WSM Project 2 — The Lemur Project

1102 Web Search and Mining,
Computer Science,
National ChengChi University.

Available Toolkits

- **Lemur**
- **Indri**

C++

- Lucene
- Terrier
- Galago
- Okapi http://okapi.opentag.com
- Haystack http://haystacksearch.org
- …

You are allowed to apply any toolkit to Project 2.

You can find more related links from wm5 website.

Available Toolkits

- **Lemur**
- **Indri**

We focus the two in this presentation.

- Lucene
- Terrier
- Galago
- Okapi
- Haystack
- …

http://www.lemurproject.org/

Lemur 4.12

Indri 5.x

Available Toolkits

- **Lemur**
- **Indri**
- Lucene
- Terrier
- Galago
- Okapi
- Haystack
- …

We focus the two in this pr

http://www.lemurp

Lemur 4.12

Indri 5.x

Available Toolkits

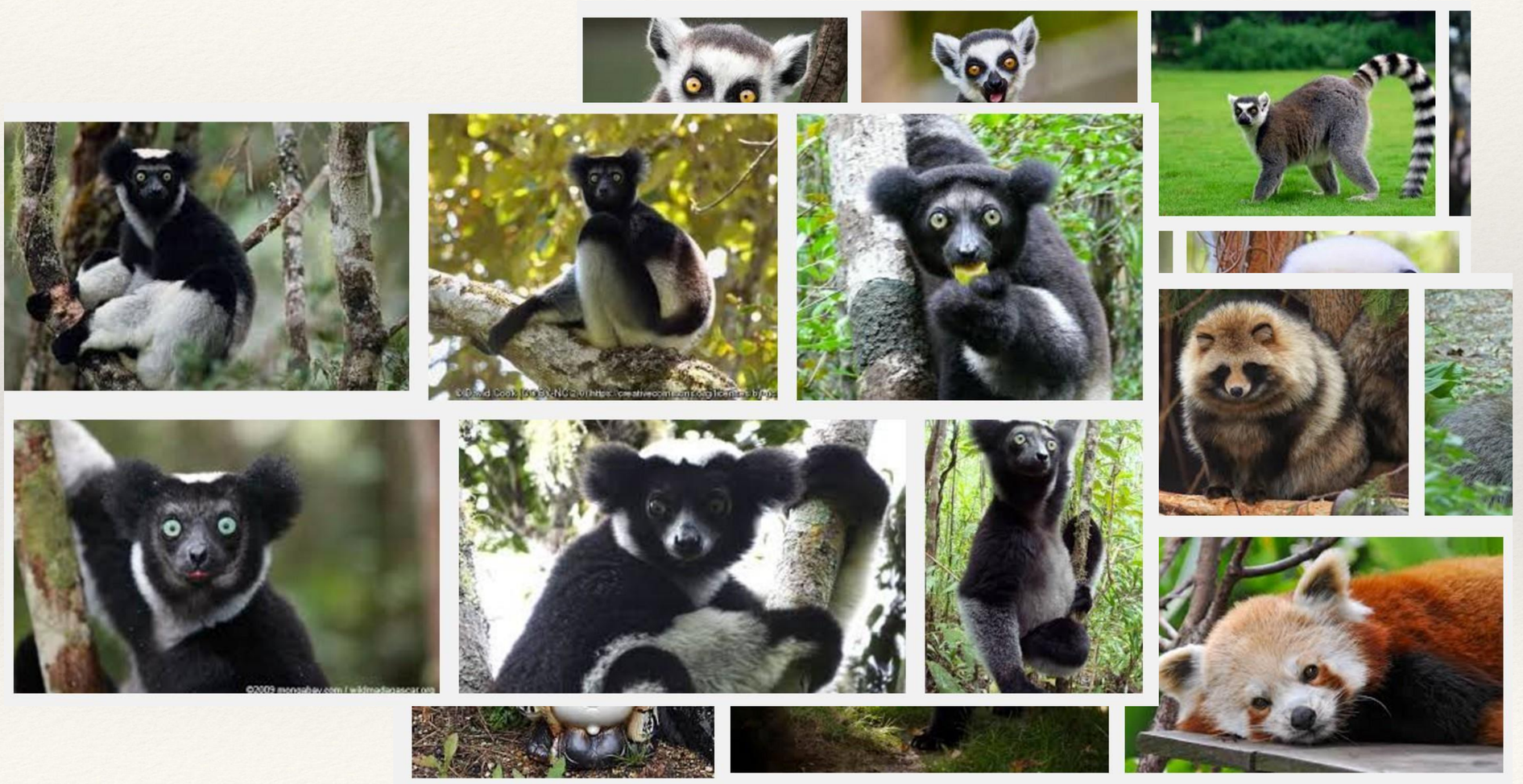
- **Lemur**
- **Indri**
- Lucene
- Terrier
- Galago
- Okapi
- Haystack
- …

We focus the two in this pr

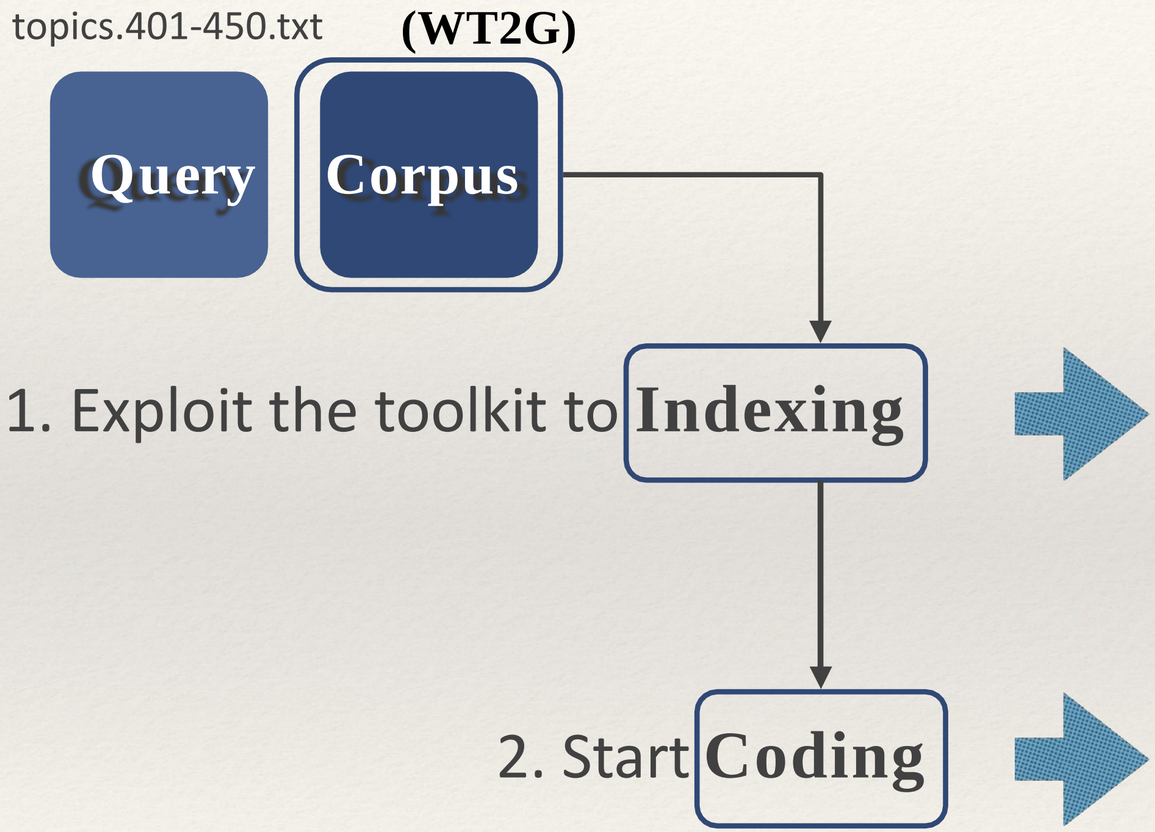
http://ww

Available Toolkits

- **Lemur**
- **Indri**
- Lucene
- Terrier
- Galago
- Okapi
- Haystack
- …



Your Task



- ./BuildIndex — lemur
- ./IndriBuildIndex — Indri

- Develop your own program **or**
- Code with API

implement the retrieval system

Documentation and Support

- **Lemur API**
  — Lemur Toolkit Documentation
- **Indri API**
  — Automatically generated source code documentation.
- **Lemur Project Wiki**
  — Wiki pages of documentation for the Lemur Project software, including the Indri Search Engine.

Here is a list of all namespaces with brief descriptions:

Indri	
indri	Namespaces within the **indri** system
indri::api	**Indri** API classes for interacting with **indri** collections
indri::atomic	Atomic actions for **thread** support
indri::collection	Document manager and ancillary **collection** components
indri::file	Filesystem interaction and file-based storage classes
indri::index	**Index** construction and interaction components
indri::infnet	Inference **net** and inference **net** node classes
indri::lang	**Indri query** language nodes and support classes
indri::net	**Indri** network components
indri::parse	File input, parsing, stemming, and stopping classes
indri::parse::CharClass	
indri::query	**Indri query** processing and scoring components
indri::server	**Indri query server** classes
indri::thread	**Thread** and threading support classes
indri::utility	Utility classes for **indri** components
indri::xml	XML support classes

Installation —Lemur4.12

- To make the lemur toolkit library and applications
  1. go to top level lemur directory.
  2. run `make` (or gmake)

- To clean the lemur toolkit (remove everything but the source)
  1. go to top level lemur directory.
  2. run `make clean` (or gmake clean)

- To install the lemur toolkit library and applications.
  1. follow the instructions above for configuring and making the toolkit library and applications.
  2. go to top level lemur directory.
  3. run `make install` (or gmake install)

Compiling and Installing on Linux and Mac OS X

Cygwin ← **As to Windows, please refer to Wiki Pages.**

Installation

welcome discussion

鮑聖文

8 April at 18:12

在Ubuntu上面要編譯lemur4.1.2，試了一些方法之後，得到了一個不錯的解法

**** this is for lemur ****

sudo apt-get update
sudo apt-get upgrade
sudo apt-get install gcc-4.4 g++-4.4

cd LEMUR_DIR
export CC=gcc-4.4
export CXX=g++-4.4
./configure
make -jn (n等於你電腦的核心數，例如 -j8)

sudo make install (假如要安裝到系統的話)

然後indri用最新版的gcc或者clang都可以過

有任何問題的話，請不吝提出，說不定我這個步驟成功只是走運w

Installation —Indri5.x

- To configure **indri**
  a. go to top level **indri** directory.
  b. run `configure` to generate MakeDefns and Makefile.app. `configure` accepts following arguments:
     - `--prefix=<install-directory>` Specifies the base directory for installation. Default is /usr/local.
     - `--enable-assert` Enable assert statements in the code. Default is disabled.
     - `--enable-java` compiles and installs the swig generated java wrappers. Default is disabled.
     - `--enable-php` compiles and installs the swig generated php wrappers. Default is disabled.
     - `--enable-csharp` compiles and installs the swig generated C# wrappers. Default is disabled.
     - `--with-javahome=<path>` Path to JAVAHOME for compiling the swig generated shared library.
     - `--with-php-config=<path>` Path to php-config binary. Only required if php-config is not on the path.
     - `--with-swig=<path>` Path to swig binary. Only required if the wrapper interfaces are changed.
     - `--with-site-seed=<hostname>` Hostname to use as the seed for building a site search index.

- To make the **indri** library and applications
  a. go to top level **indri** directory.
  b. configure **indri**
  c. run `make` (or gmake)

- To clean the **indri** source tree (remove everything but the source)
  a. go to top level **indri** directory.
  b. run `make clean` (or gmake clean)

- To install the **indri** library and applications.
  a. go to top level **indri** directory.
  b. configure **indri**
  c. run `make` (or gmake)
  d. run `make install` (or gmake install)

Compiling and Installing on Linux and Mac OS X

As to Windows, please refer to Wiki Pages.

Indexing —Data Format

- **Lemur**
  - TREC Text
  - **TREC Web**
  - HTML
- **Indri**
  - TREC Text
  - **TREC Web**
  - Plain Text
  - DOC
  - PPT
  - …

(Option 1) Develop your own program

similar to Project 1

VSM model / LM moel

(1) Stemming & Removing Stop Words & Indexing / Load **Inverted List**

(2) Transfer Queries into a Vector / Compute Probability by given condition

(3) Transfer Documents into Vectors

(4)

a. Calculate the Similarity between the Query Vector and the Document Vectors

b. Calculate the Probability based on given Query & Document

(5) Rank the Documents according to the Similarity scores

(Option 2) or using API

Programming with the Indri API

- [Using the API to Write Your Own Application]
- [Example Applications in C++]
- [Creating your own Parser]

(wiki page)

1. **Copy Makefile.app from the top level lemur directory to the directory with your application's source code. Edit the file and fill in values for the following:**
   OBJS -- list of each of the object files needed to build your application.
   PROG -- name for your application.

2. **Use make -f Makefile.app to build your application.**

Building the Index

```
<parameters>
    <index>/path/to/outputIndex</index>

    <corpus>
      <path>/path/to/collection1/</path>
      <class>trecweb</class>
    </corpus>
</parameters>
```

BuildIndex Parameters

<stemmer>
<stopper>
…

Basic usage:

IndriBuildIndex <parameter_file>

Building the Index

```
<parameters>
    <index>/path/to/outputIndex</index>

    <corpus>
      <path>/path/to/collection1/</path>
      <class>trecweb</class>
    </corpus>
</parameters>
```

```
4:51: Closed /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B06
4:51: Opened /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B19
4:51: Documents parsed: 246609 Documents indexed: 246609
4:51: Closed /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B19
4:51: Opened /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B32
4:52: Documents parsed: 246929 Documents indexed: 246929
4:52: Closed /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B32
4:52: Opened /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B12
4:52: Documents parsed: 247004 Documents indexed: 247004
4:52: Closed /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B12
4:52: Opened /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B38
4:52: Documents parsed: 247156 Documents indexed: 247156
4:52: Closed /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B38
4:52: Opened /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B37
4:52: Documents parsed: 247219 Documents indexed: 247219
4:52: Closed /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B37
4:52: Opened /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B34
4:53: Documents parsed: 247491 Documents indexed: 247491
4:53: Closed /tmp2/cmchen/Indri-code/homework/WT2G/Wt08/B34
4:53: Closing index
5:13: Finished
```

BuildIndex Parameters

<stemmer>

<stopper>

…

DumpIndex

Command	Argument(s)	Description
term (t)	Term text	Print inverted list for a term
termpositions (tp)	Term text	Print inverted list for a term, with positions
fieldpositions (fp)	Field name	Print inverted list for a field, with positions
expressionlist (e)	Expression	Print inverted list for an Indri expression, with positions
xcount (x)	Expression	Print count of occurrences of an Indri expression
documentid (di)		Field, Value
documentname (dn)	Document ID	Print the text representation of a document ID
documenttext (dt)	Document ID	Print the text of a document
documentdata (dd)	Document ID	Print the full representation of a document
documentvector (dv)	Document ID	Print the document vector of a document
invlist (il)	(None)	Print the contents of all inverted lists
vocabulary (v)	(None)	Print the vocabulary of the index
stats (s)	(None)	Print statistics for the Repository

DumpIndex —Demo

```
cmchen@clip2 [01:58:00] [/tmp2/cmchen/lemur-4.12/LJ_Project/Data/LJ_TMM/index_article]
-> % dumpindex .
dumpindex <repository> <command> [ <argument> ]*
These commands retrieve data from the repository:
    Command              Argument       Description
    term (t)             Term text      Print inverted list for a term
    termpositions (tp)   Term text      Print inverted list for a term, with positions
    fieldpositions (fp)  Field name     Print inverted list for a field, with positions
    expressionlist (e)   Expression     Print inverted list for an Indri expression, with positions
    xcount (x)           Expression     Print count of occurrences of an Indri expression
    dxcount (dx)         Expression     Print document count of occurrences of an Indri expression
    documentid (di)      Field, Value   Print the document IDs of documents having a metadata field matching this value
    documentname (dn)    Document ID    Print the text representation of a document ID
    documenttext (dt)    Document ID    Print the text of a document
    documentdata (dd)    Document ID    Print the full representation of a document
    documentvector (dv)  Document ID    Print the document vector of a document
    invlist (il)         None           Print the contents of all inverted lists
    vocabulary (v)       None           Print the vocabulary of the index
    stats (s)                           Print statistics for the Repository
These commands change the data inside the repository:
    compact (c)          None           Compact the repository, releasing space used by deleted documents.
    delete (del)         Document ID    Delete the specified document from the repository.
    merge (m)            Input indexes  Merges a list of Indri repositories together into one repository.
```

DumpIndex —Example

- **Status**

```
Repository statistics:
documents:       247491
unique terms:    1525847
total terms:     261143893
```

- **Inverted List**

<word> <total appear times> <# of document>

```
bluebar 1 1
        147298 1 110
bluebell 5 4
        12390 1 605
        44105 1 72
        60223 1 67
        221286 2 520 601
blueberry 289 254
        1163 1 751
        1544 1 289
        1939 2 265 273
        2110 1 847
        2649 1 208
        3704 1 1098
```

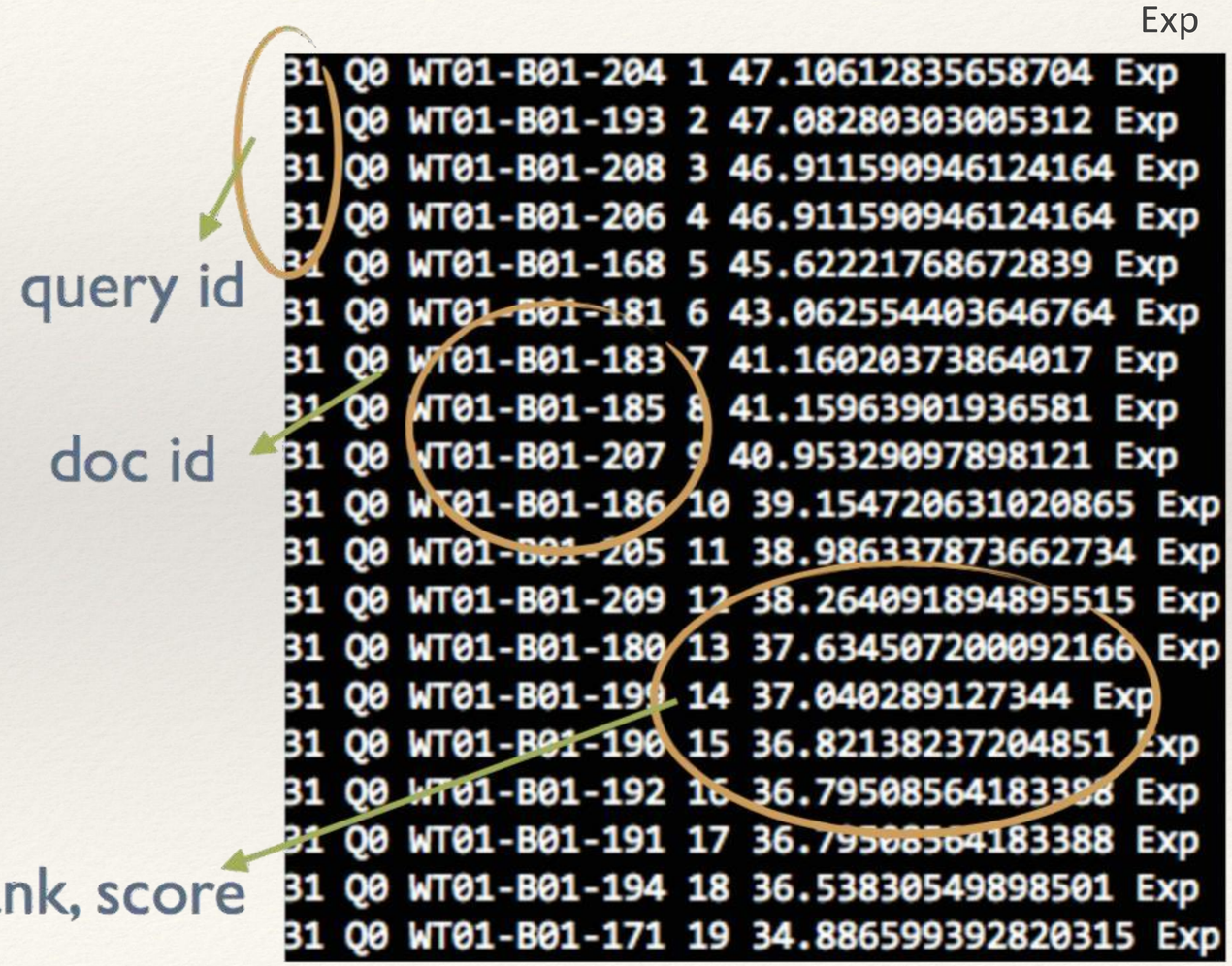
<docID> <appear times> <position>

Project 2—Document Retrieval

- **Query**

```
<top>

<num> Number: 401
<title> foreign minorities, Germany

<desc> Description:
What language and cultural differences impede the integration
of foreign minorities in Germany?

<narr> Narrative:
A relevant document will focus on the causes of the lack of
integration in a significant way; that is, the mere mention of
immigration difficulties is not relevant.  Documents that discuss
immigration problems unrelated to Germany are also not relevant.

</top>
```

- **Returned Documents**



Project 2 —Evaluation

- **Evaluation** **./trec_eval [reference answer] [your prediction]**

```
..[chih-mingchen@MacBook-Pro] - [~/wsm/lemur-4.12/mine] - [Fri Apr 25, 09:57]
..[$] <()> ./trec_eval qrels.401-430.txt my_returned_list

Queryid (Num):       20
Total number of documents over all queries
    Retrieved:    20000
    Relevant:       959
    Rel_ret:        667
Interpolated Recall - Precision Averages:
    at 0.00       0.6114
    at 0.10       0.4519
    at 0.20       0.3737
    at 0.30       0.3395
    at 0.40       0.2716
    at 0.50       0.2204
    at 0.60       0.1070
    at 0.70       0.0675
    at 0.80       0.0476
    at 0.90       0.0142
    at 1.00       0.0000
Average precision (non-interpolated) for all rel docs(averaged over queries)
                  0.2032
```

```
Precision:
  At    5 docs:   0.4000
  At   10 docs:   0.3650
  At   15 docs:   0.3233
  At   20 docs:   0.3075
  At   30 docs:   0.2617
  At  100 docs:   0.1600
  At  200 docs:   0.1103
  At  500 docs:   0.0580
  At 1000 docs:   0.0334
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.2521
```

Elasticsearch

https://docs.google.com/presentation/d/187_HYbCv1-iZj9Ez3g3VX0FKaaOjq2AgI8gI4Ejd98k/edit?usp=sharing

Any Question?