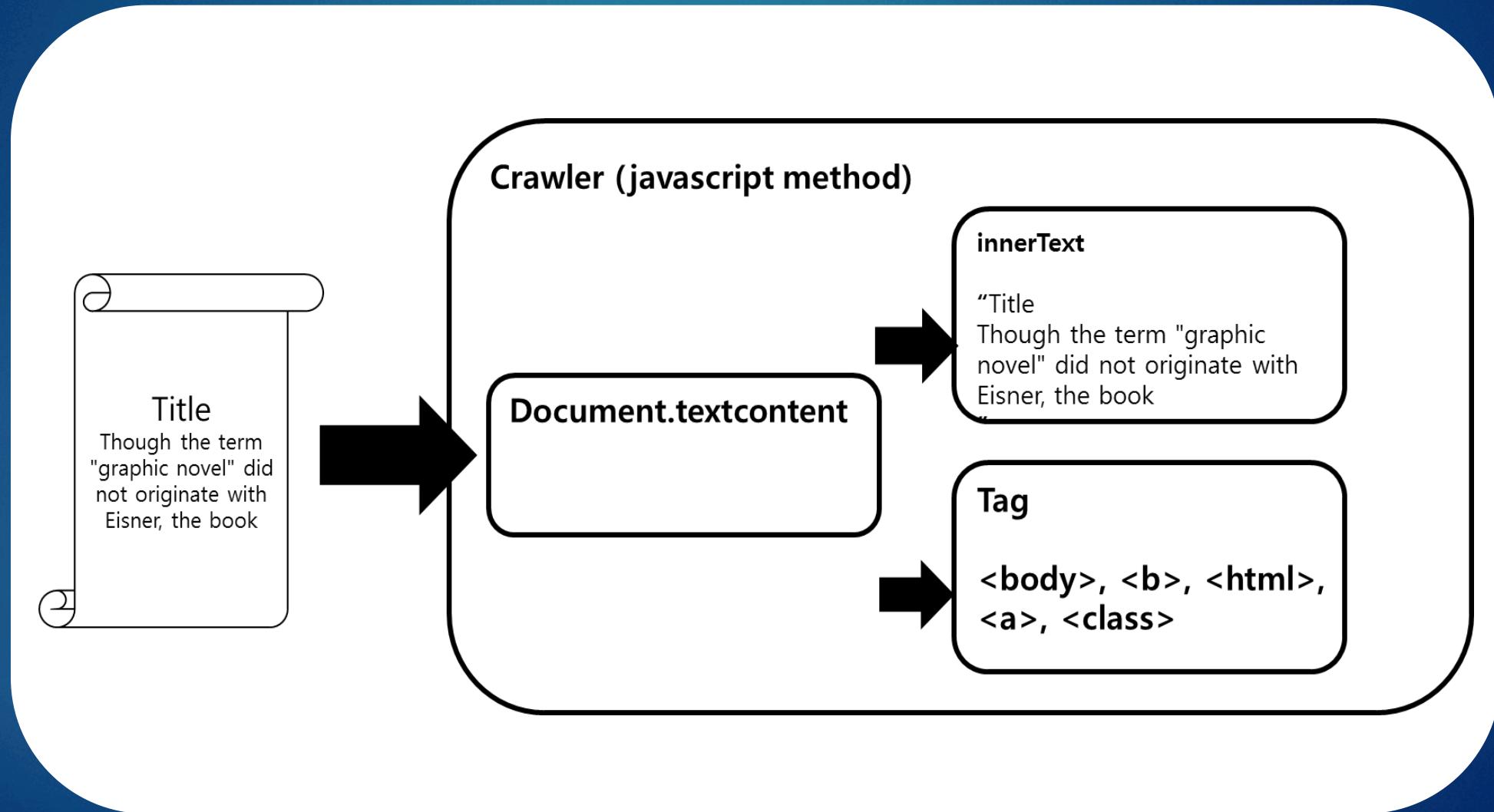
Web crawler

Bag-of-word generator

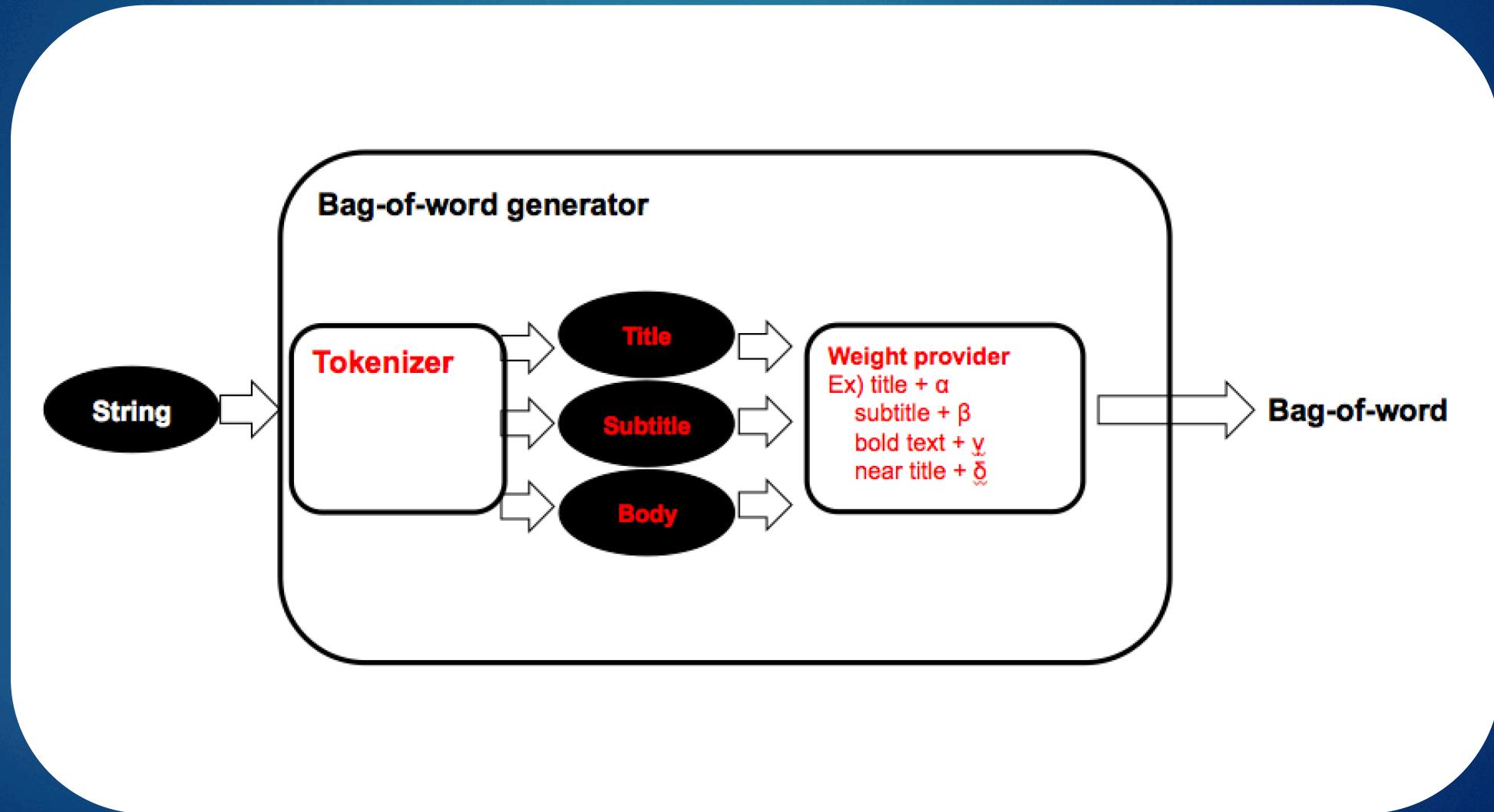
Visualizer



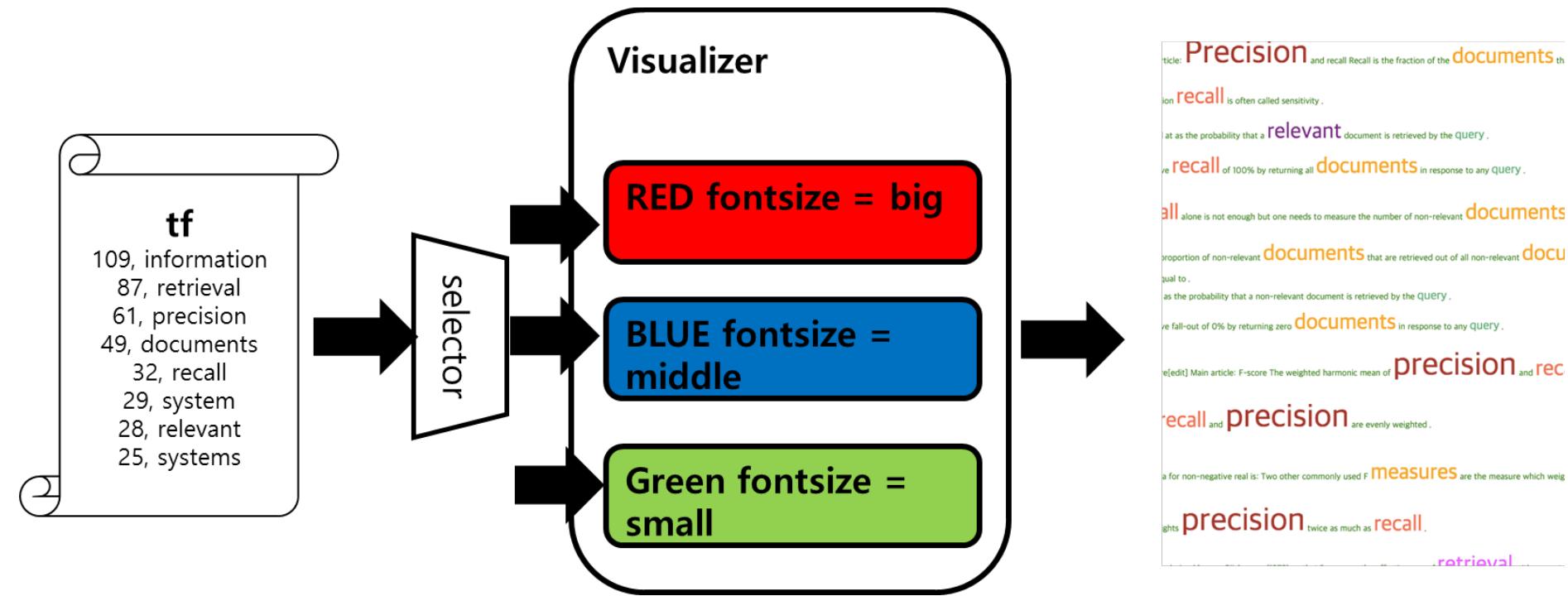
Architecture



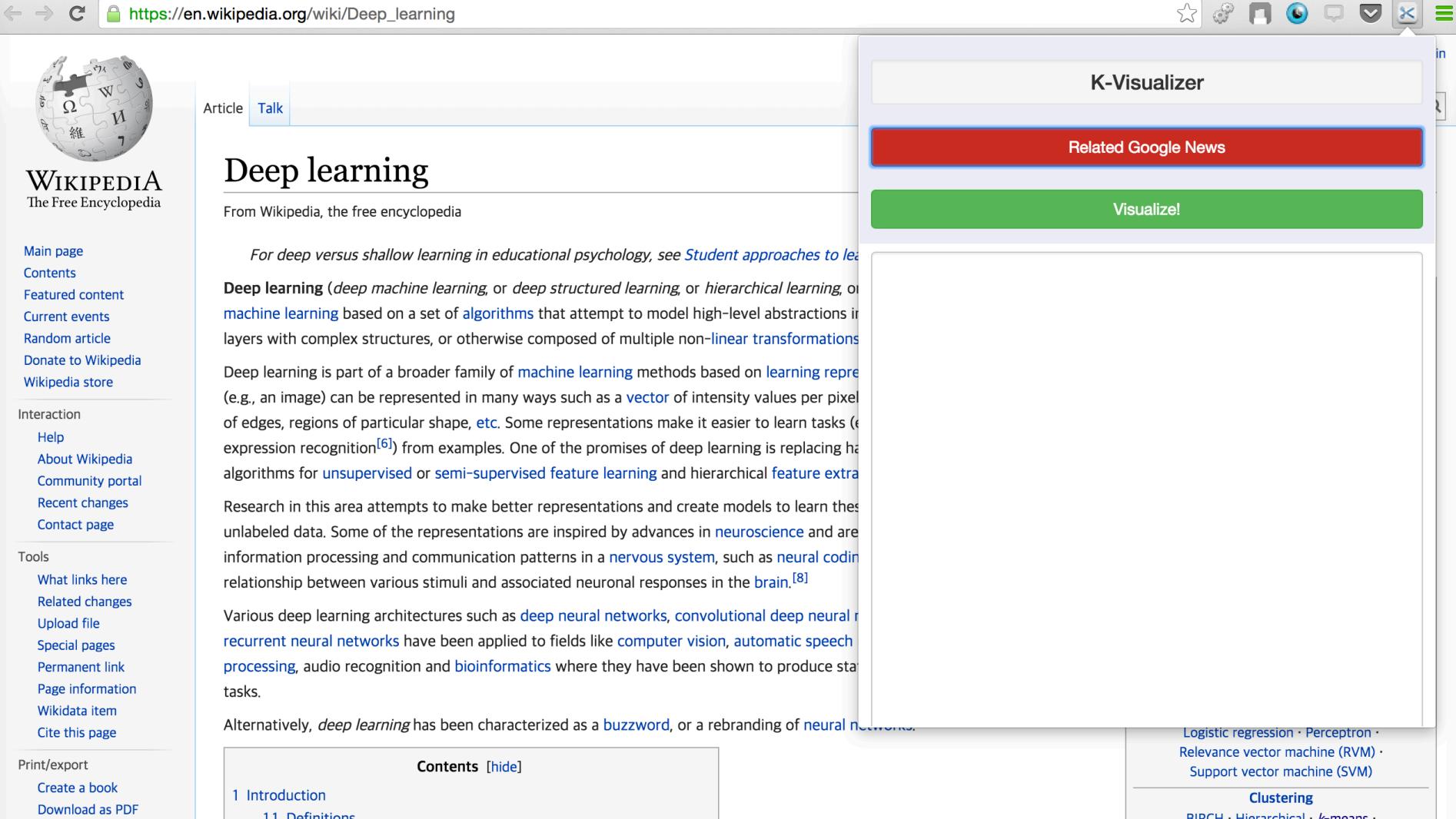
Architecture



Architecture



Chrome Plugin



The screenshot shows a Chrome browser window with the URL https://en.wikipedia.org/wiki/Deep_learning. The main content is the Wikipedia article on Deep learning. A sidebar on the right is titled "K-Visualizer". It features a red bar labeled "Related Google News" and a green button labeled "Visualize!". The rest of the sidebar is currently empty.

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF

Article Talk

Deep learning

From Wikipedia, the free encyclopedia

For deep versus shallow learning in educational psychology, see [Student approaches to learning](#).

Deep learning (*deep machine learning*, or *deep structured learning*, or *hierarchical learning*, or *deep belief networks*) is a subset of *machine learning* based on a set of [algorithms](#) that attempt to model high-level abstractions in data by learning multiple levels of representation using [hierarchical layers](#) with complex structures, or otherwise composed of multiple non-linear transformations of its input.

Deep learning is part of a broader family of *machine learning* methods based on *learning representations* (e.g., an image) can be represented in many ways such as a *vector* of intensity values per pixel, or a *vector* of edges, regions of particular shape, *etc.* Some representations make it easier to learn tasks (e.g., *object recognition* and *image captioning*) from examples. One of the promises of deep learning is replacing hand-engineered features with automatically learned features. These algorithms for *unsupervised* or *semi-supervised feature learning* and hierarchical *feature extraction* have been applied to fields like *computer vision*, *automatic speech recognition*, *audio recognition* and *bioinformatics* where they have been shown to produce state-of-the-art results.

Research in this area attempts to make better representations and create models to learn these representations directly from unlabeled data. Some of the representations are inspired by advances in *neuroscience* and are based on the way biological nervous systems process and communicate information. The relationship between various stimuli and associated neuronal responses in the *brain*.^[8]

Various deep learning architectures such as *deep neural networks*, *convolutional deep neural networks*, *recurrent neural networks* have been applied to fields like *computer vision*, *automatic speech processing*, *audio recognition* and *bioinformatics* where they have been shown to produce state-of-the-art results.

Alternatively, *deep learning* has been characterized as a *buzzword*, or a rebranding of *neural networks*.

Contents [hide]

1 Introduction
1.1 Definitions

Logistic regression · Perceptron ·
Relevance vector machine (RVM) ·
Support vector machine (SVM)

Clustering
BIRCH · Hierarchical · *k*-means ·

How?

There are many keyword extraction techniques...

How?

Such as..

1. Word Frequency Analysis
2. Word Co-Occurrence Relationships
3. Frequency-Based Single Document Keyword Extraction
4. Content-Sensitive Single Document Keyword Extraction
5. Keyword Extraction Using Lexical Chains
6. Keyphrase Extraction Using Bayes Classifier

How?

We can use..

1. Word Frequency Analysis
2. ~~Word Co-Occurrence Relationships~~
3. ~~Frequency-Based Single Document Keyword Extraction~~
4. ~~Content-Sensitive Single Document Keyword Extraction~~
5. ~~Keyword Extraction Using Lexical Chains~~
6. ~~Keyphrase Extraction Using Bayes Classifier~~

How?

Because ..

1. Web-based
2. Single-document

Word Frequency Analysis

- 1) Clean the text
(Remove punctuations and stop words)
- 2) Tokenize the text
- 3) Find the TF (term frequency) for each unique token
- 4) Sort each token in descending count order

Feature1 - Frequency Weighting

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

Article [Talk](#)

Read [Edit](#) [View history](#) [Search](#)

Information retrieval

From Wikipedia, the free encyclopedia

Information retrieval (IR) is the activity of obtaining [information](#) resources relevant to an information need from a collection of information resources. Searches can be based on [metadata](#) or on [full-text](#) (or other content-based) indexing.

Automated information retrieval systems are used to reduce what has been called "[information overload](#)". Many universities and [public libraries](#) use IR systems to provide access to books, journals and other documents. [Web search engines](#) are the most visible IR applications.

Contents [hide]

- 1 Overview
- 2 History
- 3 Model types
 - 3.1 First dimension: mathematical basis
 - 3.2 Second dimension: properties of the model
- 4 Performance and correctness measures
 - 4.1 Precision
 - 4.2 Recall
 - 4.3 Fall-out
 - 4.4 F-score / F-measure
 - 4.5 Average precision
 - 4.6 Precision at K
 - 4.7 R-Precision
 - 4.8 Mean average precision
 - 4.9 Discounted cumulative gain
 - 4.10 Other measures
 - 4.11 Visualization

Information science

General aspects

Information access · Information architecture
Information management
Information retrieval
Information seeking · Information society
Knowledge organization · Ontology · Taxonomy
Philosophy of information
Science, technology and society

Related fields and sub-fields

Bibliometrics · Categorization
Censorship · Classification
Computer data storage · Cultural studies
Data modeling · Informatics
Information technology
Intellectual freedom
Intellectual property · Memory
Library and information science
Preservation · Privacy
Quantum information science

[Information science portal](#)

K-Visualizer

Related Google News

Visualize!

- 159, information
- 123, retrieval
- 109, precision
- 49, documents
- 45, model
- 44, recall
- 40, average
- 36, measures
- 29, system
- 28, relevant
- 27, management
- 25, systems
- 25, edit
- 24, the
- 23, query
- 23, search
- 22, computer
- 22, software
- 20, performance
- 20, science

1 appearance = weight 1

Feature2 – Meta Tag Weighting

▶ HTML

```
<HTML>  
<TITLE> </TITLE>  
<a href=""> IR </a>  
<b> Prof. Choo </b>
```



Feature2 – Meta Tag Weighting

$\langle \text{TITLE} \rangle \xleftarrow{\hspace{1cm}} + \alpha$

$\langle a \rangle \xleftarrow{\hspace{1cm}} + \beta$

$\langle b \rangle \xleftarrow{\hspace{1cm}} + \gamma$

Why Meta Tag Weighting?

- ▶ Problem
 1. Topic Modeling of single page is under **sparse text** constraint
 2. No viable method exists for constructing Term-Document Matrix

Why Meta Tag Weighting?

- ▶ Problem

Difficult to Extract Keywords

IS under **sparse text** constraint

2. No viable method exists for constructing Term-Document Matrix

Why Meta Tag Weighting?

- ▶ Need
 - ‘Additional’ or ‘Hidden’ information

Tags can provide additional
semantic information

Why Meta Tag Weighting?

Mining “Hidden Phrase” Definitions from the Web

Informing Science Journal Information Sciences Volume 6, 2003 H. Liu¹, and

HTML Tags as Extraction Cues for Web Page Description Construction

*Timothy C. Craven
The University of Western Ontario, London, Ontario, Canada*

craven@uwo.ca

Abstract

Using four previously identified samples of Web pages containing meta-tagged descriptions, the value of meta-tagged keywords, the first 200 characters of the body, and text marked with common HTML tags as extracts helpful for writing summaries was estimated by applying two measures: density of description words and density of two-word description phrases. Generally, titles and keywords showed the highest densities. Parts of the body showed densities not much different from the body as a whole: somewhat higher for the first 200 characters and for text tagged with "center" and "font"; somewhat lower for text tagged with "a"; not significantly different for "table" and "div". Evidence of non-random clumping of description words in the body of some pages nevertheless suggests that further pursuit of automatic passage extraction methods from the body may be worthwhile. Implications of the findings for aids to summarization, and specifically the TexNet32 package, are discussed.

Keywords : HTML; extracting; metadata; summarization; computer software; World Wide Web.

a	70%-78%
body	87%-95%
br	61%-71%
center	46%-51%
div	27%-34%
font	59%-69%
head	88%-97%
img	75%-78%
meta	80%-88%
p	69%-78%
script	28%-36%
table	61%-69%
td	60%-67%
title	88%-96%
tr	60%-67%

Table 1: Frequencies of most common tags

Feature3 – Distance Weighting (Feature 1)

- ▶ Weight on words within ε of `<title>`,

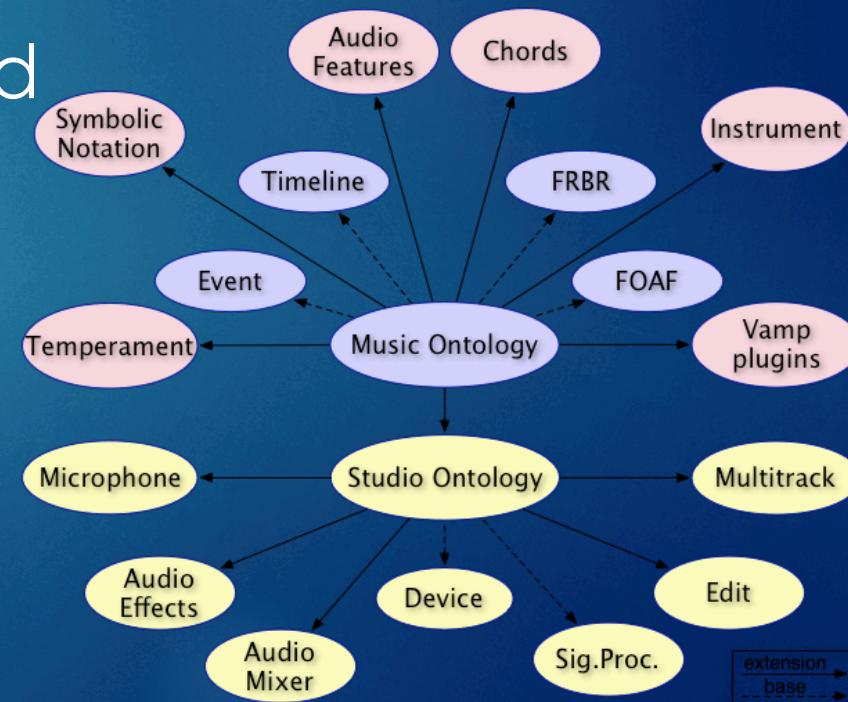
i.e.

$$|x - \langle title \rangle| < \varepsilon$$

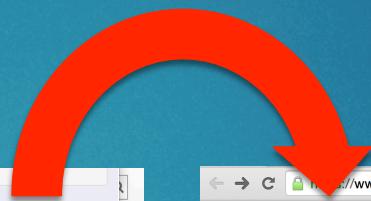
where $|x - \langle title \rangle|$ is distance from
`<title>` tag

Feature4 – Similarity Weighting

- ▶ Weight based on ‘Ontology based Similarity score’
 - 1) Extract words from innerText of <title>
 - 2) Measure similarity score of selected words with words in the text
 - 3) Add weight on the words with high similarity score



Feature 5 – Google News Recommendation



The screenshot shows a Wikipedia article titled "Deep learning". The page content discusses deep learning as a type of machine learning. It mentions that deep learning involves multiple layers of non-linear transformations and can be used for tasks like image recognition. The page also links to various sub-topics such as neural networks, convolutional neural networks, and recurrent neural networks.

The screenshot shows a Google search results page for the query "deep learning neural networks". The top result is a news article from "The Next Platform" titled "Wider Net Cast Over Deep Learning On GPUs". Below it are other news articles, including one from VentureBeat about an open-source deep learning framework and another from Live Science about Google's new AI system. The results are presented in a standard Google search layout with news, images, and other types of content.

Evaluation



Test K-Visualizer

The screenshot shows a web browser window with multiple tabs open. The active tab is the English Wikipedia page for "Information retrieval". The sidebar on the right is a "K-Visualizer" tool.

K-Visualizer

- Related Google News
- Visualize!

The main content area displays the Wikipedia article on Information retrieval, which includes a table of contents and several sections of text.

Table of Contents:

- 1 Overview
- 2 History
- 3 Model types
 - 3.1 First dimension: mathematical basis
 - 3.2 Second dimension: properties of the model
- 4 Performance and correctness measures
 - 4.1 Precision
 - 4.2 Recall
 - 4.3 Fall-out
 - 4.4 F-score / F-measure
 - 4.5 Average precision
 - 4.6 Precision at K
 - 4.7 R-Precision
 - 4.8 Mean average precision
 - 4.9 Discounted cumulative gain
 - 4.10 Other measures
 - 4.11 Visualization
- 5 Timeline
- 6 Awards in the field
- 7 See also
- 8 References
- 9 Further reading
- 10 External links

1. Only with term frequency

Information retrieval

Information

Information management Information retrieval Information seeking technology and society Related fields and sub-fields Bibliometrics - Categorization Censorship - Classification Computer science Intellectual freedom Intellectual property Memory Library and information retrieval

relevant information information retrieval

Automated information retrieval systems

Many universities and public libraries use information retrieval systems to provide access to books journals and other documents.

109, information
87, retrieval
61, precision
49, documents
32, recall
29, system
28, relevant
25, systems
25, edit
25, management
24, query
22, software
22, computer
21, model
19, search
18, models
16, average
16, measure
16, document
16, ext

Precision and recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

In binary classification recall is often called sensitivity.

So it can be looked at as the probability that a relevant document is retrieved by the query.

It is trivial to achieve recall of 100% by returning all documents in response to any query.

Therefore recall alone is not enough but one needs to measure the number of non-relevant documents also for example by computing the precision.

Fall-out[edit] The proportion of non-relevant documents that are retrieved out of all non-relevant documents available. In binary classification fall-out is closely related to specificity and is equal to .

It can be looked at as the probability that a non-relevant document is retrieved by the query.

It is trivial to achieve fall-out of 0% by returning zero documents in response to any query.

F-score / F-measure[edit] Main article: F-score The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is: This is also known as the measure because recall and precision are evenly weighted.

The general formula for non-negative real is: Two other commonly used F MEASURES are the measure which weights recall twice as much as precision and the measure which weights precision twice as much as recall.

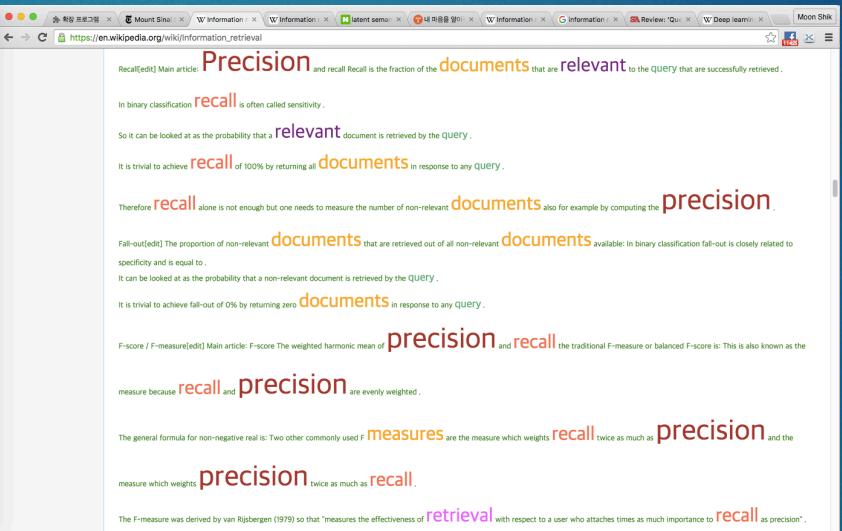
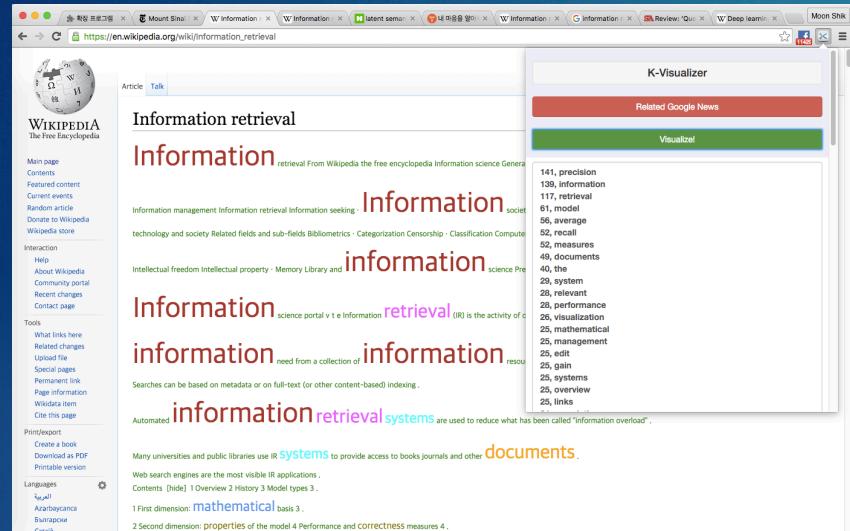
The F-measure was derived by van Rijsbergen (1979) so that "measures the effectiveness of retrieval with respect to a user who attaches times as much importance to recall as precision".

It is based on van Rijsbergen's effectiveness measure.

Their relationship is: where F-measure can be a better single metric when compared to precision and recall both precision and recall give different

109, information
87, retrieval
61, precision
49, documents
32, recall
29, system
28, relevant
25, systems
25, edit
25, management
23, query
22, software
22, computer
21, model
19, search
18, models
16, average
16, measure
16, document
16, ext
15, science
15, data
14, published
14, user
14, relevance
14, text
14, analysis
13, network
13, computing
13, evaluation
12, based
12, measures
11, conference
11, score
10, set
10, indexing
10, vector
9, terms
9, rank
9, machine

2. With tag weight(30,20,15)



141, precision
139, information
117, retrieval
61, model
56, average
52, recall
52, measures
49, documents
40, the
29, system
28, relevant
28, performance
26, visualization
25, mathematical
25, management
25, edit
25, gain
25, systems
25, overview
25, links
24, cumulative
24, history
24, discounted
23, reading
23, types
23, field
23, basis
23, references
23, query
23, external
23, properties
22, correctness
22, timeline
22, second
22, awards
22, software
22, computer
20, mean
20, and
20, also

3. With tag weight(50,20,15)

Information retrieval

Information retrieval From Wikipedia the free encyclopedia Information science General

Information management Information retrieval Information seeking

Information technology and society Related fields and sub-fields Bibliometrics - Categorization Censorship - Classification Computer

Intellectual freedom Intellectual property - Memory library and information

Information retrieval (IR) is the activity of

information need from a collection of

Automated information retrieval systems are used to reduce what has been called "information overload".

Many universities and public libraries use systems to provide access to books journals and other documents.

Web search engines are the most visible IR applications.

Contents [hide] 1 Overview 2 History 3 Model types 4

1 First dimension: mathematical basis 3

2 Second dimension: properties of the model 4 Performance and CORRECTNESS measures 4.

Precision and recall Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

In binary classification recall is often called sensitivity.

So it can be looked at as the probability that a relevant document is retrieved by the query.

It is trivial to achieve recall of 100% by returning all documents in response to any query.

Therefore recall alone is not enough but one needs to measure the number of non-relevant documents also for example by computing the precision.

Fall-out[edit] Main article: Fall-out The proportion of non-relevant documents that are retrieved out of all non-relevant documents available. In binary classification fall-out is closely related to specificity and is equal to .

It can be looked at as the probability that a non-relevant document is retrieved by the query.

It is trivial to achieve fall-out of 0% by returning zero documents in response to any query.

F-score / F-measure[edit] Main article: F-score The weighted harmonic mean of precision and recall

measure because recall and precision are evenly weighted.

The general formula for non-negative real is: Two other commonly used measures are the measure which weights recall twice as much as precision and the measure which weights precision twice as much as recall.

The F-measure was derived by van Rijsbergen (1979) so that "measures the effectiveness of retrieval with respect to a user who attaches twice as much importance to recall as precision".

159, information
141, precision
137, retrieval
61, model
56, average
52, recall
52, measures
49, documents
40, the
29, system
28, relevant
28, performance
26, visualization
25, mathematical
25, management
25, edit
25, gain
25, systems
25, overview
25, links
24, cumulative
24, history
24, discounted
23, reading
23, types
23, field
23, basis
23, references
23, query
23, external
23, properties
22, correctness
22, timeline
22, second
22, awards
22, software
22, computer
20, mean
20, and
20, also

4. With tag weight(30,12,10)

Information retrieval

Information retrieval is a field of study that deals with the problem of retrieving information from large collections of data. It is a sub-field of computer science and information science.

Information retrieval is concerned with the development of algorithms and systems for the retrieval of information from digital documents. It is used in a wide range of applications, including search engines, document management systems, and information retrieval systems.

The term "information retrieval" was first coined by Peter M. Morris in 1968. Since then, the field has grown significantly, with many new techniques and applications being developed.

Information retrieval is often used in conjunction with other fields, such as machine learning and natural language processing, to improve the performance of information retrieval systems.

Information retrieval is a complex field, and there are many challenges involved in developing effective information retrieval systems. One of the main challenges is dealing with the large amounts of data that are available, and ensuring that the retrieved information is relevant and useful to the user.

Information retrieval is a rapidly growing field, and there is a great deal of research and development currently underway.

Information retrieval

Information retrieval is a field of study that deals with the problem of retrieving information from large collections of data. It is a sub-field of computer science and information science.

Information retrieval is concerned with the development of algorithms and systems for the retrieval of information from digital documents. It is used in a wide range of applications, including search engines, document management systems, and information retrieval systems.

The term "information retrieval" was first coined by Peter M. Morris in 1968. Since then, the field has grown significantly, with many new techniques and applications being developed.

Information retrieval is often used in conjunction with other fields, such as machine learning and natural language processing, to improve the performance of information retrieval systems.

Information retrieval is a complex field, and there are many challenges involved in developing effective information retrieval systems. One of the main challenges is dealing with the large amounts of data that are available, and ensuring that the retrieved information is relevant and useful to the user.

Information retrieval is a rapidly growing field, and there is a great deal of research and development currently underway.

139, information
117, retrieval
109, precision
49, documents
45, model
44, recall
40, average
36, measures
29, system
28, relevant
25, edit
25, management
25, systems
24, the
23, query
22, software
20, computer
20, performance
19, search
18, models
18, visualization
17, gain
17, mathematical
17, overview
17, links
16, measure
16, discounted
16, ext
16, history
16, document
16, cumulative
15, external
15, reading
15, references
15, field
15, science
15, properties
15, basis
15, types
15, data

5. With distance weight(3)

The screenshot shows a Wikipedia article on 'Information retrieval'. A sidebar on the right contains a 'K-Visualizer' tool. The visualization displays the frequency of various terms related to information retrieval. The top of the chart lists: 127, information; 91, retrieval; 61, precision; 40, documents; 32, recall; 29, system; 28, relevant; 27, management; 25, edit; 25, systems; 23, query; 22, software; 22, computer; 21, search; 21, model; 16, models; 17, relevance; 16, document; 16, extrema; and 16, average.

The screenshot shows the same Wikipedia page for 'Information retrieval'. The content focuses on the concepts of precision and recall. It defines precision as the fraction of retrieved documents that are relevant to the query, and recall as the fraction of all relevant documents that are retrieved. The text also discusses the F-score, which is the weighted harmonic mean of precision and recall. The page includes several links to other related topics like 'F-measure', 'IR applications', and 'Model types'.

127, information
91, retrieval
61, precision
49, documents
32, recall
29, system
28, relevant
27, management
25, edit
25, systems
23, query
22, software
22, computer
21, search
21, model
18, models
17, science
16, document
16, ext
16, average
16, measure
15, data
14, analysis
14, published
14, relevance
14, user
14, text
13, network
13, computing
13, evaluation
12, measures
12, based
11, conference
11, score
10, indexing
10, society
10, set
10, vector
9, terms
9, queries

6. With distance weight(5)

The screenshot shows a search results page for "information retrieval" on Wikipedia. The top navigation bar includes tabs for Article, Talk, and Related Google News. The main content area features a large image of the Earth from space. Below the image, the title "Information retrieval" is displayed, followed by a summary: "retrieval From Wikipedia the free encyclopedia Information science General information management Information retrieval Information seeking". A large red box highlights the word "Information". To the right, there is a sidebar titled "K-Visualizer" with a "Visualized" button. The search results list includes terms such as "information", "retrieval", "relevant", "indexing", "systems", and "documents". A sidebar on the right lists related topics like "information", "retrieval", "relevant", "indexing", "systems", and "documents".

Recall[edit] Main article: **Precision** and recall Recall is the fraction of the **documents** that are **relevant** to the **query** that are successfully retrieved.

In binary classification **recall** is often called sensitivity.

So it can be looked at as the probability that a **relevant document** is retrieved by the **query**.

It is trivial to achieve **recall** of 100% by returning all **documents** in response to any **query**.

Therefore **recall** alone is not enough but one needs to **measure** the number of non-relevant **documents** also for example by **computing** the **precision**.

Fall-out[edit] The proportion of non-relevant **documents** that are retrieved out of all non-relevant **documents** available: In binary classification fall-out is closely related to specificity and is equal to .

It can be looked at as the probability that a non-relevant **document** is retrieved by the **query**.

It is trivial to achieve fall-out of 0% by returning zero **documents** in response to any **query**.

F-score / F-measure[edit] Main article: F-Score The weighted harmonic mean of **precision** and **recall**: the traditional F-measure or balanced F-score is: This is also known as the **measure** because **recall** and **precision** are evenly weighted.

The general formula for non-negative real is: Two other commonly used F **measures** are the **measure** which weights **recall** twice as much as **precision** and the **measure** which weights **precision** twice as much as **recall**.

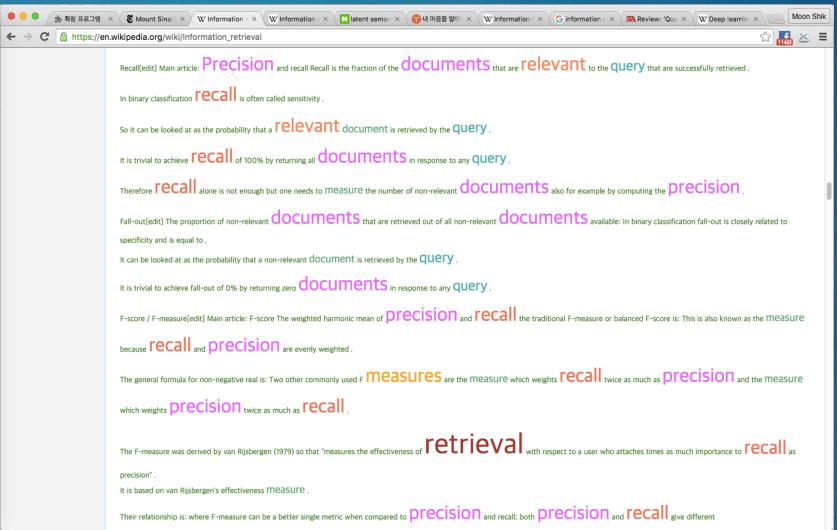
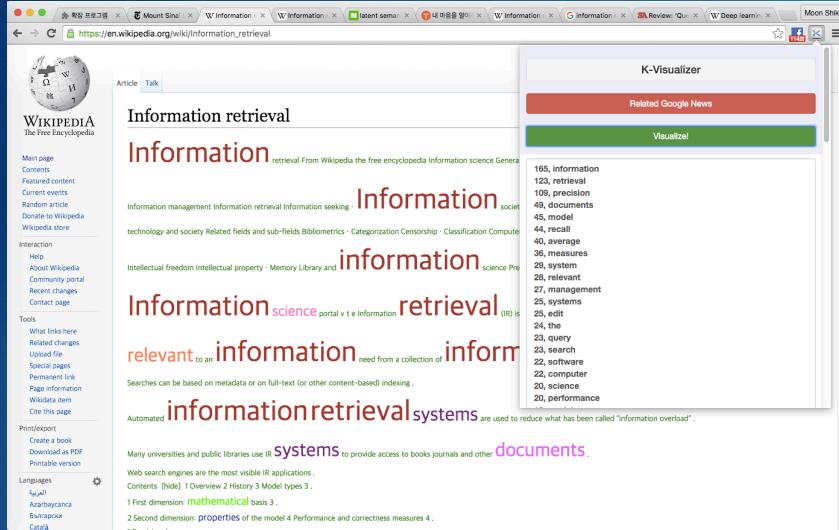
The F-measure was derived by van Rijsbergen (1979) so that "measures the effectiveness of retrieval with respect to a **USER** who attaches twice as much importance to **recall** as precision".

It is based on van Rijsbergen's effectiveness **measure**.

This relationship in which F-measures can be a better single metric when compared to **precision** and recall both **precision** and **recall** are different

145, information
95, retrieval
61, precision
49, documents
32, recall
29, management
29, system
28, relevant
25, systems
25, edit
23, query
23, search
22, computer
22, software
21, model
19, science
18, models
16, average
16, document
16, ext
16, measure
15, data
14, text
14, user
14, published
14, analysis
14, relevance
13, network
13, evaluation
13, computing
12, society
12, measures
12, based
11, score
11, conference
10, knowledge
10, set
10, vector
10, wikipedia
10, indexing

7. With tag and distance weight (30, 12, 10, 5, 3)



165, information
123, retrieval
109, precision
49, documents
45, model
44, recall
40, average
36, measures
29, system
28, relevant
27, management
25, systems
25, edit
24, the
23, query
23, search
22, software
22, computer
20, science
20, performance
18, models
18, visualization
17, overview
17, links
17, mathematical
17, gain
16, measure
16, cumulative
16, document
16, ext
16, discounted
16, history
15, reading
15, field
15, properties
15, basis
15, types
15, references
15, data
15, external

Analysis with AHP(Analytic Hierarchy Process)

(1) 가중치 산정 결과

	Factor 01	Factor 02	Factor 03	Factor 04	Factor 05	Factor 06	Factor 07
Weight	0.129	0.131	0.127	0.167	0.133	0.136	0.176

(2) 비교 행렬

	Factor 01	Factor 02	Factor 03	Factor 04	Factor 05	Factor 06	Factor 07
Factor 01		0.802	1.213	0.711	0.899	1.139	0.702
Factor 02			0.918	0.759	0.942	1.049	0.669
Factor 03				0.847	0.91	0.857	0.787
Factor 04					1.299	1.176	0.961
Factor 05						0.882	0.752
Factor 06							0.806
Factor 07							

Weight of Factor 07 is largest

→ tag and distance weight increased satisfaction level

Best combination?

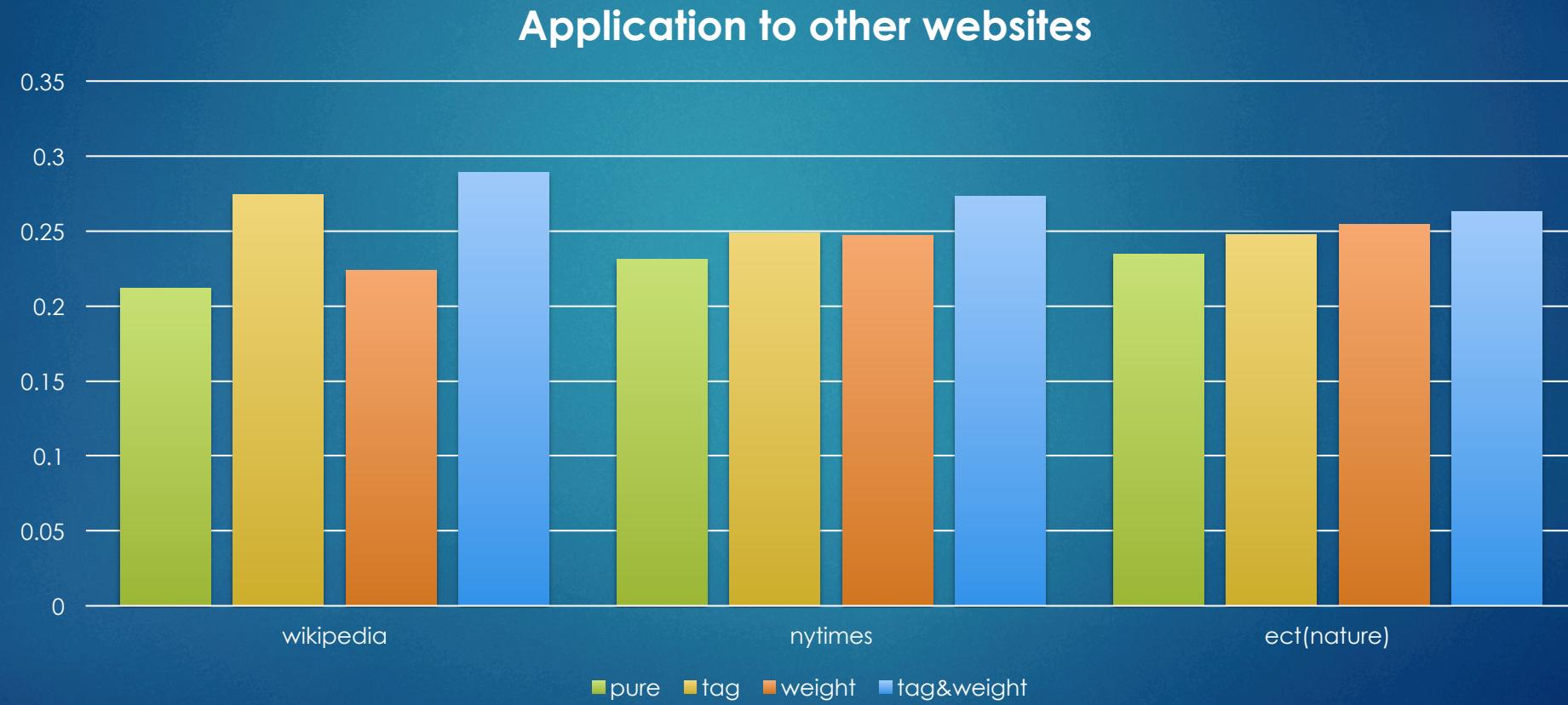
With Meta Tag And Distance Weighting

Precision and recall. Recall is the fraction of the documents that are relevant to the query that are successfully retrieved. In binary classification recall is often called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore recall alone is not enough but one needs to measure the number of non-relevant documents also for example by computing the precision. Fall-out[edit] The proportion of non-relevant documents that are retrieved out of all non-relevant documents available: In binary classification fall-out is closely related to specificity and is equal to . It can be looked at as the probability that a non-relevant document is retrieved by the query. It is trivial to achieve fall-out of 0% by returning zero documents in response to any query. F-score / F-measure[edit] Main article: F-score The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is: This is also known as the measure because recall and precision are evenly weighted. The general formula for non-negative real is: Two other commonly used F measures are the measure which weights recall twice as much as precision and the measure which weights precision twice as much as recall. retrieval The F-measure was derived by van Rijsbergen (1979) so that "measures the effectiveness of retrieval with respect to a user who attaches times as much importance to recall as precision". It is based on van Rijsbergen's effectiveness measure. Their relationship is: where F-measure can be a better single metric when compared to precision and recall; both precision and recall give different

Without Meta Tag And Distance Weighting

Precision and recall. Recall is the fraction of the documents that are relevant to the query that are successfully retrieved. In binary classification recall is often called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore recall alone is not enough but one needs to measure the number of non-relevant documents also for example by computing the precision. Fall-out[edit] The proportion of non-relevant documents that are retrieved out of all non-relevant documents available: In binary classification fall-out is closely related to specificity and is equal to . It can be looked at as the probability that a non-relevant document is retrieved by the query. It is trivial to achieve fall-out of 0% by returning zero documents in response to any query. F-score / F-measure[edit] Main article: F-score The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is: This is also known as the measure because recall and precision are evenly weighted. The general formula for non-negative real is: Two other commonly used F measures are the measure which weights recall twice as much as precision and the measure which weights precision twice as much as recall. retrieval The F-measure was derived by van Rijsbergen (1979) so that "measures the effectiveness of retrieval with respect to a user who attaches times as much importance to recall as precision". It is based on van Rijsbergen's effectiveness measure. Their relationship is: where F-measure can be a better single metric when compared to precision and recall; both precision and recall give different

Comparitive Result



Wikipedia is ..

Not logged in Talk Contributions Create account Log in

Deep learning

From Wikipedia, the free encyclopedia

For deep versus shallow learning in educational psychology, see [Student approaches to learning](#)

Deep learning (*deep machine learning*, or *deep structured learning*, or *hierarchical learning*, or sometimes *DL*) is a branch of [machine learning](#) based on a set of [algorithms](#) that attempt to model high-level abstractions in data by using multiple processing layers with complex structures, or otherwise composed of multiple non-linear transformations.

Deep learning is part of a broader family of [machine learning](#) methods based on [learning representations](#) of data. An observation (e.g., an image) can be represented in many ways such as a [vector](#) of intensity values per pixel, or in a more abstract way as a set of edges, regions of particular shape, etc. Some representations make it easier to learn tasks (e.g., face recognition or facial expression recognition^[6]) from examples.

Machine learning and data mining

Problems

- Classification · Clustering · Regression ·
- Anomaly detection · Association rules ·
- Reinforcement learning · Structured prediction ·
- Feature engineering · Feature learning ·
- Online learning · Semi-supervised learning ·
- Unsupervised learning · Learning to rank ·
- Grammar induction

Supervised learning
(classification · regression)

Decision trees · Ensembles (Bagging, Boosting, Random forest) · k -NN · Linear regression · Naive Bayes · Neural networks · Logistic regression · Perceptron · Relevance vector machine (RVM) · Support vector machine (SVM)

Clustering

PIRGH: Unsupervised / machine learning

Elements Network Sources Timeline Profiles »

height:1.4em;font-size:88%>...</table>

<p>Deep learning
" ("<i>deep machine learning</i>" , or "<i>deep structured learning</i>" , or "<i>hierarchical learning</i>" , or sometimes "<i>DL</i>") is a branch of "machine learning" based on a set of "algorithms" that attempt to model high-level abstractions in data by using multiple processing layers with complex structures, or otherwise composed of multiple non-linear transformations" ."

>^{>...}</sup>

html body #content #bodyContent div#mw-content-text.mw-content-ltr p a

Styles Event Listeners DOM Breakpoints Properties

Filter

element.style { }

@media screen a:visited { color: #0b0080; }

@media screen a { text-decoration: none; color: #0645ad; background: none; }

a:-webkit-any-link { user agent stylesheet color: -webkit-link; text-decoration: underline; cursor: auto; }

Inherited from p

margin - border - padding - auto x auto - - -

background-attachment: scroll; background-clip: border-box; background-color: #rgba(0, 0, 0, 0); background-image: none;

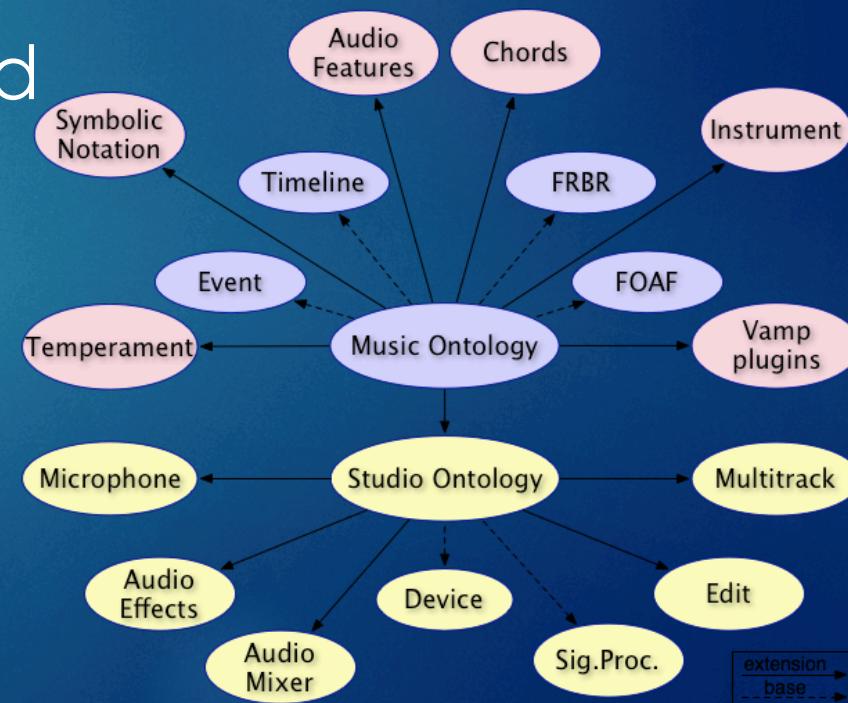
Show inherited

Future Works



Feature4 – Similarity Weighting

- ▶ Weight based on ‘Ontology based Similarity score’
 - 1) Extract words from innerText of <title>
 - 2) Measure similarity score of selected words with words in the text
 - 3) Add weight on the words with high similarity score





Thank you!

Reference

1. Ontology based Similarity Measure in Document Ranking

1. <http://disi.unitn.it/~p2p/RelatedWork/Matching/pxc387774.pdf>

2. Hypertext Categorization using Hyperlink Patterns and Meta Data

1. <http://users.softlab.ntua.gr/facilities/public/AD/Text%20Categorization/ghani01hypertext.pdf>

3. Mining “Hidden Phrase” Definitions from the Web

4. ‘Best Keyword Extraction Algorithms’ – Quora

1. <https://www.quora.com/What-are-the-best-keyword-extraction-algorithms-for-natural-language-processing-and-how-can-they-be-implemented-in-Python>

Reference

5. Survey of keyword extraction techniques

<http://www.cs.unm.edu/~pdevineni/papers/Lott.pdf>

6. HTML tags as Extraction Cues for Web Page Description Construction

1. <http://inform.nu/Articles/Vol6/v6p001-012.pdf?q=powerful-writing-in-30-words-or-less>