

BlackLab: a researcher's best friend

Jan Niestadt, INL

*github.com/INL/BlackLab
[@BlacklabINL](https://twitter.com/BlacklabINL)*



retrieval engine

retrieval engine (**woof!**)

retrieval engine **for annotated text.**

multi-input retrieval engine for
annotated text.

multi-input, **multi-query-language**,
retrieval engine for annotated text.

easy, multi-input, multi-query-
language, retrieval engine for
annotated text.

easy, multi-input, multi-query-
language, **complex** retrieval engine for
annotated text.

easy, multi-input, multi-query-
language, complex retrieval **and**
analysis engine for annotated text.

fast, scalable, easy, multi-input, multi-
query-language, complex retrieval and
analysis engine for annotated text.

open source, fast, scalable, easy,
multi-input, multi-query-language,
complex retrieval and analysis engine
for annotated text.

open source, fast, scalable, easy,
multi-input, multi-query-language,
complex retrieval and analysis engine
for annotated text **written in Java.**

open source, fast, scalable, easy,
multi-input, multi-query-language,
complex retrieval and analysis engine
for annotated text written in Java **using**
Apache Lucene.

Actively developed, open source, fast, scalable, easy, multi-input, multi-query-language, complex retrieval and analysis engine for annotated text written in Java using Apache Lucene.

Actively developed, open source, fast, scalable, easy, multi-input, multi-query-language, complex retrieval and analysis engine for annotated text written in Java using Apache Lucene with a bright future.

Actively developed, open source, fast,
scalable, easy, multi-input, multi-
query, multi-language, complex retrieval and
analysis engine for annotated text
written in Java using Apache Lucene
with a bright future.

useful and fun search tool

verb like a adj noun

“His heart was spinning like a broken compass.”

TEI

Alto

Any XML

Any text

PageXML

FoLiA

Sketch
Engine

SRU/CQL



Your own

Lucene

CWB CQP

QL

Google-like

Simple word query:

“labrador”

Regular expression:

“.+rad.*”

Sequences of words:

“quick” “black” “labrador”

Word properties:

[pos = "adj"] [word != "yellow"] "labrador"

Token-level regex operators:

[pos = "adj"]+ "labrador"

"quick" "black"? "labrador"

"b.+"{2,} "labrador"

Matchall operator:

```
“quick” []{1,3} “labrador” // distance  
[]{3} // tri-grams
```

XML tag search:

`<ne> “Lassie” </ne>`

`“Fido” within <ne/>`

`<ne/> containing “Rover”`

More on the way:

1:[] “dog” “and” 2:[] “cat” & 1.word = 2.word

(matches “black dog and black cat”)

Want more ways to query?

Let us know!



Per Hit Per Document Hits grouped Documents grouped

Total hits: 32955
Total pages: 660

Prev 1 2 3 4 5 6 7 8 9 10 11 ... Next Toggle titles

Left context	Hit text	Right context	Lemma	Part of speech
... politiek of tenminste met vergaderen.	Hond	De mensen van de kerk ...	hond	N(soort, ev, basis, zijd, stan)
... hij zin heeft om de	hond	te schoppen als hij de ...	hond	N(soort, ev, basis, zijd, stan)
... van der Kolk De lachende	hond	Uitg. Veen, 158 biz., fl.24,90 ...	hond	N(soort, ev, basis, zijd, stan)
... Goed, er komt eens een	hond	aan je staan snuffelen. Dan ...	hond	N(soort, ev, basis, zijd, stan)
... lafaard opleverde. Daarna wilde geen	hond	meer naar accordeons luisteren. Amerika ...	hond	N(soort, ev, basis, zijd, stan)
... zeggen dat Yasser Arafat een	hond	is die moet worden vermoord ...	hond	N(soort, ev, basis, zijd, stan)
... zeggen dat Yasser Arafat een	hond	is die moet worden vermoord ...	hond	N(soort, ev, basis, zijd, stan)
... eens tegen, als tegen een	hond	." Ze gaf de man tips ...	hond	N(soort, ev, basis, zijd, stan)
... huis zou verlaten om de	hond	uit te laten. En dat ...	hond	N(soort, ev, basis, zijd, stan)
... dat terwijl wij helemaal geen	hond	hebben! " Relatie ministers met politie ...	hond	N(soort, ev, basis, zijd, stan)
... Rol Maurice en de Helpende	Hond	Op de parkeerplaats van de ...	hond	N(soort, ev, basis, zijd, stan)
... heel lief is. 'DE HELPENDE	HOND	' staat er onder. Ik sta ...	hond	N(soort, ev, basis, zijd, stan)
... geven. Had ik maar zo'n	hond	. Dan kon hij naast mij ...	hond	N(soort, ev, basis, zijd, stan)
... rolstoel te ontpoepen. De Helpende	Hond	pakt iedere avond mijn tas ...	hond	N(soort, ev, basis, zijd, stan)
... sinds twee jaar al een	hond	. Ook een Labrador, en ook ...	hond	N(soort, ev, basis, zijd, stan)
... meneer vertellen dat we die	hond	niet zo in een hete ...	hond	N(soort, ev, basis, zijd, stan)
... unieke verkiezingen", vindt M. de	Hond	van het bureau Inter/View. Hij ...	hond	N(soort, ev, basis, zijd, stan)
... en D66 gaan bij De	Hond	te rade. Hij maakt na ...	hond	N(soort, ev, basis, zijd, stan)
... bureau de plank mis. De	Hond	voorspelde dat 55 procent van ...	hond	N(soort, ev, basis, zijd, stan)
... een rol kunnen spelen". De	Hond	verklaart het verschil uit het ...	hond	N(soort, ev, basis, zijd, stan)
... opdagen. Dat gebeurde volgens De	Hond	omdat erg veel kiezers toch ...	hond	N(soort, ev, basis, zijd, stan)
... partijen hebben hier volgens De	Hond	van geprofiteerd. Schild noemt dit ...	hond	N(soort, ev, basis, zijd, stan)
... noemt dit een schijnredenering: „De	Hond	moet niet met de opkomst ...	hond	N(soort, ev, basis, zijd, stan)
... vast te stellen", zegt De	Hond	. Schild is het hiermee niet ...	hond	N(soort, ev, basis, zijd, stan)
... methode is actueler", zegt De	Hond	„Je ziet dat de uitslag ...	hond	N(soort, ev, basis, zijd, stan)
... nauwkeurig te voorspellen", zegt De	Hond	. Het échte verschil zit vooral ...	hond	N(soort, ev, basis, zijd, stan)
... stemhokje alsnog GroenLinks kiest. De	Hond	hanteert een ander correctie-mechanisme. Inter/View ...	hond	N(soort, ev, basis, zijd, stan)
... de grootste wordt", zegt De	Hond	„Daarbij komt dat bewegingen in ...	hond	N(soort, ev, basis, zijd, stan)

300:37: dat dit es des grauen seggnen, alsoe aiset hier voignet in desen brieue.
300:38: Dat es dat, ic jan sal verwaren ende berechten **de man**, ende al therseep
300:39: van putte, ende dat dartoe hort alse bailliu, ende dat bi florens rade van
300:40: henengowe, ende bi siere stieringhen. also dat van allen ghevalle, van
300:41: alre verbornessen, ende van alre verscijntnessen, ofte van alre
300:42: besterfnessen, cleine ende groet, niet vte ghenomen, florens hebben sal
300:43: ende nemen ouer al dene helt, ende ic dandre te goeder rekeninghen,
300:44: ouer minen cost ende ouer mine pine, sonder die boeten van seuen ende tuintech
300:45: scellinghen iof van beneden, die sullen sijn jans allene. neware
300:46: van den sekeren renten so sal mijn joncvrowe berte ghelden niclaus van

301:1: putte rechte scult, ende bewijsde almoessene van dien die ierst vallen.
301:2: ende alse die vergouden sijn, so sal florens, mijn joncvrowe berte
301:3: niclaus suster, ende icke jan, nemen elkerlijc van ons tderdendeel
301:4: euenghelijc van dien renten te goeder rekeninghen, , tote der tijd dat
301:5: niclaus kinder sullen hebben hare jaer. Vord sone salic jan de lude no
301:6: **de man** van putte te ghenen orlogheliken dinghen leeden, hetne ware
301:7: ombe mijns selues goet, iof mijns selues recht te behoudene, sonder bi
301:8: florens wille. Ware oec dat sake dat florens, **de man**, iof de lude van
301:9: putte ieweren leeden wilde, iof ghebieden, dar soudic hem
301:10: ongheuensdelijc, ende met trowen, vorderlijc, ende ghereet toe wesen,
301:11: alse dicke, alse hijs begbert, na mire macht, het neware dat ghinghe
301:12: ieghen so na mine maghe, dat ict met eren niet doen ne mochte, so ne
301:13: soudic hem niet dienen met mijns selues liue, maer met den luden,
301:14: ende met den mannen des lands van putte, ende met allen die ic dartoe
301:15: bringhen moghte met trowen. Ende ne mochtic de lude ende **de man**
301:16: dar toe niet hebben, ende icker met trowen toe pijnde, so soudic hem
301:17: dienen met mijns selues liue, ende met dien dat ic daer toe bringhen
301:18: mochte. ende darbi onbegrepen te sine. Ware oec dat sake dat ic des niet
301:19: ne dade ghetrowelike, so verghie ic ende verkenne, ende vertie mi alles
301:20: rechts van der voghedien des lands ende des hersceps van putte, die ic
301:21: hebbe, iof hebben moghe, ende datte te florens behoef. Hier bi sal mi
301:22: florens troestech, ghehelpech ende gheradech wesen ter
301:23: bescermenessen mijns selues, des lands, ende der lude. ende ic sal hem
301:24: weder sijn ghetrowe ende ghereet te sinen dienste met al miere macht,
301:25: sonder arghe lust in beden siden. ende mi janne niet af te doene van
301:26: der manborscepe, toter tijd dat niclaus kinder mondich worden. Vord
301:27: alse tide alse die scult, ende die almoessene vergouden sal sijn, soe
301:28: sullen die renten ioncvrowe berte van putte, florens, ende jan, euenghelijc
301:29: delen harelijc tderdendeel, toter tijd dat niclaus kinder mondich

Tip:

You can use these arrows to cycle through the hits in this document.

(205)

Europifche voetgangers en twintig ruiters, onderfteund van eenen kleinen hoop Indiäanen, onder het bevel van *Guakanahari*. Een verbaazend onderscheid! Maar 't geen deeze handvol Europeäanen ontbrak aan de meenigte, wierd by hen vergoed door hunne krygskunde, hunne wapenen, hunne paarden en hunne honden.

KAREL. Hunne honden?

VADER. Ja, KAREL! men had een koppel groote honden medegebragt, om de arme naakte Indiäanen, even als het wild, daarmede te jaagen.

LOTJE. Foei, die leelyke menfchen!

VADER. Dus was het gevaar aan beide zyden even groot, en het fond zeer te duchten, welk eene uitkomst die veldflag zou hebben.

KOLUMBUS verkoos tot het fchroomelyke tooneel, 't welk nu fond geöpend te worden, den tyd van den nacht, om dat hy hoopte, dat de duifternis den fchrik der Indiäanen by eenen onverwachten aanval zou vermeerderen. Nadat het duifter was geworden, en hy zyn klein heir onder zynen broeder *Bartholomeüs*, den Kaufchik *Guakanahari* en zich-zelfen had ver-



Groot Molechaser Corpus

Home Instituut voor Nederlandse Lexicologie CLARIN

... Thalia , honderd meter verderop, een	hond	aansloeg. Steffie tilde haar hoofd ...	hond	N(soort, ev, basis, zijd, stan)
... is de eigenaar van een	hond	aansprakelijk voor de schade aan ...	hond	N(soort, ev, basis, zijd, stan)
... en de man met de	hond	aansprakelijk zijn. U als bestuurder ...	hond	N(soort, ev, basis, zijd, stan)
... op het ogenblik dat de	hond	aanstalten maakt om het voedsel ...	hond	N(soort, ev, basis, zijd, stan)
... flanken te strelen. Indien de	hond	aanstalten maakt om terug te ...	hond	N(soort, ev, basis, zijd, stan)
... melk met mineraalwater. Maakt de	hond	aanstalten te gaan overgeven, dan ...	hond	N(soort, ev, basis, zijd, stan)
... moest komen dat de zieke	hond	aanstalten zou maken om bij ...	hond	N(soort, ev, basis, zijd, stan)
... moest komen dat de zieke	hond	aanstalten zou maken om bij ...	hond	N(soort, ev, basis, zijd, stan)
... moment hadden we wellicht Koreaanse	hond	aanvaard. We kloppen ons vrolijk ...	hond	N(soort, ev, basis, zijd, stan)
... stier, de teef die de	hond	aanvaardde, de kleine vrouwtjeskatten jankend ...	hond	N(soort, ev, basis, zijd, stan)
... de opvolger van Man bijt	hond	aanvaardt de VRT geen voorstellen ...	hond	N(soort, ev, basis, zijd, stan)
... liet de andere agent zijn	hond	aanvallen, waardoor hij de agressieveling ...	hond	N(soort, ev, basis, zijd, stan)
... dan de beet van de	hond	. Aanvallende honden kunnen altijd gevaarlijk ...	hond	N(soort, ev, basis, zijd, stan)
... gebeuren dat ze onverhoeds de	hond	aanvalt als deze per ongeluk ...	hond	N(soort, ev, basis, zijd, stan)
... is niet nodig dat de	hond	aanvalt opdat de schapen respect ...	hond	N(soort, ev, basis, zijd, stan)
... straathonden die een goed verzorgde	hond	aanvielen, kennelijk iemands huisdier. Het ...	hond	N(soort, ev, basis, zijd, stan)
... OM, op verzoek van De	Hond	, aanvullend sporenonderzoek, onder andere naar ...	hond	N(soort, ev, basis, zijd, stan)
... gedurende de groei van de	hond	aanwezig, dan zal zich een ...	hond	N(soort, ev, basis, zijd, stan)
... dus te opvallend voor de	hond	aanwezig. De nieuwste generatie elektronische ...	hond	N(soort, ev, basis, zijd, stan)
... kwam. Er bleek alleen een	hond	aanwezig die met de telefoon ...	hond	N(soort, ev, basis, zijd, stan)
... de spookkamer was ook de	hond	aanwezig, een bijna één jaar ...	hond	N(soort, ev, basis, zijd, stan)
... op ieder erf een waakzame	hond	aanwezig. Er liep niemand op ...	hond	N(soort, ev, basis, zijd, stan)
... Swarts, als enige agent met	hond	aanwezig, ergerde zich aan de ...	hond	N(soort, ev, basis, zijd, stan)
... net een klant met een	hond	aanwezig, wat volgens de wet ...	hond	N(soort, ev, basis, zijd, stan)
... vingers. Er kwam een mankende	hond	aanzetten. Had ik 't zo ...	hond	N(soort, ev, basis, zijd, stan)
... goed aan te zetten. D.w.z.	hond	aanzetten met een commando ZOEK ...	hond	N(soort, ev, basis, zijd, stan)
... mens de maag van de	hond	aanzetten tot het afscheiden van ...	hond	N(soort, ev, basis, zijd, stan)
... besmetting gebeurt door een dierenbeet (hond	, aap,...). De vaccinatie bestaat uit ...	hond	N(soort, ev, basis, zijd, stan)
... fenolgroep (creosoot, Dettol) kan de	hond	aardig afrekenen; de kat kan ...	hond	N(soort, ev, basis, zijd, stan)
... op alleen grasland. Zodra onze	hond	aardig door heeft wat speuren ...	hond	N(soort, ev, basis, zijd, stan)



Per Hit Per Document Hits grouped Documents grouped

Group by word left

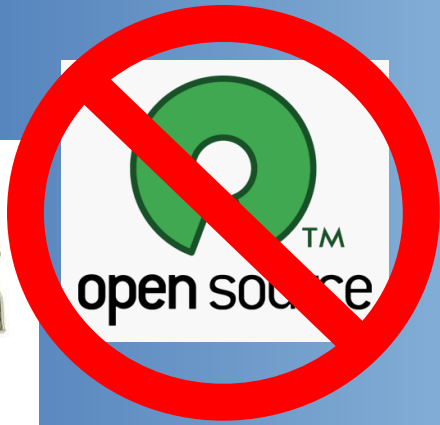
Group	Hits
de	9969
een	5590
De	2175
zijn	1288
bijt	794
geen	700
uw	581
die	502
Een	471
mijn	441
hun	395
haar	355
jonge	354
en	341
je	305
gebeten	238
onze	166
grote	162
Geen	147
derde	147
geslagen	133
zwarte	130
andere	123
dode	117
dolle	108
Brakke	106
met	103

Works in progress:

- Random sampling
- Collocations
- More cool tools planned!

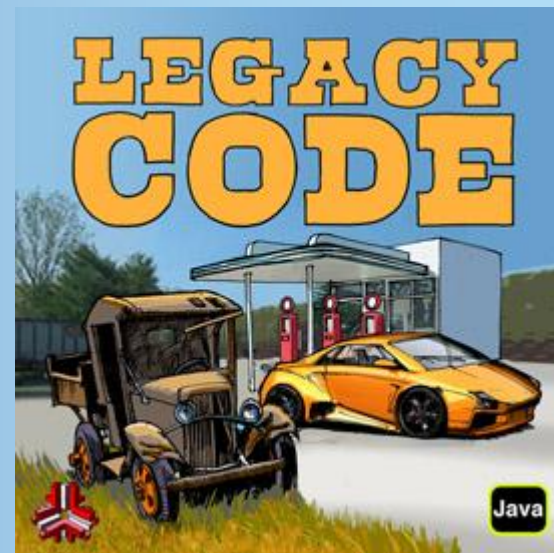
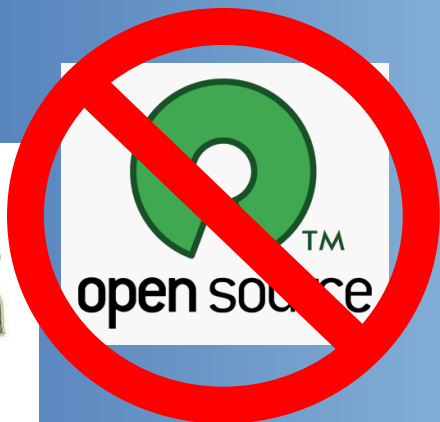


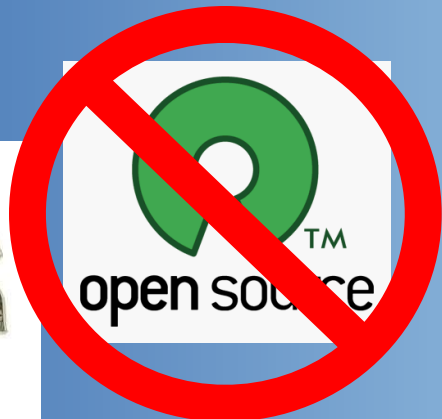
INL SCHATKAMER VAN
DE NEDERLANDSE TAAL

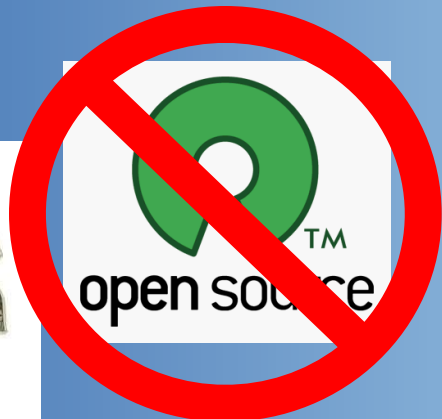


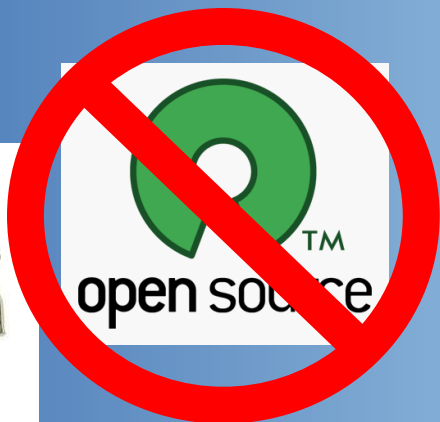
INL

INL SCHATKAMER VAN
DE NEDERLANDSE TAAL



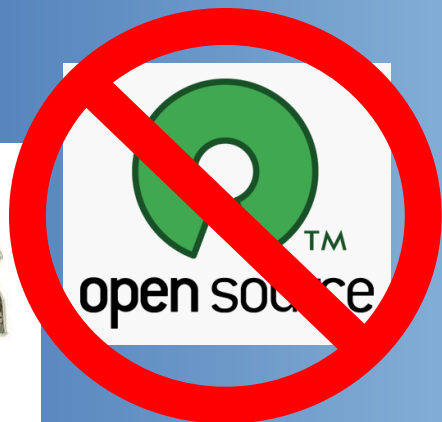






**LEGACY
CODE**





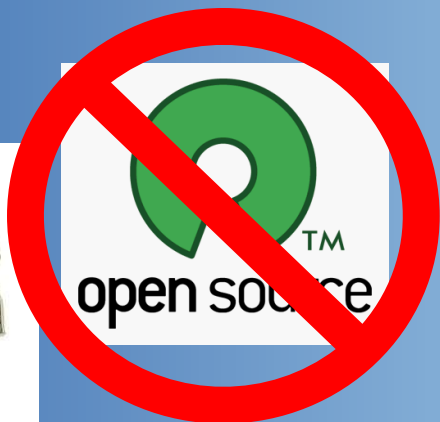
**LEGACY
CODE**



Delivery 1 Delivery 2 Delivery 3



Incremental plan



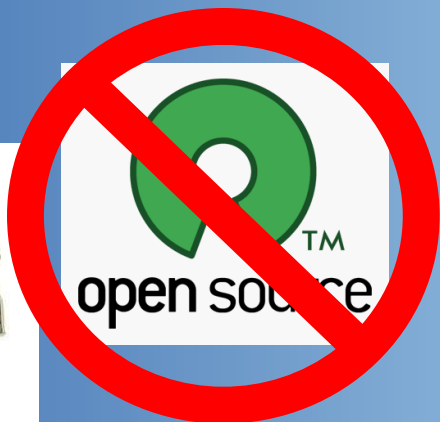
**LEGACY
CODE**



Delivery 1 Delivery 2 Delivery 3



Incremental plan



open source



Java

LEGACY
CODE



Delivery 1 Delivery 2 Delivery 3



Incremental plan

Fluorene



400M word corpus, single machine

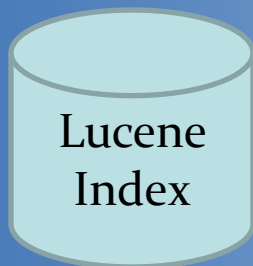


INL SCHATKAMER VAN
DE NEDERLANDSE TAAL

Search / operation	# Results	Search time
“stad”	± 190,000	< 1 s
“de” “stad”	± 120,000	2 s
“de” “sta.*”	± 370,000	6 s
“de” “grote”? “stad”	± 120,000	5 s
“die.*”{2,}	± 29,000	10 s
“water” / group by word left	± 140,000	< 1 s
“boven” / sort by left context	± 160,000	< 1 s
“onder” / group by word right	± 780,000	6 s

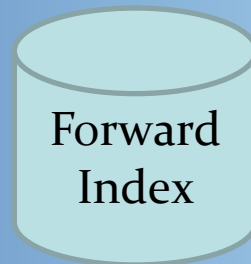
Scalability / speed:

- Lucene: ++
- Sorting/grouping: ++
- Displaying results: +
- Distributed: +/- (future)



Lucene
Index

Finding hits
(SpanQuery classes,
many custom)



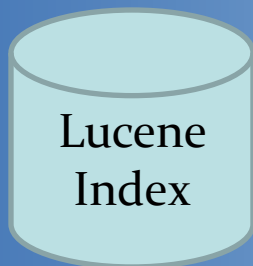
Forward
Index

Fast sorting /
grouping on
context;
collocations

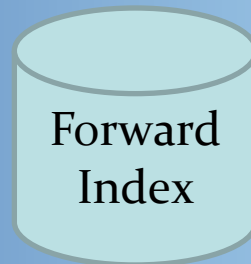


Content
Store

Displaying
results



Finding hits
(SpanQuery classes,
many custom)



Fast sorting /
grouping on
context;
collocations



Displaying
results

Idea: integrate these two to improve speed
and reduce disk/memory usage

Apache
Solr  ?

The image shows the Apache Solr logo, which consists of the word 'Apache' in a smaller font above the word 'Solr' in a larger, bold font. To the right of 'Solr' is a sunburst icon made of multiple colored segments (orange, yellow, red) radiating from a central point. A large black question mark is positioned to the right of the sunburst.

Improving Access to Text

IMPACT



INL SCHATKAMER VAN
DE NEDERLANDSE TAAL



INL SCHATKAMER VAN
DE NEDERLANDSE TAAL



NEDER lab

11010001
01001011
11101001
00110101



the
research
Dutch
Nederlab
Laboratory
in
informatics
and
language
culture
change
for

1



INL SCHATKAMER VAN
DE NEDERLANDSE TAAL

1



2

Data in supported
format (ask us!)

1



2

Data in supported
format (ask us!)

3

Index it



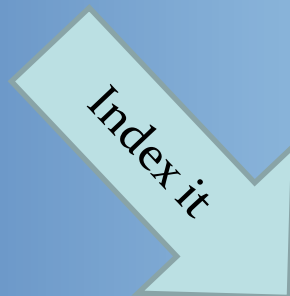
1



2

Data in supported
format (ask us!)

3



4

BlackLab index;
Test with
QueryTool

```
public class BlackLabDemoServlet extends HttpServlet {
    Searcher searcher;

    @Override
    protected void doGet(HttpServletRequest request, HttpServletResponse response) throws
        try {
            // Parse and execute query
            String query = request.getParameter("query");
            TextPattern pattern = CorpusQueryLanguageParser.parse(query);
            Hits hits = searcher.find(pattern);

            // Output hits with a little context
            ServletOutputStream out = response.getOutputStream();
            for (Hit hit: hits) {
                Concordance conc = hits.getConcordance(hit);
                out.println(conc.left + "<b>" + conc.hit + "</b>" + conc.right + "<br/>");
            }
        } catch (Exception e) {
            throw new ServletException(e);
        }
    }

    @Override
    public void init(ServletConfig config) throws ServletException {
        try {
            searcher = new Searcher(new File("/home/jan/testindex"));
        } catch (Exception e) {
            throw new ServletException(e);
        }
    }

    @Override
    public void destroy() {
        searcher.close();
    }
}
```

```
public class BlackLabDemoServlet extends HttpServlet {
    Searcher searcher;

    @Override
    protected void doGet(HttpServletRequest request, HttpServletResponse response) throws

    // Parse and execute query
    String query = request.getParameter("query");
    TextPattern pattern = CorpusQueryLanguageParser.parse(query);
    Hits hits = searcher.find(pattern);

    // Output hits with a little context
    ServletOutputStream out = response.getOutputStream();
    for (Hit hit: hits) {
        Concordance conc = hits.getConcordance(hit);
        out.println(conc.left + "<b>" + conc.hit + "</b>" + conc.right + "<br/>");
    }


}

@Override
public void init(ServletConfig config) throws ServletException {
    try {
        searcher = new Searcher(new File("/home/jan/testindex"));
    } catch (Exception e) {
        throw new ServletException(e);
    }
}

@Override
public void destroy() {
    searcher.close();
}
}
```

A green highway sign with white text and an arrow. The sign is mounted on a metal structure against a blue sky background. The text on the sign reads 'The Future' in a large, white, sans-serif font. Below it, in a smaller white font, is 'NEXT EXIT' followed by a white arrow pointing upwards and to the right.

The Future

NEXT EXIT 



Thank you!
Questions?

Jan.Niestadt@inl.nl
Twitter: @BlackLabINL
GitHub: /INL/BlackLab