Exploiting Citation Graphs in Large Corpora to Improve Relevance on Broad Queries

Haystack 2023

Marc Morissette
Director of Technology
lexum

Who are we?

Lexum is a software company that designs and operates online legal information delivery products. We specialize in the management and the publishing of legal information over the internet.

CanLII, the **Can**adian **L**egal **I**nformation **I**nstitute, is a non-profit organization founded in 2001 by the Federation of Law Societies of Canada on behalf of its 14-member law societies. Its mandate is to provide efficient and open online access to judicial decisions and legislative documents.

CanLII – Content

Legislation

Statutes & Regulations

85,000 (legislations)

1,450,000 (sections)

Case Law

Court & Tribunal Decisions

3,100,000

Legal Commentary

Books, Law Review Articles, Practice Manuals

100,000

lexum

CanLII – Users

Lawyers

Legal Researchers

"joint submission" and "crown" /p "repudiate”

“Nuisance” and protest and “180(1)”

causation /10 death and remov! /5 ventilator

(“Misrepresentation” /p (“did not read” or “failed to read”)) and “specific performance”

The Public

All walks of life

Joint custody distance

Terminated for alleged marijuana impairment

Trump tower

lexum

Broad Queries

pension benefits divorce

insurance dispute resolution

disability discrimination

fair dealing research

bankrupt discharge

strike vote validity

request for access to records

hearsay admissibility

costs awards factors

privacy invasion

residential eviction enforcement

sale agreement cancellatio

constructive dismissal

freedom of expression

recognizance to keep the peace

charter 2b

aboriginal land titles

custody assessment

pipeda s 7

Broad Queries

pension benefits divorce

insurance dispute resolution

disability discrimination

fair dealing research

bankrupt discharge

strike vote validity

request for access to records

hearsay admissibility

costs awards factors

privacy invasion

residential eviction enforcement

sale agreement cancellation

constructive dismissal

recognizance to keep the peace

freedom of expression

charter 2b

aboriginal land titles

custody assessment

pipeda s 7

How Do You Handle Broad Queries?

Legislation

Statutes & Regulations

85,000 (legislations)

1,450,000 (sections)

Case Law

Court & Tribunal Decisions

3,100,000

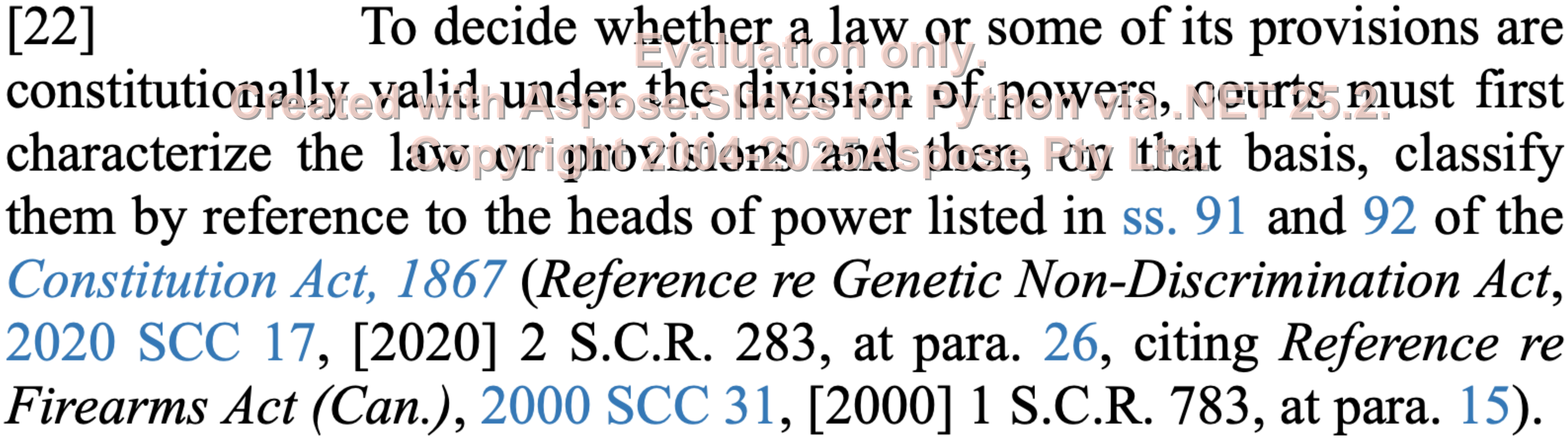
Legal Commentary

Books, Law Review Articles, Practice Manuals

100,000

lexum

Idea: Identify "Authoritative" Documents Via Citations



[22] To decide whether a law or some of its provisions are constitutionally valid under the division of powers, courts must first characterize the law or provisions and then, on that basis, classify them by reference to the heads of power listed in ss. 91 and 92 of the *Constitution Act, 1867* (*Reference re Genetic Non-Discrimination Act*, 2020 SCC 17, [2020] 2 S.C.R. 283, at para. 26, citing *Reference re Firearms Act (Can.)*, 2000 SCC 31, [2000] 1 S.C.R. 783, at para. 15).

Idea: Identify “Authoritative” Documents

Legislation

Statutes & Regulations

1,450,000 sections

cited 51,000,000 times (avg 35 per section)

Case Law

Court & Tribunal Decisions

3,100,000 decisions

cited 14,600,000 times (avg 4.7 per decision)

Legal Commentary

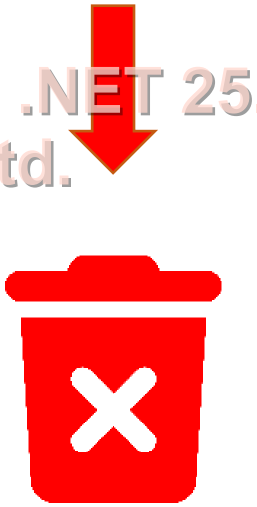
Books, Law Review Articles, Practice Manuals

100,000 documents

lexum

Intuition: Sort By *Citations*

BM25 score



sort: citedCount

Sort By *Popularity*: Does it Work?

- Navigational searches: yes
- On short highly representative fields (keywords): maybe
- For everything else

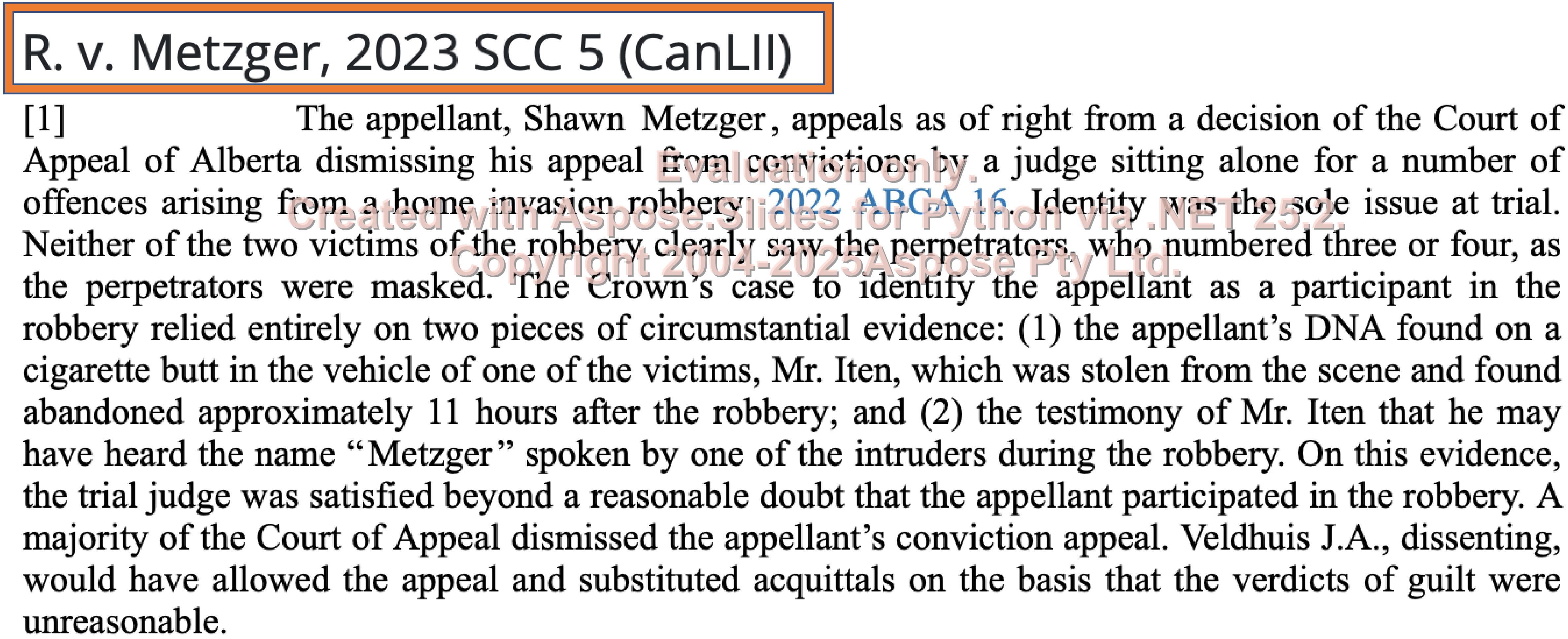
Legislation (riots)

Act, or that person’s deputy,

who receives notice that, at any place within the jurisdiction of the person, twelve or more persons are unlawfully and riotously assembled together shall go to that place and, after approaching as near as is safe, if the person is satisfied that a riot is in progress, shall command silence and thereupon make or cause to be made in a loud voice a proclamation in the following words or to the like effect:

Her Majesty the Queen charges and commands all persons being assembled immediately to disperse and peaceably to depart to their habitations or to their lawful business on the pain of being guilty of an offence for which, on conviction, they may be sentenced to imprisonment for life. GOD SAVE THE QUEEN.

Case Law

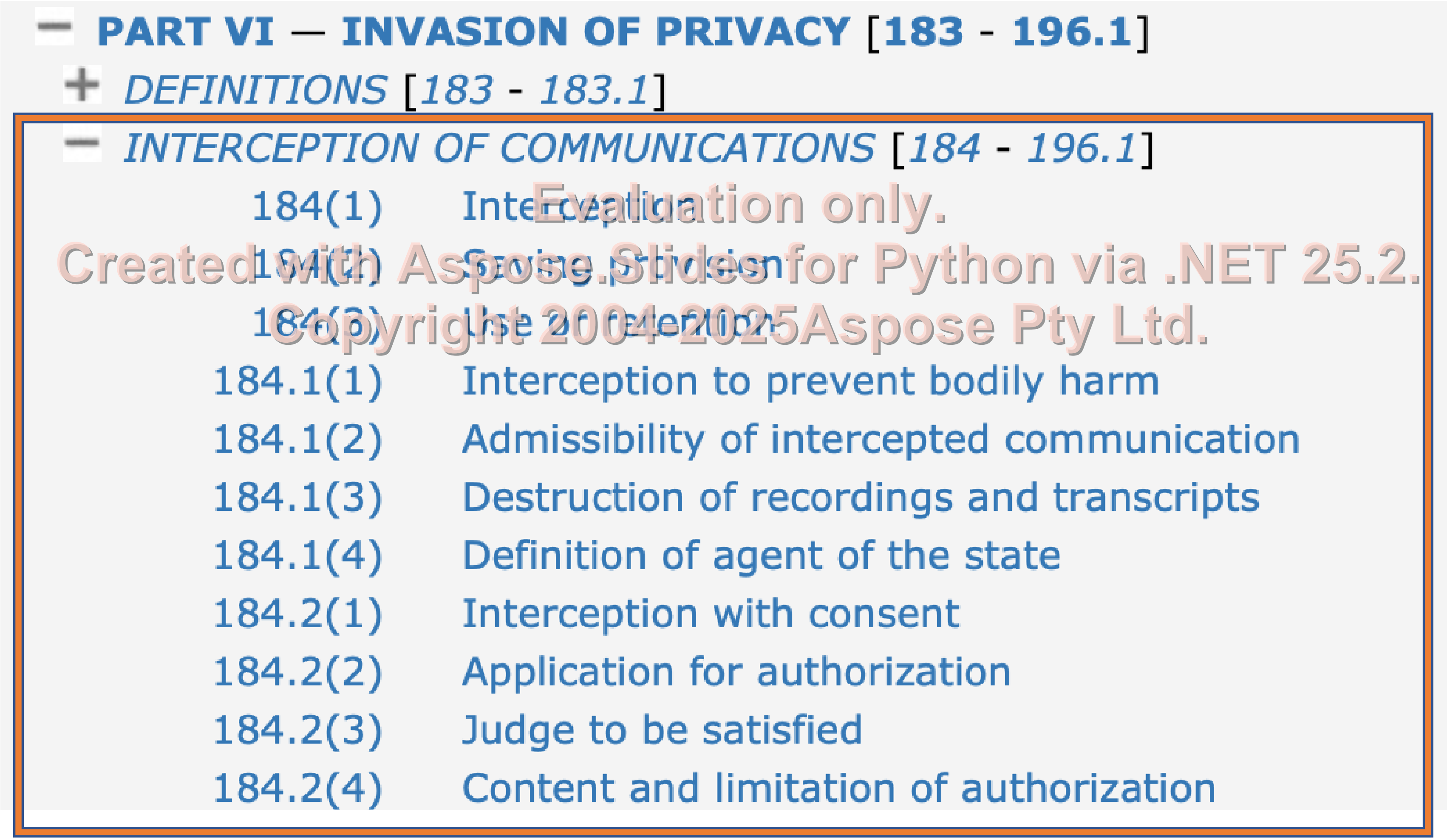


R. v. Metzger, 2023 SCC 5 (CanLII)

[1] The appellant, Shawn Metzger, appeals as of right from a decision of the Court of Appeal of Alberta dismissing his appeal from convictions by a judge sitting alone for a number of offences arising from a home invasion robbery: 2022 ABCA 16. Identity was the sole issue at trial. Neither of the two victims of the robbery clearly saw the perpetrators, who numbered three or four, as the perpetrators were masked. The Crown's case to identify the appellant as a participant in the robbery relied entirely on two pieces of circumstantial evidence: (1) the appellant's DNA found on a cigarette butt in the vehicle of one of the victims, Mr. Iten, which was stolen from the scene and found abandoned approximately 11 hours after the robbery; and (2) the testimony of Mr. Iten that he may have heard the name "Metzger" spoken by one of the intruders during the robbery. On this evidence, the trial judge was satisfied beyond a reasonable doubt that the appellant participated in the robbery. A majority of the Court of Appeal dismissed the appellant's conviction appeal. Veldhuis J.A., dissenting, would have allowed the appeal and substituted acquittals on the basis that the verdicts of guilt were unreasonable.

lexum

Is There a Better Way?



PART VI — INVASION OF PRIVACY [**183 - 196.1**]

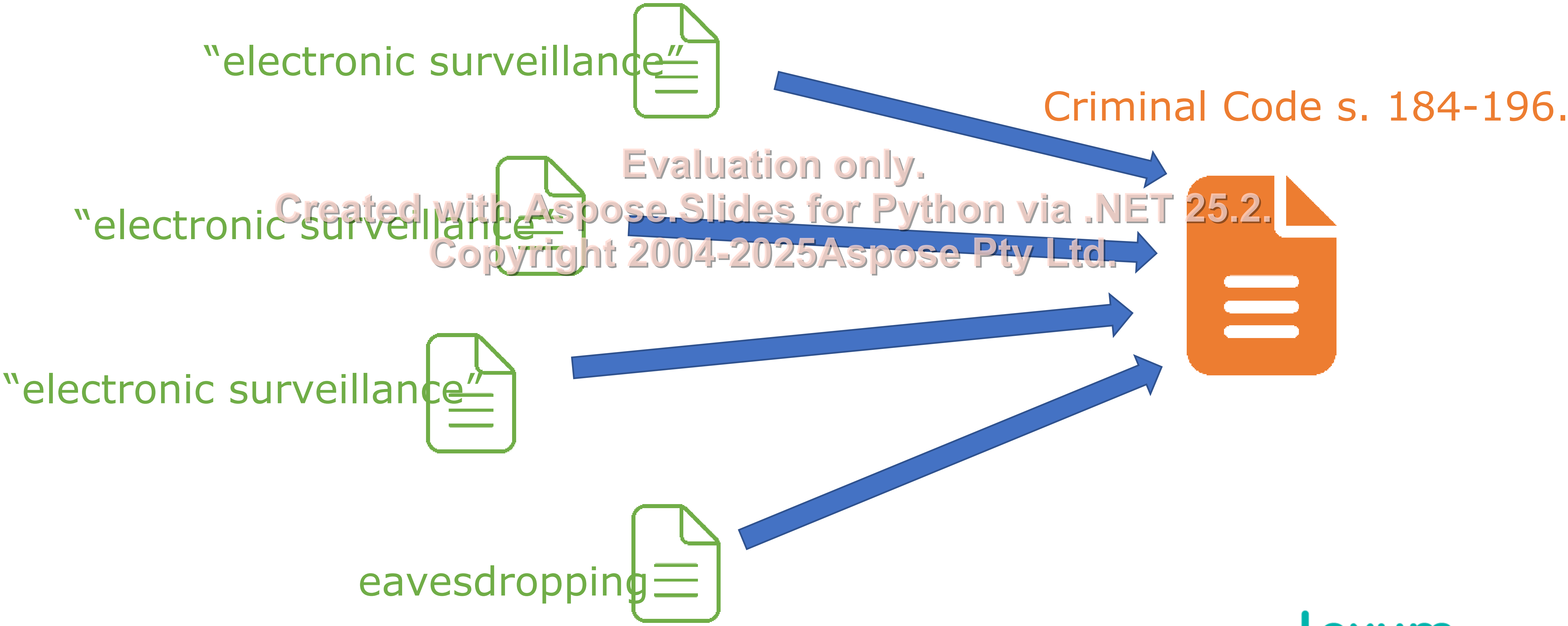
DEFINITIONS [*183 - 183.1*]

INTERCEPTION OF COMMUNICATIONS [*184 - 196.1*]

184(1)	Interception
184(2)	Saving provision
184(3)	Use or retention
184.1(1)	Interception to prevent bodily harm
184.1(2)	Admissibility of intercepted communication
184.1(3)	Destruction of recordings and transcripts
184.1(4)	Definition of agent of the state
184.2(1)	Interception with consent
184.2(2)	Application for authorization
184.2(3)	Judge to be satisfied
184.2(4)	Content and limitation of authorization

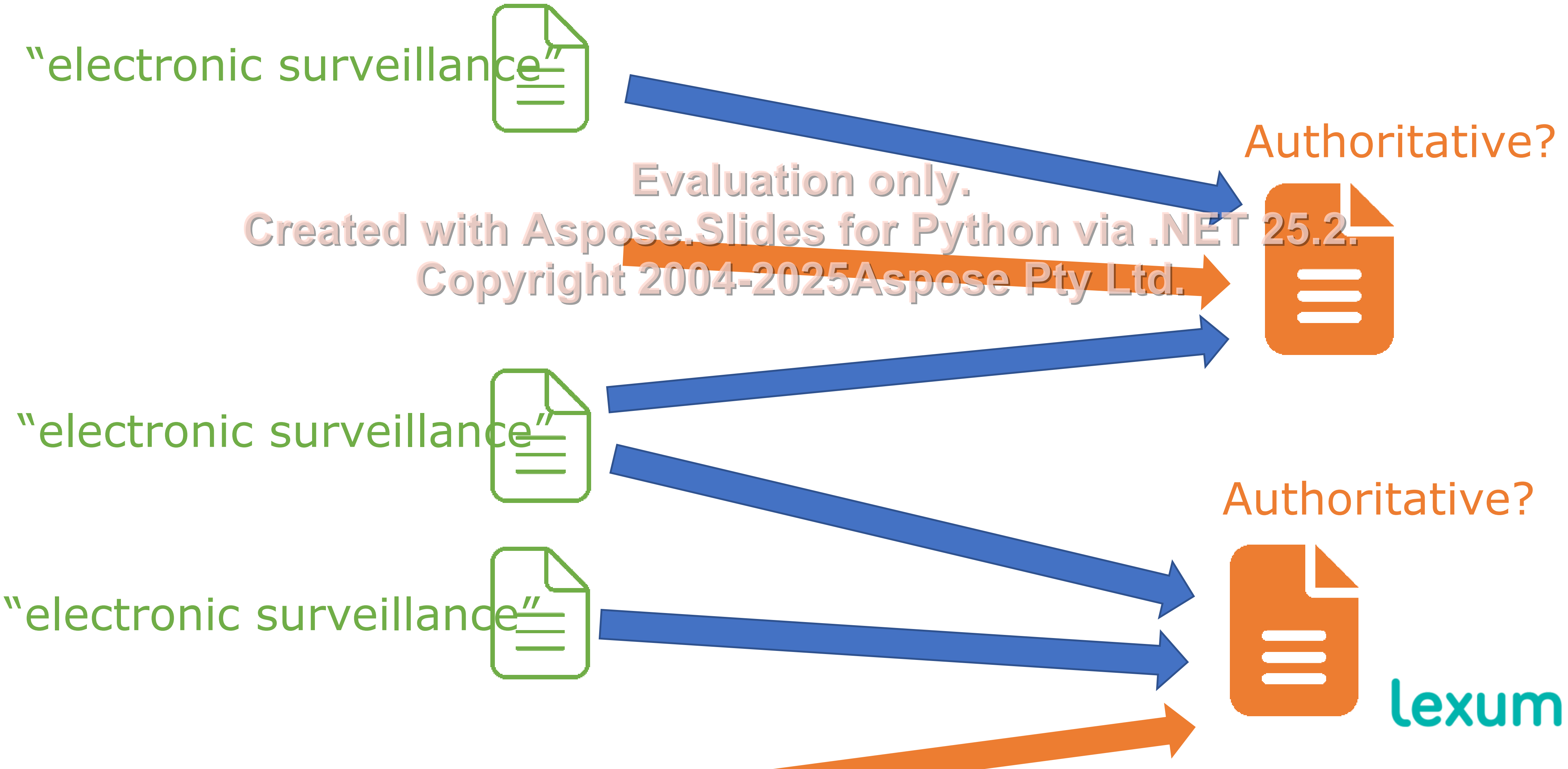
lexum

Citation Graph


“electronic surveillance”
“electronic surveillance”
“electronic surveillance”
eavesdropping
Criminal Code s. 184-196.

lexum

Thinking of An Algorithm


"electronic surveillance"
"electronic surveillance"
"electronic surveillance"
Authoritative?
Authoritative?

lexum

Algorithm

1. **A** = Set of nodes (docs) that contain the search terms
2. L(**A**,*) = All edges departing from **A** (citations)
3. **B** = All nodes pointed to by L(**A**,*) (cited documents)
4. L(*,**B**) = All edges pointing to **B** (for stat. analysis)
5. For each b in **B**, calculate some

lexum

The Math

- For every eligible document b in **B**, we want to know whether b is authoritative
- We have two competing hypotheses that we will test
  - H_0 = b is not authoritative
  - H_1 = b is authoritative
- thus,
  - H_0 = incoming links should contain few documents in **A**
  - H_1 = incoming links should contain lots of documents in **A**

The Math

- Both hypotheses follow a multinomial distribution where each incoming citation is a trial
- A Bernouilli distribution $f(t|, p_h)$ where
  - t = |L(*, b)|
  - p_0 = |**A**|/|**E**| where **E** is the set of documents eligible to cite b
  - p_1 = R*G where R and G are constants and
    - R is the query’s expected recall metric
    - G is the ratio of edges L(*, b) that are relevant

The Math

- The next step is to test the hypothesis using either
  - A one-sided statistical test on H_1
  - The Neyman-Pearson Lemma on H_0 and H_1

- Perform the statistical test using the desired confidence interval, say 95%, and we’re done!

WRONG!

WRONG!

- 6,000 documents mention *electronic surveillance*
  - Together, they cite 12,000 documents = |B|
- We expect, maybe 1 to 20 authoritative docs (true positive)

Base Rate Fallacy

- 95% confidence means 5% error rate * 11,990 = 600 false positives
- You want at least 99.99%+ confidence in such

Integrating into Solr (or OpenSearch)

1. **A** = The set of nodes that contain the search terms **Search**
2. L(**A**,*) = All edges departing from **A** (citations) **DocValues**
3. **B** = All nodes pointed to by L(**A**,*) (cited documents) **ord values**
4. L(*,**B**) = All edges pointing to **B** (for stat. analysis) **Invariant**
5. For each b in **B**, calculate some probabilities **Facet**

lexum

Integrating into Solr (or OpenSearch)

- Implementation mostly resides in a “simple” AggValueSource plugin
    - A facet aggregation

- Easy approximation: Semantic Knowledge Graph
    - Originally developed for query expansion
    - Facet function: *relatedness* in Solr and *significant_terms* in OpenSearch
    - Math is different but a good approximation with minCount > 10

Merging With Regular Results

- Authoritative results are transparently merged in with traditional search results
  - Union mode: added to results if missing – legislations
  - Intersection mode: reranking only – case law

- Implemented using a Solr SearchComponent
  - Probabilities are converted to a parameterized score

lexum

Demo Queries – Electronic surveillance

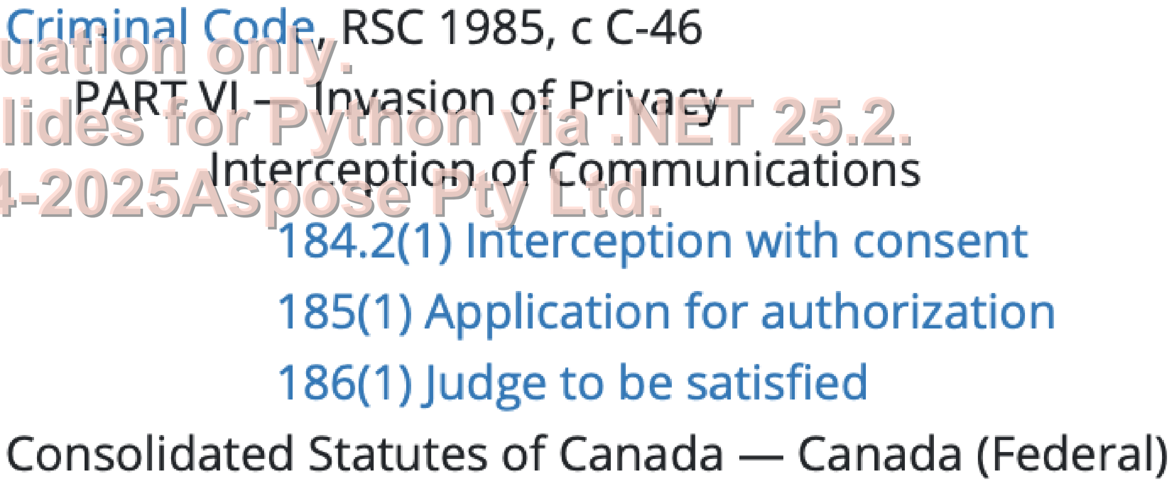
Successful Queries

- *electronic surveillance*
- *electronic interception*
- *wiretap*
- *audio interception*

Unsuccessful Queries

- *eavesdropping*

Result



Criminal Code, RSC 1985, c C-46

PART VI — Invasion of Privacy

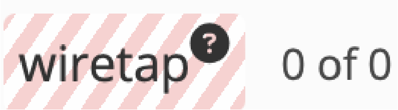
Interception of Communications

184.2(1) Interception with consent

185(1) Application for authorization

186(1) Judge to be satisfied

Consolidated Statutes of Canada — Canada (Federal)



wiretap ? 0 of 0

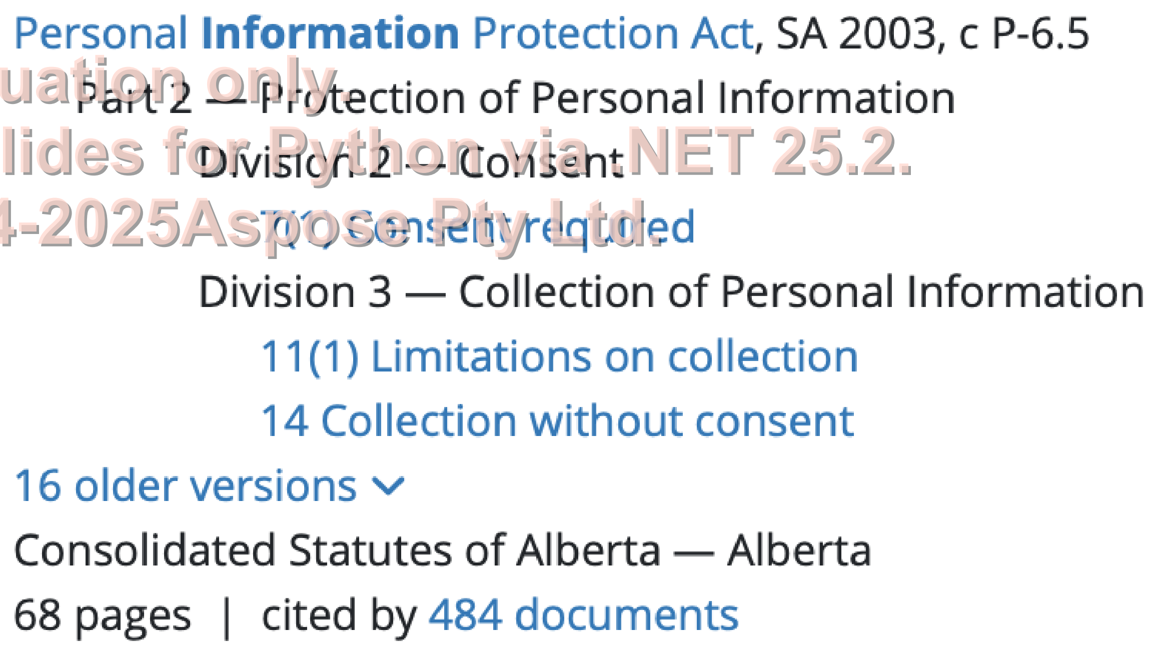
lexum

Demo Queries – Personal Information

Successful Queries

- *personally identifiable information*
- *pipeda s7*
- *consent to record information*
- *consent to collect information*

Result



Personal **Information** Protection Act, SA 2003, c P-6.5

Part 2 — Protection of Personal Information

Division 2 — Consent

7(1) Consent required

Division 3 — Collection of Personal Information

11(1) Limitations on collection

14 Collection without consent

16 older versions ⌄

Consolidated Statutes of Alberta — Alberta

68 pages | cited by 484 documents

lexum

Results vary

Evaluations

- No user engagement data yet
- Subject matter expert evaluated effect of Citation Graph on 132 queries:
  - 27% of results improved
  - 12% of results worsened
  - 60% unchanged
- We have improved integration since then

lexum

Lessons Learned

- Simple algorithms (BM25) are intuitive to users
  - It gives them a measure of control

- If you rate result engagement from 0 to 10, returning a document where none of the queried words appear rates as a solid **-10**.

This term could not be found in the current document. However, your query appears disproportionally often in documents citing this document, which is considered by the search engine a strong indication of relevance.

Lessons Learned

- Correctly adjust **E**: the set of documents eligible to cite b
  - Time bias
  - Large jurisdiction bias

lexum

Lessons Learned

- Math: apply some smoothing edge cases
  - Disable on small **A** and small **B**

- Maximize precision at the expense of recall
  - Avoid systematic errors
  - Req: *2002 SCC 1* vs
    doc text: *2002 SCC 45*, *[2002] 1 RCS 13*

lexum

Thank you!

Marc Morissette
morissette@lexum.com

lexum