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)

Implementation

Architecture

Smartphone

From Wikipedia, the free encyclopedia

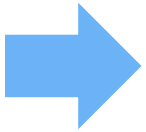
"Smartphones" redirects here. For the song by Trey Songz, see SmartPhones (song).
Not to be confused with Feature phone.


An Apple iPhone 5S

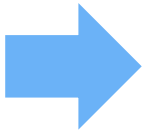
A **smartphone** or **smart phone** is a mobile phone with an advanced mobile operating system which combines features of a personal computer operating system with other features useful for mobile or handheld use.[1][2][3] They typically combine the features of a cell phone with those of other popular 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.

Keyword Visualizer

Web crawler

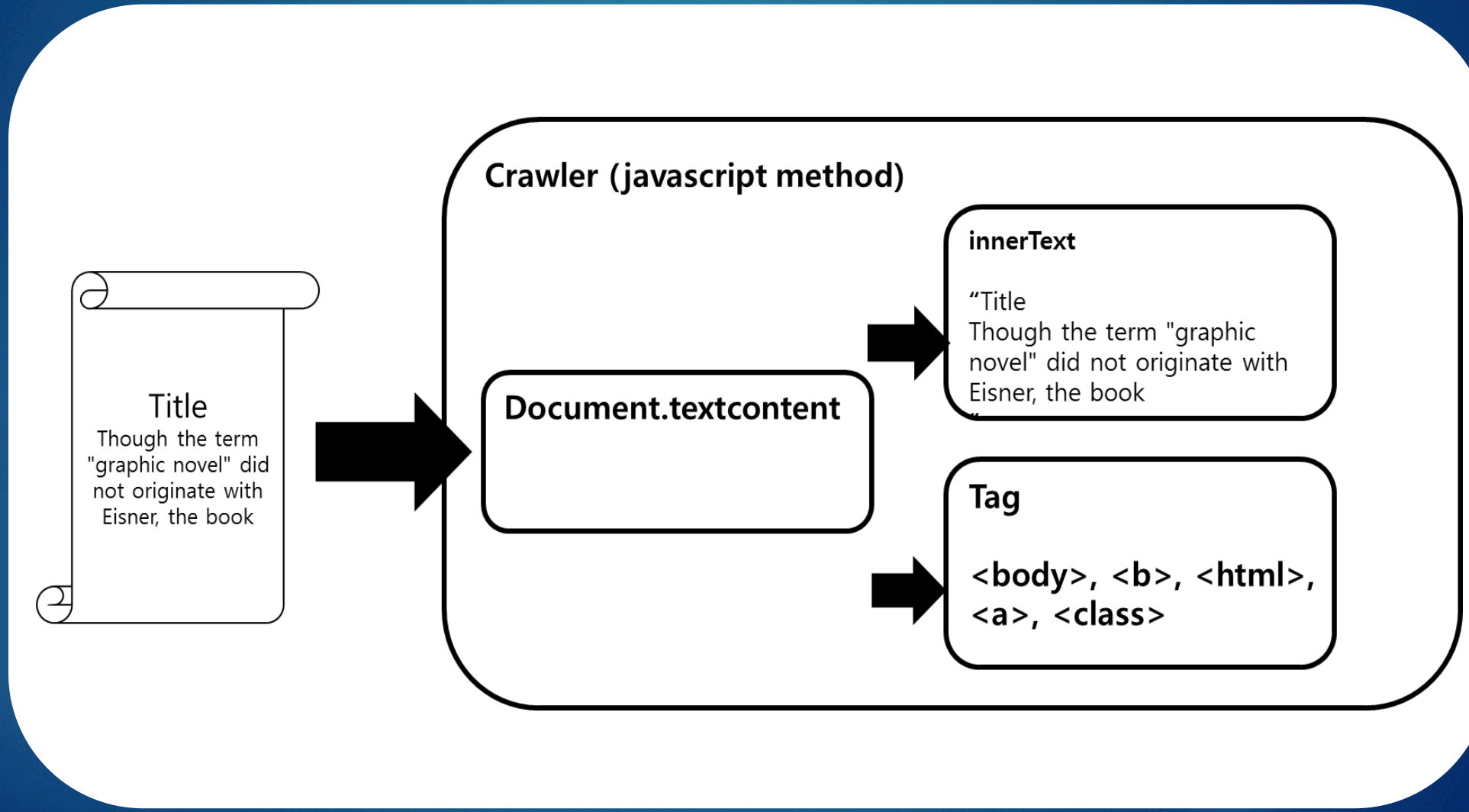


Bag-of-word generator

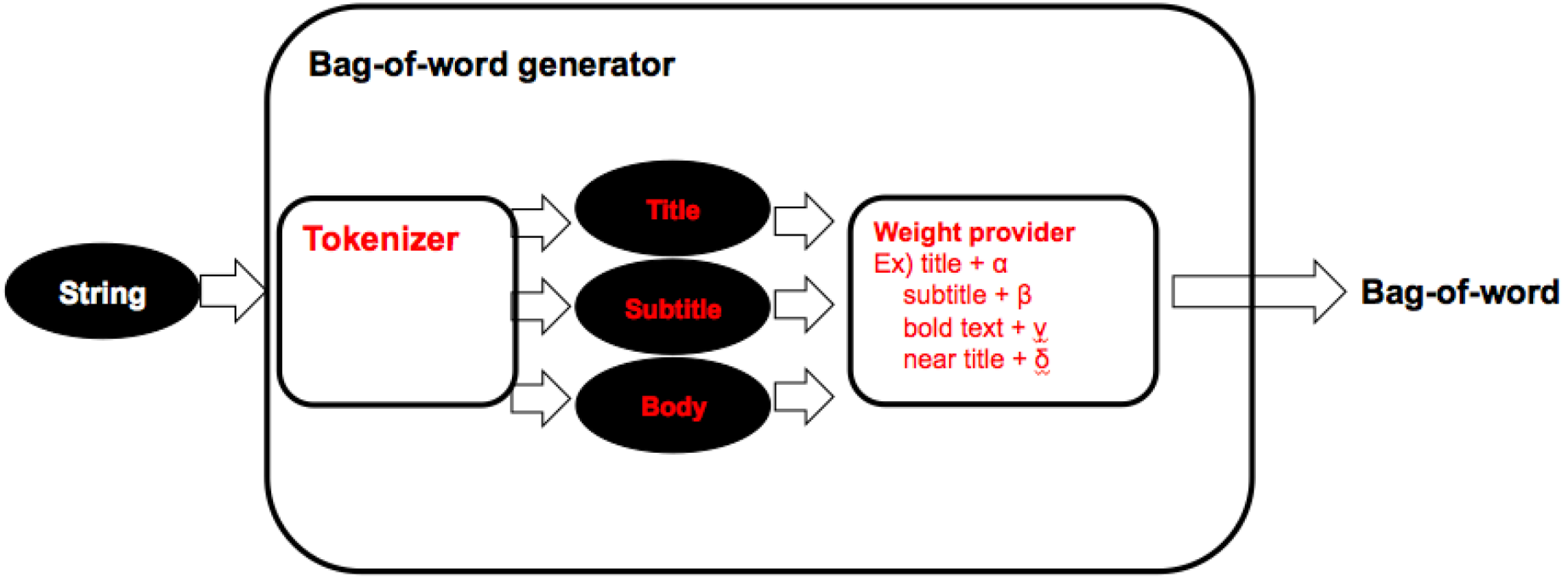


Visualizer

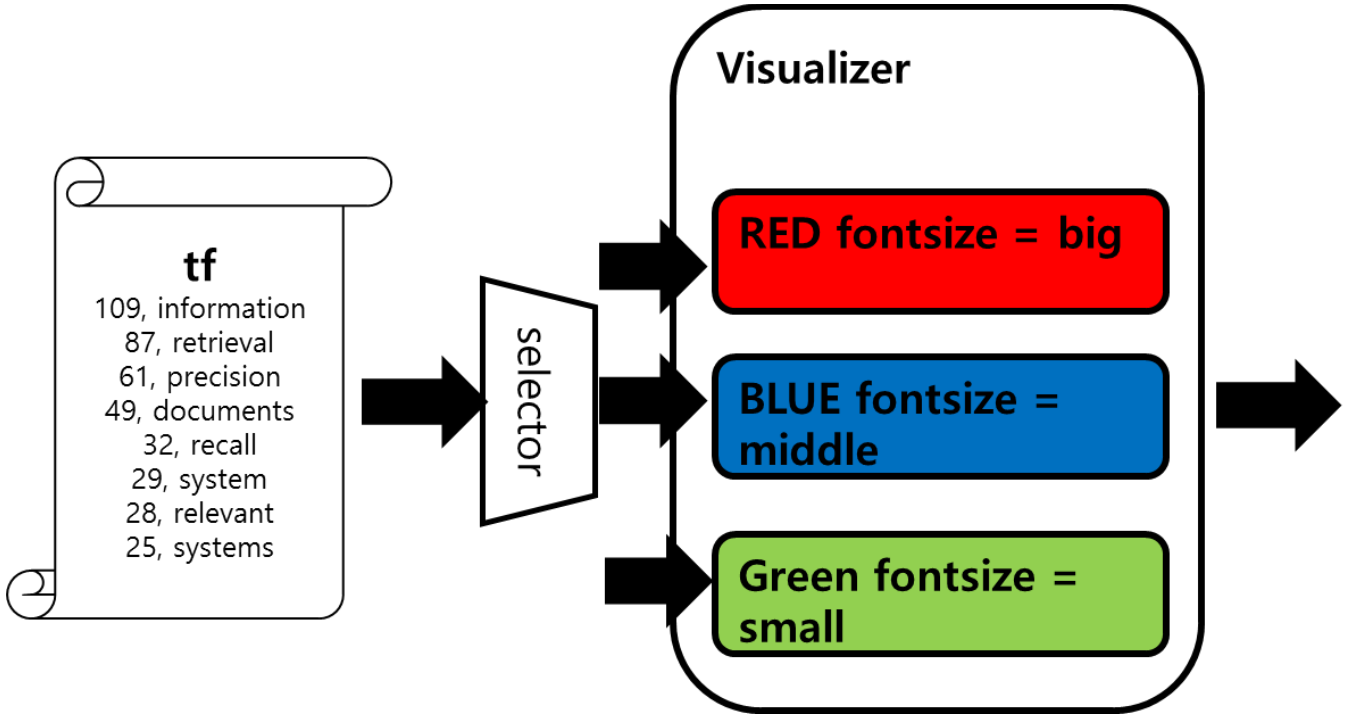
Architecture



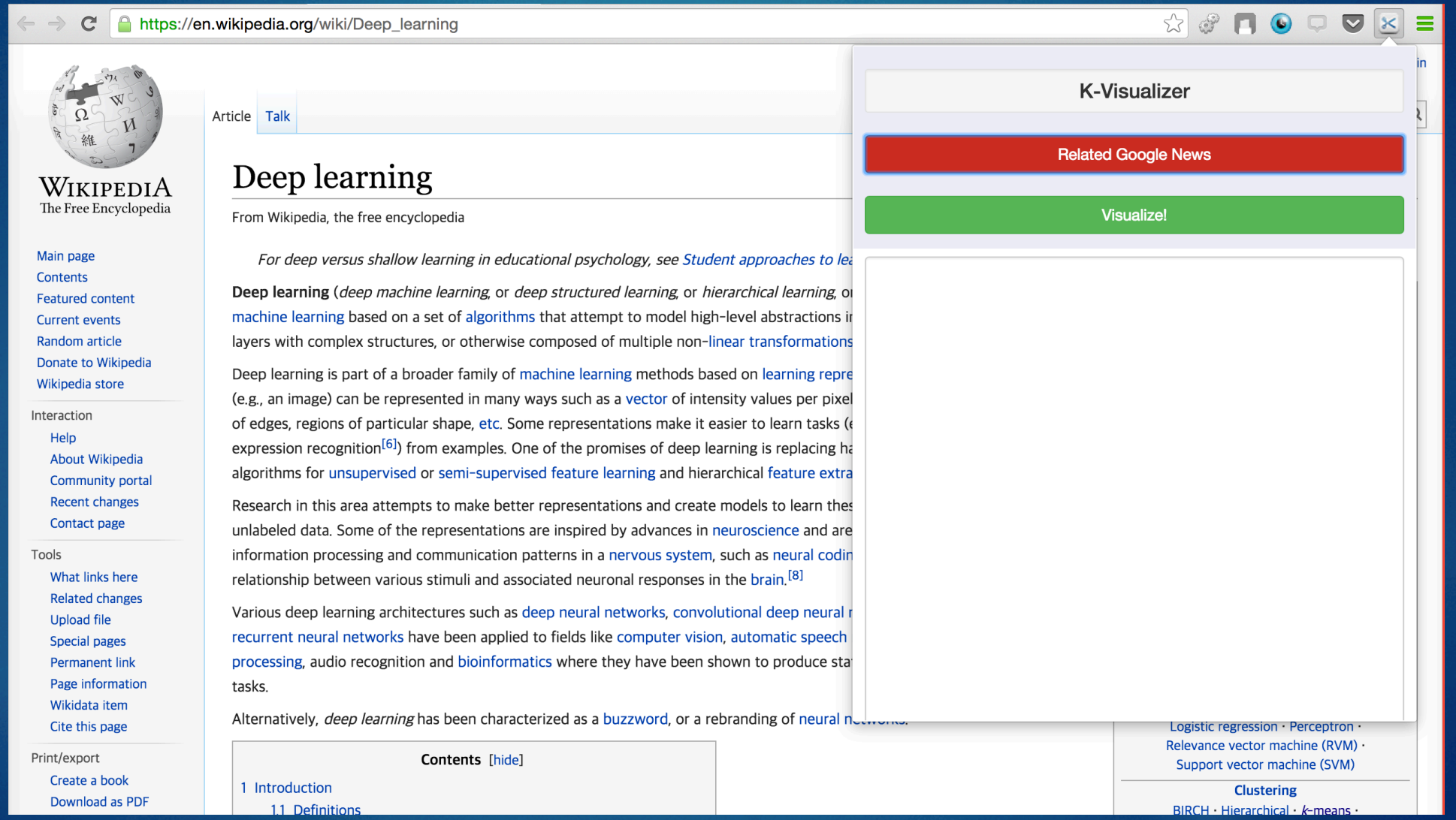
Architecture



Architecture



Chrome Plugin



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

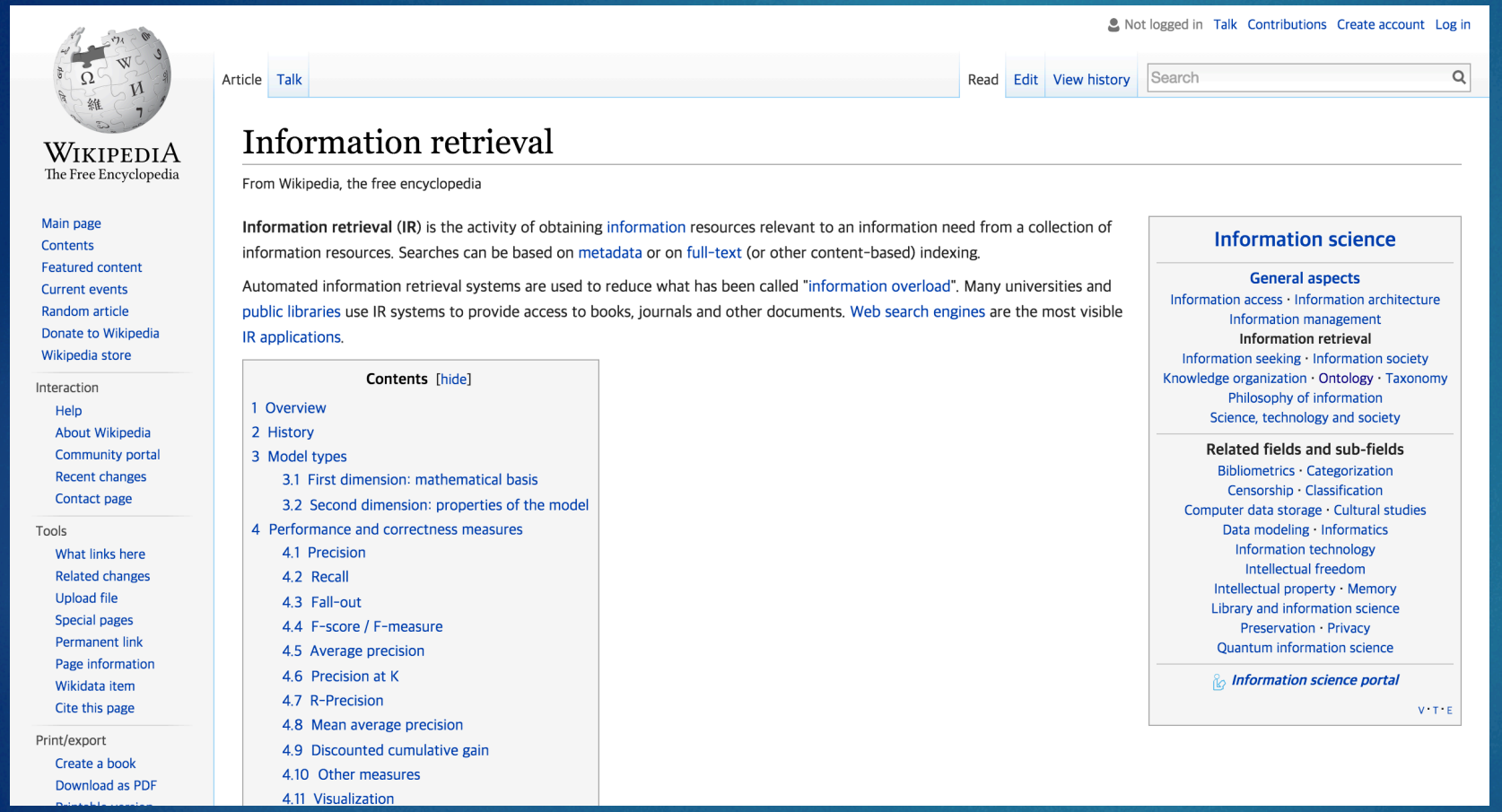
   (Remove punctuations and stop words)

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

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

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

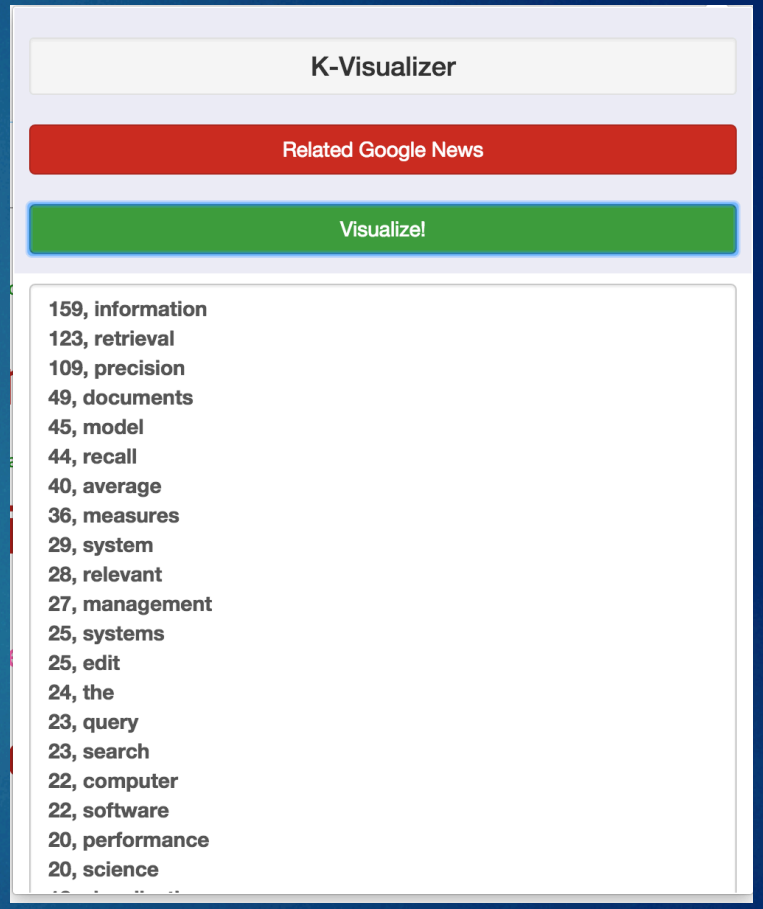
Information science

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

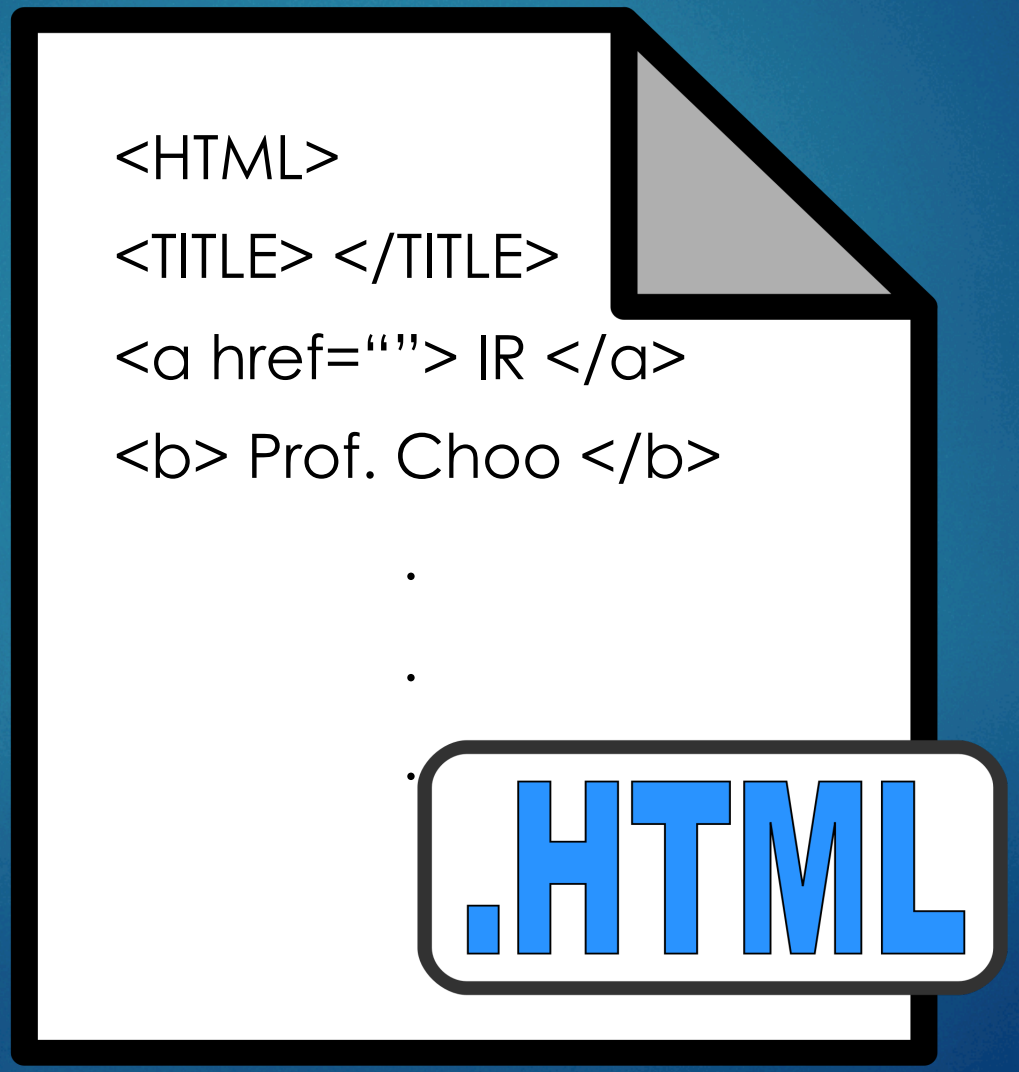
V · T · E



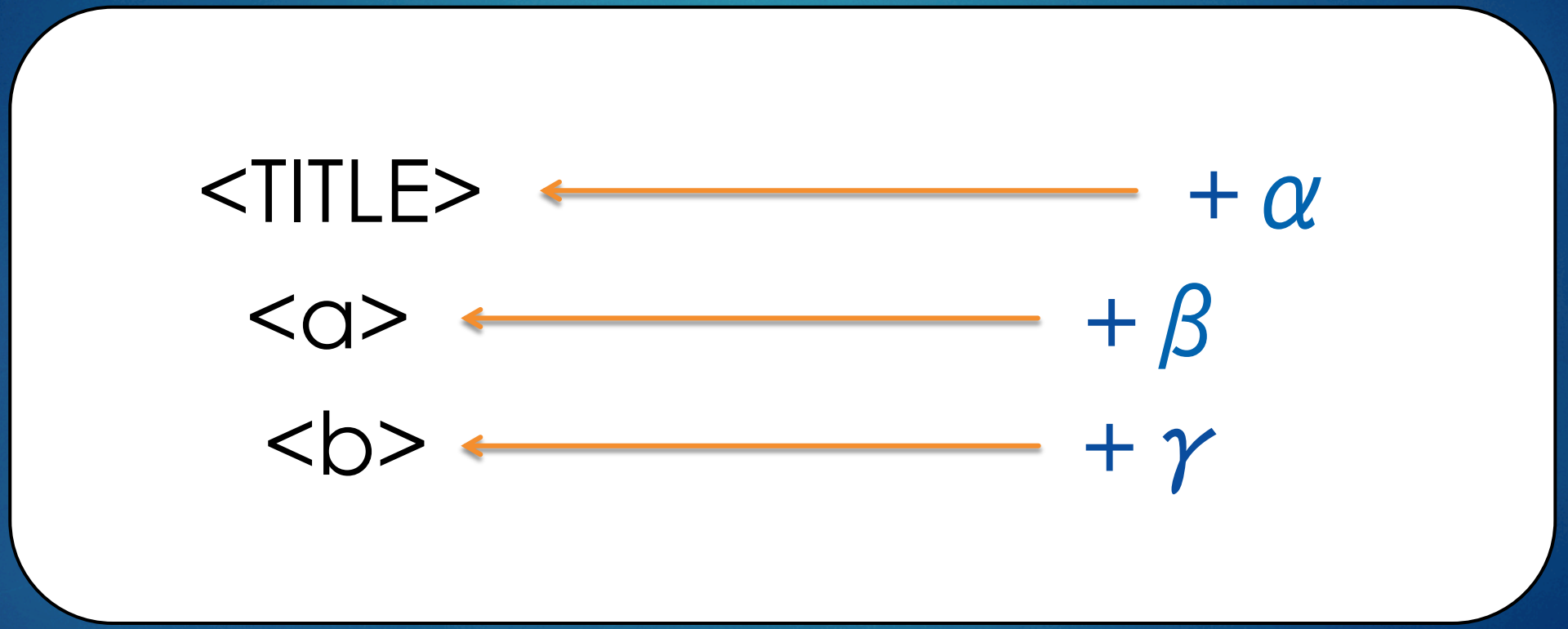
1 appearance = weight 1

Feature2 – Meta Tag Weighting

- HTML



Feature2 – Meta Tag Weighting



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

H. Liu[1], and

ulcu,

earch on
l titles of
lacement
e META
ithm that

Informing Science Journal | Information Sciences | Volume 6, 2003

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>,

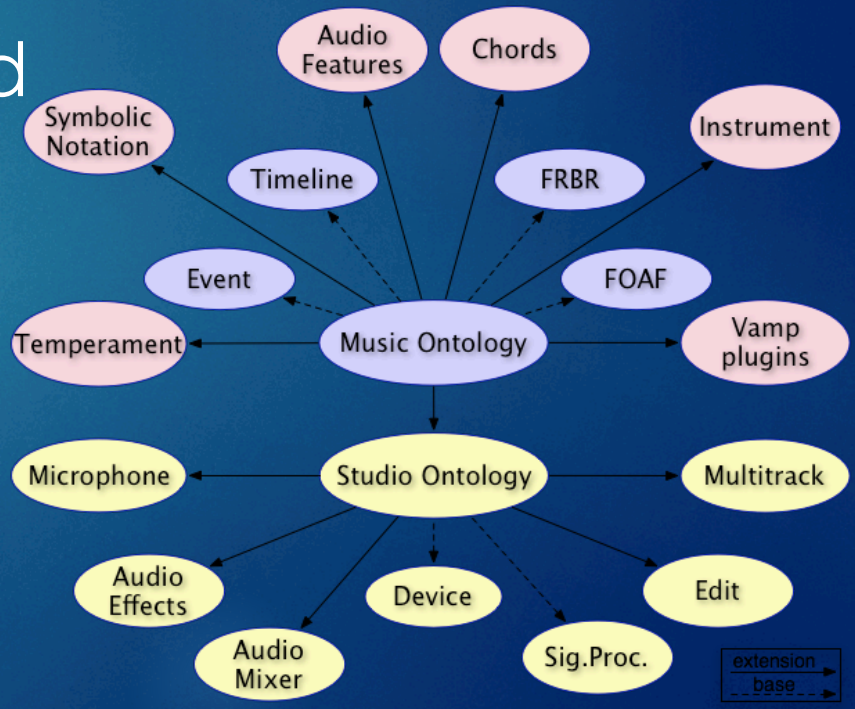
i.e.

$$|x - <title>| < \varepsilon$$

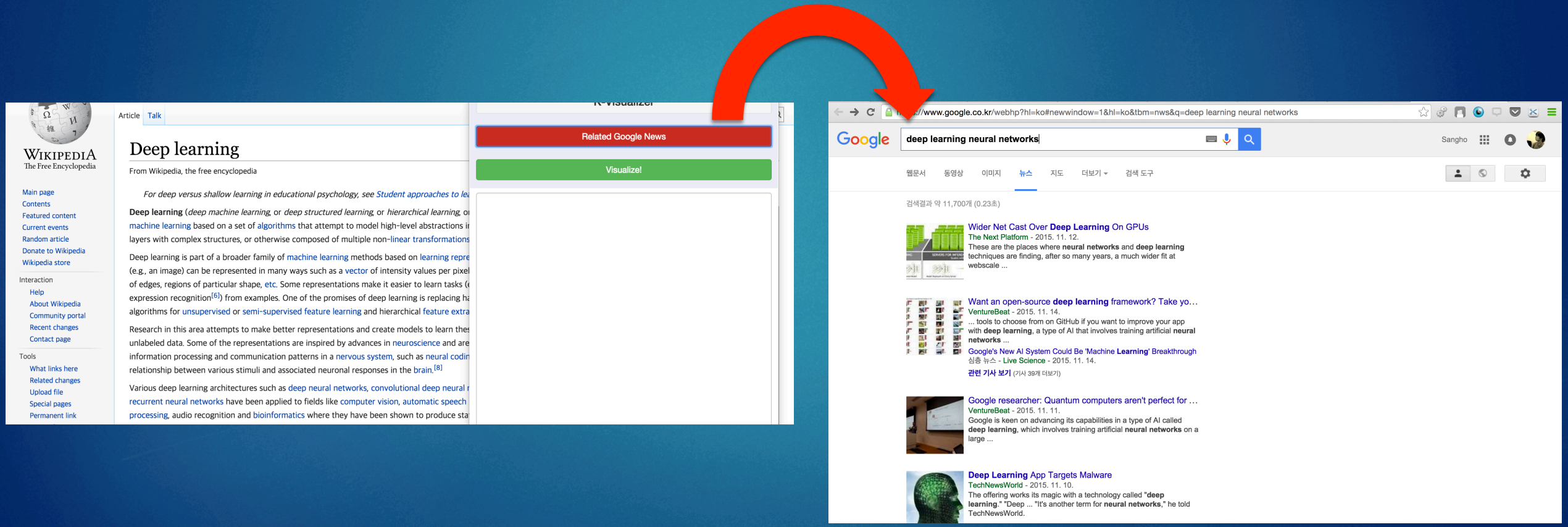
where $|x - <title>|$ is distance from <title> tag

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

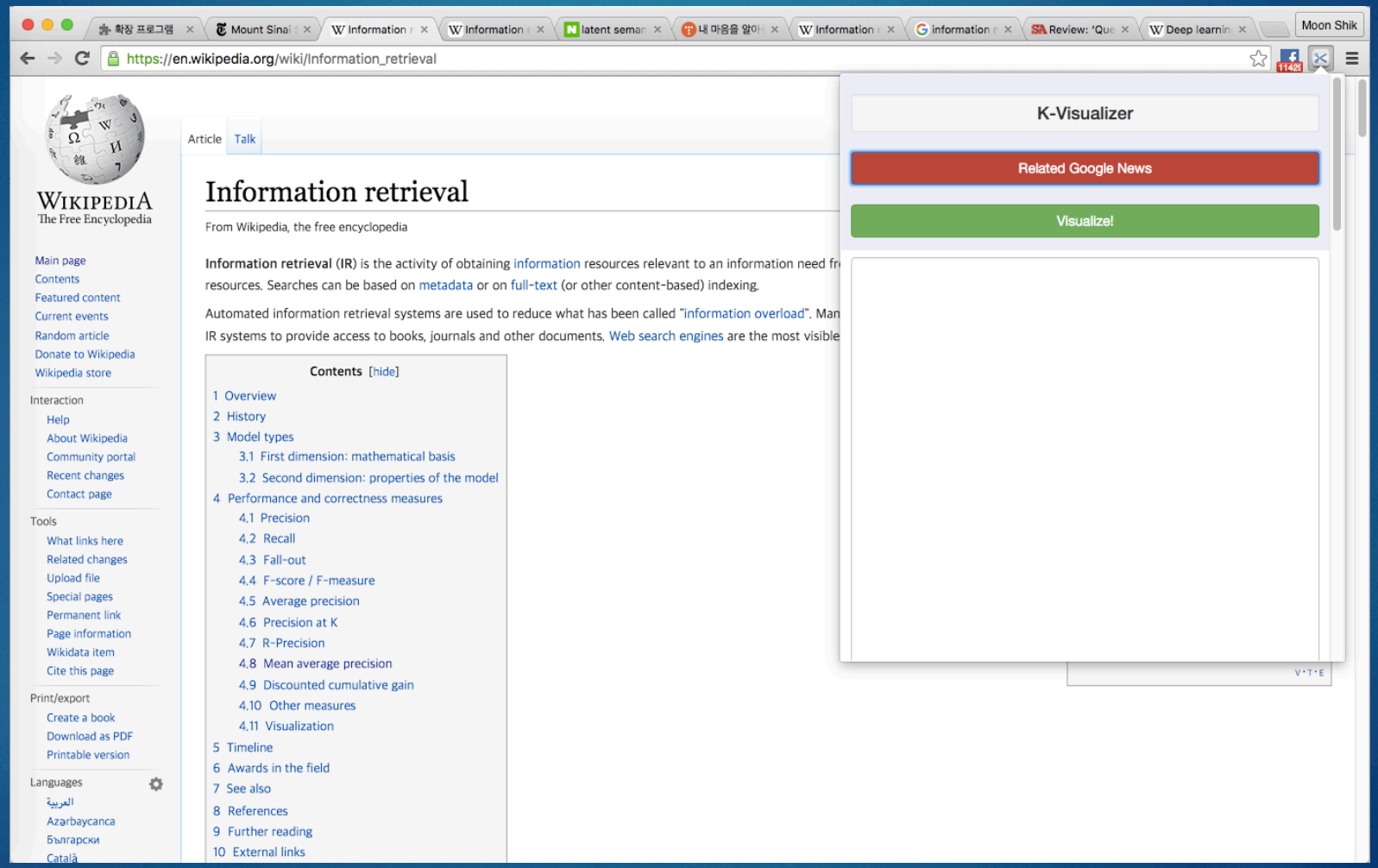


Feature 5 – Google News Recommendation

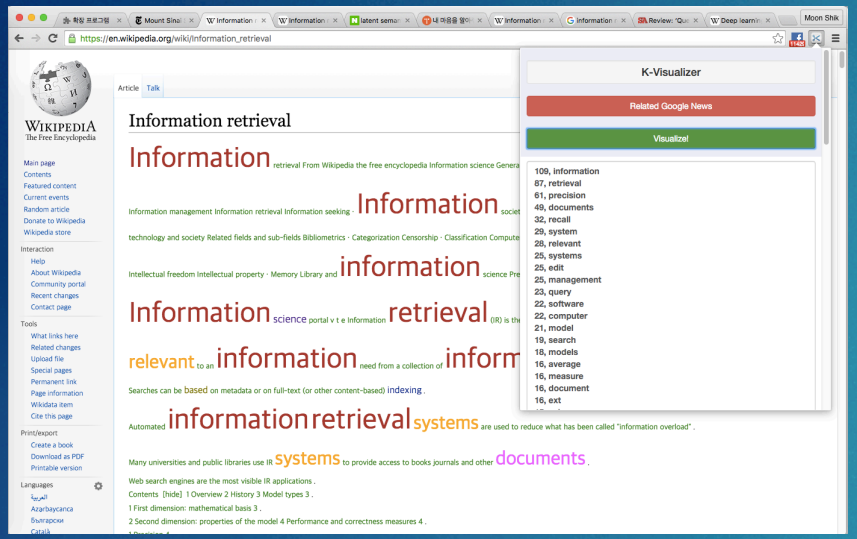


Evaluation

Test K-Visualizer

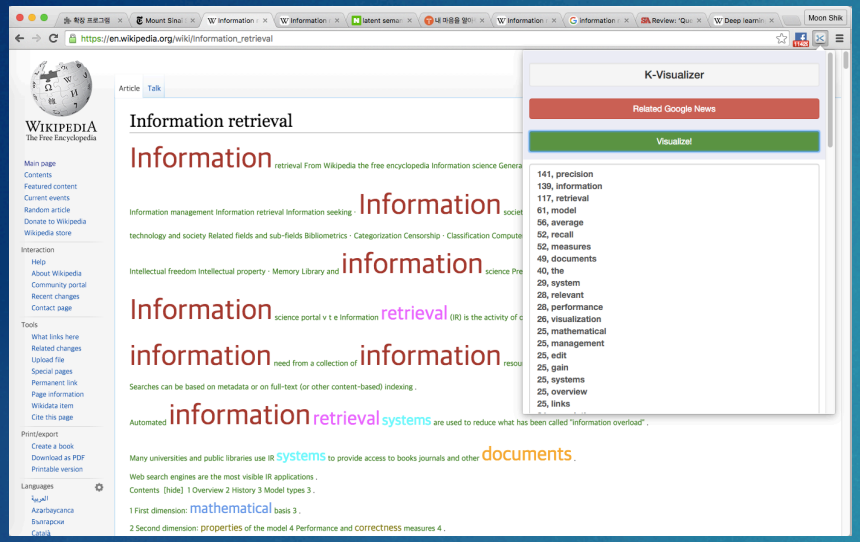


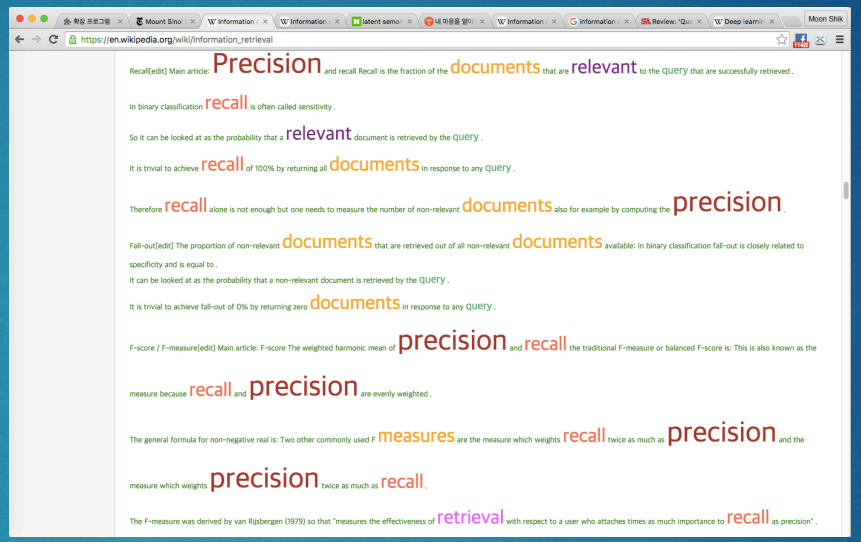
1. Only with term frequency



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

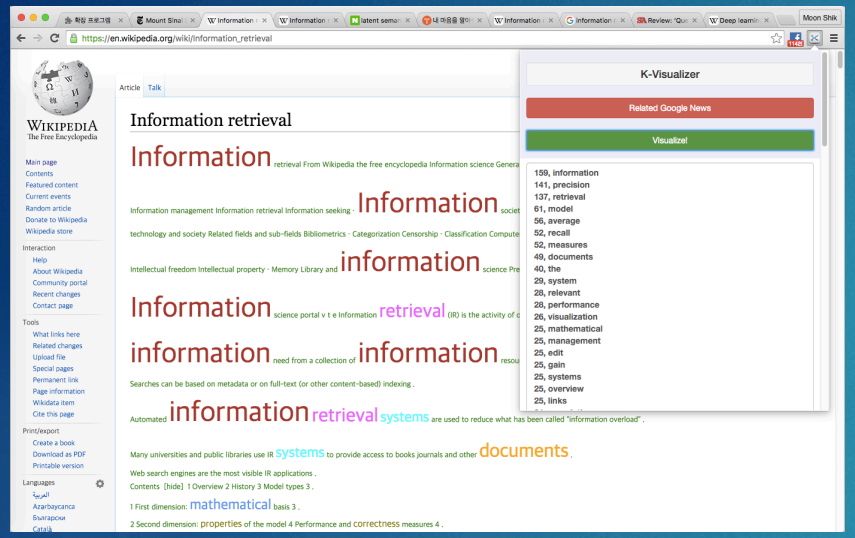
2. With tag weight(30,20,15)

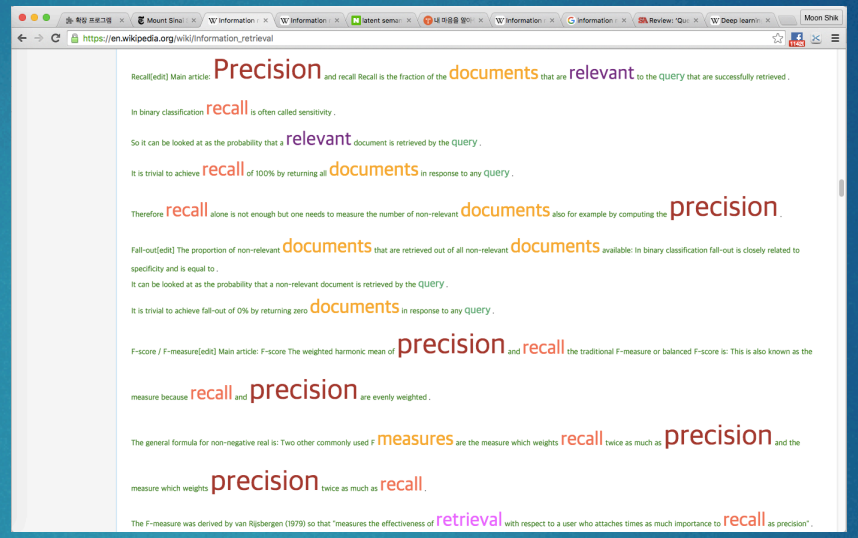




141, precision
139, information
117, retrieval
61, model
56, average
52, recall
52, measures
49, documents
40, the
29, system
28, relevant
28, performance
26, visualization
25, mathematical
25, management
25, edit
25, gain
25, systems
25, overview
25, links
24, cumulative
24, history
24, discounted
23, 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

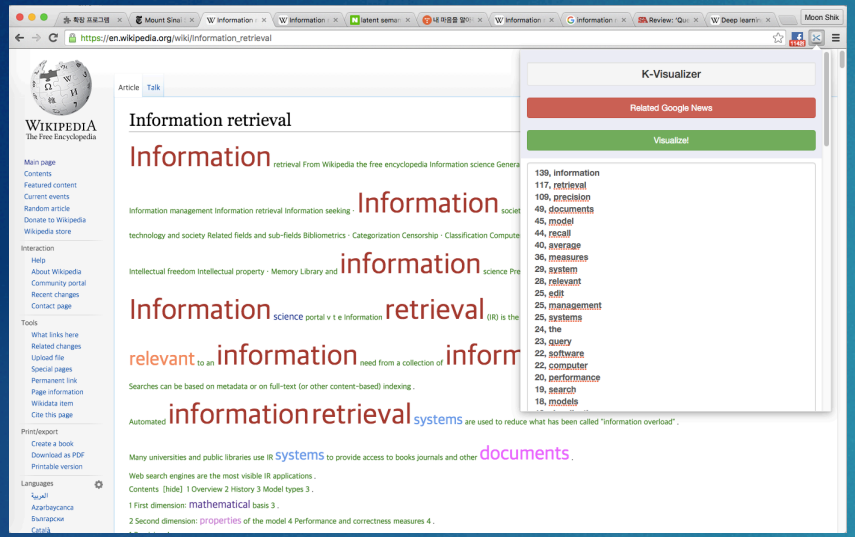
3. With tag weight(50,20,15)

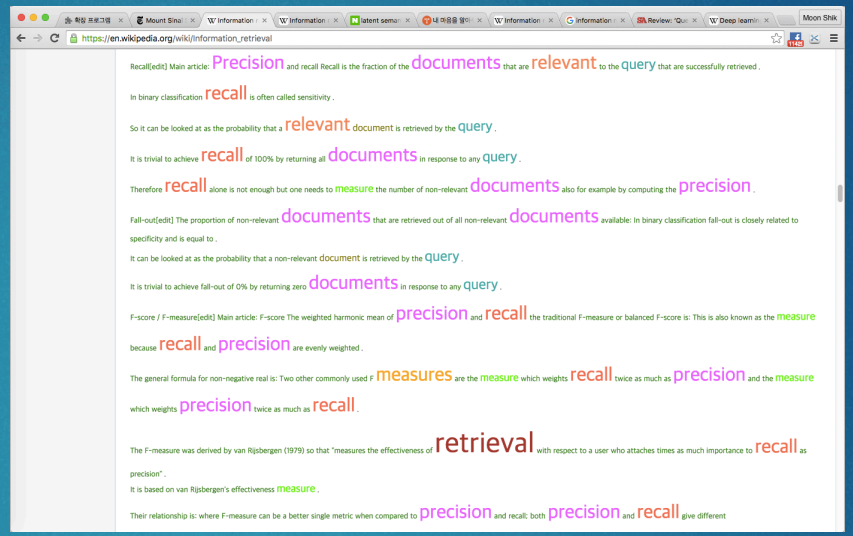




159, information
141, precision
137, retrieval
61, model
56, average
52, recall
52, measures
49, documents
40, the
29, system
28, relevant
28, performance
26, visualization
25, mathematical
25, management
25, edit
25, gain
25, systems
25, overview
25, links
24, cumulative
24, history
24, discounted
23, 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

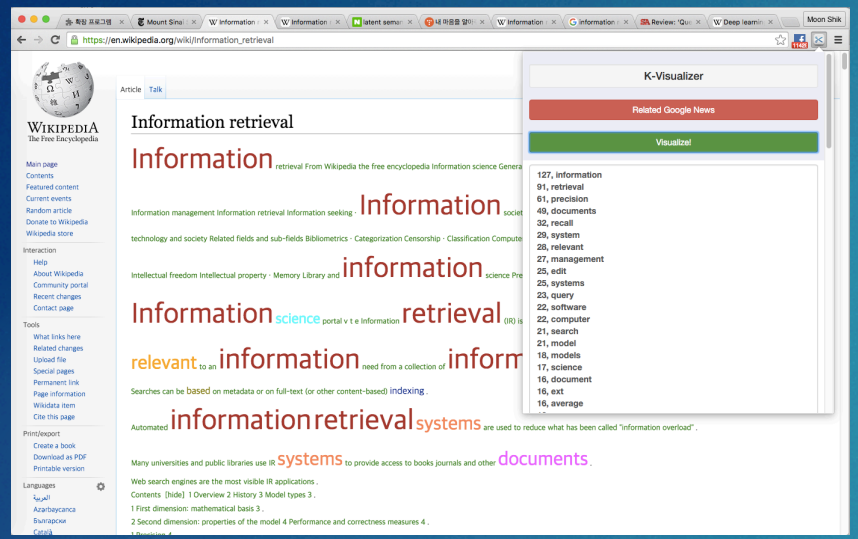
4. With tag weight(30,12,10)





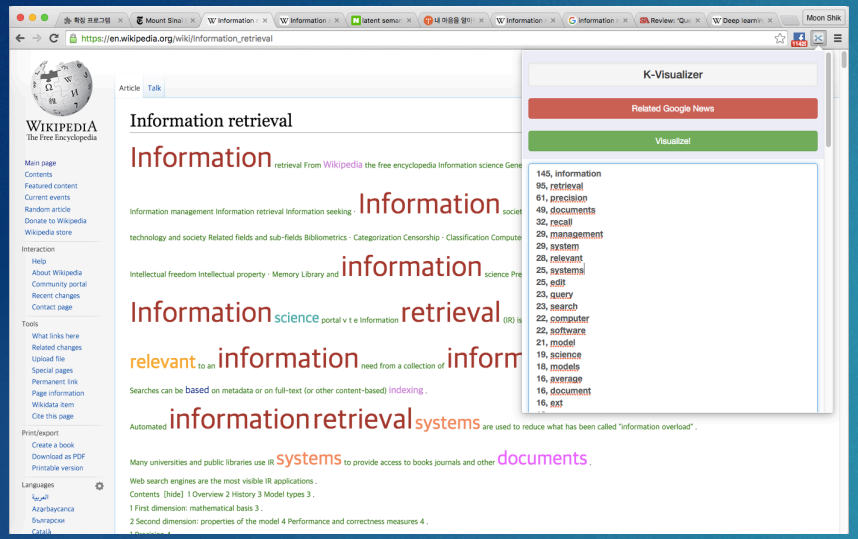
139, information
117, retrieval
109, precision
49, documents
45, model
44, recall
40, average
36, measures
29, system
28, relevant
25, edit
25, management
25, systems
24, the
23, query
22, software
22, computer
20, performance
19, search
18, models
18, visualization
17, gain
17, mathematical
17, overview
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)



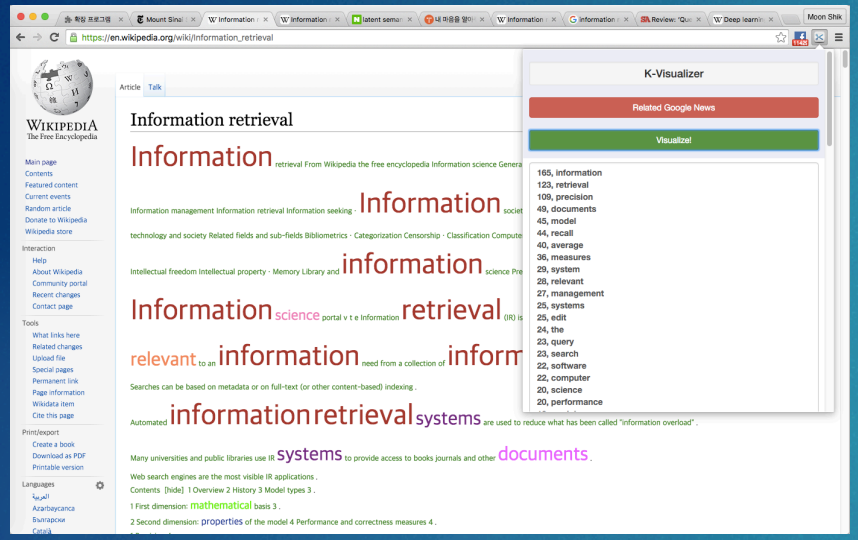
127, information
91, retrieval
61, precision
49, documents
32, recall
29, system
28, relevant
27, management
25, edit
25, systems
23, query
22, software
22, computer
21, search
21, model
18, models
17, science
16, document
16, ext
16, average
16, measure
15, data
14, analysis
14, published
14, relevance
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

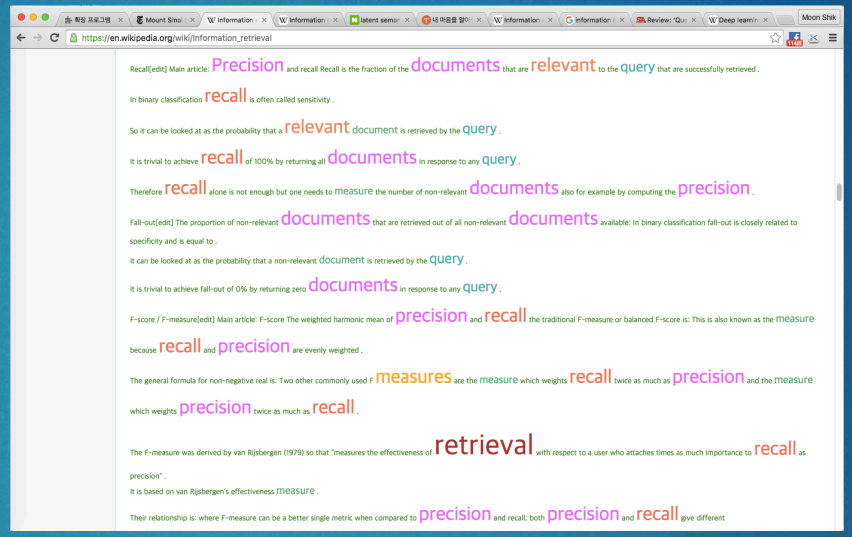
6. With distance weight(5)



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

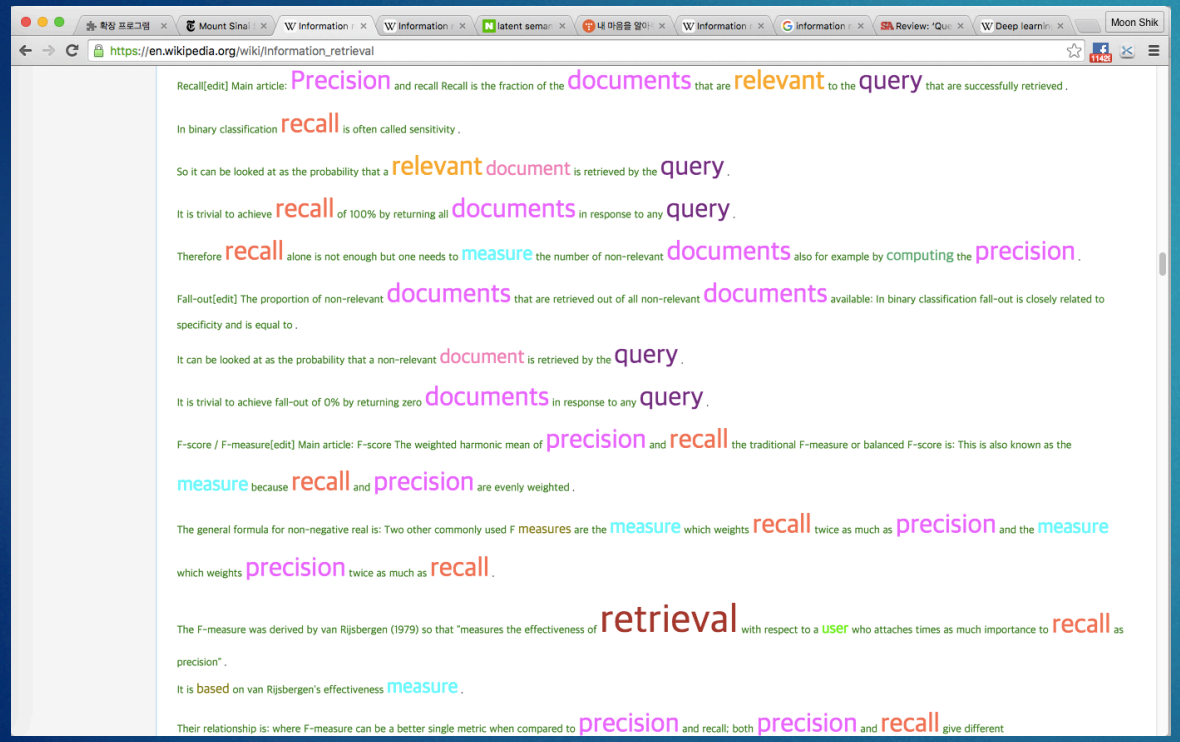
	Factor 01	Factor 02	Factor 03	Factor 04	Factor 05	Factor 06	Factor 07
Factor 01		0.802	1.213	0.711	0.899	1.139	0.702
Factor 02			0.918	0.759	0.942	1.049	0.669
Factor 03				0.847	0.91	0.857	0.787
Factor 04					1.299	1.176	0.961
Factor 05						0.882	0.752
Factor 06							0.806
Factor 07							

Weight of Factor 07 is largest

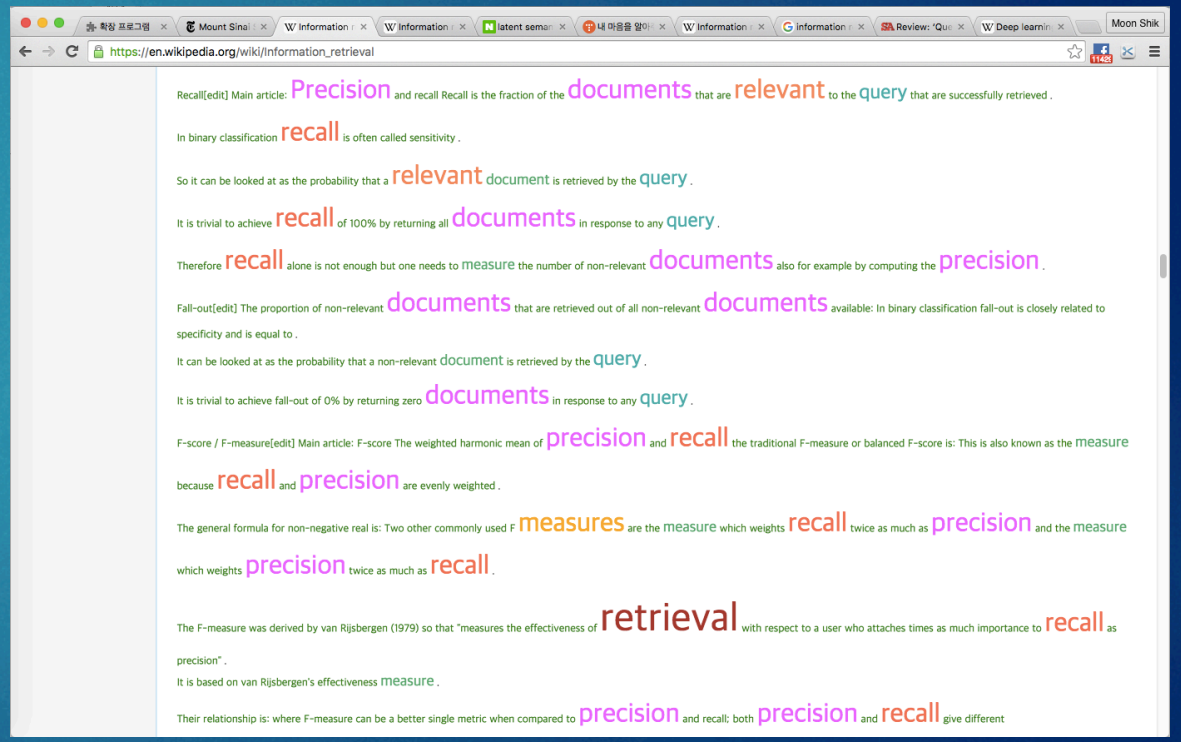
→ tag and distance weight increased satisfaction level

Best combination?

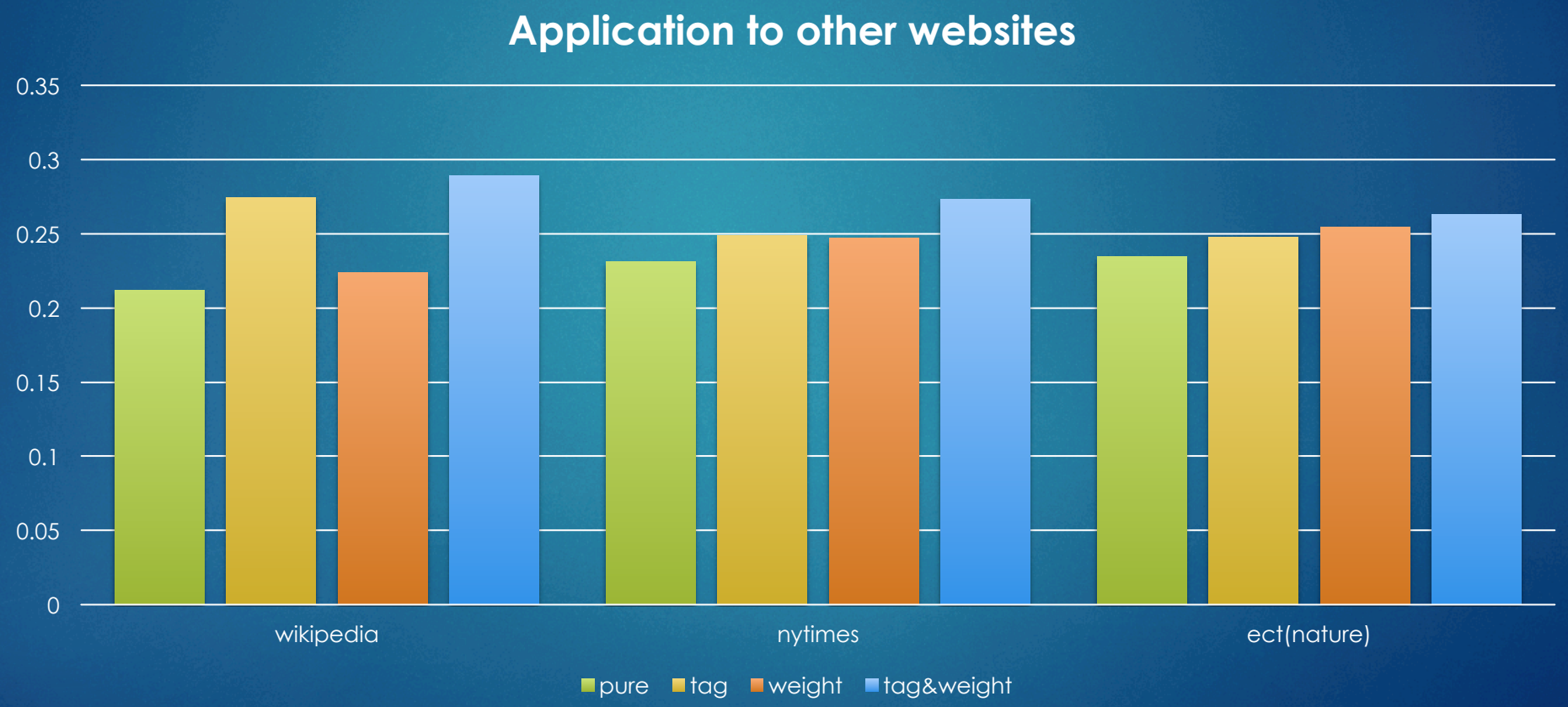
With Meta Tag And Distance Weighting



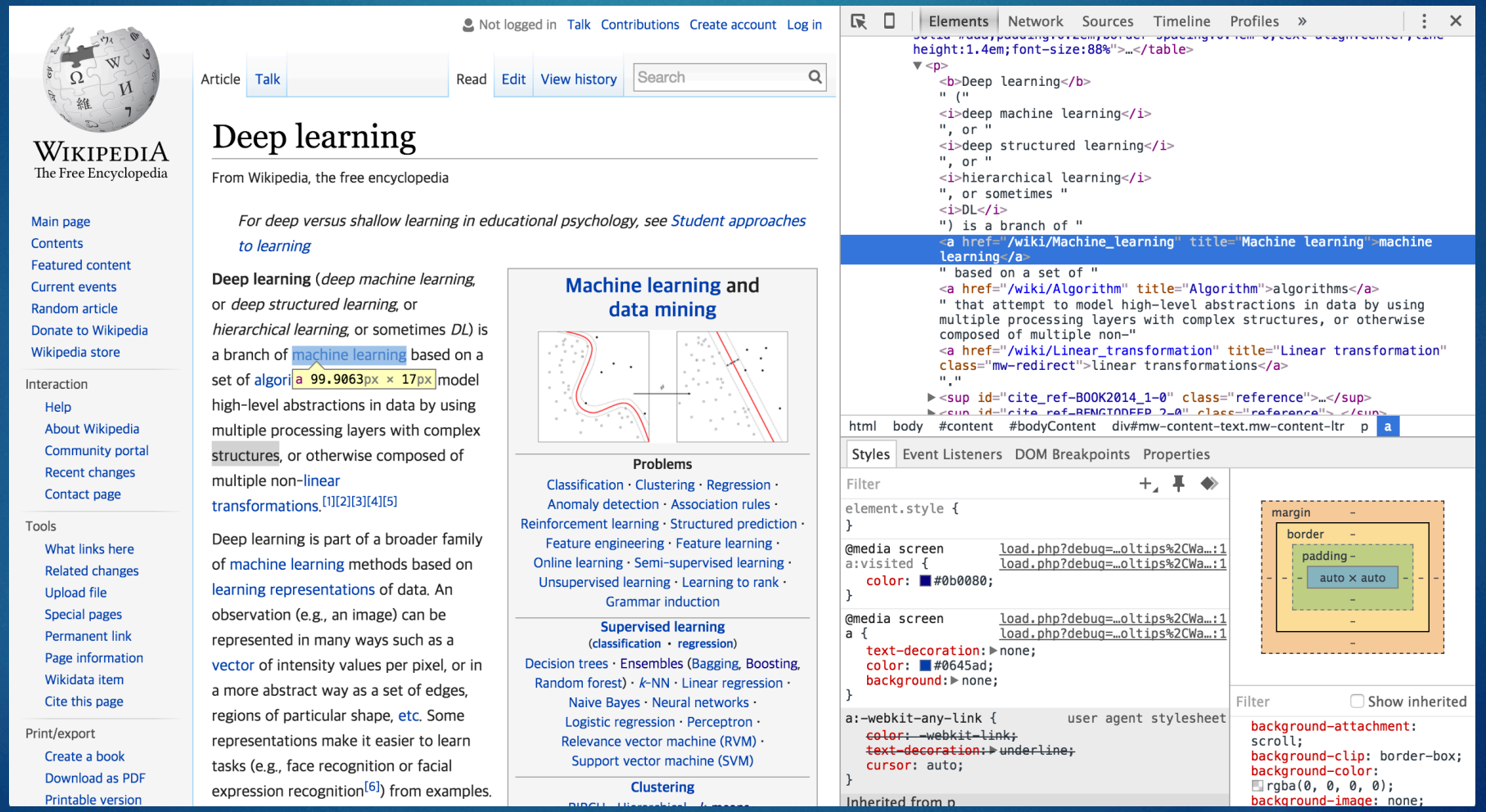
Without Meta Tag And Distance Weighting



Comparitive Result



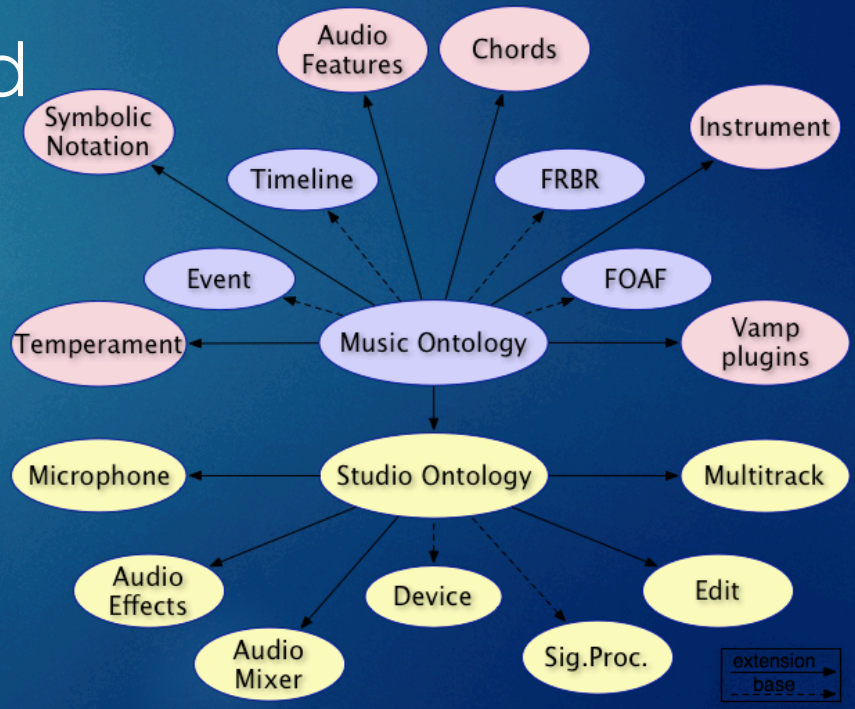
Wikipedia is ..



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