

# OCR correction based on document level knowledge

T. Nartker, K. Taghva, R. Young, J. Borsack, and A. Condit  
UNLV/Information Science Research Institute, Box 4021  
4505 Maryland Pkwy, Las Vegas, NV USA 89154-4021

## ABSTRACT

For over 10 years, the Information Science Research Institute (ISRI) at UNLV has worked on problems associated with the electronic conversion of archival document collections. Such collections typically have a large fraction of poor quality images and present a special challenge to OCR systems. Frequently, because of the size of the collection, manual correction of the output is not affordable. Because the output text is used only to build the index for an information retrieval (IR) system, the accuracy of non-stopwords is the most important measure of output quality. For these reasons, ISRI has focused on using document level knowledge as the best means of providing automatic correction of non-stopwords in OCR output. In 1998, we developed the MANICURE [1] post-processing system that combined several document level corrections. Because of the high cost of obtaining accurate ground-truth text at the document level, we have never been able to quantify the accuracy improvement achievable using document level knowledge. In this report, we describe an experiment to measure the actual number (and percentage) of non-stopwords corrected by the MANICURE system. We believe this to be the first quantitative measure of OCR conversion improvement that is possible using document level knowledge.

**Keywords:** OCR correction, non-stopword accuracy, retrieval from noisy documents

## 1. INTRODUCTION

Although the electronic conversion of archival document collections is time consuming and expensive, it has become clear that much archival information will never become useful until it is available electronically. As page reading OCR technologies have improved, such conversion operations have become more and more practical [2]. Nevertheless, because archived documents usually contain a large fraction of poor quality images, even the best OCR systems generate many conversion errors. We believe that the common context of document conversion operations provides the opportunity to exploit document level and collection level knowledge to correct these recognition errors. MANICURE was designed to accept the document text output from an OCR engine and to perform operations on it that would improve each documents retrievability. The operations applied by MANICURE include spell checking and correction based on a collection level dictionary and confusion matrix that are built dynamically. In fact, the MANICURE system is currently being used by the U.S. Department of Energy to aid in the conversion of a large collection of documents.

Over the last 5 years, ISRI has conducted many experiments that demonstrate the benefits of using MANICURE to correct OCR errors but has never before been able to quantify the improvements produced with respect to character or word accuracy. This study is our first attempt to measure and compare the accuracy of OCR output with the accuracy of MANICURE output in converting a set of test documents.

In Section 2 below, we discuss performance measures for conversion systems and focus on the metrics most appropriate for archival conversion operations. In Section 3, we present the set of documents used in this test and discuss the preparation of ground-truth characters. We also discuss the creation of multiple image copies of each document to assess the effects of image quality. In Section 4 we describe the experiments conducted and we present the results in Section 5. Finally, Section 6 presents the conclusions we draw from this study.

## 2. MEASURING THE ACCURACY OF TEXTUAL OUTPUT

When measuring the accuracy of textual output from a conversion system, it is important to determine what accuracy measure is most appropriate. A standard measure is "character accuracy" which measures the correctness of every

ASCII character on each page. Character accuracy is defined as the number of total characters minus the number of character errors, divided by the total number of characters.

$$\text{Character Accuracy} = \frac{\text{Total Characters} - \text{Character Errors}}{\text{Total Characters}}$$

Character errors are the sum of character insertions, deletions, and substitutions that are necessary to convert an output character string into the exact ground-truth string.

Because IR systems typically ignore much document text, character accuracy is not the most appropriate measure of output accuracy. For example, IR systems ignore commonly occurring words, called stopwords (such as “the” & “and”) [3]. They also ignore punctuation, most numeric digits, and all stray marks. Thus, misspelled stopwords and incorrect numbers are not errors that affect retrieval performance. Even spurious characters (delete operations), although considered errors in terms of character accuracy, will not affect retrievability. Therefore, it is clear that word accuracy is a better measure than character accuracy. In fact, in terms of retrievability of documents from an IR system, non-stopword accuracy is a better measure of conversion accuracy. Non-stopword accuracy is defined as follows:

$$\text{Non-stopword accuracy} = \frac{\text{Total non-stopwords} - \text{Number of incorrect non-stopwords}}{\text{Total non-stopwords}}$$

Although these two accuracies provide valuable measures of conversion for our application, we decided to add two more. We were also interested in the number of times any single non-stopword was misspelled in the output every time it occurred. Thus our question was, “What is the percentage of non-stopwords that are spelled correctly at least once in the document?” We call this the “unique non-stopword” accuracy. We also measured the minimum number of corrections required to have a single correct occurrence of the unique non-stopword in the document.

### 3. TEST DOCUMENTS

The most difficult part of measuring the performance of document correction systems, such as MANICURE, is to obtain the ground-truth text for a set of typical documents. The cost of retyping (and checking for correctness) any set of even 5 or 10 documents is extremely high. To solve this problem, we obtained a set of 17 documents (selected from a document collection we are converting) for which we had Microsoft Word source files. The title, authors, and number of pages for each of these documents are shown in Appendix A.

#### 3.1 Ground-truth text

Using a modified version of the program *pstotext* [4], we prepared a text version of the Word document. We manually removed all drawings, tables, equations, photographs, and maps from these text files. Also, we manually added “don’t care” characters (~) to non-ASCII characters and characters such as subscripts and superscripts. The text remaining included only title pages, table of content pages, reference pages, and main paragraph text. Table 1 shows the number of non-stopwords, the number of characters in non-stopwords, the number of unique non-stopwords, and the number of characters in unique non-stopwords in each document. Overall, more than 1.3 million characters were in this test set.

<b>Table 1. Number of Non-stopwords and Characters in the 17 Document Sample</b>				
Document Number	Total Number of Non-Stopwords	Number of Characters in Non-Stopwords	Number of Unique Non-Stopwords	Characters in Unique Non-Stopwords
1	6974	56611	912	7880
2	9641	80684	1720	15199
3	9595	79777	1796	16064
4	5572	47005	1088	9674
5	4115	33855	778	6641
6	12379	101435	2131	18691
7	4381	37387	769	7033
8	13920	113193	2465	21669
9	6318	51210	1193	10743
10	5112	44172	1175	10750
11	7792	62586	1307	11738
12	36713	302763	4198	39005
13	8440	70020	1611	14221
14	4523	37402	959	8568
15	14247	120653	1883	16925
16	8441	69501	1338	12360
17	6320	52870	907	7874
Average for all documents	9675.47	80066.12	1542.94	13825.59

We also parsed this text to remove all punctuation, most of the digits, and all stopwords<sup>1</sup>. A concerted effort was made to retain “collection related words,” even those containing digits. Thus, the text remaining contained only English non-stopwords (and project related non-stopwords that were not in a normal dictionary) and formed the basis for computing accuracies.

### 3.2 Test images

Although the cost of generating this ground-truth data was reasonable, tests of OCR output accuracy from these images would not produce results that were typical for archival conversion operations. Images printed from Microsoft Word and scanned would be much higher quality than is typical for the conversion of legacy collections.

For this reason, we chose to make a set of successive photocopies of each of the 17 documents and to compute accuracies for each generation. In fact, we decided to repeat accuracy tests for a series of copies of all documents beginning with very high quality images and progressing to 4<sup>th</sup> generation photocopies. First, we extracted images from native Microsoft Word documents that were never printed or scanned. Because they were completely free of defects associated with either the printing or scanning process, we refer to these as “original” images.

Second, we printed a hardcopy of each document from Microsoft Word and scanned these pages. We call this version of the documents “generation 0.” Four successive photocopies were made and scanned beginning with generation 0 and ending with a 4<sup>th</sup> generation copy. Care was taken to use the same photocopy engine to produce all successive copies. Thus, for each of the 17 documents, we computed accuracies for each of 6 different images of each document. Although we felt that the 3<sup>rd</sup> or 4<sup>th</sup> generation photocopy would be the best approximation of the average quality of images in archival conversion, we decided to explore what could be learned by measuring conversion accuracies over a wide range of image qualities.

<sup>1</sup> The stopwords removed were the Brown Corpus list of 450 stopwords.

## 4. TEST METHODOLOGY

We obtained a copy of the “Developers Kit 2000” version 10.0 distributed by the Scansoft Corporation. Also, we constructed an accuracy-measurement program <sup>2</sup> to compare an output text stream and a ground-truth text stream and to compute the average accuracies for each document in the stream. This program measures the average non-stopword accuracy (and the average accuracy of the characters of non-stopwords) for all non-stopwords (and for “unique” non-stopwords) for each input document.

Although only 17 documents were involved, six different images of each document were tested. The first image was the “original” image extracted from Microsoft Word. The second image was the first-printed and scanned image (i.e., generation 0). The third through sixth images are the first through fourth generation photocopies of generation 0 of these documents. Thus, the print quality of this series of images varies from “no printer and no scanner induced distortion” to 4<sup>th</sup> generation photocopy of the first printed and scanned image.

The test we conducted began by submitting all 6 images of all 17 documents to the SDK2000 OCR system <sup>3</sup> and then submitting the OCR output to the accuracy-measurement program. Appendix B shows the average OCR output accuracies measured for all of these images for each document. Next, we submitted the OCR output to the MANICURE system. MANICURE output was then submitted to the accuracy-measurement program. Appendix C shows the average MANICURE output accuracies measured for all of these images for each document. Both Appendix B and C also show the average of these averages for all 17 documents.

## 5. TEST RESULTS

To assess the potential for correcting OCR output using document level knowledge, we focus on the average of the averages for all 17 documents, especially of all non-stopwords. First, we will look at the average character accuracies for characters in non-stopwords.

### 5.1 Accuracy of characters in non-stopwords

The average character accuracies for all 17 documents are shown in Table 2 below. The top two rows show accuracies from the raw OCR output. The bottom two rows show accuracies from MANICURE output. The first and third rows show accuracies for all non-stopwords and the second and fourth rows show accuracies for unique non-stopwords.

<b>System</b>	<b>Character Accuracy of</b>	<b>Orig.</b>	<b>gen 0</b>	<b>gen 1</b>	<b>gen 2</b>	<b>gen 3</b>	<b>gen 4</b>
RAW OCR OUTPUT	All non-stopwords	99.37	99.40	99.20	98.67	97.89	97.79
	Unique non-stopwords	99.65	99.58	99.46	99.22	98.93	98.80
MANICURE OUTPUT	All non-stopwords	99.50	99.47	99.30	98.83	98.16	98.06
	Unique non-stopwords	99.66	99.57	99.44	99.27	98.93	98.86

Table 2 shows that the character accuracy of all non-stopwords from the MANICURE output is slightly better than the raw OCR output. This improvement is more profound for higher generation photocopies (i.e., 97.79% for raw OCR and 98.06% for MANICURE for the fourth generation copy). These results show that the MANICURE post-processing system does improve the character accuracy of “all” non-stopwords in a document but does not significantly change the accuracy of “unique” non-stopwords.

<sup>2</sup> This program was modeled after a set of accuracy measurement tools developed at ISRI in the 1990's [5].

<sup>3</sup> The MTX engine within SDK2000 was used for this experiment.

## 5.2 Accuracy of non-stopwords

Analysis of the accuracy of non-stopwords shows improvement that is more significant.

<b>System</b>	<b>Word Accuracy Of</b>	<b>Orig.</b>	<b>gen 0</b>	<b>gen 1</b>	<b>gen 2</b>	<b>gen 3</b>	<b>gen 4</b>
RAW OCR OUTPUT	All non-stopwords	97.44	97.03	96.45	95.05	92.78	91.91
	Unique non-stopwords	97.89	97.56	97.28	96.77	95.64	95.46
MANICURE OUTPUT	All non-stopwords	98.01	97.54	97.23	96.15	94.64	94.14
	Unique non-stopwords	98.74	98.36	98.15	97.65	96.61	96.51

The word accuracy improvement of MANICURE output over raw OCR output for all non-stopword output ranges from 0.51% for generation 0 to 2.55% for generation 4. The accuracy improvement of MANICURE output over raw OCR output for unique non-stopword output ranges from 0.80% for generation 0 to 1.05% for generation 4.

## 6. CONCLUSIONS

First, note that the benefit of using document level knowledge to correct OCR output is uniformly positive for all versions of the test documents. It is important to keep in mind that word (and non-stopword) accuracies are always lower than character accuracies. Because one incorrect character in a word causes the whole word to be in error, and because character errors tend to be spread among different words, character accuracies are uniformly higher than word accuracies.

Table 4 shows the benefit of applying document level knowledge (specifically the MANICURE system) in correcting OCR output. A 1% improvement in word accuracy over 97% correct words corresponds to eliminating 1/3 of the word errors generated by the OCR system. In fact, the percentage of incorrect words corrected varied from 17.2 to 27.5

	<b>Orig.</b>	<b>gen 0</b>	<b>gen 1</b>	<b>gen 2</b>	<b>gen 3</b>	<b>gen 4</b>
Non-stopword accuracy from OCR	97.44	97.03	96.45	95.05	92.78	91.91
Non-stopword accuracy from MANICURE	98.01	97.54	97.23	96.15	94.64	94.14
% improvement	0.57	0.51	0.78	1.10	1.86	2.23
% of incorrect words corrected		17.2	22.0	22.2	25.8	27.5

Perhaps most important, these results show that this improvement increases as page quality decreases.

## REFERENCES

1. Kazem Taghva, Allen Condit, Julie Borsack, John Kilburg, Changshi Wu, and Jeff Gilbreth, "The MANICURE document processing system", *Proc. IS&T/SPIE 1998 Intl. Symp. On Electronic Imaging Science and Technology*, San Jose, CA, January 1998.
2. Kazem Taghva, Julie Borsack, and Allen Condit, "Evaluation of model-based retrieval effectiveness with OCR text", *ACM Transactions on Information Systems*, **14**(1):63-93, January 1996.
3. Gerald Salton, *Automatic Text Processing*, Addison-Wesley, New York, 1989.
4. Paul McJones and Andrew Birrell, *pstotext*, Digital Equipment Corp. Systems Research Center, <http://research.compaq.com/SRC/virtualpaper/pstotext.html>.
5. Stephen Rice and Thomas Nartker, "The ISRI Analytic Tools for OCR Evaluation", UNLV/Information Science Research Institute, TR-96-02, August 1996.

## APPENDIX A.

**Title, author's, and number of pages for each of the 17 documents.**

<b>Doc. No.</b>	<b>Document Title</b>	<b>Author(s)</b>	<b>Num Pgs</b>
1	Plutonium Can-In-Canister - Design Basis Event Analysis	unknown	71
2	Analysis of Mechanisms for Early Waste Package Failure	unknown	64
3	Engineered Barrier System Features, Events, and Processes and Degradation Modes Analysis	G. Barr, O. Lev	71
4	Effects of Fault Displacement on Emplacement Drifts	Fei Duan	63
5	Transfer Coefficient Analysis	De (Wesley) Wu	38
6	Longevity of Emplacement Drift Ground Support Materials	David Tang	80
7	Physical and Chemical Environmental Abstraction Model	E. James Nowak	32
8	Clad Degradation – FEP's Screening Arguments	E. Siegmann	69
9	Summary of Analytical Methods and Processes Used in the Design of Uncanistered Spent Nuclear Fuel Waste Packages	unknown	65
10	Saturated Zone Colloid-Facilitated Transport	Andrew Wolfsberg	42
11	Design Analysis for the Defense High-Level Waste Disposal Container	unknown	69
12	Analysis of Geo-chemical Data for the Unsaturated Zone	J. Fabryka-Martin	156
13	Input Parameter Values for External and Inhalation Radiation Exposure Analysis	Kurt Raufenstrauch	54
14	Data Qualification Report: Water Chemistry and Infiltration Rate Data for Use on the Yucca Mountain Project	TRW Systems	36
15	Biosphere Dose Conversion Factors for Postulated Post-closure Extrusive Igneous Event	unknown	150
16	DSNF and Other Waste Form Degradation Abstraction	Thomas Thorton	47
17	Inventory Abstraction	Guy Ragan	58

## APPENDIX B.

### Average accuracy of ALL non-stopwords in OCR output

Document Number	Accuracy of All Non-stopwords						Accuracy of Characters in Non-stopwords					
	orig	gen0	gen1	gen2	gen3	gen4	orig	gen0	gen1	gen2	gen3	gen4
1	97.75	97.33	95.80	95.14	93.20	87.81	99.49	99.45	98.21	98.02	97.43	96.43
2	97.37	97.26	96.57	96.11	93.53	88.39	99.18	99.51	99.38	99.29	98.65	97.07
3	98.25	96.96	96.27	93.35	88.14	89.84	99.30	99.34	99.05	97.66	96.53	96.78
4	98.49	98.80	98.60	97.43	95.73	95.82	99.57	99.73	99.57	99.17	98.64	98.75
5	97.81	97.50	97.28	94.51	92.93	92.47	99.60	99.56	99.58	98.58	97.85	98.17
6	97.96	97.58	97.26	95.89	94.14	93.26	99.48	99.42	99.43	98.99	98.38	98.37
7	97.67	96.74	96.53	95.48	93.04	92.67	99.65	99.39	99.37	98.99	98.25	98.44
8	98.02	97.64	97.35	96.22	94.85	93.53	99.48	99.46	99.44	99.20	98.73	98.23
9	98.42	97.67	97.29	95.43	92.85	92.04	99.74	99.55	99.54	98.68	97.57	97.84
10	98.63	98.63	98.44	97.52	94.70	93.84	99.76	99.83	99.79	99.63	98.80	98.78
11	98.47	98.14	97.86	96.97	95.66	95.38	99.60	99.57	99.55	99.17	99.01	99.02
12	96.07	95.99	95.56	93.36	91.17	90.68	98.97	98.98	98.91	97.66	97.02	97.03
13	96.82	96.33	96.14	95.43	92.54	91.99	99.29	99.18	99.24	99.08	97.85	97.57
14	97.10	98.61	96.46	95.87	92.13	92.42	98.59	99.72	98.56	98.26	96.12	96.34
15	97.54	96.44	95.91	94.61	92.50	91.86	99.46	99.28	99.01	98.67	97.70	97.73
16	98.98	97.96	97.18	95.79	95.11	94.79	99.74	99.56	99.45	99.17	98.89	98.97
17	91.14	89.94	89.07	86.80	85.06	85.63	98.47	98.31	98.24	97.20	96.69	96.93
<b>Average of all Documents</b>	<b>97.44</b>	<b>97.03</b>	<b>96.45</b>	<b>95.05</b>	<b>92.78</b>	<b>91.91</b>	<b>99.37</b>	<b>99.40</b>	<b>99.20</b>	<b>98.67</b>	<b>97.89</b>	<b>97.79</b>

### Average accuracy of UNIQUE non-stopwords in OCR output

Document Number	Accuracy of Unique Non-stopwords						Accuracy of Characters in Unique Non-stopwords					
	orig	gen0	gen1	gen2	gen3	gen4	orig	gen0	gen1	gen2	gen3	gen4
1	99.45	99.01	98.68	98.57	97.37	96.49	99.89	99.77	99.00	98.97	98.73	98.34
2	98.95	98.72	98.60	98.14	97.27	95.76	99.74	99.75	99.78	99.64	99.44	99.18
3	98.46	98.11	97.51	96.68	91.42	93.90	99.59	99.59	99.29	99.02	97.69	98.34
4	98.53	99.08	98.44	98.35	97.33	97.79	99.66	99.81	99.58	99.72	99.20	99.45
5	99.36	98.97	98.84	98.07	97.94	97.69	99.85	99.82	99.77	98.80	99.61	98.71
6	98.64	97.51	97.65	96.95	96.20	95.96	99.46	99.23	99.27	99.13	98.82	98.51
7	99.09	99.09	98.83	98.44	97.92	97.14	99.80	99.82	99.80	98.92	99.69	98.66
8	98.30	97.53	97.81	96.92	97.12	96.55	99.47	99.28	99.38	99.24	99.29	99.10
9	98.91	98.07	97.90	97.65	96.31	96.06	99.76	99.66	99.67	99.27	98.73	98.86
10	99.06	99.32	99.15	98.55	96.77	96.51	99.77	99.85	99.80	99.68	99.19	99.00
11	98.47	98.47	98.01	98.24	96.79	97.17	99.52	99.54	99.47	99.48	99.29	99.34
12	85.02	84.83	84.59	83.64	82.35	81.87	99.42	99.40	99.36	98.60	98.28	98.17
13	98.32	98.01	97.83	97.08	95.78	95.72	99.52	99.34	99.50	99.33	98.76	98.92
14	98.64	98.75	97.39	96.87	96.14	95.62	99.79	99.72	98.81	98.93	97.41	97.61
15	98.19	97.61	97.82	97.19	96.55	96.02	99.57	99.47	99.48	99.31	99.28	99.11
16	98.58	98.06	97.68	97.09	97.31	96.86	99.56	99.47	99.42	99.26	99.32	99.27
17	98.13	97.46	97.02	96.58	95.26	95.70	99.67	99.40	99.43	99.36	99.11	98.98
<b>Average of all Documents</b>	<b>97.89</b>	<b>97.56</b>	<b>97.28</b>	<b>96.77</b>	<b>95.64</b>	<b>95.46</b>	<b>99.65</b>	<b>99.58</b>	<b>99.46</b>	<b>99.22</b>	<b>98.93</b>	<b>98.80</b>

**APPENDIX C.**

**Average accuracy of ALL non-stopwords in MANICURE output**

Document Nmber	Accuracy of All Non-stopwords						Accuracy of Characters in Non-stopwords					
	orig	gen0	gen1	gen2	gen3	gen4	orig	gen0	gen1	gen2	gen3	gen4
1	99.00	98.52	97.19	96.62	95.87	92.83	99.80	99.69	98.43	98.26	97.85	97.09
2	98.68	98.36	98.03	97.80	96.44	92.57	99.49	99.70	99.64	99.57	99.09	97.68
3	98.49	97.47	97.07	94.66	91.71	91.27	99.36	99.40	99.17	97.82	97.10	97.02
4	99.37	99.34	99.01	98.19	96.61	97.29	99.76	99.79	99.70	99.25	98.74	98.78
5	98.13	97.67	97.64	95.07	94.43	94.51	99.68	99.59	99.64	98.66	98.04	98.32
6	98.64	98.02	98.04	96.90	95.69	95.72	99.63	99.46	99.51	99.12	98.59	98.68
7	98.56	97.38	97.58	96.62	95.39	95.50	99.86	99.42	99.39	99.01	98.45	98.66
8	98.88	98.20	98.25	97.54	96.52	95.50	99.70	99.57	99.60	99.42	99.00	98.46
9	99.03	98.13	98.12	96.47	94.52	94.30	99.88	99.62	99.61	98.83	97.81	98.09
10	98.69	98.81	98.73	98.34	95.97	95.34	99.76	99.85	99.83	99.73	98.82	98.46
11	98.64	98.33	98.77	97.81	96.95	96.71	99.71	99.66	99.70	99.35	99.25	99.24
12	96.94	96.70	96.57	95.12	93.75	93.61	99.21	99.18	99.14	98.22	97.76	97.77
13	97.39	96.60	96.77	96.17	94.25	93.99	99.30	99.09	99.18	99.06	97.99	97.73
14	97.66	98.96	96.99	96.60	93.30	93.74	98.72	99.77	98.63	98.35	96.25	96.64
15	98.09	97.28	96.81	95.80	94.22	94.13	99.57	99.44	99.15	98.86	97.95	98.09
16	98.89	97.99	97.64	96.74	96.54	96.52	99.63	99.45	99.43	99.21	99.05	99.16
17	91.16	90.41	89.62	88.02	86.79	86.90	98.46	98.39	98.30	97.35	96.97	97.09
<b>Average of all Documents</b>	<b>98.01</b>	<b>97.54</b>	<b>97.23</b>	<b>96.15</b>	<b>94.64</b>	<b>94.14</b>	<b>99.50</b>	<b>99.47</b>	<b>99.30</b>	<b>98.83</b>	<b>98.16</b>	<b>98.06</b>

**Average accuracy of UNIQUE non-stopwords in MANICURE output**

Document Nmber	Accuracy of Unique Non-stopwords						Accuracy of Characters in Unique Non-stopwords					
	orig	gen0	gen1	gen2	gen3	gen4	orig	gen0	gen1	gen2	gen3	gen4
1	99.67	99.34	99.01	99.01	97.81	97.26	99.94	99.85	99.05	99.05	98.79	98.49
2	99.42	99.13	99.07	98.72	97.62	96.40	99.84	99.80	99.84	99.57	99.35	99.23
3	98.34	98.16	97.75	96.68	92.30	94.26	99.59	99.60	99.31	99.01	97.73	98.46
4	99.36	99.45	98.90	98.62	97.70	98.44	99.84	99.86	99.67	99.70	99.18	99.47
5	99.61	99.10	98.97	97.94	98.20	98.07	99.91	99.83	99.79	98.77	99.59	98.75
6	98.69	97.65	97.75	97.28	96.39	96.39	99.46	99.25	99.27	99.12	98.83	98.84
7	99.48	99.22	99.09	98.57	98.31	97.79	99.89	99.84	99.83	99.82	99.69	99.62
8	98.50	97.73	98.09	97.40	97.24	96.96	99.53	99.32	99.37	99.22	99.24	99.10
9	99.25	98.49	98.41	98.07	96.65	96.23	99.82	99.74	99.64	99.32	98.76	98.83
10	99.06	99.32	99.15	98.72	97.19	96.68	99.79	99.85	99.81	99.70	99.10	98.23
11	98.70	98.39	98.24	98.39	97.02	97.32	99.63	99.58	99.55	99.55	99.37	99.37
12	97.40	97.02	96.78	95.86	94.88	94.50	99.28	99.21	99.15	98.73	98.54	98.33
13	97.58	97.33	97.21	96.65	95.10	95.22	99.31	99.11	99.27	99.20	98.60	98.74
14	98.96	98.75	97.50	97.08	96.35	95.93	99.87	99.73	98.82	98.97	97.48	97.75
15	98.19	97.61	97.82	97.24	96.76	96.49	99.54	99.43	99.45	99.29	99.21	99.17
16	98.28	97.91	97.61	96.86	97.23	96.79	99.35	99.32	99.15	99.10	99.20	99.13
17	98.13	97.46	97.13	97.02	95.59	95.92	99.61	99.40	99.45	99.44	99.21	99.06

<b>Average of all Documents</b>	<b>98.74</b>	<b>98.36</b>	<b>98.15</b>	<b>97.65</b>	<b>96.61</b>	<b>96.51</b>	<b>99.66</b>	<b>99.57</b>	<b>99.44</b>	<b>99.27</b>	<b>98.93</b>	<b>98.86</b>
---	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------	--------------