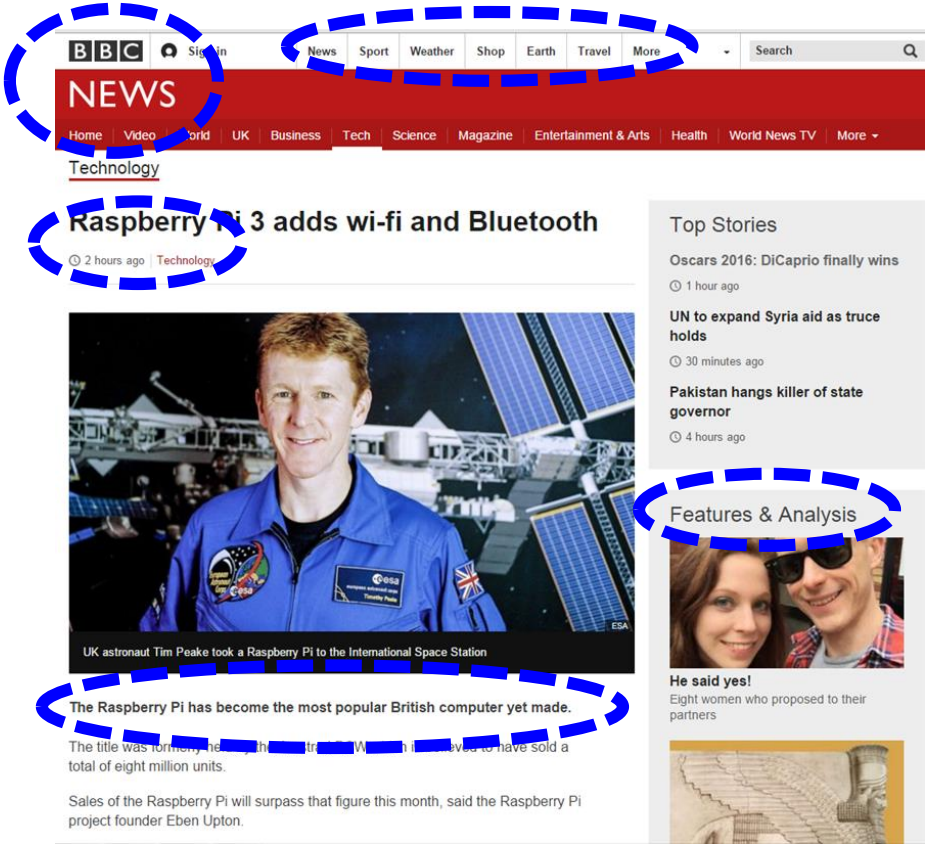
DOM-based Keyword Extraction from Web Pages

Himat Shah, Mohammad Rezaei, **Pasi Fränti**

20.12.2019

Challenges in web pages

1. Lack of standard structure
2. Irrelevant text: Adds, Tags, Styling
3. Variety of languages
4. High variation of text length



Limitations of existing solutions

1. Supervised methods
   - Intensive human work for labeling
   - Needs updates
   - Needs large training corpus
   - Difficult to support all domains and languages
2. NLP components
   - Time consuming
   - Difficult to support all languages

D-rank

Proposed method

- Scoring candidate keywords based on position: URL, title, headers, hyperlinks
- Stop words based on language detection
- Term frequency + popularity in Wikipedia

Candidate keywords

1. Webpage Text

- HTML (Tags)
- CSS + Scripts

2. Language detection

- Stop Words

```
<em class="orb-hilight">
 Copyright © 2019 BBC.
</em>

<p>The BBC is not responsible for the content of external sites.</p>
```

English

Copyright © 2019 BBC. **The** BBC **is not** responsible **for the** content **of** external sites

Stop Words

Finnish

Chinese

English

German

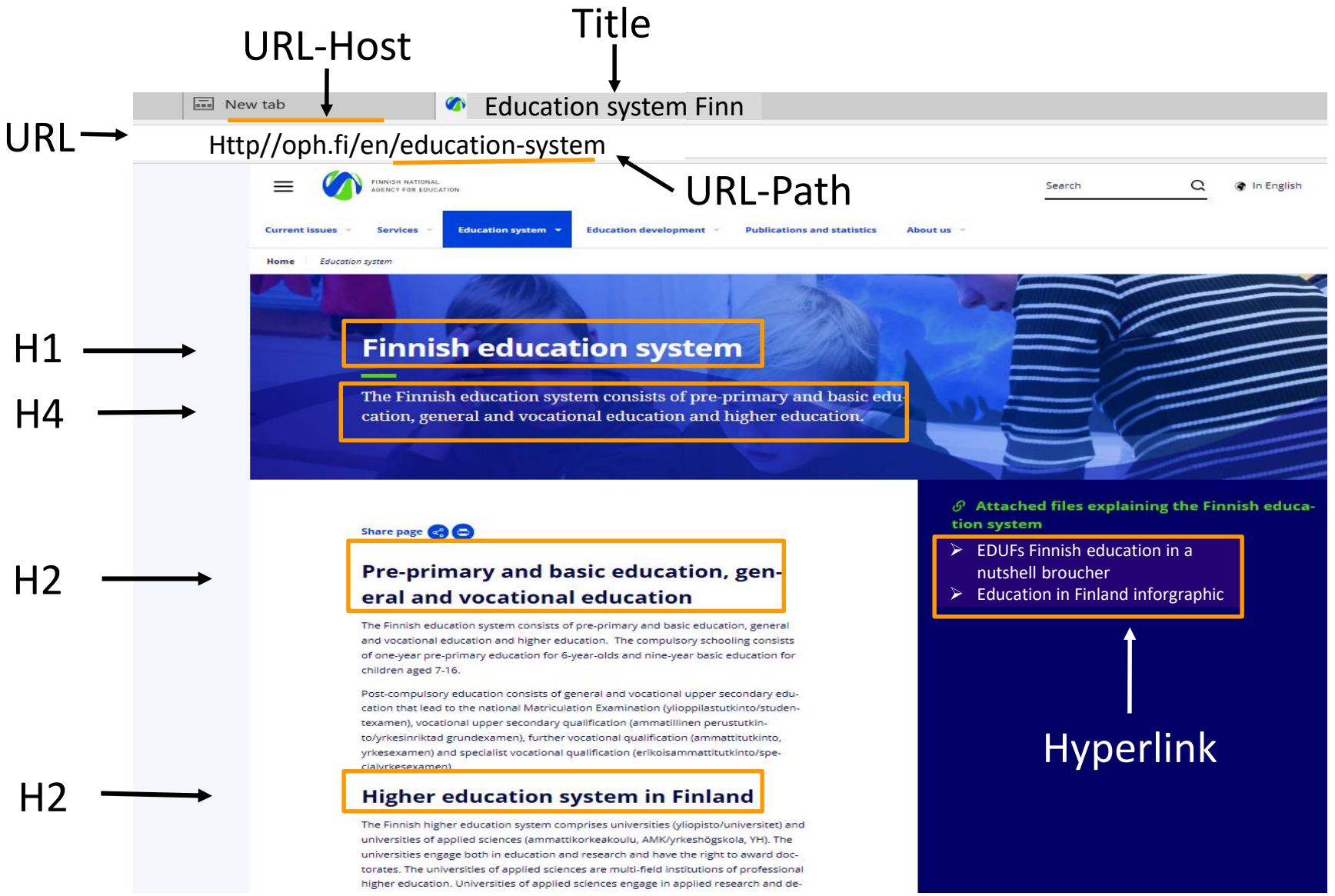
Candidates:

Copyright BBC

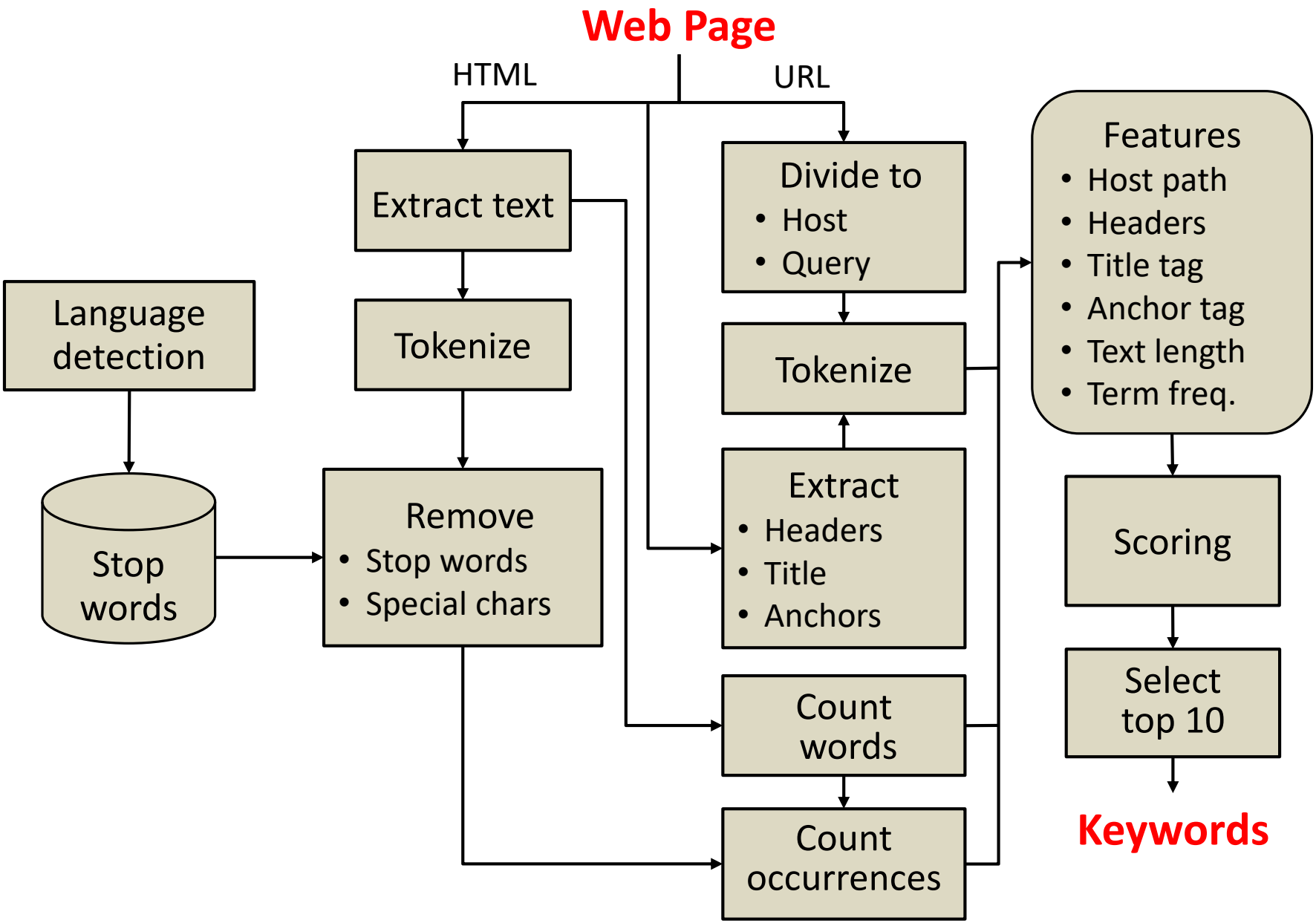
responsible content

external sites

Important positions in web page



D-rank method



Scoring the words

Position of word	Score
H1	6
H2	5
H3	4
H4 H5 H6	2
Title	5
URL host	5
URL path and query	4
Hyperlink (anchor)	2

Scoring the position in title

First or last word in Title tag or Meta tag: **5**

<**title**>BBC-Homepage </**title**>

<**meta** name="title" content="BBC-Homepage" /**meta**>

Candidates:

- BBC $1 \times 5 = \mathbf{5}$
- Homepage $1 \times 5 = \mathbf{5}$
- Video
- News

Scoring-position in web link

Host	Path	Query
https:// www.bbc.com/	england/london/	uk=? the-news-video
5	4	4

Candidates:

- BBC — 1× 5 = 5
- Homepage
- Video — 1× 4 = 4
- News — 1× 4 = 4

Popularity among header and anchor tags

<h1 id="page-title">BBC Homepage</h1>

<h2 >Accessibility links</h2>

<h2 class="module__title">News</h2>

<h3 class="title"> A really simple guide to the UK general election </h3>

Home

Candidates:

- BBC $1 \times 6 = 6$
- Homepage $1 \times 6 = 6$
- Home $1 \times 2 = 2$
- News $1 \times 5 = 5$

Frequency Weight

Feature extraction

Features	Value
URL	http//www.bbc.com
Url-Host	bbc
URL-Path	-
Text length	1080
Title	bbc-homepage
H1	bbc homepage
H2	accessibility links news sport london weather editors picks latest business news technology
H3	hour marathon mark the weird bathroom habits of the west gen
H4-h6	-
Anchor	homepage skip to content accessibility help bbc account notifications home news sport weather iplaye

Scoring term frequency

- Score = 0.2 × TF IF $N > 50$
- Score = 0.5 × TF IF $N \leq 50$

Scoring candidate keywords

	Headers					Anchor		URL	Score	
Candidate Words	TF	H1	H2	H3	H4-6	A	Title	Path & Query	TF	Final
BBC	17	1	1	1	0	1	1	1	0	3.4
Homepage	3	1	0	0	0	1	1	0	0	0.6
World	12	0	1	1	0	1	0	0	0	2.4
Video	7	0	1	1	0	1	0	0	0	1.4
News	6	0	1	1	0	1	0	0	0	1.2
Pictures	6	0	1	1	0	1	0	0	0	1.2
London	3	0	1	1	0	1	0	0	0	0.6
Made	3	0	1	1	0	1	0	0	0	0.6

Using Wikipedia for elimating unimportant words

- Eliminate common words
- Select top 20 highest scored candidates
- Select 10 lowest according to Wikipedia score.

Word frequencies in Wikipedia

Words	Final score	Wiki count	Words	Final score	Wiki count
BBC	30.4	15390	Sport	7.6	137260
Homepage	13.4	1957	Accessibility	7.6	8942
World	13.4	1918571	Weather	7.6	99530
Video	12.4	466293	Technology	7.6	317555
News	12.2	344310	UK	7.4	348679
Pictures	12.2	89721	Time	7.4	2513534
London	11.6	617139	Latest	7.4	51912
Made	11.6	1840000	Languages	7.4	167073
Around	11.6	916691	Earth	7.2	176749
Business	8.6	568900	politics	7.2	167994

Experiments

Hard evaluation

- PRECISION = $\frac{True\ Positive}{True\ positive + False\ Negative}$
- RECALL = $\frac{True\ Positive}{True\ positive + False\ Positive}$
- F-SCORE $= 2 \times \frac{Precision \times Recall}{Precision + Recall}$

Exact match

GT :	student	**university**	tution	opportunities	
Extracted Keywords :	studies	**university**	lecture	chances	free

TP = 1 [university]

FP = 4 [studies, free, chances, lecture]

FN = 3 [students, tuition, opportunities]

TN = ∞ - (1 + 3 + 4) [∞ stands for all words in document]

Too harsh

Exact match

GT :	**student**	**university**	tution	opportunities
Extracted Keywords :	**studies**	**university**	lecture	chances	free

TP = 2 [studxxx, university]

FP = 3 [free, chances, lecture]

FN = 2 [tuition, opportunities]

TN = ∞ - (2 + 3 + 2) [∞ stands for all words in document]

Soft evaluation

http://cs.uef.fi/sipu/soft/stringsim/

GT : student university tution opportunities

Extracted Keywords : studies university lecture chances free

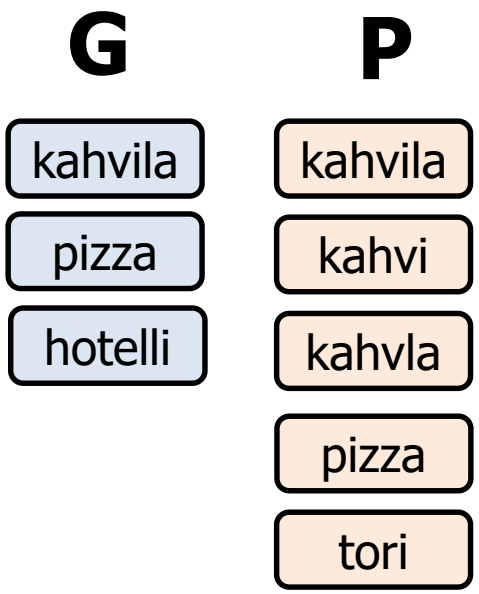
TP = 2.02

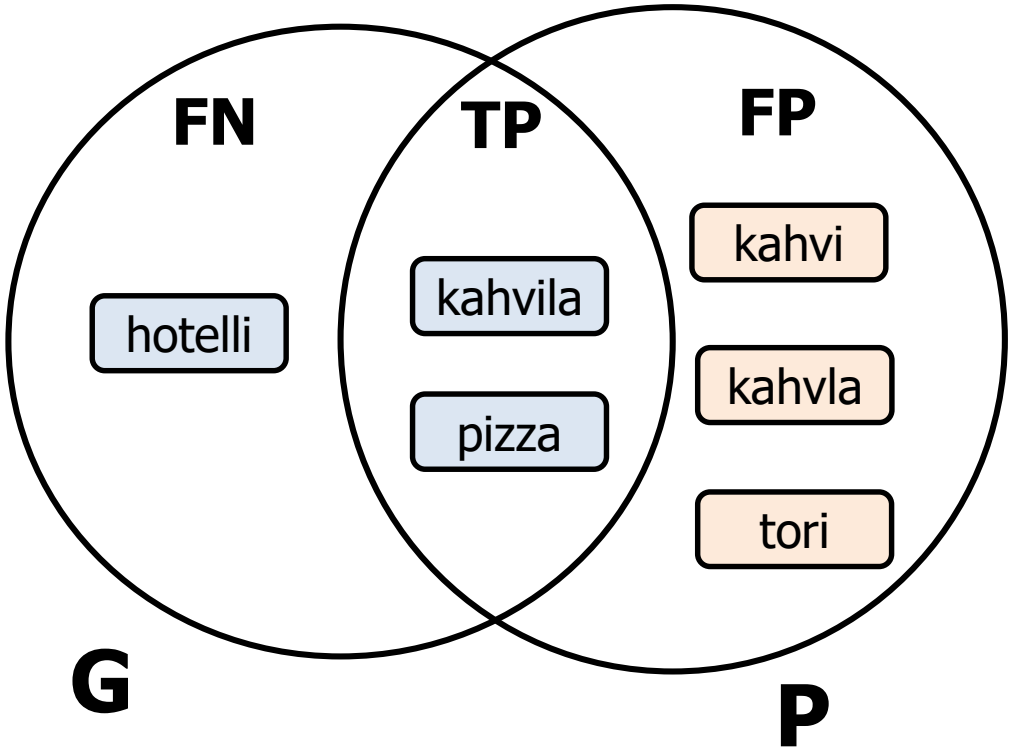
FP = 5 – 2.02

FN = 4 – 2.02

```
Tokens A: [students, university, tuition, opportunities]
Tokens B: [studies, university, lecture, chances, free]

                university      students       tuition opportunities           SUM       INVERSE
    university          1.00          0.20          0.20          0.15          1.55          0.64
      students          0.20          1.00          0.25          0.23          1.68          0.59
       tuition          0.20          0.25          1.00          0.38          1.83          0.55
 opportunities          0.15          0.23          0.38          1.00          1.77          0.57

 CARDINALITY  : 2.35

              chances  university     studies     lecture        free         SUM     INVERSE
     chances        1.00        0.10        0.29        0.00        0.14        1.53        0.65
  university        0.10        1.00        0.20        0.10        0.10        1.50        0.67
     studies        0.29        0.20        1.00        0.14        0.14        1.77        0.56
     lecture        0.00        0.10        0.14        1.00        0.14        1.39        0.72
        free        0.14        0.10        0.14        0.14        1.00        1.53        0.65

 CARDINALITY  : 3.26

                university      students       tuition opportunities       chances    university       studies
    university          1.00          0.20          0.20          0.15          0.10          1.00          0.20
      students          0.20          1.00          0.25          0.23          0.13          0.20          0.63
       tuition          0.20          0.25          1.00          0.38          0.00          0.20          0.29
 opportunities          0.15          0.23          0.38          1.00          0.23          0.15          0.38
       chances          0.10          0.13          0.00          0.23          1.00          0.10          0.29
    university          1.00          0.20          0.20          0.15          0.10          1.00          0.20
      studies          0.20          0.63          0.29          0.38          0.29          0.20          1.00
      lecture          0.10          0.13          0.14          0.23          0.00          0.10          0.14
         free          0.10          0.13          0.00          0.15          0.14          0.10          0.14

 CARDINALITY  : 3.59
 UNION        : 3.59
 INTERSECTION : 2.02


Jaccard + Levenshtein : 0.56
```

Example





Two out of 3 are correct, but some penalty for duplicate errors.

Precision = 2/3 ~ 55 %
Recall = 2/3 ~ 66 %
Fscore = 60 %

Results

Methods compared

Process	Cl-Rank	H-Rank	D-Rank
Webpage (HTML)	Extract Text	Extract Text	Extract Text + DOM Features
Language	English	English	Any Language
NLP	Stemming + Lemmatization	Stemming + Lemmatization	-
Term Frequency	Based on frequency	Based on frequency	Based on position
Linguistic	POS + Nouns	POS + Nouns, Adjectives, Verbs	-
WordNet	Synonyms for words	Synonyms for words	-
Machine learning	Unsupervised Clustering	Unsupervised Clustering	-

Data sets statistics

Dataset	Language	Size	Keywords	Locations
Guardian	English	421	5 - 15	guardian.com
MACWorld	English	212	5 - 10	macworld.com
Mopsi Services	Finnish	414	1 - 10	cs.uef.fi/mopsi
German	German	100	5 - 20	-

Guardian dataset

Hard Measures

Method	Precision	Recall	F-score
D-rank	0.24	0.37	0.29
Cl-rank	0.26	0.40	0.31
H-rank	0.20	0.29	0.23
TF	0.20	0.24	0.22

Soft Measures

Method	Precision	Recall	F-score
D-rank	0.51	0.73	0.59
Cl-rank	0.54	0.75	0.62
H-rank	0.46	0.63	0.51
TF	0.52	0.63	0.57

MACWorld dataset

Hard Measures

Method	Precision	Recall	F-score
D-rank	0.32	0.24	0.27
Cl-rank	0.26	0.20	0.22
H-rank	0.22	0.37	0.27
TF	0.16	0.18	0.17

Soft Measures

Method	Precision	Recall	F-score
D-rank	0.58	0.46	0.50
Cl-rank	0.56	0.42	0.47
H-rank	0.48	0.68	0.55
TF	0.34	0.38	0.36

Mopsi services

Hard Measures			
Method	Precision	Recall	F-score
D-rank	0.19	0.20	0.19
Cl-rank	-	-	-
H-rank	-	-	-
TF	0.11	0.26	0.12

Soft Measures			
Method	Precision	Recall	F-score
D-rank	0.42	0.45	0.43
Cl-rank	-	-	-
H-rank	-	-	-
TF	0.29	0.70	0.36

German dataset

Hard Measures			
Method	Precision	Recall	F-score
D-rank	0.26	0.27	0.23
Cl-rank	-	-	-
H-rank	-	-	-
TF	0.19	0.15	0.13

Soft Measures			
Method	Precision	Recall	F-score
D-rank	0.52	0.58	0.58
Cl-rank	-	-	-
H-rank	-	-	-
TF	0.37	0.32	0.29

Summary of results

F-score (%)

Method	Guardian		MACWorld		Mopsi		German	
	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft
D-rank	29	59	**27**	50	**19**	**43**	**23**	**58**
Cl-rank	**31**	**62**	22	47	-	-	-	-
H-rank	23	51	**27**	**55**	-	-	-	-
TF	22	57	17	36	12	36	13	29

Conclusions

- Language and domain independent method for keyword extraction
- Position of the word in the page URL, title, Hyperlinks, headers and frequency of the words used as a features
- Fast and effective w/o need for NLP

END