

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
- its change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite

Headers

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

e.g.: Differences Between Cell Phones

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 don't think there's Cell phone is very useful and useful for many other.

What is Cell Phone?
There are many living under or roofs, very do know what a Cell phone is. But for those who don't, here's what it means. A mobile phone just comes as a

Differences Between Cell Phones

The cell phone itself has an ongoing evolution. You will need to shell out money for this and one and you want an upgrade. Your over personal usage, may be you want a different one, and, obviously you want one the sales assistant behind the desk!

The use of "handheld" was not recommended by the Radio Committee Association in a statement in November 2002 as it believed that exposure may increase the risk of cancer. The European Commission's Environment and Industry and the French Agency for the Environment and Sustainable Development issued similar statements in 2003. Professor Georges Lemaire, a French epidemiologist, has also warned that mobile phones emit radio waves that may affect the brain and interfere with heart.

Overall Health Risks
Many scientific studies have investigated possible health effects of mobile phone radiation. These studies are occasionally measured by some organizations. The World Health Organization (WHO) has conducted a review of all available evidence on mobile phones and concluded that there was no link in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCHER). It concludes from the available evidence that no significant health effect has been demonstrated from mobile phone radiation at normal reception 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



Query ty

KW

Intrinsic

Header

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 Agence française de sécurité sanitaire environnemental 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

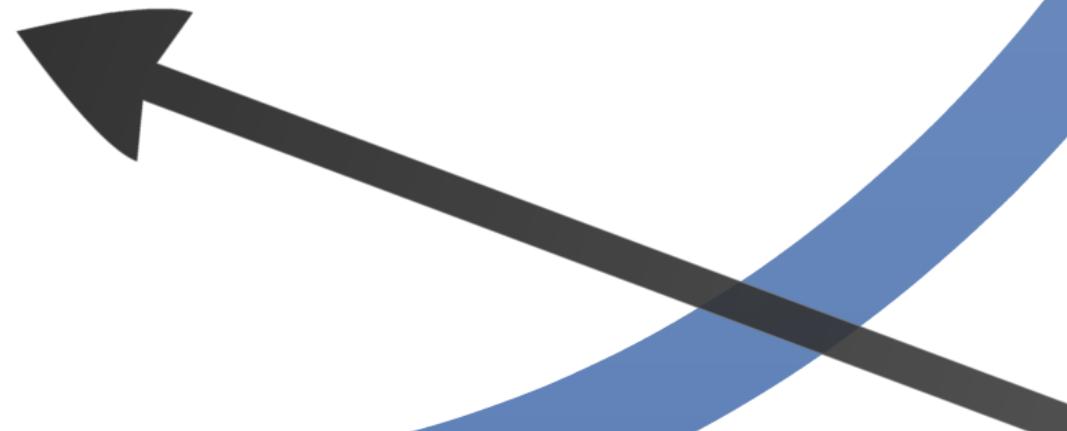
6

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:

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

e.g.: mobile phone signal prepaid carriers



Query ty

KW

Intrinsic

Header

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 Agence française de sécurité sanitaire environnemental 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:

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
- its change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite

Headers

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

e.g.: Differences Between Cell Phones

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 live without a cell phone is very useful and popular among others.

What is Cell Phone?
Everyone is living under or near very often what is cell phone is. But do you know about, how it works? A mobile phone just comes at a

Differences Between Cell Phones

The cell phone itself has an ongoing evolution. You will need to shell out money for this and see what you want an upgrade. Your over personal usage may lead to some damage, but it's not a problem, just, choose you want over the sales assistant (below the details).

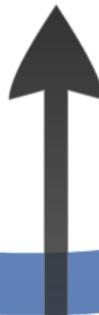
The use of "handheld" was not recommended by the Radio Committee Association in a statement in November 2002 as it is believed that exposure may increase the risk of cancer. The European Commission's Environment and Industry and Energy (ECE) Agency has issued a circular on similar environmental issues involving handheld devices in 2003. Professor Georges Lemaire, a member of the European Parliament, has also raised concerns about the radio waves from mobile phones in and outside the home.

Overall Health Risks
Many scientific studies have investigated possible health effects of mobile phone radiations. These studies are occasionally measured by some parameters such as SAR (Specific Absorption Rate). This parameter was established in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENHR). It concludes from the available evidence that no significant health effect has been demonstrated from mobile phone radiation at normal reception 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
- its 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 française de la sécurité 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

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

6

Based on average word

Frequency Class

- its 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
- its change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite

Headers

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

e.g.: Differences Between Cell Phones

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 live without a cell phone is very useful and popular among others.

What is Cell Phone?
Everyone is living under or near very often what is cell phone is. But do you know about, how it works? A mobile phone just comes at a

Differences Between Cell Phones

The cell phone itself has an ongoing evolution. You will need to shell out money for this and see what you want an upgrade. Your over personal usage may lead to some damage, but it's not a problem, just, choose you want over the sales assistant (below the details).

The use of "handheld" was not recommended by the Radio Committee Association in a statement in November 2002 as it is believed that exposure may increase the risk of cancer. The European Commission's Environment and Industry and Energy (ECE) Agency has issued a circular on similar environmental issues involving handheld devices in 2003. Professor Georges Lemaire, a member of the European Parliament, has also raised concerns about the radio waves from mobile phones in and outside the home.

Overall Health Risks
Many scientific studies have investigated possible health effects of mobile phone radiations. These studies are occasionally measured by some parameters such as SAR (Specific Absorption Rate). This parameter was established in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENHR). It concludes from the available evidence that no significant health effect has been demonstrated from mobile phone radiation at normal reception 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 introduced 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

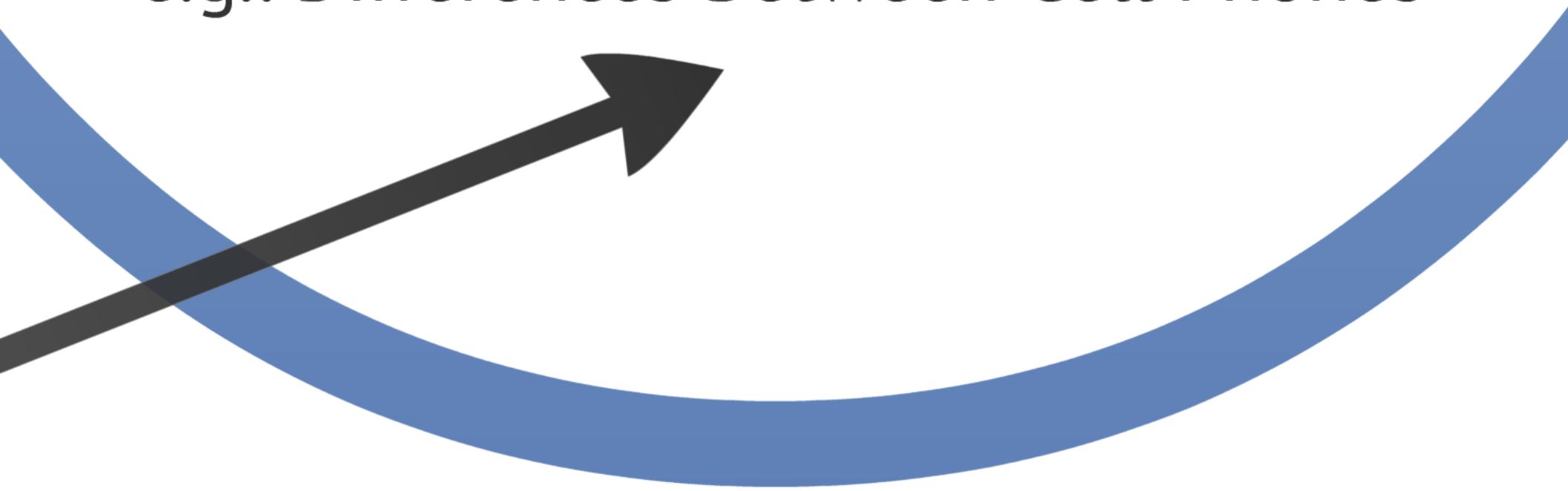
The cell phone itself isn't an ongoing expense. You
money for this only once unless you want an upgrade.
taste, style and need should determine what cell phone
sales assistant behind the desk).

...

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

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
- its change indicates suspicious passage

e.g.: in professor lawrie challis clipping ferrite

Headers

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

e.g.: Differences Between Cell Phones

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 don't think there's Cell phone is very useful and useful for many other.

What is Cell Phone?
There are many living under or roofs, very do know what a Cell phone is. But for those who don't, here's what it means. A mobile phone just comes as a

Differences Between Cell Phones

The cell phone itself has an ongoing evolution. You will need to shell out money for this and one and you want an upgrade. Your over personal usage, may be a reason to determine what, and, where you want over the sales assistant (below the desk).

The use of "handheld" was not recommended by the Radio Committee Association in a statement in November 2002 as it is believed that exposure may increase the risk of cancer. The European Commission's Environment and Industry and Energy (ECE) Agency has issued a call for a similar environmental impact assessment (EIA) to 2003. Professor Georges Lemaire, a member of the ECE committee, has also raised concerns about the radio waves from mobile phones in and outside the home.

Overall Health Risks
Many scientific studies have investigated possible health effects of mobile phone radiations. These studies are occasionally measured by some parameters such as the number of calls, the time spent on the phone, etc. This was published in 2007 by the European Commission Scientific Committee on Emerging and Newly Identified Health Risks (SCENHR). It concludes from the available evidence that no significant health effect has been demonstrated from mobile phone radiation at normal reception 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%

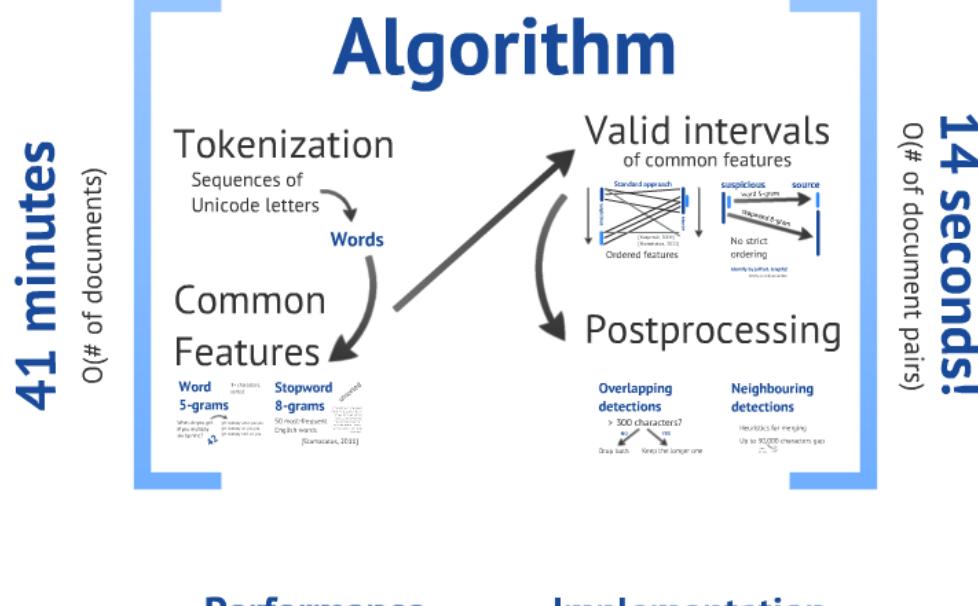
Three Way Search Engine with Queries Multi-feature Document Comparison

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

**Faculty of Informatics
Masaryk University**



Detailed Comparison



Nak

AS.SS

t
y
b
c
PREZI

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?



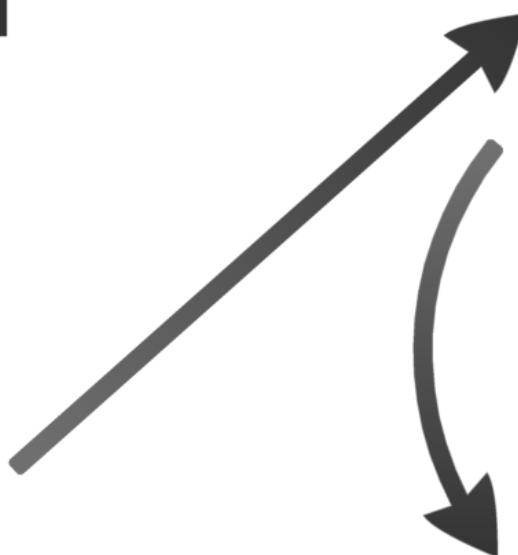
get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

Stopword 8-grams

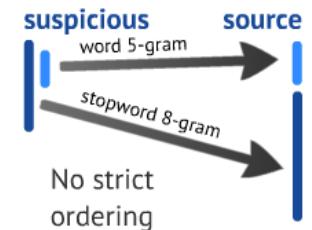
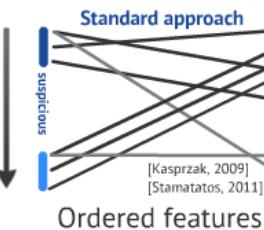
unsorted

50 most-frequent
English words

[Stamatatos, 2011]



Valid intervals of common features



Identify by (offset, length)!
Not by an ordinal number

Postprocessing

Overlapping detections

> 300 characters?

NO



Drop both

YES



Keep the longer one

Neighbouring detections

Heuristics for merging
Up to 30,000 characters gap



Based on
for recall

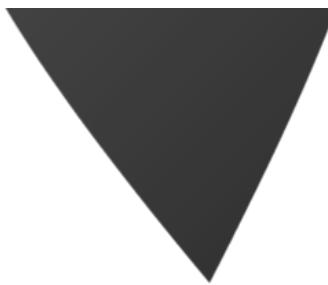
Tokenization

Sequences of
Unicode letters



Words





Words



Features



Word 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

3+ characters,
sorted

Stopword 8-grams

50 most-frequent
English words

[Stamatatos, 2011]

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

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

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

3+ characters,
sorted

Stopword 8-grams

50 most-frequent
English words

[Stamatatos, 2011]

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

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**

50 most-frequent

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

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?



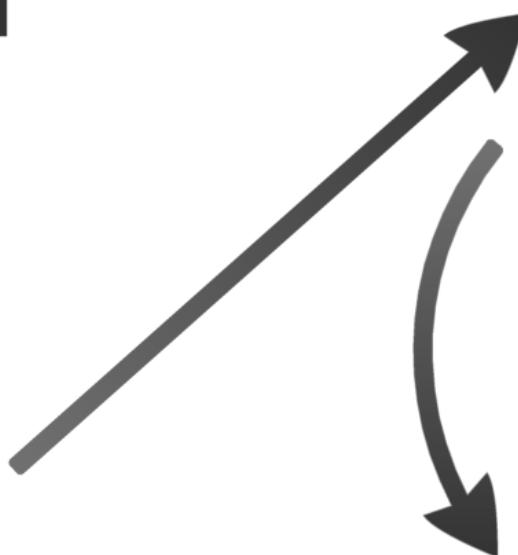
get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

Stopword 8-grams

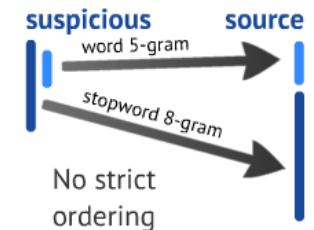
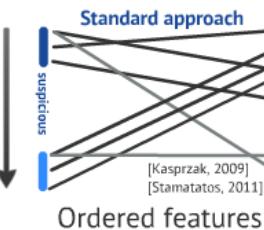
unsorted

50 most-frequent
English words

[Stamatatos, 2011]



Valid intervals of common features



Identify by (offset, length)!
Not by an ordinal number

Postprocessing

Overlapping detections

> 300 characters?

NO

Drop both

YES

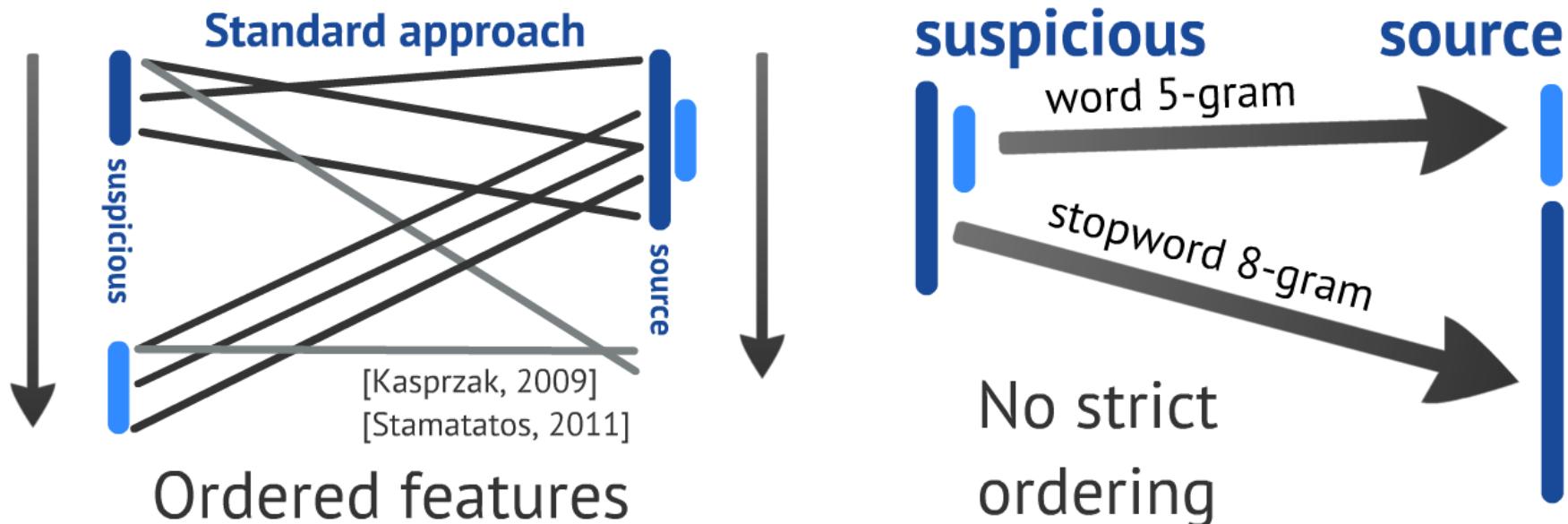
Keep the longer one

Neighbouring detections

Heuristics for merging
Up to 30,000 characters gap

Based on
for recall

Valid intervals of common features



Identify by (offset, length)!

Not by an ordinal number

Standard approach



Standard approach



Ordered features

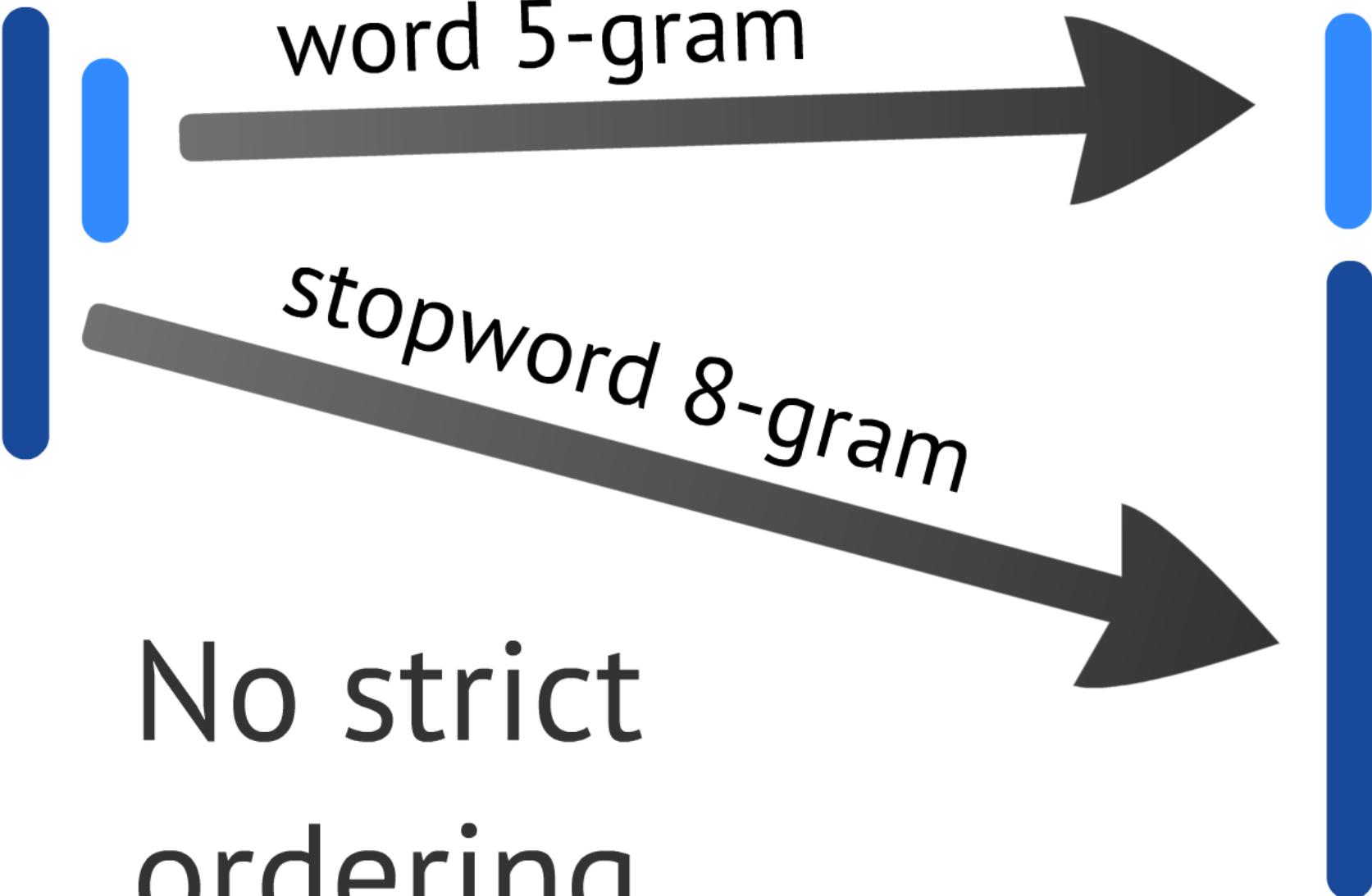
suspicious

source

word 5-gram

stopword 8-gram

No strict
ordering

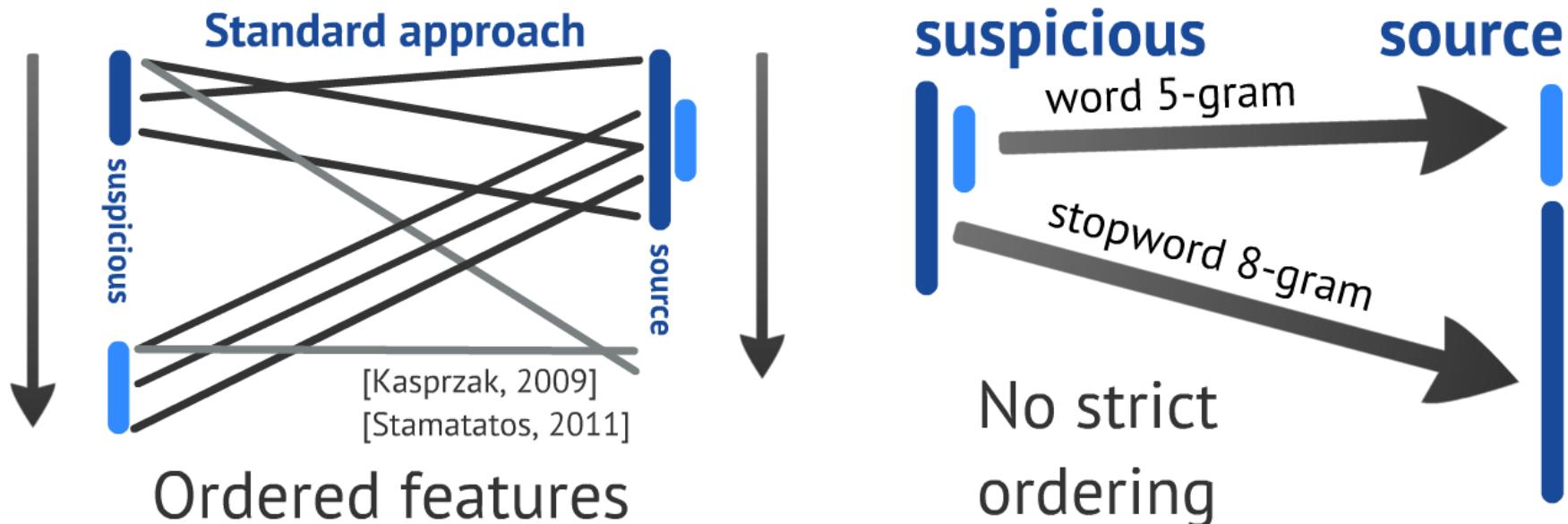


No strict ordering

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?



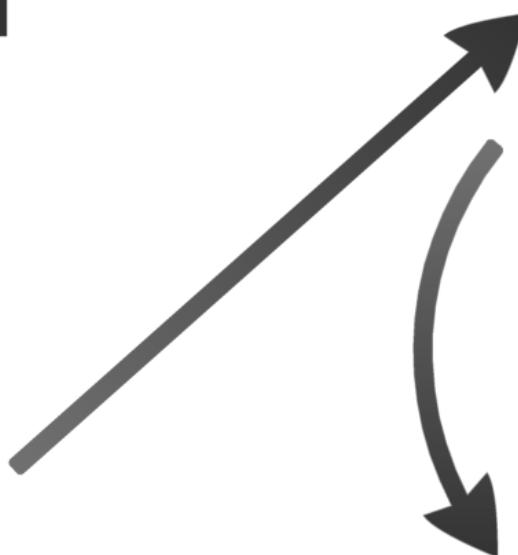
get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

Stopword 8-grams

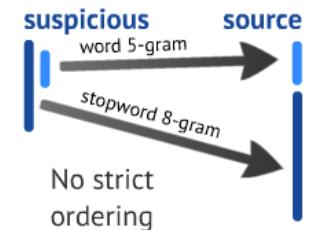
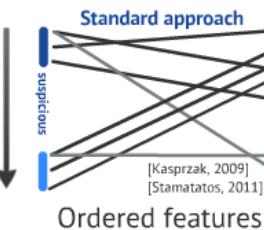
unsorted

50 most-frequent
English words

[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

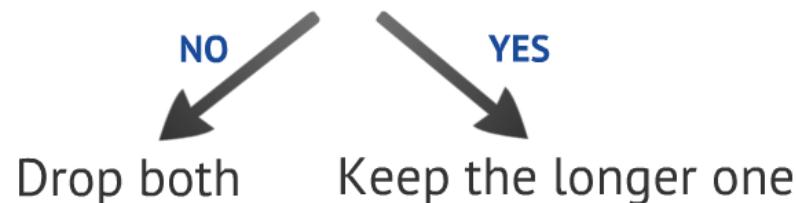
Up to 30,000 characters gap



Postprocessing

Overlapping detections

> 300 characters?



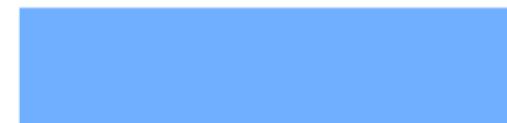
Neighbouring detections

Heuristics for merging

Up to 30,000 characters gap

biased
for recall

*that's
several
pages!*



Overlapping detections

> 300 characters?

NO

YES

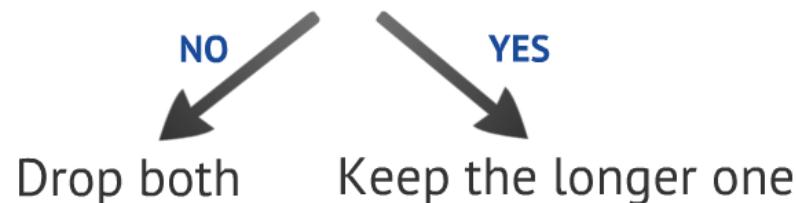
Drop both

Keep the longer one

Postprocessing

Overlapping detections

> 300 characters?



Neighbouring detections

Heuristics for merging

Up to 30,000 characters gap

biased
for recall

*that's
several
pages!*



Neighbouring detections

Heuristics for merging

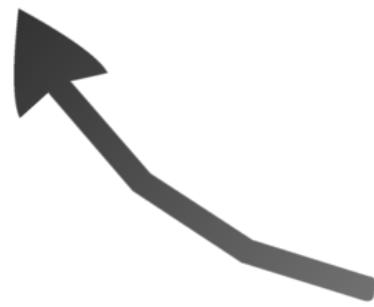
Up to 30,000 characters gap

biased
for recall

*that's
several
pages!*

biased
for recall

30,000



*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?



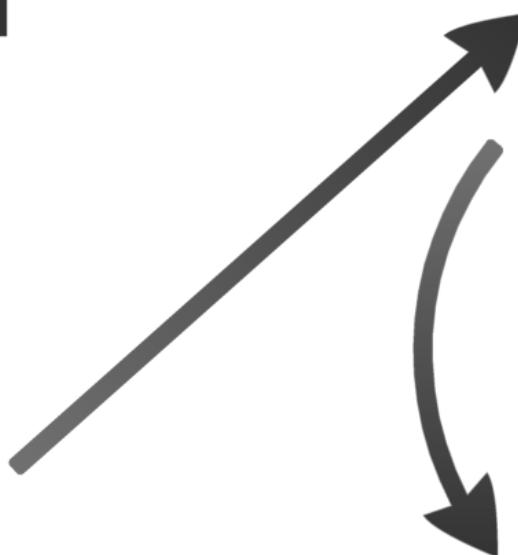
get-multiply-what-you-you
get-multiply-six-you-you
get-multiply-nine-six-you

Stopword 8-grams

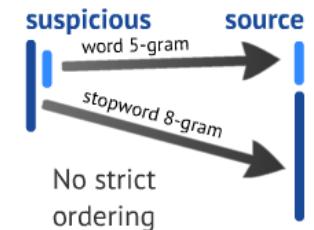
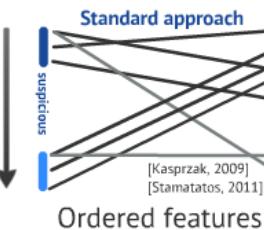
unsorted

50 most-frequent
English words

[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

Based on
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

What do you get
if you multiply
six by nine?

42

3+ characters,
sorted

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

Stopword
8-grams

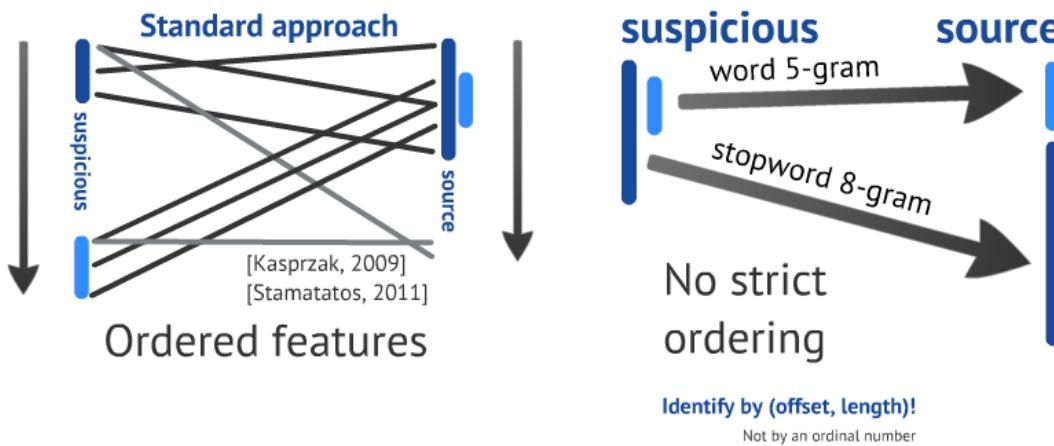
50 most-frequent
English words

unsorted

n't the of and a in to is was
it for with he be on as 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]

Valid intervals of common features



Postprocessing

Overlapping detections

> 300 characters?

NO VFC

Neighbouring detections

Heuristics for merging

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

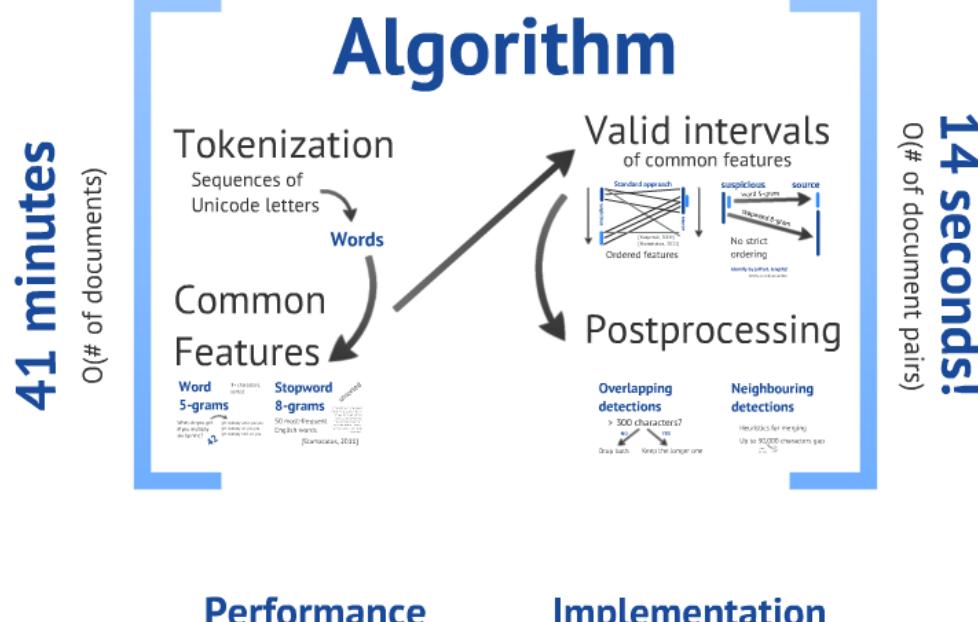
14 seconds!

Implementation

Pure Perl
669 lines of code



Detailed Comparison



Nak

AS.SS

t
y
b
c
PREZI

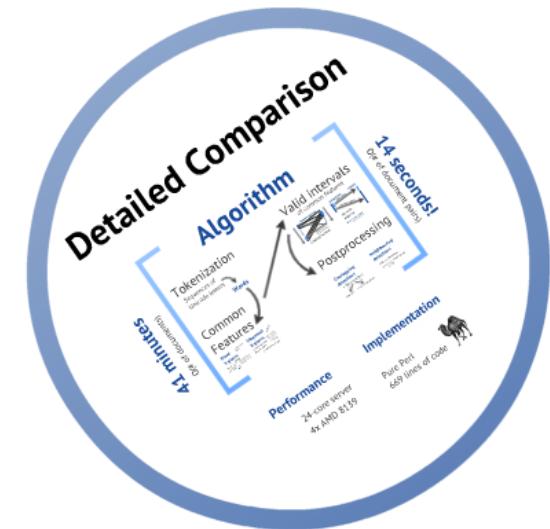
Conclusions



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

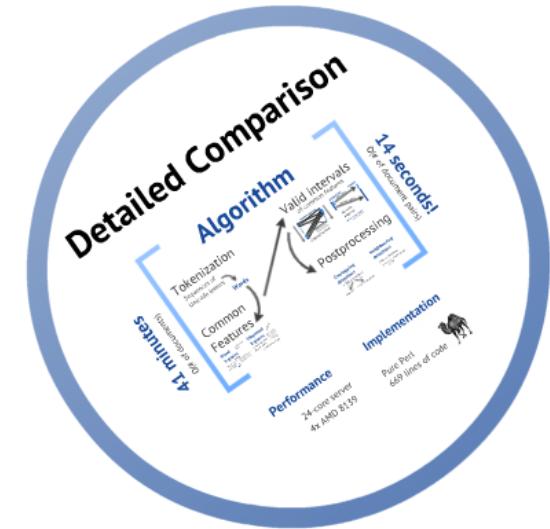
Conclusions



Three Way
Search Engine with
Multi-feature
Document
Queries
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)