Approximate Search for Keywords in Handwritten Text Images

José Andrés, Alejandro H. Toselli, Enrique Vidal

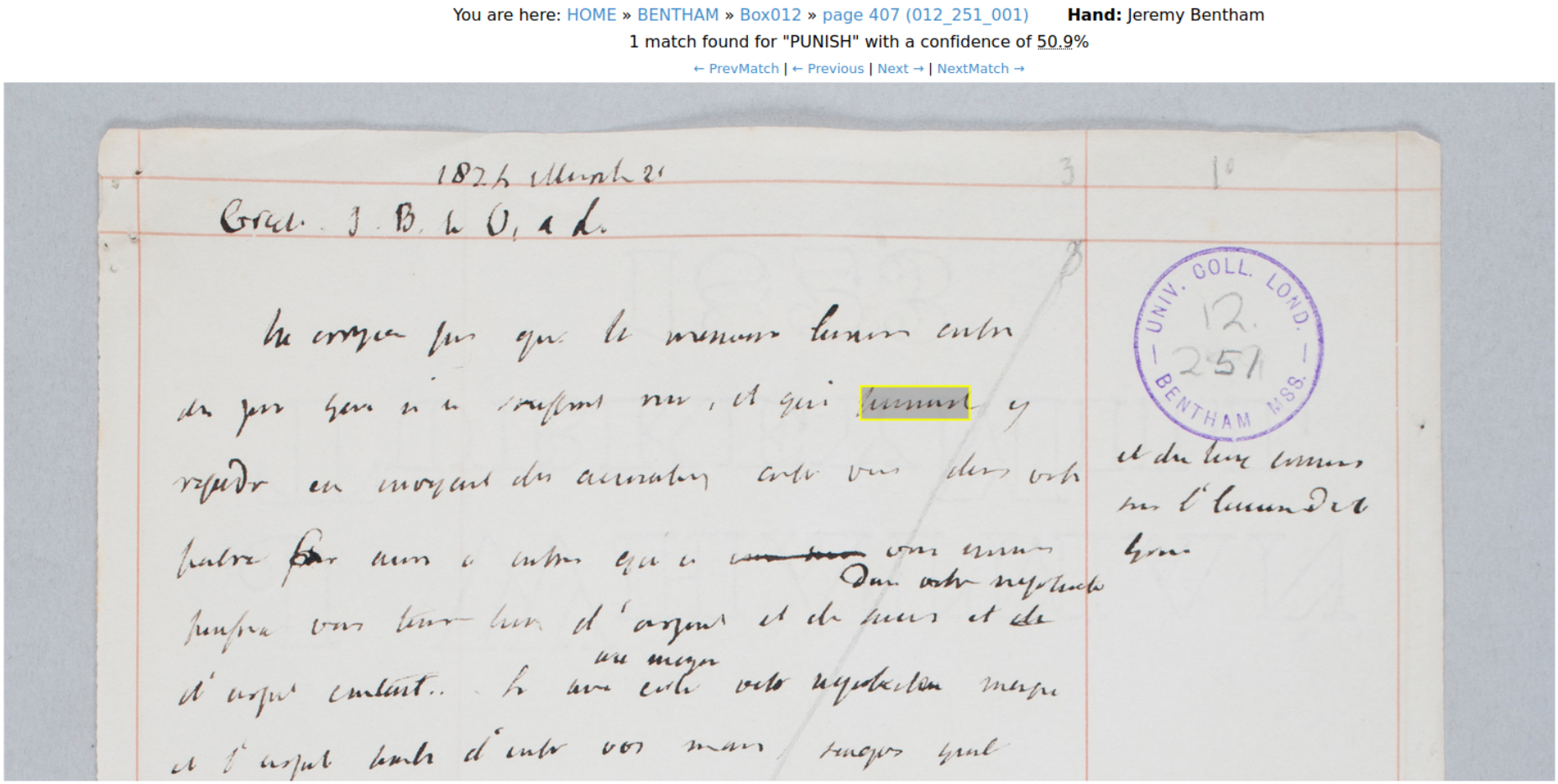
Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València, Spain

May 24th, 2022

Introduction

Introduction

Keyword search



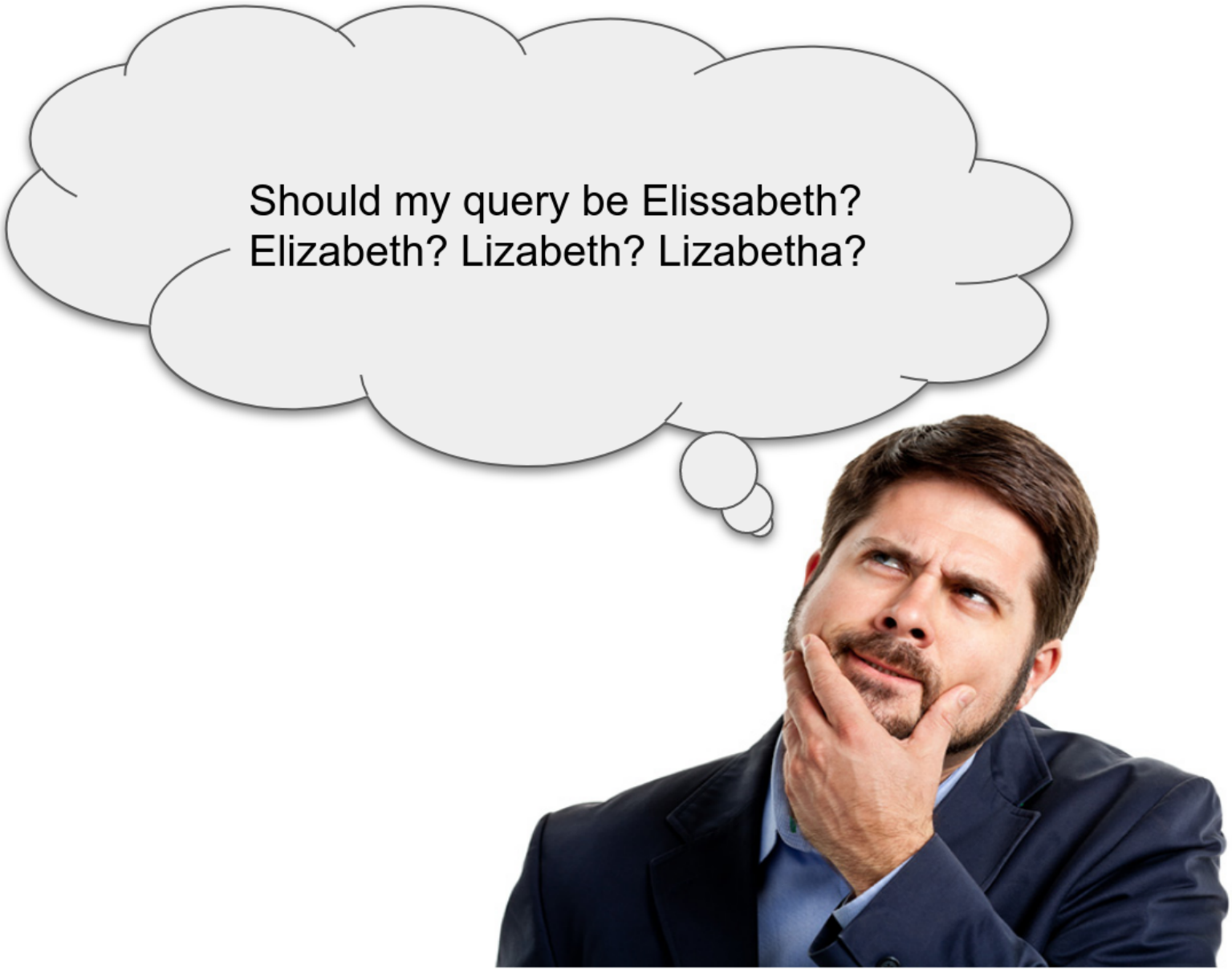
It allows to search free text over the images!

However...

Sometimes, users don't know what to search:

- Languages evolve through time
- Inaccurate system hypotheses

This fact sometimes makes difficult querying



Approximate search

Therefore, we are interested in queries that allow **approximate word spelling**.

A single-word approximate-spelling query is given by a base word, along with an indication that some flexibility is allowed.

Notation: $base\_word{\sim}flexibility$

Levenshtein distance

Types of character errors:

1. Insertion: Lisabeth $\rightarrow$ **E**lisabeth (Insert character "E" at the beginning)
2. Deletion: Eli**s**sabeth $\rightarrow$ Elisabeth (Deletion of the character "s")
3. Substitution: Elissabet**h** $\rightarrow$ Elissabet**a** (substitution of "h" for "a")

Implementation details

The number of strings that could match the query "aptitude∼3" employing an alphabet of 26 characters is larger than 750000.

Instead, we can only consider the pseudo-words which appear in the PrIx (as the relevant probability of the other is clearly 0!)

Implementation details

How does a query on our system work? Two phases:

1. **Offline**: Store all the pseudo-words that are found in the PrIx into a DAWG.
2. **Online**: Calculate the Levenshtein distance between the query and the pseudo-words stored in the DAWG.

Implementation details: online phase

How do we calculate efficiently the distance between the base word query and all the pseudo-words stored in the DAWG?

Calculate the Levenshtein distance from each prefix of a pseudo-word to all the possible prefixes of the query base word.

This fact allows avoiding repeating calculations and pruning the search if possible.

Evaluation protocols

Two different evaluation criterias:

1. **Objective**: TP if a query matches a pseudo-word of a PrIx spot and one of the words of the GT transcript of the text line geometrically associated to that spot. Otherwise FP.
2. **Subjective**: Same constraints as in the objective criteria and the word found in the GT transcripts must be semantically related to the query (base) word. Otherwise FP.

Dataset statistics

Table: Statistics of the test set defined for the Bentham experimental dataset and for the query sets Q1, Q2 and Q3, used in the evaluation experiments.

Test set		Q1	Q2	Q3
Pages	357	357	253	352
Lines	12 363	12 080	896	3 070
Running words	89 870	87 070	965	3 629
Unique words (Lexicon)	6 988	6 953	861	650

Q1: all the words that appear in the GT and are larger than 1 char.
Q2: all the words that appear in Q1 and whose relevance probability is 0.
Q3: set of words chosen manually by a user to assess the semantic similitude.

Retrieval performance

Table: Exact and approximate-spelling retrieval performance (mAP) for "objective" and "subjective" evaluation protocols. 95% confidence intervals are never larger than 0.03.

Query set	Q1 (6 953 words)	Q2 (861 words)	Q3 (650 words)	
Evaluation criteria	Objective	Objective	Objective	Subjective
Exact (baseline)	0.76	0.00	0.78	0.78
Approx. $d_0 = 1$	0.81	0.40	0.83	0.81
Approx. $d_0 = 2$	0.85	0.67	0.84	0.65

Computational performance

Table: Dataset size, memory usage (MB) and single-query response time (milliseconds) of approximate-spelling search.

Dataset	Bentham GT	Bentham Full	Scale factor
Number of images	357	89 911	251.9
Running words	89 870	25 487 932	283.6
Number of unique pseudo-words ($\|S\|$)	5 951 009	37 172 635	6.2
Memory usage (MB)	192	1 602	8.3
Query time (ms) for $d_0 = 1$	0.7	1.3	1.9
Query time (ms) for $d_0 = 2$	10.5	24.7	2.4

Illustrative queries examples

Let's see some illustrative queries:

Government∼1
Judicial∼2

Search interface available at: http://prhlt-kws.prhlt.upv.es/bentham/

Illustrative queries examples

Let's see some illustrative queries:

Government∼1
Judicial∼2
Sever∼2

Search interface available at: http://prhlt-kws.prhlt.upv.es/bentham/

Conclusions

- Approximate search **might** help the users to retrieve relevant information that wouldn't be easily found.
- It has been developed with reasonable memory consumption and time performance.

Conclusions

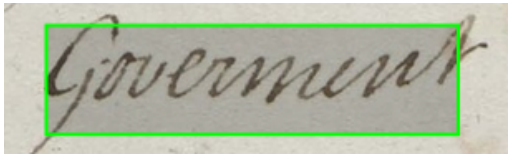
Approximate Search for Keywords in Handwritten Text Images

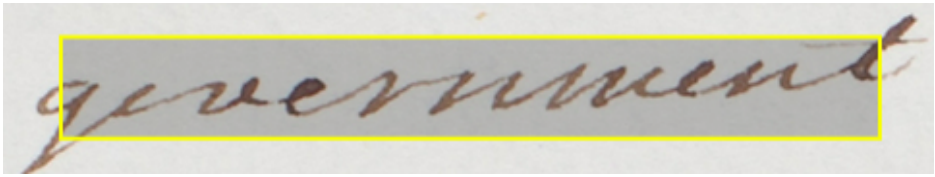
José Andrés, Alejandro H. Toselli, Enrique Vidal

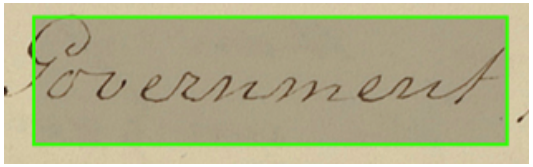
Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València, Spain

May 24th, 2022

Illustrative retrieval examples

a) 

b) 

c) 

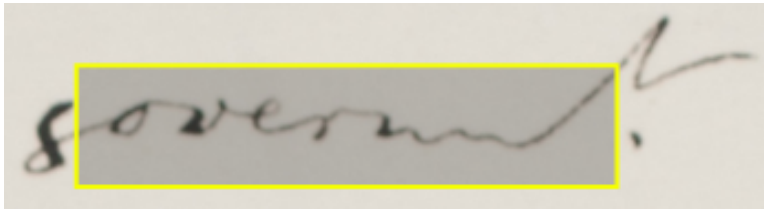
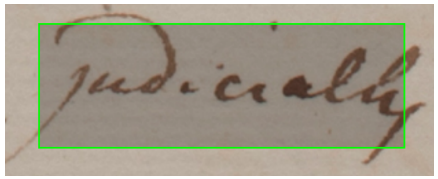
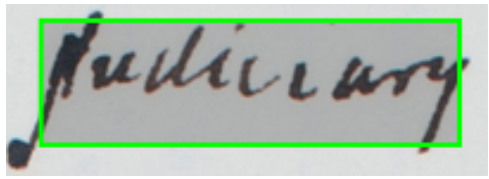
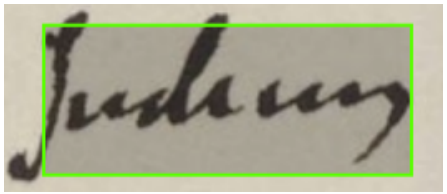
d) 

Figure: *Objective* and *subjective* TPs retrieved by "government∼1". The GT for all these spots is "government". In the case a) the PrIx provides a high RP hypothesis for "goverment", in b) for "gevernment", in c) for "sovernment", and in d) for "governent".

Illustrative retrieval examples

a) 

b) 

c) 

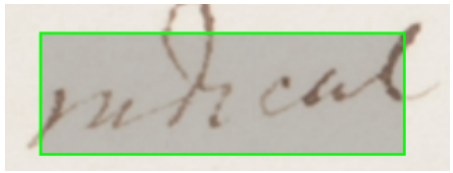
d) 

Figure: *Objective* and *subjective* TPs retrieved by "judicial∼2". In the case a) the PrIx provides a high RP hypothesis for "judicially", in b) for "judiciary", in c) for "judicing", and in d) for "judical". *subjective* TPs.

Illustrative retrieval examples

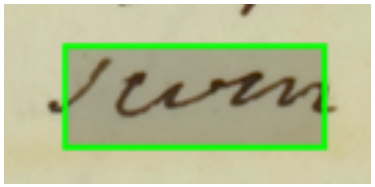
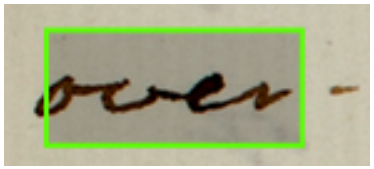
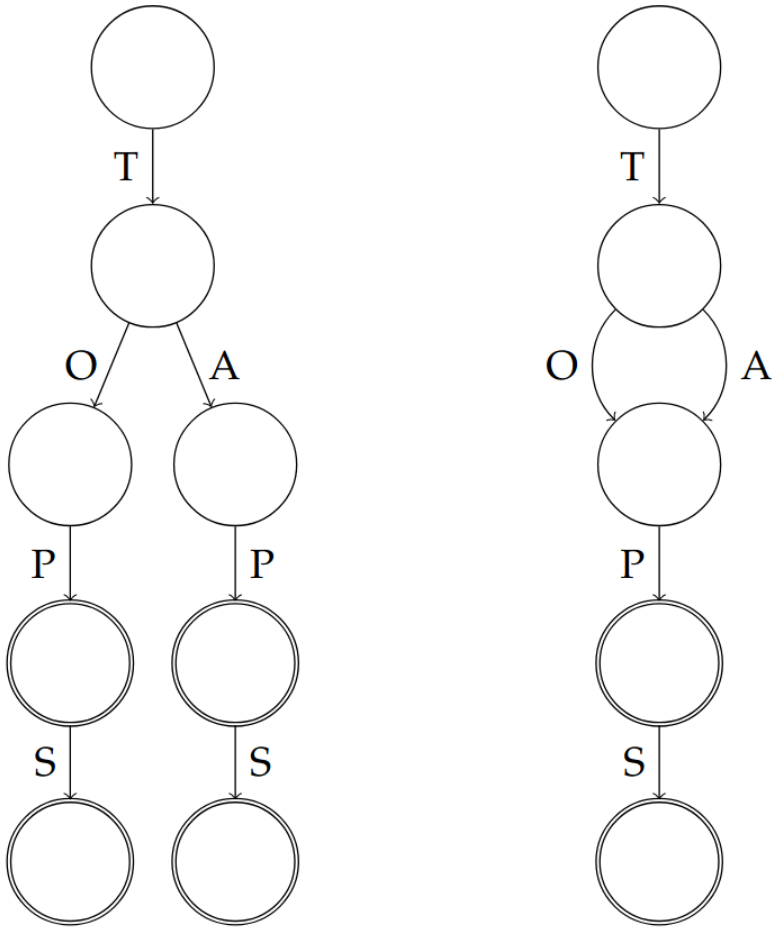
a)  b) 

Figure: *Objective* TPs but *subjective* FPs retrieved by the query "sever∼2". The hypotheses with highest RP are "seven" for a) and "over" for b). Both are *objective* TPs because they are within $d_0 = 2$ edit distance from "sever" and match the corresponding GT words. However, both spots are *subjective* FPs, because they are not only different form "sever", but also semantically unrelated with it.

DAWG

Data structure that eliminates prefix, interfix and suffix redundancy.



Online search

		T	A	P	S
	0	1	2	3	4
T	1	0	1	2	3
O	2	1	1	2	3
P	3	2	2	1	2

		T	A	P	S
	0	1	2	3	4
T	1	0	1	2	3
O	2	1	1	2	3
P	3	2	2	1	2
S	4	3	3	2	1

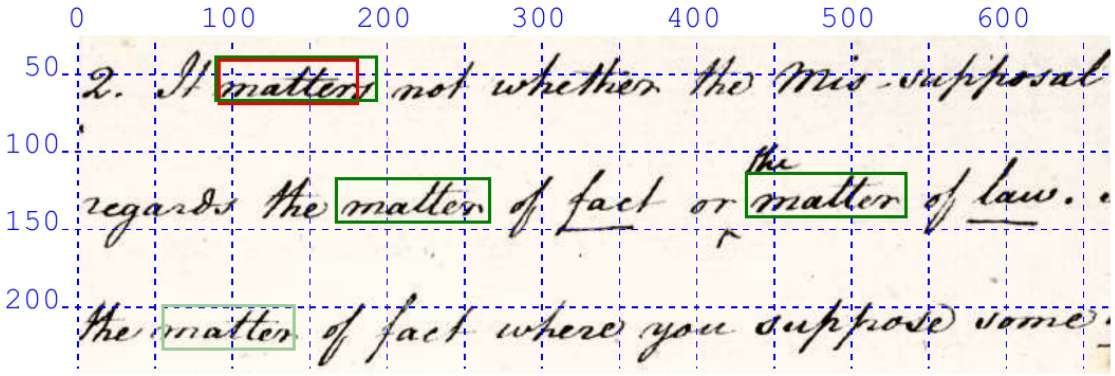
Only the last row changes!!

Computational performance

Table: Dataset size, memory usage (MB) and single-query response time (milliseconds) of approximate-spelling search.

Dataset	Bentham GT	Bentham Full	Scale factor
Number of images	357	89 911	251.9
Running words	89 870	25 487 932	283.6
Number of unique pseudo-words ($\|S\|$)	5 951 009	37 172 635	6.2
Number of characters in S	55 654 799	355 514 400	6.4
DAWG size (edges)	3 158 261	24 818 936	7.8
Memory usage (MB)	192	1 602	8.3
Query time (ms) for $d_0 = 1$	0.7	1.3	1.9
Query time (ms) for $d_0 = 2$	10.5	24.7	2.4

Prlx example




```
# pageID="Bentham-071-021-002-part"
#    keyword confid    bounding box
#
           2  0.929    1  36  20 31
          21  0.064    1  36  24 31
          IT  0.982   33  36  27 31
          IF  0.012   33  36  26 31
     MATTERS  0.989   77  36  99 31
      MATTER  0.011   77  36  93 31
         NOT  0.999  216  36   7 31
     WHETHER  1.000  256  36  99 31
         THE  0.997  389  36  33 31
MIS-SUPPOSAL  1.000  455  36 193 31

         THE  0.927  430  88  30 31
          HE  0.056  434  88  25 31
         ...  ...      ...     ...
```

```
REGARDS  0.857    5 115  84 31
REWARDS  0.138    5 115  90 31
    THE  0.993  110 115  43 31
 MATTER  0.998  160 115  93 31
     OF  0.996  271 115  23 31
   FACT  0.999  306 115  49 31
     OR  0.973  377 115  37 31
     ON  0.021  377 115  42 31
 MATTER  0.990  425 116 100 31
     OF  0.995  542 115  25 31
     BY  0.407  575 115  30 31
    ANY  0.175  575 115  55 31
    ...  ...      ...     ...
    LAW  0.032  575 115  36 31
    LAY  0.031  575 115  55 31
    ...  ...      ...     ...
    PAY  0.012  575 115  59 31
```

```
      THE  0.990    1 198  28 31
   MATTER  0.934   61 198  64 31
       OF  0.988  141 198  28 31
     FAST  0.367  182 198  62 31
      FAR  0.186  182 198  36 31
      ...  ...      ...     ...
     FACT  0.017  182 198  46 31
       AS  0.142  200 198  29 31
      HAS  0.022  200 198  29 31
    WHERE  0.992  255 198  90 31
      YOU  0.761  365 198  45 31
     YOUR  0.030  365 198  47 31
     GOES  0.064  372 198  45 31
  SUPPOSE  0.975  429 198 120 31
SUPPOSED  0.024  429 198 125 31
     SOME  0.834  570 198  78 31
  SOONER  0.016  576 198  83 31
      ONE  0.109  580 198  65 31
       ME  0.022  620 198  22 31
```