

Three Way Search Engine with Queries **Multi-feature Document Comparison**

**Šimon Suchomel and Jan Kasprzak
with Michal Brandejs**



Faculty of Informatics
Masaryk University



Three types of queries

Keywords

- KW extraction based on tf-idf
- KW enriched with collocations
- non positionable
- unconditionally executed
- 4 KW queries per doc on average
- covering the whole document

e.g.: mobile phone signal prepaid carriers

Intrinsic plagiarism

- positionable
- conditionally executed
- covering the suspicious passage
- based on Average Word Frequency Class
- it's change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite

Headers

- naive headers detection
- positionable
- conditionally executed
- covering the part introduces by the header

e.g.: Differences Between Cell Phones

Suspicious document

All About Cell Phone

More and more people have been using cell phones nowadays. In fact, some people even consider it as a necessity and we can't leave them. Cell phone is very useful and helpful for many after all.

What is Cell Phone?

Mostly people are using under a cord, you do know what a cell phone is. But for the uninitiated, here's what it means. A mobile phone (also known as a

cell phone) is a portable telephone.

The cell phone itself is an expensive device. You will need to shell out money for this only once if you want an upgrade. Your own personal bank, wife and kids should determine what cell phone you want (but the sales assistant behind the desk).

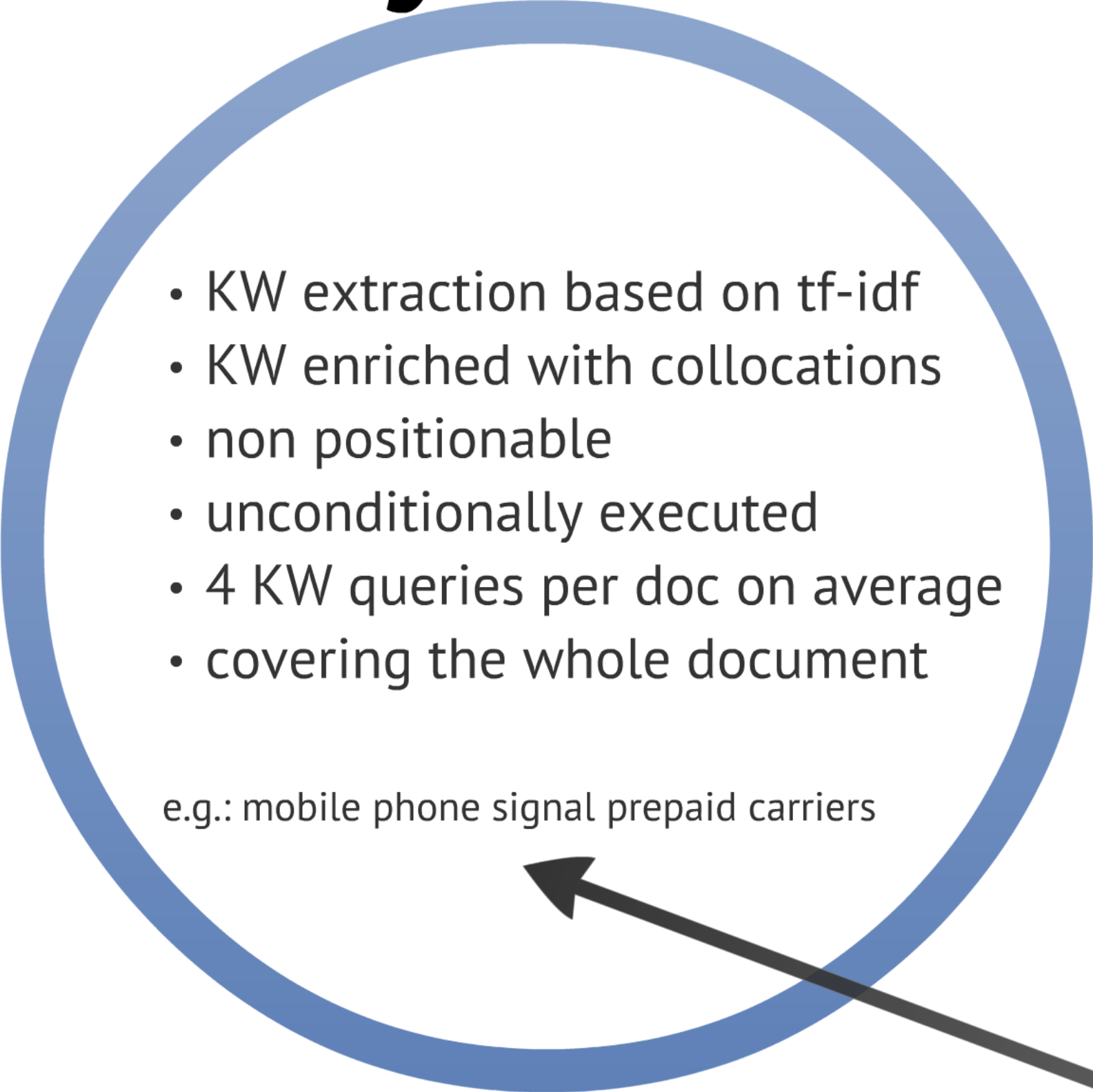
The use of "hand-free" was not recommended by the British Consumers' Association in a statement in November 2005 as they believed that exposure was increased. However, reassurance for the British Department of Trade and Industry and others for the French Agence Française de Sécurité Sanitaire advised that the increased risk of cancer is not a concern.

Learn about cell phones and their use in the world.

Overall Health Risks

Many scientific studies have investigated possible health effects of mobile phone radiation. These studies are occasionally reviewed by some scientific committees to assess overall risk. The most recent assessment was published in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENHR). It concludes from the available research that no significant health effects have been demonstrated from mobile phone radiation at normal exposure levels.

Query type	Extracted	Omitted	Similarities portion
KW	4.16	N/A	72.5%
Intrinsic	5.16	2.35	24.3%
Headers	10.34	4.75	3.2%

- 
- KW extraction based on tf-idf
 - KW enriched with collocations
 - non positionable
 - unconditionally executed
 - 4 KW queries per doc on average
 - covering the whole document

e.g.: mobile phone signal prepaid carriers

Suspicious document

All About Cell Phone

More and more people have been using cell phone nowadays. In fact, some people even consider it as a necessity and we can't blame them. Cell phone is very useful and helpful for many after all.

What is Cell Phone?

Unless you're living under a rock, you do know what a cell phone is. But for the uninitiated, here's what it means. A mobile phone (also known as a

...

Differences Between Cell Phones

The cell phone itself isn't an ongoing expense. You will need to shell out money for this only once unless you want an upgrade. Your own personal taste, style and need should determine what cell phone you want (not the sales assistant behind the desk).

...

The use of "hands-free" was not recommended by the British Consumers' Association in a statement in November 2000 as they believed that exposure was increased. However, measurements for the (then) UK Department of Trade and Industry and others for the French l Agence fran aise de s curit sanitaire environmental showed substantial reductions. In 2005 Professor Lawrie Challis and others said clipping a ferrite bead onto hands-free kits stops the radio waves travelling up the wire and into the head.

4

6

Overall Health Risks

Many scientific studies have investigated possible health effects of mobile phone radiations. These studies are occasionally reviewed by some scientific committees to assess overall risks. The most recent assessment was published in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR). It concludes from the available research that no significant health effect has been demonstrated from mobile phone radiation at normal exposure levels:

Query ty

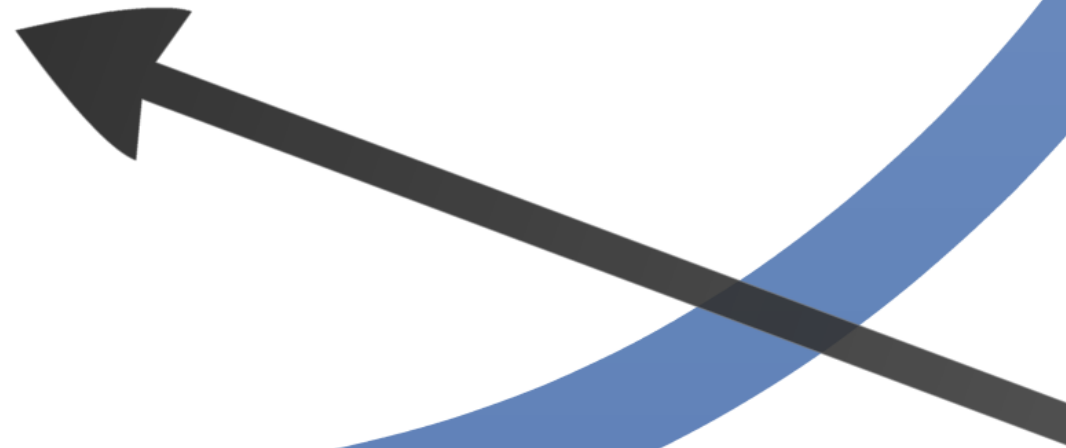
KW

Intrinsic

Header

- unconditionally executed
- 4 KW queries per doc on average
- covering the whole document

e.g.: mobile phone signal prepaid carriers



Suspicious document

All About Cell Phone

More and more people have been using cell phone nowadays. In fact, some people even consider it as a necessity and we can't blame them. Cell phone is very useful and helpful for many after all.

What is Cell Phone?

Unless you're living under a rock, you do know what a cell phone is. But for the uninitiated, here's what it means. A mobile phone (also known as a

...

Differences Between Cell Phones

The cell phone itself isn't an ongoing expense. You will need to shell out money for this only once unless you want an upgrade. Your own personal taste, style and need should determine what cell phone you want (not the sales assistant behind the desk).

...

The use of "hands-free" was not recommended by the British Consumers' Association in a statement in November 2000 as they believed that exposure was increased. However, measurements for the (then) UK Department of Trade and Industry and others for the French l Agence française de s curit sanitaire environmental showed substantial reductions. In 2005 Professor Lawrie Challis and others said clipping a ferrite bead onto hands-free kits stops the radio waves travelling up the wire and into the head.

4

6

Overall Health Risks

Many scientific studies have investigated possible health effects of mobile phone radiations. These studies are occasionally reviewed by some scientific committees to assess overall risks. The most recent assessment was published in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR). It concludes from the available research that no significant health effect has been demonstrated from mobile phone radiation at normal exposure levels:

Query ty

KW

Intrinsic

Header

Three types of queries

Keywords

- KW extraction based on tf-idf
- KW enriched with collocations
- non positionable
- unconditionally executed
- 4 KW queries per doc on average
- covering the whole document

e.g.: mobile phone signal prepaid carriers

Intrinsic plagiarism

- positionable
- conditionally executed
- covering the suspicious passage
- based on Average Word Frequency Class
- it's change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite

Headers

- naive headers detection
- positionable
- conditionally executed
- covering the part introduces by the header

e.g.: Differences Between Cell Phones

Suspicious document

All About Cell Phone

More and more people have been using cell phones nowadays. In fact, some people even consider it as a necessity and we can't leave them. Cell phone is very useful and helpful for many after all.

What is Cell Phone?

Mostly people are using under a cord, you do know what a cell phone is. But for the uninitiated, here's what it means. A mobile phone (also known as a

cell phone) is a portable telephone.

The cell phone itself is an expensive device. You will need to shell out money for this only once a day you want an upgrade. Your own personal bank, like and need should determine what cell phone you want (but the sales assistant behind the desk).

The use of "hand-free" was not recommended by the British Consumers' Association in a statement in November 2005 as they believed that exposure was increased. However, reassessment by the French Department of Trade and Industry and others for the French Agency for Food and Drug Safety and Control in 2005 found no significant increase in the risk of cancer.

Learn about cell phones and their use in the world.

Overall Health Risks

Many scientific studies have investigated possible health effects of mobile phone radiation. These studies are occasionally reviewed by some scientific committees to assess overall risk. The most recent assessment was published in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENHR). It concludes from the available research that no significant health effects have been demonstrated from mobile phone radiation at normal exposure levels.

Query type	Extracted	Omitted	Similarities portion
KW	4.16	N/A	72.5%
Intrinsic	5.16	2.35	24.3%
Headers	10.34	4.75	3.2%

- positionable
- conditionally executed
- covering the suspicious passage
- based on Average Word Frequency Class
- it's change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite



...

Differences Between Cell Phones

The cell phone itself isn't an ongoing expense. You will need to shell out money for this only once unless you want an upgrade. Your own personal taste, style and need should determine what cell phone you want (not the sales assistant behind the desk).

...

The use of "hands-free" was not recommended by the British Consumers' Association in a statement in November 2000 as they believed that exposure was increased. However, measurements for the (then) UK Department of Trade and Industry and others for the French l Agence fran aise de s curit sanitaire environmental showed substantial reductions. In 2005 Professor Lawrie Challis and others said clipping a ferrite bead onto hands-free kits stops the radio waves travelling up the wire and into the head.

4

6

Overall Health Risks

Many scientific studies have investigated possible health effects of mobile phone radiations. These studies are occasionally reviewed by some scientific committees to assess overall risks. The most recent assessment was published in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR). It concludes from the



based on Average word
Frequency Class

- it's change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite



Three types of queries

Keywords

- KW extraction based on tf-idf
- KW enriched with collocations
- non positionable
- unconditionally executed
- 4 KW queries per doc on average
- covering the whole document

e.g.: mobile phone signal prepaid carriers

Intrinsic plagiarism

- positionable
- conditionally executed
- covering the suspicious passage
- based on Average Word Frequency Class
- it's change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite

Headers

- naive headers detection
- positionable
- conditionally executed
- covering the part introduces by the header

e.g.: Differences Between Cell Phones

Suspicious document

All About Cell Phone

More and more people have been using cell phones nowadays. In fact, some people even consider it as a necessity and we can't leave them. Cell phone is very useful and helpful for many after all.

What is Cell Phone?

Mostly people are using under a cell, you do know what a cell phone is. But for the uninitiated, here's what it means. A mobile phone (also known as a

cell phone) is a portable telephone.

The cell phone itself is an expensive device. You will need to shell out money for this only once if you want an upgrade. Your own personal bank, life and health should determine what cell phone you want (not the sales assistant behind the desk).

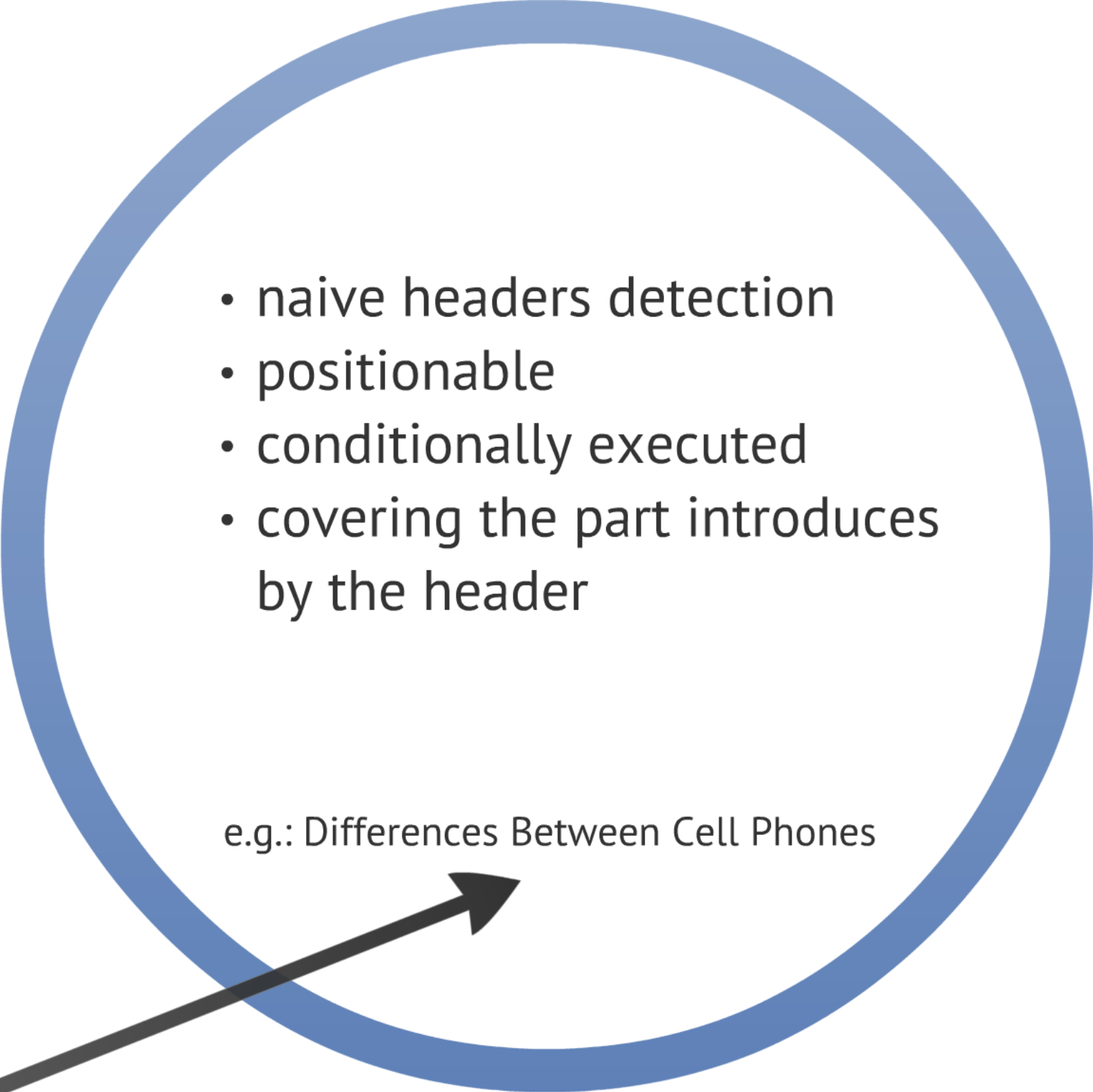
The use of "hand-free" was not recommended by the British Consumers' Association in a statement in November 2005 as they believed that exposure was increased. However, reassessment by the French Department of Trade and Industry and others for the French Agency for Food and Drug

regulation in 2007 found no significant health effects from the use of mobile phones and others used during a study. (source: healthline.com)

Overall Health Risks

Many scientific studies have investigated possible health effects of mobile phone radiation. These studies are occasionally reviewed by some scientific committees to assess overall risk. The most recent assessment was published in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENHR). It concludes from the available research that no significant health effects have been demonstrated from mobile phone radiation at normal exposure levels.

Query type	Extracted	Omitted	Similarities portion
KW	4.16	N/A	72.5%
Intrinsic	5.16	2.35	24.3%
Headers	10.34	4.75	3.2%

- 
- naive headers detection
 - positionable
 - conditionally executed
 - covering the part introduces by the header

e.g.: Differences Between Cell Phones



people even consider it as a necessity and we can't
is very useful and helpful for many after all.

What is Cell Phone?

Unless you're living under a rock, you do know what
for the uninitiated, here's what it means. A mobile
...

Differences Between Cell Phones

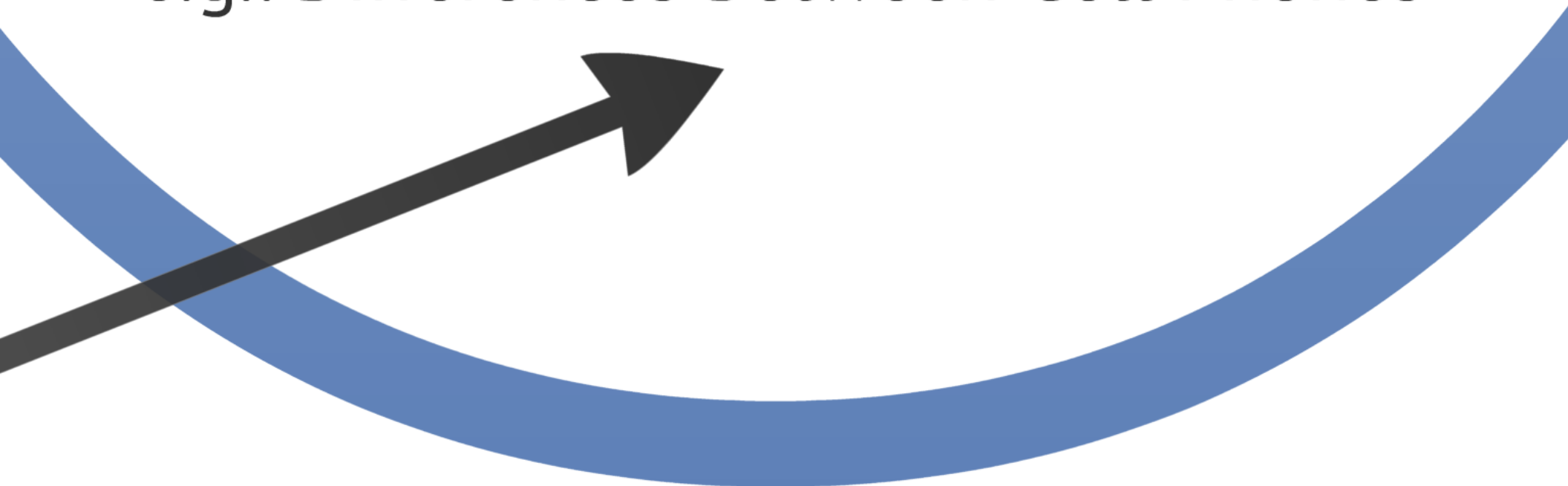
The cell phone itself isn't an ongoing expense. You
money for this only once unless you want an upgrade
taste, style and need should determine what cell phone
sales assistant behind the desk).

...

The use of "hands-free" was not recommended by the

covering the part introduced
by the header

e.g.: Differences Between Cell Phones



Query type	Extracted	Omitted	Similarities portion
KW	4.16	N/A	72.5%
Intrinsic	5.16	2.35	24.3%
Headers	10.34	4.75	3.2%

Three types of queries

Keywords

- KW extraction based on tf-idf
- KW enriched with collocations
- non positionable
- unconditionally executed
- 4 KW queries per doc on average
- covering the whole document

e.g.: mobile phone signal prepaid carriers

Intrinsic plagiarism

- positionable
- conditionally executed
- covering the suspicious passage
- based on Average Word Frequency Class
- it's change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite

Headers

- naive headers detection
- positionable
- conditionally executed
- covering the part introduces by the header

e.g.: Differences Between Cell Phones

Suspicious document

All About Cell Phone

More and more people have been using cell phones nowadays. In fact, some people even consider it as a necessity and we can't leave them. Cell phone is very useful and helpful for many after all.

What is Cell Phone?

Mostly people are using under a cord, you do know what a cell phone is. But for the uninitiated, here's what it means. A mobile phone (also known as a

cell phone) is a portable telephone.

The cell phone itself is an expensive device. You will need to shell out money for this only once if you want an upgrade. Your own personal bank, wife and kids should determine what cell phone you want (but the sales assistant behind the desk).

The use of "hand-free" was not recommended by the British Consumers' Association in a statement in November 2005 as they believed that exposure was increased. However, reassessment by the French Department of Trade and Industry and others for the French Agency for Food, Drugs and Cosmetics in 2005 found no significant increase in the risk of cancer.

Learn about cell phones and their use in the world.

Overall Health Risks

Many scientific studies have investigated possible health effects of mobile phone radiation. These studies are occasionally reviewed by some scientific committees to assess overall risk. The most recent assessment was published in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENHR). It concludes from the available research that no significant health effects have been demonstrated from mobile phone radiation at normal exposure levels.

Query type	Extracted	Omitted	Similarities portion
KW	4.16	N/A	72.5%
Intrinsic	5.16	2.35	24.3%
Headers	10.34	4.75	3.2%

Three Way Search Engine with Queries **Multi-feature Document Comparison**

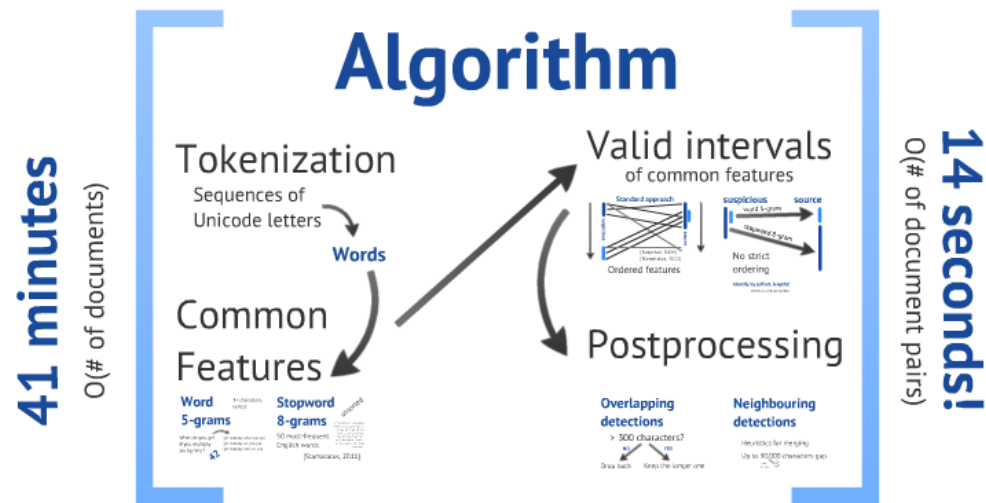
**Šimon Suchomel and Jan Kasprzak
with Michal Brandejs**



Faculty of Informatics
Masaryk University



Detailed Comparison



Performance

24-core server
4x AMD 8139

Implementation

Pure Perl
669 lines of code



Algorithm

Tokenization

Sequences of
Unicode letters

Words

Common Features

**Word
5-grams**

3+ characters,
sorted

What do you get
if you multiply
six by nine?
get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

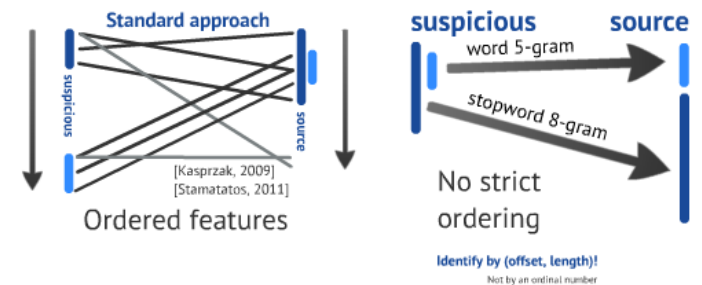
**Stopword
8-grams**

50 most-frequent
English words
[Stamatatos, 2011]

get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

unsorted
the the off and a in to is was
it for with he has on that the
up able to use not the this
from but had which use
they an in have we their
been has have will would
her there can all as if who
what said

Valid intervals of common features



Postprocessing

**Overlapping
detections**

> 300 characters?

NO Drop both
YES Keep the longer one

**Neighbouring
detections**

Heuristics for merging

Up to 30,000 characters gap

stand
for recall

Tokenization

Sequences of
Unicode letters



Words





Words



Features



Word 5-grams

3+ characters,
sorted

What do you get
if you multiply
six by nine?

42

get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

Stopword 8-grams

unsorted

50 most-frequent
English words

n't the of and a in to is was
it for with he be on i that by
at you 's are not his this
from but had which she
they or an were we their
been has have will would
her there can all as if who
what said

[Stamatatos, 2011]

Word

3+ characters,
sorted

5-grams

What do you get
if you multiply
six by nine?

42



get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

Features



Word 5-grams

3+ characters,
sorted

What do you get
if you multiply
six by nine?

42

get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

Stopword 8-grams

unsorted

50 most-frequent
English words

n't the of and a in to is was
it for with he be on i that by
at you 's are not his this
from but had which she
they or an were we their
been has have will would
her there can all as if who
what said

[Stamatatos, 2011]

Stopword

8-grams

50 most-frequent English words

unsorted

n't the of and a in to is was
it for with he be on i that by
at you 's are not his this
from but had which she
they or an were we their
been has have will would
her there can all as if who
what said

[Stamatatos, 2011]

Common Features



Word
5-grams

3+ characters,
sorted

Stopword
8-grams

unsorted

50 most-frequent

n't the of and a in to is was
it for with he be on i that by
at you 's are not his this
from but had which she
they or an were we their

Algorithm

Tokenization

Sequences of
Unicode letters

Words

Common Features

**Word
5-grams**

3+ characters,
sorted

What do you get
if you multiply
six by nine?
get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

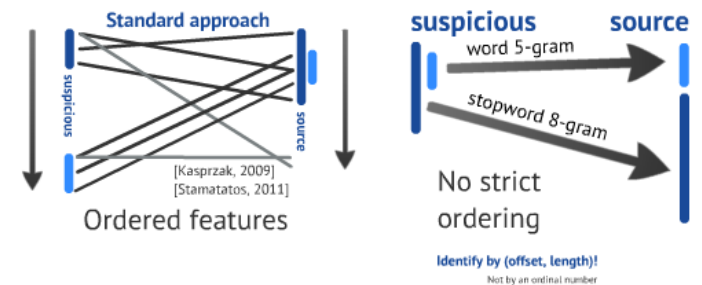
**Stopword
8-grams**

50 most-frequent
English words
[Stamatatos, 2011]

50 most-frequent
English words
[Stamatatos, 2011]

unsorted
the the off and a in to is was
it for with he have on that by
up go to a one not the this
from but had which use
they an all have we their
been has have will would
her there can all as if who
what said

Valid intervals of common features



Postprocessing

**Overlapping
detections**

> 300 characters?

NO → Drop both
YES → Keep the longer one

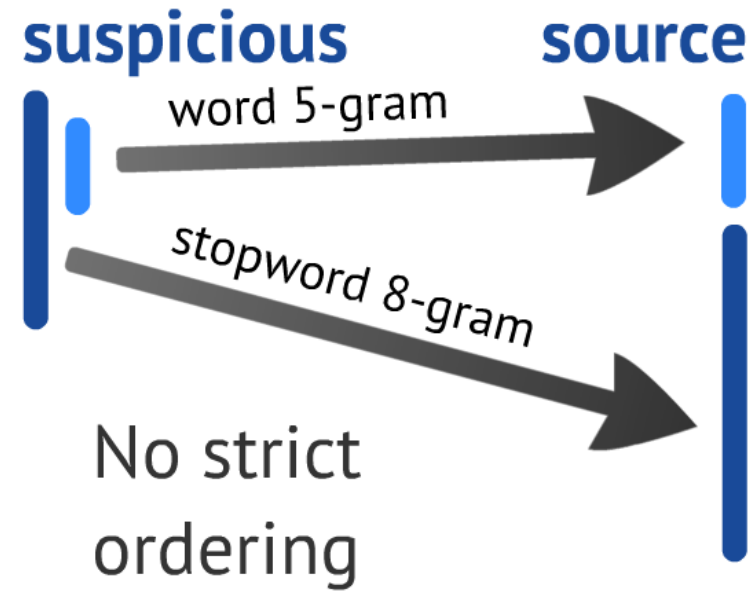
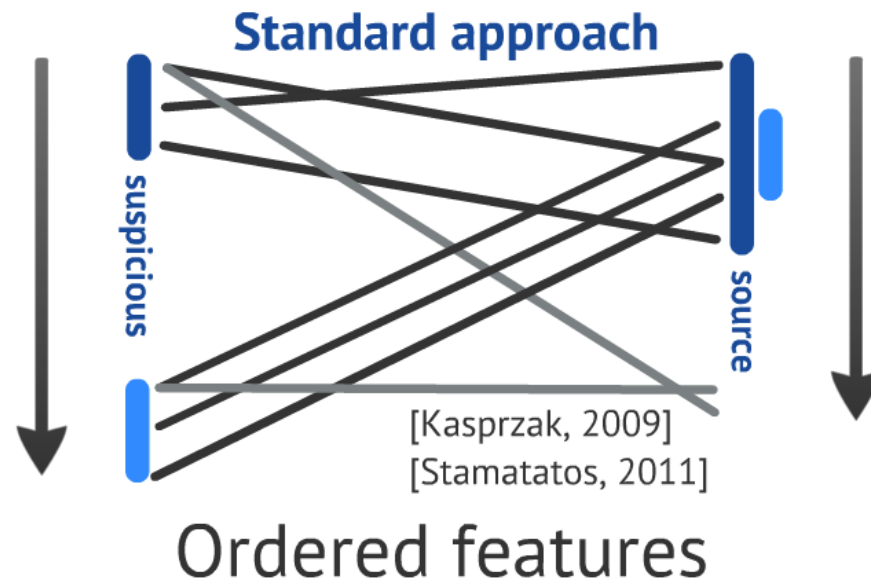
**Neighbouring
detections**

Heuristics for merging

Up to 30,000 characters gap

stand
for recall

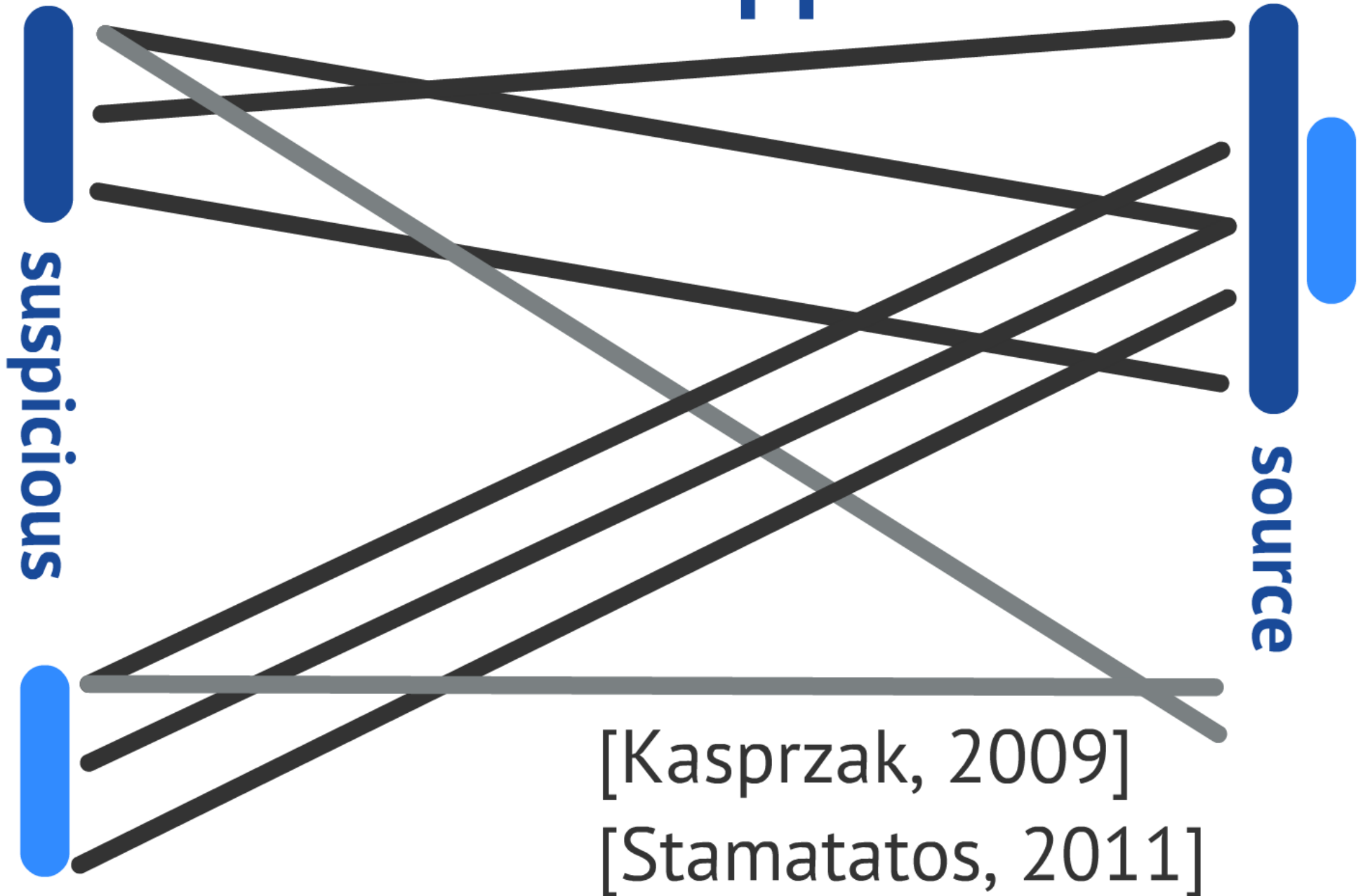
Valid intervals of common features



Identify by (offset, length)!

Not by an ordinal number

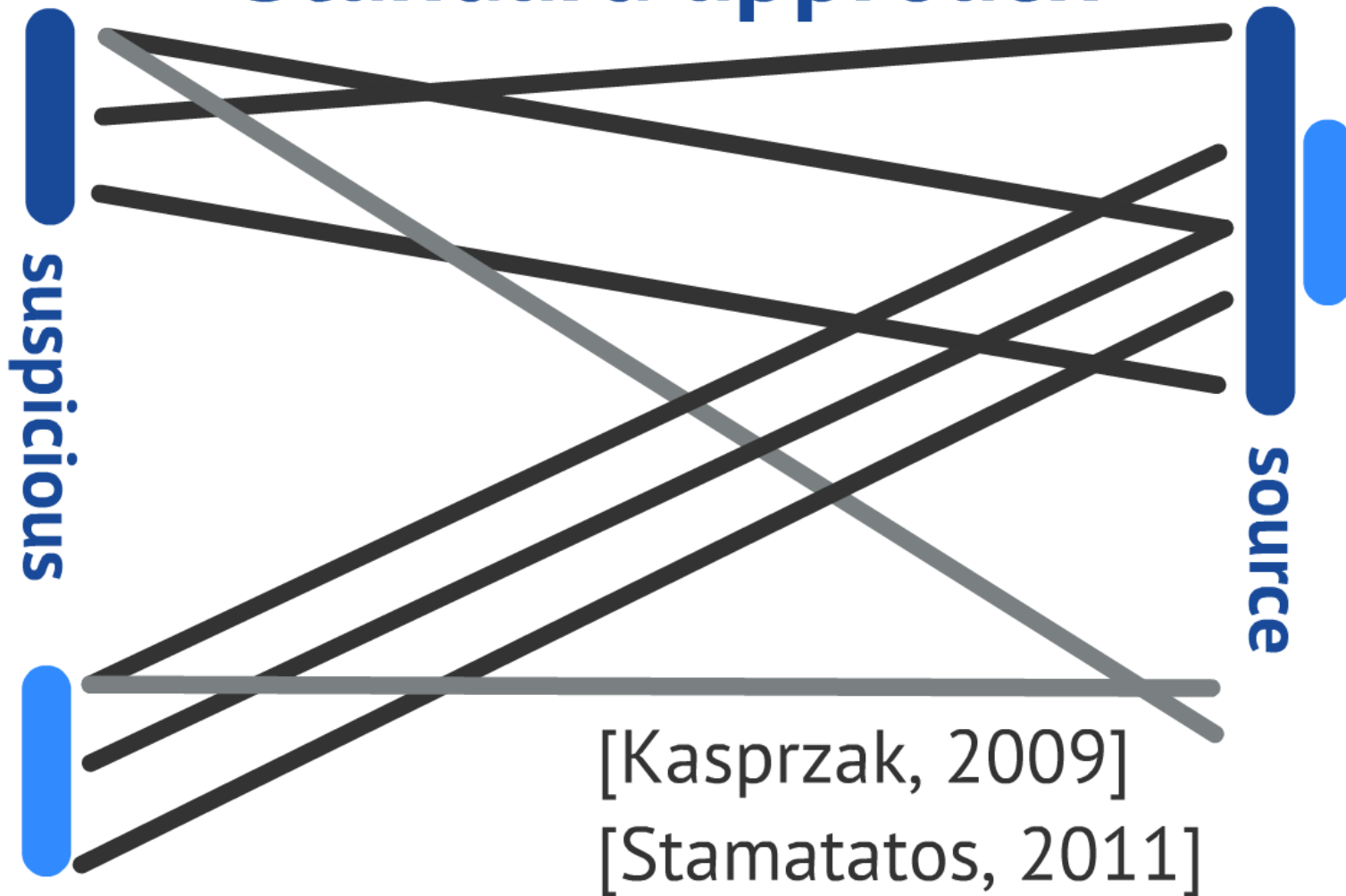
Standard approach



[Kasprzak, 2009]

[Stamatatos, 2011]

Standard approach



Ordered features

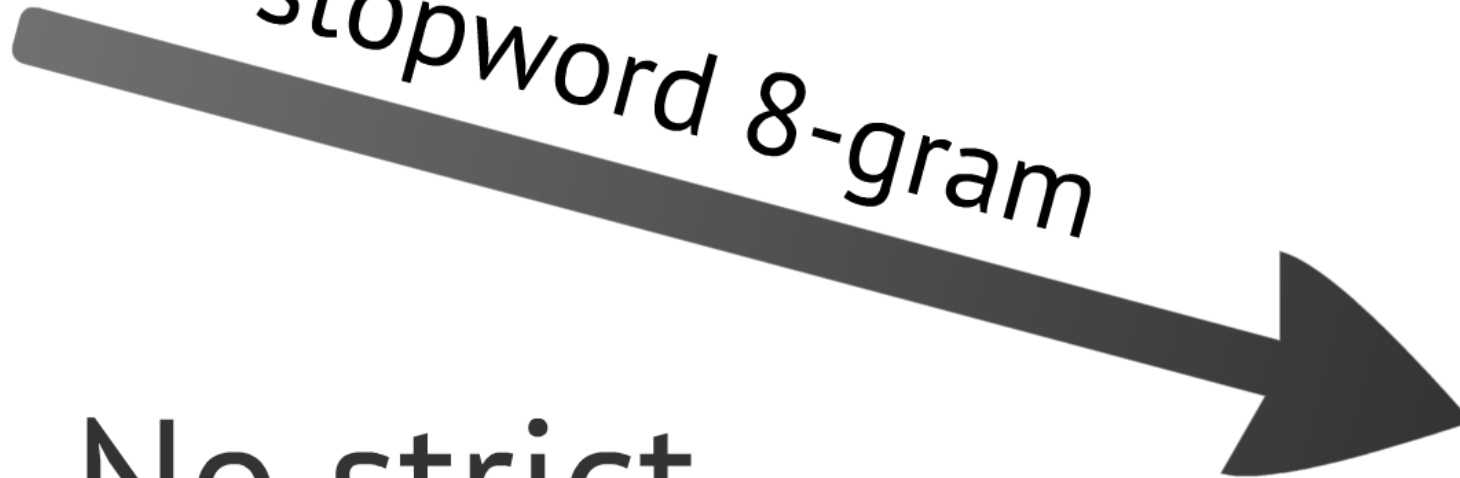
suspicious

source

word 5-gram



stopword 8-gram



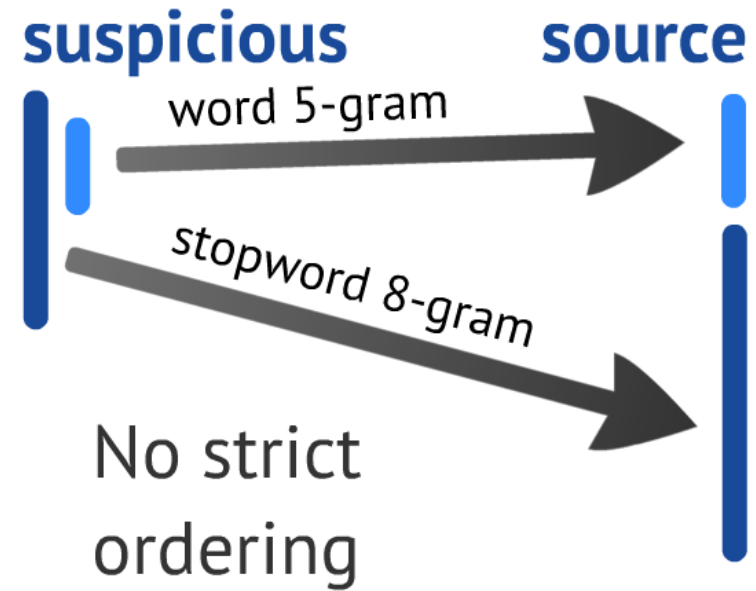
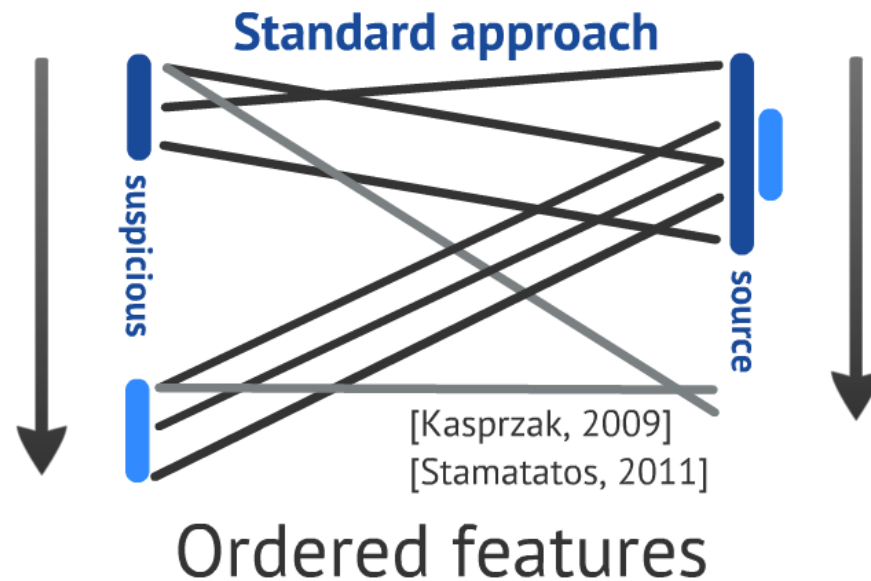
No strict
ordering

No strict ordering

Identify by (offset, length)!

Not by an ordinal number

Valid intervals of common features



Identify by (offset, length)!

Not by an ordinal number

Algorithm

Tokenization

Sequences of
Unicode letters

Words

Common Features

**Word
5-grams**

3+ characters,
sorted

What do you get
if you multiply
six by nine?
get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

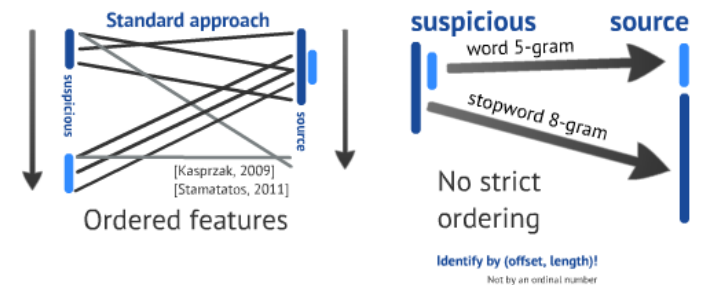
**Stopword
8-grams**

50 most-frequent
English words
[Stamatatos, 2011]

get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

unsorted
the off and a in to is was
at for with he has on that by
up go to a one not the this
from but had which use
they an in have we their
been has have will would
her there can all as of who
what said

Valid intervals of common features



Postprocessing

**Overlapping
detections**

> 300 characters?

NO Drop both
YES Keep the longer one

**Neighbouring
detections**

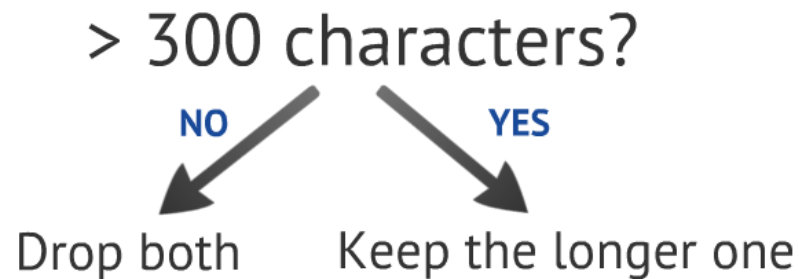
Heuristics for merging

Up to 30,000 characters gap

stand
for recall

Postprocessing

Overlapping detections



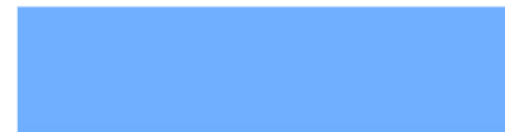
Neighbouring detections

Heuristics for merging

Up to 30,000 characters gap

biased
for recall

that's
several
pages!



Overlapping detections

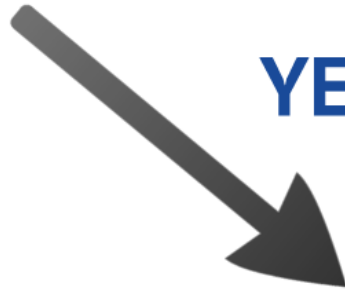
> 300 characters?

NO



Drop both

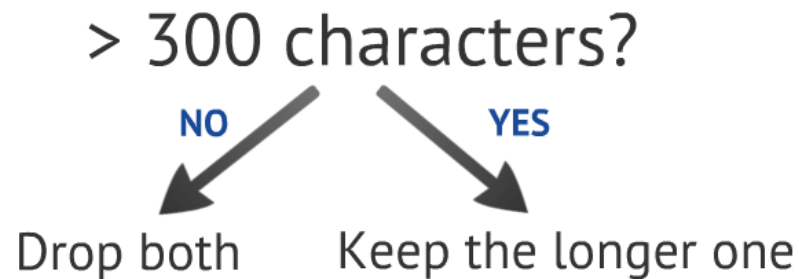
YES



Keep the longer one

Postprocessing

Overlapping detections



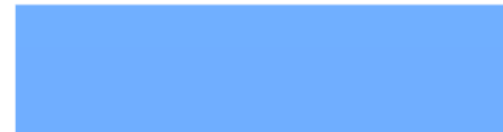
Neighbouring detections

Heuristics for merging

Up to 30,000 characters gap

biased
for recall

that's
several
pages!



Neighbouring detections

Heuristics for merging

Up to 30,000 characters gap

biased
for recall

*that's
several
pages!*

30,000

biased
for recall

*that's
several
pages!*



Algorithm

Tokenization

Sequences of
Unicode letters

Words

Common Features

**Word
5-grams**

3+ characters,
sorted

What do you get
if you multiply
six by nine?
get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

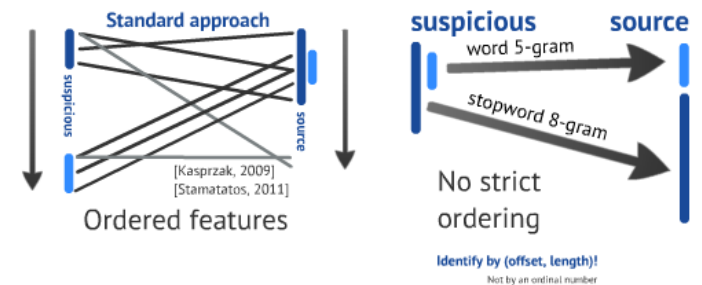
**Stopword
8-grams**

50 most-frequent
English words
[Stamatatos, 2011]

get-the-off-and-a-in-to-is-was
at-for-with-the-had-on-that-by
up-into-a-one-not-the-into
from-but-had-which-one
there-in-on-never-one-there
been-has-have-well-would
her-there-can-all-as-if-who
seem-just

[Stamatatos, 2011]

Valid intervals of common features



Postprocessing

**Overlapping
detections**

> 300 characters?

NO Drop both
YES Keep the longer one

**Neighbouring
detections**

Heuristics for merging

Up to 30,000 characters gap

stand
for recall

Performance

24-core server

4x AMD 8139

41 minutes

$O(\# \text{ of documents})$

Tokenization

Sequences of
Unicode letters

Words

Common Features

Word
5-grams

3+ characters,
sorted

What do you get
if you multiply
six by nine?



get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

Stopword
8-grams

unsorted

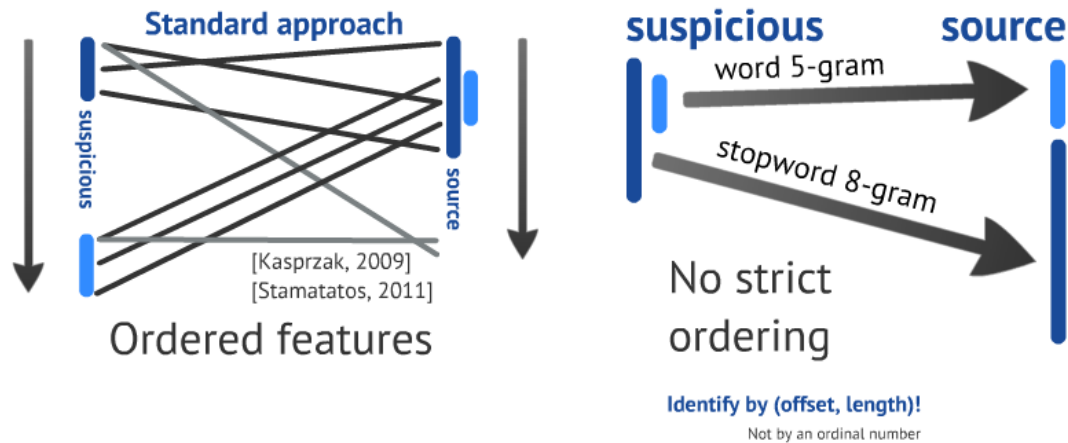
50 most-frequent
English words

n't the of and a in to is was
it for with he be on i that by
at you 's are not his this
from but had which she
they or an were we their
been has have will would
her there can all as if who
what said

[Stamatatos, 2011]

14 seconds!
 $O(\# \text{ of document pairs})$

Valid intervals of common features



Postprocessing

Overlapping
detections

> 300 characters?

no / yes

Neighbouring
detections

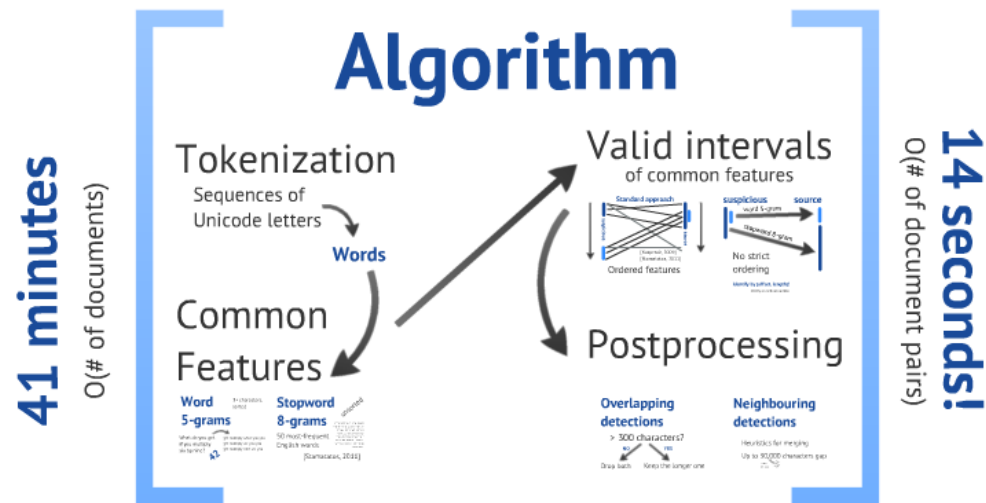
Heuristics for merging

Implementation

Pure Perl
669 lines of code



Detailed Comparison



Performance

24-core server
4x AMD 8139

Implementation

Pure Perl
669 lines of code



Conclusions



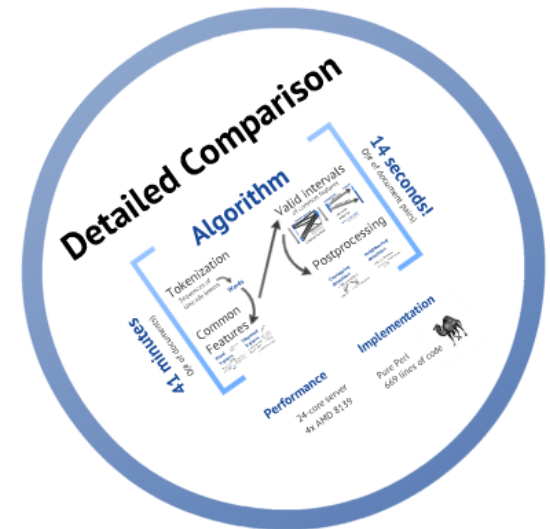
Three Way
Search Engine
Queries

Multi-feature
with Document
Comparison

Šimon Suchomel and Jan Kasprzak
with Michal Brandejs



Faculty of Informatics
Masaryk University



Narrowed search

**KW, Intrinsic, Headers
with befitting aiming
and combination**

Multiple types
of features

**Purely based on
(offset, length)**



Narrowed search

**KW, Intrinsic, Headers
with befitting aiming
and combination**



Multiple types of features

**Purely based on
(offset, length)**

Conclusions



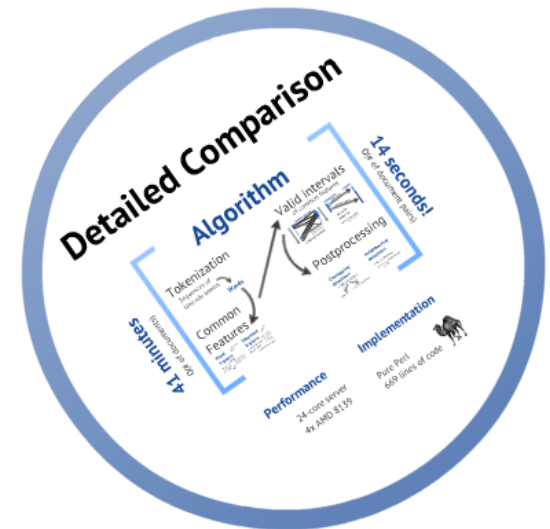
Three Way
Search Engine
Queries

Multi-feature
with Document
Comparison

Šimon Suchomel and Jan Kasprzak
with Michal Brandejs



Faculty of Informatics
Masaryk University



Narrowed search

**KW, Intrinsic, Headers
with befitting aiming
and combination**

Multiple types
of features

**Purely based on
(offset, length)**