



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/4571/>

Monograph:

Sanderson, M. (1997) Duplicate Detection in the Reuters Collection. Technical Report.
Department of Computing Science , University of Glasgow.

Technical Report (TR-1997-5) of the Department of Computing Science at the University of
Glasgow

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



Duplicate detection in the Reuters collection¹

Mark Sanderson

1 Introduction

While conducting some experiments with the Reuters collection, it was discovered that contained within it were a number of documents that were exact duplicates of each other (see Figure 1). A short study was conducted to try to discover how many such documents there were. The results of this study revealed that the notion of a duplicate document was not as simple as first thought.

The contents of this report are as follows. A brief review of previous duplicate detection research will be presented, followed by a description of the methods and results of the duplicate detection work conducted here. In addition, there is an appendix holding the document ids of the various types of duplicate found.

2 Other duplicate research

Duplicate detection does not appear to be an area of interest to IR expect perhaps in the relatively new field of data fusion. However, bibliographic databases and electronic publishing are both areas where research can be found on duplicate documents.

2.1 Bibliographic databases

In a bibliographic database, the main task is not to find exact duplicate records, rather it is to find those that refer to the same work but differ in some manner. Differences are typically due to inaccurate or inconsistent data entry. One such detection method was developed by Ridley [Ridley 92] who adopted a two stage technique. First, all records in a database were assigned a number generated from a *hashing function* that used as its input, fields of a bibliographic record. Any records that had the same hashing number were examined in greater detail in the second stage. This entailed a comparison of fields by customised processes: i.e. the author field process looked for missing initials; the title field process looked for a missing suffix. Detection techniques of this kind are supported by the work of O'Neill et al. [O'Neill 93] who manually examined duplicate bibliographic records to find which fields were most likely to differ.

1. These experiments were performed on the Reuters 22,173 collection, created by David Lewis. This has recently been replaced with a new version, the Reuters 21,578 collection containing 595 fewer documents. The results of the work reported here were re-examined for this new version and confirmed as still being valid for it. Therefore, all references made to the Reuters collection can be taken to refer to the 21,578 version. This collection can currently be found at

```

<REUTERS TOPICS="YES"
LEWISSPLIT="NOT-USED"
CGISPLIT="PUBLISHED-TESTSET"
OLDID="21689" NEWID="17066">
<DATE>24-APR-1987 07:23:50.50</DATE>
<TOPICS><D>money-
fx</D><D>dlr</D></TOPICS>
<PLACES><D>japan</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;V
&#22;&#22;&#1;f0474&#31;reute
u f BC-BANK-OF-JAPAN-INTERVE 04-24
0085</UNKNOWN>
<TEXT>&#2;
<TITLE>BANK OF JAPAN INTERVENES
IN TOKYO MARKET</TITLE>
<DATELINE> TOKYO, April 24 -
</DATELINE> <BODY>The Bank of Japan
intervened just after the Tokyo market opened
to support the dollar from falling below
140.00 yen, dealers said.
The central bank bought a moderate amount
of dollars to prevent its decline amid bearish
sentiment for the U.S. Currency, they said.
The dollar opened at a record Tokyo low of
140.00 yen against 140.70/80 in New York
and 141.15 at the close here yesterday. The
previous Tokyo low was 140.55 yen set on
April 15.
REUTER
&#3;</BODY></TEXT>
</REUTERS>

```

```

<REUTERS TOPICS="YES"
LEWISSPLIT="TEST"
CGISPLIT="TRAINING-SET"
OLDID="1682" NEWID="17041">
<DATE>23-APR-1987 20:21:46.09</DATE>
<TOPICS><D>money-
fx</D><D>dlr</D></TOPICS>
<PLACES></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;RM
&#22;&#22;&#1;f3091&#31;reute
b f BC-BANK-OF-JAPAN-INTERVE 04-23
0086</UNKNOWN>
<TEXT>&#2;
<TITLE>BANK OF JAPAN INTERVENES
IN TOKYO MARKET</TITLE>
<DATELINE> TOKYO, April 24 -
</DATELINE> <BODY>The Bank of Japan
intervened just after the Tokyo market opened
to support the dollar from falling below
140.00 yen, dealers said.
The central bank bought a moderate amount
of dollars to prevent its decline amid bearish
sentiment for the U.S. Currency, they said.
The dollar opened at a record Tokyo low of
140.00 yen against 140.70/80 in New York
and 141.15 at the close here yesterday. The
previous Tokyo low was 140.55 yen set on
April 15.
REUTER
&#3;</BODY></TEXT>
</REUTERS>

```

Figure 1. Reuters documents referring to the same event whose body texts are identical.

2.2 Electronic publishing

As electronic publishing becomes more common, the potential problems of copyright violation and of plagiarism will increase. Most efforts devised to combat these problems concentrate on attempts to prevent or at least make it difficult for people to copy electronic documents. However, the detection of duplicates or partial duplicates is another approach. Brin et al. [Brin 95] proposed a system where electronic publishers store in a centralised database, *signatures* of all their published works. A signature would in some way summarise a document. The owners of this database could continually scan other electronic document collections looking for duplicates that might violate their copyright.

The method that Brin et al. proposed for building these signatures involved the breaking up of documents into what they call chunks. They suggest that these could be sentences, paragraphs, or some form of interleaved text unit. Each chunk of a document is passed to a hashing function that produces a number (quite how this function works

is unclear from the paper). All numbers of that document are concatenated to form a signature. Detection of duplication is simply a process of comparing the hash numbers of two document signatures and looking for an unexpectedly high number of matches.

A method similar to this was adopted for the Reuters based work presented here. As only duplicate documents were of interest, the size of chunk was chosen to be a whole document, and the hashing function was a term selection method based on *idf* weights. This detection method is now described.

3 The duplicate detection for Reuters documents

During the building of an IR system [Sanderson 91], the following was noted. Performing relevance feedback based on a single document, resulted in a query composed of terms from that document alone. A retrieval based on that query almost always resulted in a document ranking whose relevance scores were distributed in the manner shown in Figure 2. The highest relevance score was assigned to the document that relevance feedback was based on. All other retrieved documents were assigned a significantly lower score. It was hypothesised that a query generated from relevance

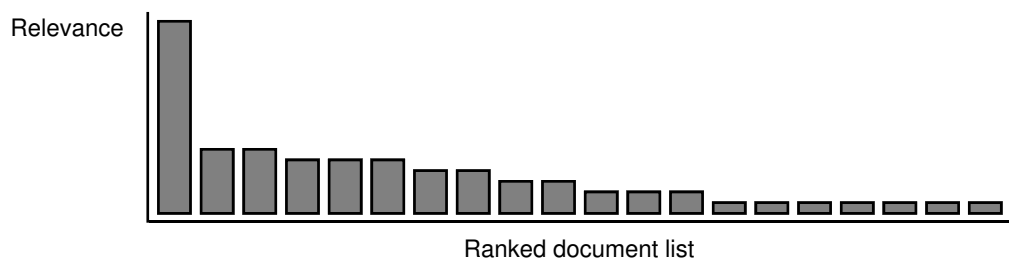


Figure 2. Relevance scores assigned to a document ranking.

feedback based on a single document would uniquely identify that document. The only exception to this would be if there was an exact duplicate of it.

It was a detection method based on this hypothesis that was tested in these experiments. It works as follows. For each individual document in a collection, generate a query using relevance feedback based on just that document², perform a retrieval and analyse any other documents with a high relevance score to discover if they are duplicates. If such a duplicate is found by this method, it is described here as one document *retrieving* another. Although this was found to work well, after some informal testing, further modifications to the method were made and they are now described³.

3.1 First modification

The first modification arose when documents such as the pair in Figure 3 were found. As can be seen, one is a longer version of the other. Unfortunately, for document pairs of this type, the shorter would retrieve the longer as a potential duplicate even though

2. It was found that queries composed of 20 terms were large enough to accurately find the duplicates.

3. Since conducting this work, Kirriemuir [Kirriemuir 95] has investigated this area and has devised a broadly similar method, although it is slightly less exhaustive in its pursuit of duplicates.

it is not. This happens because all the words in the shorter version of the document (from which relevance feedback generates a query) appear in the longer version.

<pre> <REUTERS TOPICS="NO" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="10068" NEWID="5155"> <DATE>14-MAR-1987 23:23:04.16</DATE> <TOPICS></TOPICS> <PLACES><D>yugoslavia</D></PLACES > <PEOPLE></PEOPLE> <ORGS></ORGS> <EXCHANGES></EXCHANGES> <COMPANIES></COMPANIES> <UNKNOWN> &#5;&#5;&#5;RM &#22;&#22;&#1;f0844&#31;reute r f BC-UNION-LEADERS-TOUR-YU 03- 14 0104</UNKNOWN> <TEXT>&#2; <TITLE>UNION LEADERS TOUR YUGOSLAVIA TO QUELL STRIKE</TITLE> <DATELINE> BELGRADE, March 15 - </DATELINE> <BODY>Yugoslav trade union leaders are touring the country in an attempt to quell a wave of strikes following a partial wages freeze, official sources said. Eyewitnesses in the northern city of Zagreb reported far more police on the streets than normal after the city and areas nearby experienced the biggest wave of strikes in the country in recent memory. National newspapers in Belgrade have given few details of the strikes. But Zagreb papers said thousands of workers went on strike and thousands more were threatening action over pay cuts. Official sources said there were also strikes at a Belgrade medical centre, a food factory in Sambor, and enterprises in Nis, Leskovac and Kraljevo, as well as other towns. They said national union officials were travelling throughout the country to speak to meetings in an attempt to restore calm. But trade union leaders were avoiding making statements to the press and had not made their stand on the strikes clear. Western diplomats said the strikes appeared to be spontaneous and without any unified orchestration. REUTER &#3;</BODY></TEXT> </REUTERS> </pre>	<pre> <REUTERS TOPICS="NO" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="10256" NEWID="5343"> <DATE>16-MAR-1987 09:46:11.84</DATE> <TOPICS></TOPICS> <PLACES><D>yugoslavia</D></PLACES > <PEOPLE></PEOPLE> <ORGS></ORGS> <EXCHANGES></EXCHANGES> <COMPANIES></COMPANIES> <UNKNOWN> &#5;&#5;&#5;C G T M &#22;&#22;&#1;f2044&#31;reute d f BC-UNION-LEADERS-TOUR-YU 03- 16 0120</UNKNOWN> <TEXT>&#2; <TITLE>UNION LEADERS TOUR YUGOSLAVIA TO QUELL STRIKE</TITLE> <DATELINE> BELGRADE, March 16 - </DATELINE> <BODY>Yugoslav trade union leaders are touring the country in an attempt to quell a wave of strikes following a partial wages freeze, official sources said. Eyewitnesses in the northern city of Zagreb reported far more police on the streets than normal after the city and areas nearby experienced the biggest wave of strikes in the country in recent memory. National newspapers in Belgrade have given few details of the strikes. But Zagreb papers said thousands of workers went on strike and thousands more were threatening action over pay cuts. Western diplomats said the strikes appeared to be spontaneous and without unified orchestration. Reuter &#3;</BODY></TEXT> </REUTERS> </pre>
--	--

Figure 3. Documents referring to the same event where one is a longer version of the other.

Therefore, it was decided that two documents were exact duplicates only if the first document retrieved the second and the second retrieved the first. This would hopefully avoid the type of document pair shown here. After conducting the experiments, it was realised that this modification would probably not have been necessary if the term weighting scheme, used in retrieval, had been based on within document frequencies and document length normalisation.

3.2 Second modification

The second modification occurred when the type of document pair in Figure 4 was found. As can be seen, these documents are almost identical but they refer to different events. It would appear that for a number of regular events, like the financial transactions reported in Figure 4, the Reuters staff have a standard set of templates that they use for such events. To avoid this type of document pair it was decided that potential duplicates had to be relayed within 48 hours of each other.

<pre> <REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12705" NEWID="522"> <DATE> 2-MAR-1987 11:44:41.93</DATE> <TOPICS><D>money- fx</D><D>interest</D></TOPICS> <PLACES><D>usa</D></PLACES> <PEOPLE></PEOPLE> <ORGS></ORGS> <EXCHANGES></EXCHANGES> <COMPANIES></COMPANIES> <UNKNOWN> &#5;&#5;&#5;V RM &#22;&#22;&#1;f0060&#31;reute b f BC-/FED-ADDS-RESERVES-V 03-02 0060</UNKNOWN> <TEXT>&#2; <TITLE>FED ADDS RESERVES VIA CUSTOMER REPURCHASES</TITLE> <DATELINE> NEW YORK, March 2 - </DATELINE> <BODY>The Federal Reserve entered the U.S. Government securities market to arrange 1.5 billion dlrs of customer repurchase agreements, a Fed spokesman said. Dealers said Federal funds were trading at 6-3/16 pct when the Fed began its temporary and indirect supply of reserves to the banking system. Reuter &#3;</BODY></TEXT> </REUTERS> </pre>	<pre> <REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="19586" NEWID="3164"> <DATE> 9-MAR-1987 11:49:35.16</DATE> <TOPICS><D>interest</D><D>money- fx</D></TOPICS> <PLACES><D>usa</D></PLACES> <PEOPLE></PEOPLE> <ORGS></ORGS> <EXCHANGES></EXCHANGES> <COMPANIES></COMPANIES> <UNKNOWN> &#5;&#5;&#5;V RM &#22;&#22;&#1;f0663&#31;reute b f BC-/FED-ADDS-RESERVES-V 03-09 0060</UNKNOWN> <TEXT>&#2; <TITLE>FED ADDS RESERVES VIA CUSTOMER REPURCHASES</TITLE> <DATELINE> NEW YORK, March 9 - </DATELINE> <BODY>The Federal Reserve entered the U.S. Government securities market to arrange 2.5 billion dlrs of customer repurchase agreements, a Fed spokesman said. Dealers said Federal funds were trading at 6-3/16 pct when the Fed began its temporary and indirect supply of reserves to the banking system. Reuter &#3;</BODY></TEXT> </REUTERS> </pre>
--	---

Figure 4. Documents whose body text is very similar but each refers to a different event.

4 Testing the method

To test the effectiveness of the duplicate detection method, potential duplicates of every document in the Reuters collection were retrieved and placed into one of three sets: documents pairs that appeared to be duplicates but reported different events; documents pairs where one was a longer version of the other; and documents pairs that were exact duplicates. The accuracy with which documents were placed in each set was then measured.

4.1 First set: documents that report different events

In examining each document pair in this set, the following test question was asked,

Do these documents refer to the same event?

In Table 1 we can see that 88% of pairs passed this test which indicates that the modification was effective in separating the *template documents* from the exact duplicates. The four pairs that were incorrectly assigned were exact duplicates relayed more than 48 hours apart.

Passed	29	88%
Failed	4	12%
Total	33	

Table 1. Results of the first document duplicate test.

4.2 Second set: documents where one is a longer version of the other

There were 283 pairs in this set. The test question applied while inspecting each pair was,

Do these two documents refer to the same event and is one of them a longer version of the other?

As can be seen in Table 2, only 49% of the pairs passed this test. Of those that failed, around a third were template documents like those found in Section 4.1. If a chronological test had been applied (were documents relayed within 48 hours of each other?), these would have been eliminated. The majority of the other incorrectly identified pairs were documents referring to the same event where one was a corrected version of the other. There were also a number of document pairs referring to distinct events that were relayed within a short time of each other, for example, hourly stock exchange reports. Quite how one would eliminate this type of pair without resorting to a collection specific solution is not clear.

Passed	139	49%
Failed	144	51%
Total	283	

Table 2. Results of the second document duplicate test.

4.3 Third set: documents that are exact duplicates of each other

These were document pairs that passed both modifications: each document retrieves the other, and they were relayed within 48 hours of each other. The number of pairs identified was 322. The test question applied while inspecting each pair was,

Do these documents refer to the same event and are the body texts within them identical?

As can be seen in Table 3, all but two of the document pairs passed this test. The two that failed referred to the same event but had very slight changes to their body text. These were judged to be corrections of the versions earlier document and were therefore not exact duplicates.

Passed	320	99%
Failed	2	1%
Total	322	

Table 3. Results of the final document duplicate test.

5 Conclusions

The main objective of this work was to identify exact duplicate documents in the Reuters collection. The method used to find them was highly effective, correctly identifying 320 pairs and only failing to find four. During the creation of this detection method, a number of other duplicate document types were found:

- expanded documents, where both refer to the same event, but one is a longer version of the other;
- corrected documents, where both refer to the same event, but one is a corrected version of the other;
- and template documents, where nearly identical documents refer to different events.

Tests were devised to identify these types but they were found to have variable success.

6 References

Brin 95

S. Brin, J. Davis, H. Garcia-Molina (1995). Copy detection mechanism for digital documents, in Proceedings of SIGMOD.

Kirriemuir 95

J.W. Kirriemuir & P. Willett (1995). Identification of duplicate and near-duplicate full-text records in database search outputs using hierarchic cluster analysis, in Program - automated library and information systems, 29(3): 241-256.

O'Neill 93

E.T. O'Neill, S.A. Rogers & W.M. Oskins (1993). Characteristics of duplicate records in OCLC's on-line union catalogue, in Library Resources & Technical Services, 37(1): 59-71.

Ridley 92

M.J. Ridley (1992). An expert system for quality control and duplicate detection in bibliographic databases, in Program - automated library and information systems, 26(1): 1-18.

Sanderson 91

M. Sanderson & C.J. van Rijsbergen (1991). NRT: news retrieval tool, in Electronic Publishing, EP-odd, 4(4): 205-217.

A List of duplicates

This appendix presents six tables containing pairs of document ids that passed or failed the three tests described in Section 4. The tables are presented in the order in which the tests are described in that section. The document ids are those used in the Reuters 21,578 collection.

519	11422	5344	9857	12456	1971
522	3164	6044	10859	12471	1971
522	7769	6044	9972	12495	18011
1120	11422	7025	1969	13398	8144
1125	10864	7204	8343	13799	5
3729	10859	7764	8343	13942	15580
3729	6044	7769	3164	14486	15952
3729	9972	9972	10859	14675	97
3735	522	10495	2678	15870	3334
3735	7769	12081	15710		

Table 4. List of document pairs that passed the test in Section 4.1: documents that refer to different events.

5123	5281
16090	16199
16094	16357
16624	6236

Table 5. List of document pairs that failed the test in Section 4.1. These are in fact exact duplicate documents like that passed the test in Section 4.3.

279	524	4809	5394	8235	8280	10675	10809	13213	13217	16361	16368
419	759	4995	5008	8290	8389	10927	10952	13494	13537	16383	1125
489	502	5009	5031	8440	8516	10934	10948	13512	13531	16607	16649
505	550	5037	5061	8585	8661	11172	11275	13527	13666	16937	16965
878	990	5155	5343	8588	8670	11177	11236	13613	14676	17195	17282
889	955	5156	5330	8606	8703	11292	11344	13692	13725	17201	17269
891	956	5161	5325	8688	8696	11605	11626	13814	13818	17846	17861
912	948	5163	5265	8729	8763	11881	11951	14489	14624	18066	18108
925	1022	5176	5290	9180	9256	11882	11949	14492	14640	18695	18735
1139	1145	5181	5279	9298	9323	11936	11939	14554	14572	18752	18768
1482	1516	5206	5271	9689	9797	12002	12009	14839	14957	18858	18902
1618	1637	5766	5862	9755	9976	12089	12107	15382	15525	19039	19157
1677	1734	5773	5857	9770	9821	12158	12192	15400	15503	19528	19582
2520	2538	5786	5895	9784	9848	12225	12236	15442	15460	19597	19605
2614	2631	6016	6061	9833	9891	12407	12466	15453	15486	19738	19754
3092	3103	6177	6208	9899	9910	12709	12720	15455	15503	19985	19986
3092	3122	6458	6670	10261	10377	12744	12833	15470	15549	20004	20088
3185	3202	6593	6621	10268	10375	12784	12835	15650	15658	21138	21122
3470	3522	6606	6746	10268	10406	12791	12842	15718	15738	21148	21017
3484	3583	6950	7044	10271	10379	12797	12834	15863	3314		
3832	3883	6970	7029	10297	10376	12800	12835	16103	16241		
3953	3997	7551	7595	10375	10406	12880	12919	16131	16254		
3987	4001	8074	8234	10405	10410	13034	13045	16139	16256		
4552	4595	8141	8244	10623	10767	13042	13050	16168	16544		

Table 6. List of document pairs that passed the test in Section 4.2: documents that refer to the same event, but one is a longer version of the other.

491 495	3735 1125	7769 1125	11941 12016	13551 14155	15233 6620
522 10864	3735 6046	7769 8344	12068 12339	13644 14224	15481 15471
522 1125	3735 8344	8077 8202	12081 14360	13696 14686	15560 15610
522 8344	3995 1878	8080 8198	12455 1969	13742 14442	15855 3315
626 630	4361 4385	8165 8263	12455 7025	13799 14486	15952 5
656 688	4847 4866	8344 10864	12456 7031	13799 15952	16017 16019
1125 11425	4969 5004	8344 1125	12471 7031	13827 14084	16174 16257
1125 3164	5168 5192	8344 3164	12495 5344	13834 15540	16236 13697
1300 1332	5177 5286	8872 8874	12495 9857	13840 14036	16379 16519
1627 1646	5180 5278	9690 9798	12601 12607	13840 14239	16383 10864
1629 1641	5309 5362	9696 9843	12857 12861	13840 14431	16491 16504
1773 1885	5890 5948	9896 9970	13211 13212	13875 15578	16774 16787
1979 2018	6046 10864	9915 10038	13299 14821	13921 13924	17783 17805
2185 2215	6046 1125	9926 9977	13308 13327	13992 14686	18011 5344
2883 2891	6046 3164	10201 10207	13308 13338	14036 14239	18011 9857
2952 17191	6046 522	10492 10571	13308 13369	14036 14431	18392 18394
2973 3048	6046 7769	10503 10864	13311 13330	14065 14288	18750 4293
3128 3133	6117 6126	10503 1125	13327 13338	14065 14443	18920 18930
3131 3133	6552 6679	10659 10769	13327 13369	14150 19689	18939 18928
3625 3704	6624 6634	10864 11425	13338 13369	14239 14431	19086 19149
3676 4239	7031 1971	10864 3164	13381 13389	14288 14443	19648 19803
3676 4292	7207 10864	11167 11254	13388 13392	14360 15710	19802 19808
3698 3732	7207 1125	11627 11638	13541 13622	14486 5	20411 20452
3735 10864	7769 10864	11829 11841	13545 13556	14781 14789	21355 21353

Table 7. List of document pairs that failed the test in Section 4.2.

32	55	3793	4066	6032	6066	10265	10352	13380	13382	17050	17068
230	240	4037	4126	6343	6397	10266	10353	13416	13530	17051	17071
258	425	4038	4139	6377	6393	10270	10360	13417	13534	17069	17078
264	344	4039	4129	6596	6637	10270	10392	13441	13444	17194	17304
414	421	4041	4125	6944	7024	10274	10389	13609	13652	17205	17283
415	427	4042	4128	6957	7023	10280	10282	13696	13992	17211	17270
519	1120	4044	4200	6961	7030	10308	10364	13807	13832	17216	17277
561	566	4046	4124	6978	7028	10312	10351	13839	13883	17217	17271
567	582	4052	4127	6991	7018	10333	10409	14476	14499	17224	17285
854	965	4068	4119	7204	7764	10343	10365	14613	14711	17229	17303
873	952	4070	4166	7241	7257	10360	10392	14618	14712	17230	17265
877	964	4072	4221	7521	7626	10630	10797	14654	14710	17236	17298
888	957	4073	4222	7524	7610	10661	10774	14656	14670	17240	17266
893	991	4079	4118	7527	7612	10665	10781	14659	14765	17244	17267
906	1014	4095	4116	7529	7611	10677	10761	14674	14713	17245	17274
907	946	4383	4400	7533	7592	10689	10771	14770	14931	17248	17306
911	947	4422	4441	7536	7630	10719	10763	14779	14913	17249	17272
926	942	4562	4574	7537	7633	10732	10764	14818	14952	17254	17289
1086	1089	4600	4604	7550	7594	10734	10773	14819	14951	17293	17300
1142	1155	4616	4712	7587	7614	10749	10808	14820	14904	17295	17299
1547	1559	4617	4711	8050	8051	10795	10812	14825	14905	17522	17533
1704	1712	4617	4740	8075	8201	10845	10873	14845	14908	17575	17593
1905	1974	4618	4752	8078	8256	10942	11013	14846	15042	17698	17702
1921	1973	4625	4727	8097	8189	11176	11245	14868	14901	17817	17820
1926	2354	4633	4709	8101	8183	11184	11240	14871	14902	17831	17838
1941	1972	4648	4708	8103	8184	11195	11244	15021	15022	17895	17947
1985	2015	4651	4718	8106	8186	11212	11238	15375	15452	17908	17944
2021	2023	4654	4717	8107	8187	11219	11246	15400	15455	17910	17946
2143	2158	4662	4713	8109	8188	11450	11551	15405	15439	17917	17945
2170	2200	4664	4714	8110	8192	11783	11844	15408	15456	18057	18105
2353	2386	4667	4706	8118	8195	11797	12180	15665	15426	18387	18388
2595	2599	4668	4726	8121	8196	11800	11845	15735	15740	18465	18549
2646	2655	4669	4704	8591	8657	11817	11848	15741	15748	18488	18564
2730	2734	4670	4705	8592	8662	11839	11852	15744	15750	18558	18574
2989	3070	4680	4703	8602	8663	12038	12041	15746	15747	18561	18593
3007	3043	4685	4754	8607	8658	12066	12072	15795	15373	18671	18673
3045	3079	4711	4740	8607	8710	12398	12469	15838	15872	19084	19085
3052	3066	4721	4738	8608	8676	12412	12467	16111	16189	19170	19171
3063	3071	4757	4762	8610	8672	12440	12612	16130	16215	19730	19755
3103	3122	4883	4927	8621	8664	12456	12471	16132	16183	20092	20103
3128	3131	5167	5273	8635	8667	12457	12473	16137	16181	20097	20098
3286	3379	5183	5280	8641	8674	12730	12837	16153	16184	20162	20167
3441	3553	5361	5368	8649	8680	12764	12830	16182	16186	20273	20309
3447	3530	5617	5629	8658	8710	12776	12804	16191	16239	20846	20847
3449	3528	5771	5864	8875	8900	12784	12800	16436	16494	20943	20930
3461	3526	5775	5871	9138	9216	12961	12970	16442	16505	20958	20948
3464	3525	5778	5858	9283	9382	12994	13012	16507	16511	21365	21364
3472	3520	5784	5863	9482	10374	13056	13074	16756	16785	21394	21358
3478	3519	5809	5853	9516	9529	13101	13107	16763	16790	21554	21552
3489	3555	5824	5855	9628	9657	13276	13290	16935	16957	21556	21512
3502	3670	5831	5856	9695	9776	13315	13544	16952	16967		
3614	3627	5906	6102	9697	9816	13320	13542	16981	16989		
3644	3648	5973	6099	9712	9777	13321	13543	17041	17066		
3735	3164	6000	6067	9751	9897	13365	13529	17047	17072		

Table 8. List of document pairs that passed the test in Section 4.3: exact duplicates published within 48 hours of each other.

5932 5958
14304 14308

Table 9. List of document pairs that failed the test in Section 4.3.