TREC-5 Interactive Track Report

Paul Over

National Institute of Standards, Gaithersburg, MD 20899

1 Outline

This report presents the framework for the TREC-5 interactive track experiments in detail, cites the results but refers the reader to the site reports for their analysis, and discusses unexpected disagreements between relevance and aspectual assessment performed by the NIST assessors.

- Goals
- Experimental Design
- Execution and Results
- Review of Assessor Disagreements
- References
- Appendix A - Task specification
- Appendix B - General instructions and topic text
- Appendix C - Reporting results to NIST
- Appendix D - Instructions for NIST aspect assessors
- Appendix E - Summary evaluation
- Appendix F - Supporting material for consistency review

2 Goals

The high-level goal of the Interactive Track in TREC-5 was the investigation of searching as an interactive task by examining the process as well as the outcome. As more people use or procure interactive search systems the ability to understand their relative utility within a given context becomes more important. In order to build better interactive systems much more needs to be known about how humans interact with such systems in filling their various needs for information. To these ends an experiment was designed, including:

1. a task for the searchers participating in the experiment
2. agreements on data recording and formats for results reporting
3. measures

3 Experimental Design

3.1 Task

A minimum of four searchers per site were each to perform six searches on the TREC-5 adhoc collection, drawing topics from a set of 12 especially chosen adhoc topics modified for use in the interactive track. Three of the searches were to be on the site's experimental system and three on a control system, the purpose of which was to aid in comparing systems across sites. See Appendix A for details on the experimental design.

The topics described an information need with many aspects - an aspect being roughly one of many possible answers to a question which the topic in effect posed. For example, topic 254 asks the question "What kinds of medications or procedures other than heart surgery have been used to treat heart ailments ?" Searchers found aspects/answers such as angioplasty, lasers, stents, etc. in the documents of the collection. Language was added to each topic to make the meaning of aspect for that topic clearer. See Appendix B for the general instructions to the searchers and the text of the topics.

The task of the interactive searchers was to save documents, which, taken together, covered as many different aspects of the topic as possible in the 20 minutes allowed per search. The number of documents saved was not important. Participants were not penalized for saving documents that covered a given aspect more than once.

The task posed special challenges to participating systems and searchers. Beyond the usual goal of finding documents about a particular topic, it would seem to favor systems which also support the searcher in efficiently:

1. isolating and comparing material in a document about different aspects of the topic: "the document pertains to the topic but which, if any, specific aspects are covered and where?"

2. avoiding other documents with mostly the same or similar material: "once I find one or more aspects, how much help, if any, do I get in avoiding documents and document sections which cover already identified aspects?"

Two sorts of result data were to be collected. Sparse format data for each search comprised the list of documents saved and the elapsed time of the search. Rich format data recorded significant events in the course of the interaction and their timing. Sparse format data were to be the basis for the summary evaluation, which produced a triple for each search: precision, aspectual recall, and time. Rich format data were intended for analytical evaluation by the experimenters. See Appendix C for information on the format of the results reported to NIST.

3.2 Interactive Track Assessment and Evaluation at NIST

For each topic, a pool was formed containing the unique documents saved by at least one searcher for that topic regardless of site. The measures for the interactive track required two sorts of assessments: traditional relevance assessment and a new aspectual assessment.

Relevance assessment for the interactive track amounted only to judging those few documents not already judged for the main TREC task. Every document in the pool was assessed for aspects regardless of whether it had been judged relevant or not.

Aspectual assessment required new procedures. Where possible (there were two exceptions) the assessor who had performed the relevance assessment was assigned to do the aspectual analysis. For each topic, the aspect assessor was asked to:

1. read the topic carefully

2. read each of the documents to be judged and gradually:

   (a) create a list of the aspects found somewhere in the documents

   (b) select and record a short phrase describing each aspect found

   (c) determine which documents contain which aspects

   (d) bracket each aspect in the text of the document in which it was found (as this assessment task was new, the assessors worked with paper and pencil)

The NIST assessors were given the same topic information as the interactive track searchers except that the paragraph that begins "Please save at least one document" was systematically replaced with a section called "Answers", for example in the case of topic 256i:

> Please save at least one document that identifies EACH DIFFERENT criticism of this trend. If one document discusses several criticisms, then you need not save other

documents that repeat those aspects, since your goal is to identify the different criticisms that have been made.

was be replaced by

Answers: each different criticism of this trend.

In dealing with the assessors, the term “aspect” was replaced with the term “answer.” The assessors seemed to have little difficulty understanding the basic task. The questions that did arise concerned mainly the granularity of the answers/aspects, e.g., “should I lump these 3 low-level answers together and count them as 1?” See Appendix D for details on the instructions given to the NIST interactive track assessors.

3.3 Summary Evaluation

For each search (by a given participant for a given topic at a given site), NIST:

1. used the submitted list of selected documents and the assessor's aspect-document mapping for the topic to calculate the percentage of aspects covered by the submitted documents.

2. calculated the precision for the submitted document set using the standard relevance judgements

3. took the elapsed clock time for the search from the submitted data

3.4 Analytical Evaluation

No analytical evaluation was performed at NIST. Additional information may be available in the site reports from Rutgers University and City University, London in the TREC-6 Proceedings.

4 Execution and Results

A relatively late start, problems installing the control system and indexing the collection for its use, and in one case problems setting up the experimental system resulted in the two participating systems' not being able to execute the complete experimental design. No searches using the control system were run. Searches assigned to the control system by the design were instead performed on the experimental system.

The summary measures for each system are presented separately in Appendix E. The two sites' results are not directly comparable to each other, not only because no control was used, but also because the collections searched were different. See the site reports from City University, London and Rutgers University in the TREC-5 Proceedings for more information about the actual execution of experimental design.

Even if the experiments as run do not allow comparison of systems, exploratory analysis of the results for each system underlines the importance of dealing in any experimental design with some or all of the following:

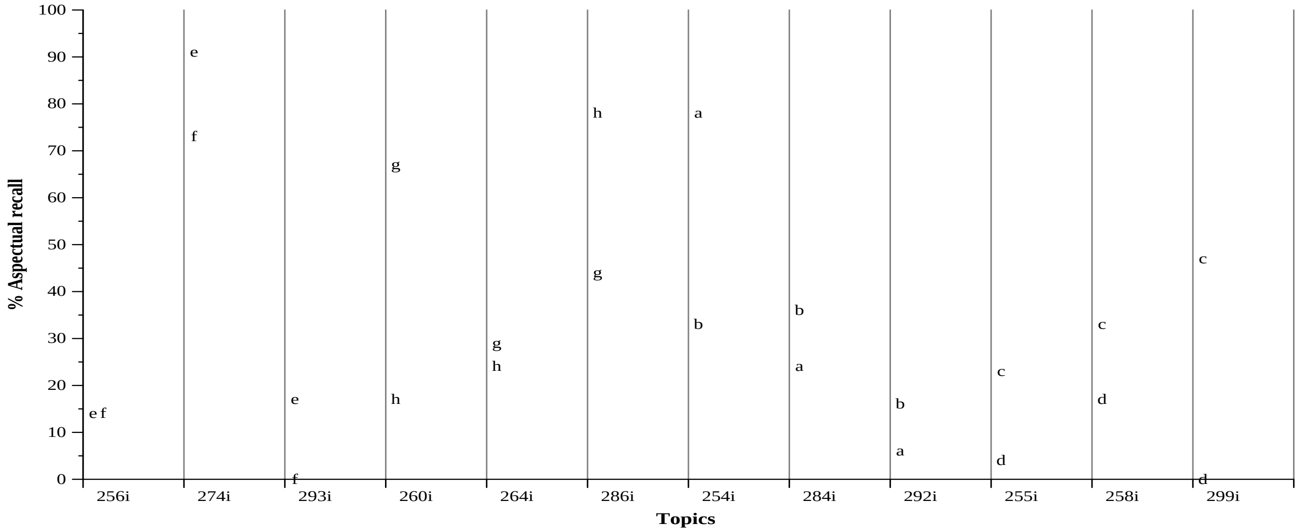
- a strong searcher effect - within a single system and for a single topic, results vary greatly for different searchers

- a strong topic effect - within a single system and for a single searcher, results vary greatly for different topics

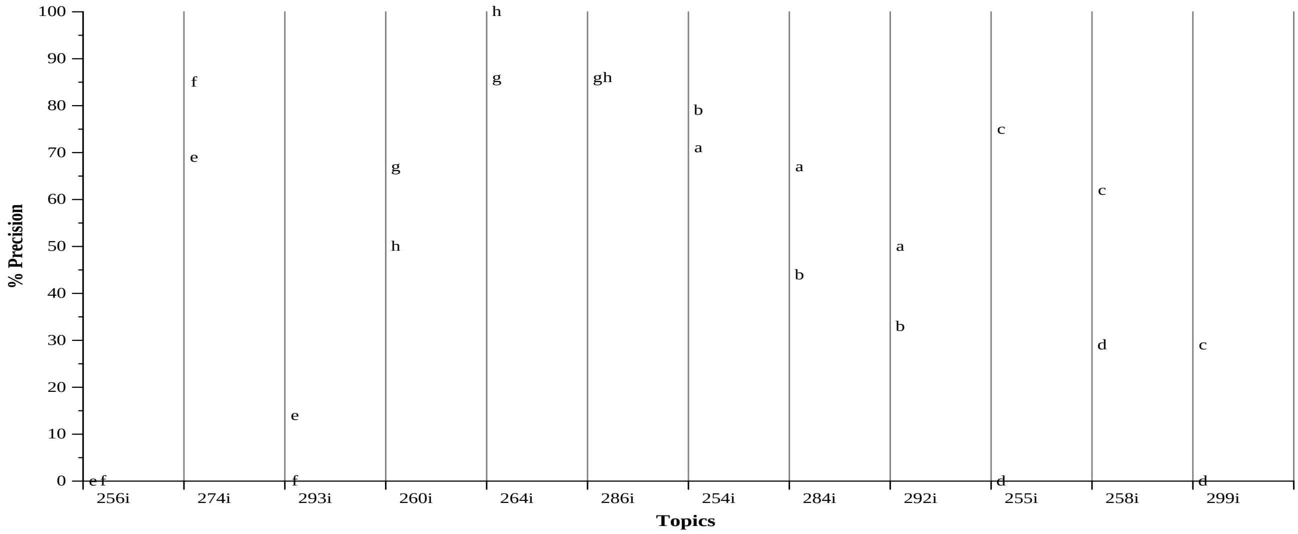
- a seacher-topic interaction - within a single system, the effect of different searchers on the results varies for different topics

Figure 1 and Figure 2 depict the 2 main measures for each system by topic and with searcher information encoded in the data points. For example in the upper graph for the results from Rutgers the second column from the left shows the results for topic 274i. Participants “d” and “m” each completed a search and saved documents which covered about 65% of the total number of aspects identified by the NIST assessors.

See for the site reports from City University, London and Rutgers University for more information about the actual execution of experimental design and discussions of an analytical nature.

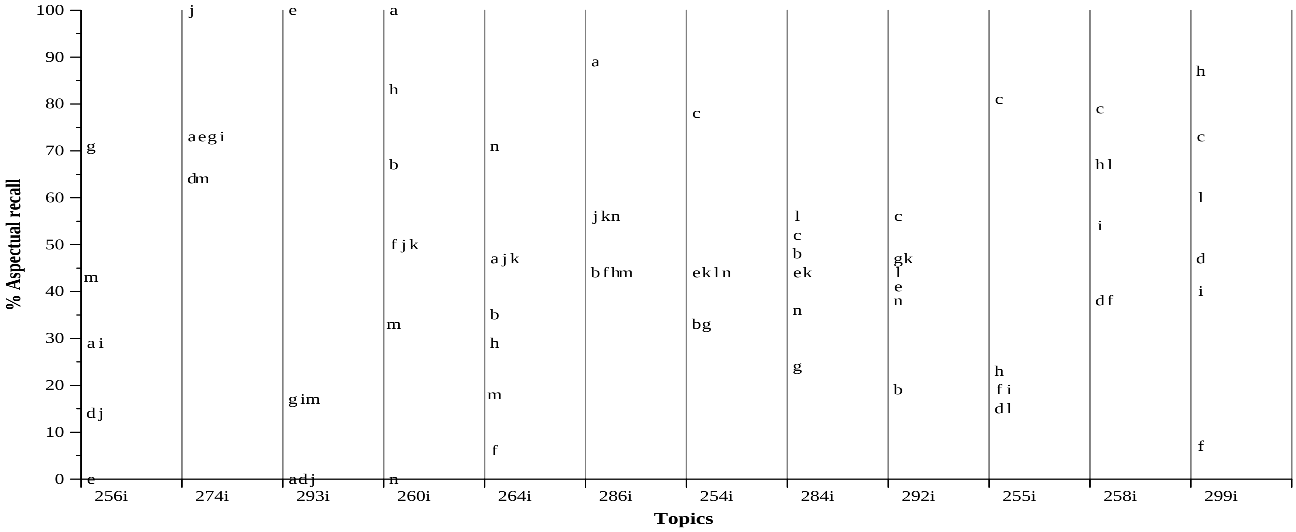


Aspectual recall for all searches by topic
Searchers: a=and b=kar c=bri d=col e=cla f=lis g=ala h=sar

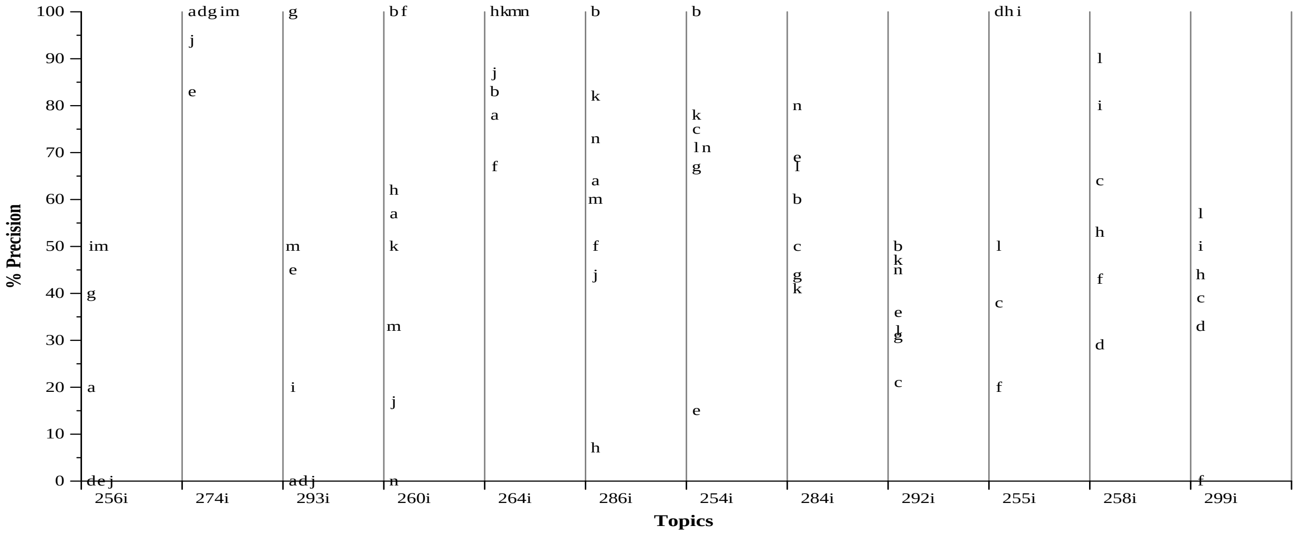


Precision for all searches by topic
Searchers: a=and b=kar c=bri d=col e=cla f=lis g=ala h=sar

Figure 1: Results from City University, London.



Aspectual recall of all searches by topic
Searchers: a=s001 b=s002 c=s003 d=s004 e=s005 f=s006 g=s007 h=s008 i=s009 j=s010 k=s011 l=s012 m=s013 n=s014



Precision of all searches by topic
Searchers: a=s001 b=s002 c=s003 d=s004 e=s005 f=s006 g=s007 h=s008 i=s009 j=s010 k=s011 l=s012 m=s013 n=s014

Figure 2: Results from Rutgers University.

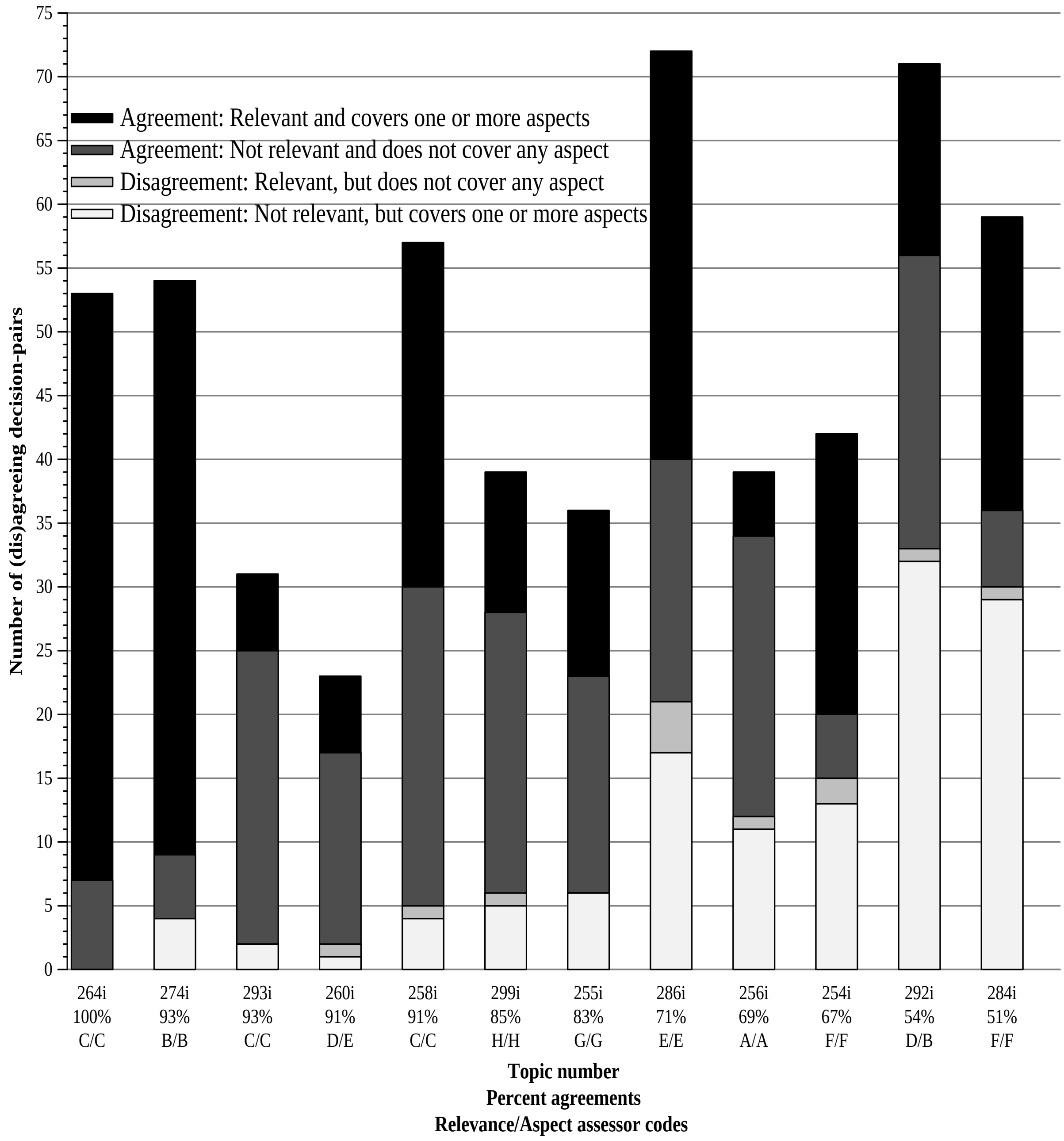


Figure 3: NIST Assessor (dis)agreements in relevance versus aspect judgements by topic.

5 Review of Assessor Disagreements

In the process of evaluating the results submitted to NIST, the assessors answered two questions for each of the 583 documents in two sessions within 2 weeks of each other.

- "Is it relevant to the topic?"
- "What, if any, aspects does it cover?"

For 10 of the 12 topics, the relevance assessor was the same person as the aspect assessor. For two topics, the relevance and aspect assessors were not the same person.

The designers of the interactive task assumed there would be a high degree of consistency between relevance and aspect assessment results. Being relevant was expected to imply covering one or more aspects and vice versa, but the TREC-5 interactive track assessments appear to violate this assumption in a significant number of cases. As part of the follow-up to TREC-5 conference, the author reviewed the TREC-5 interactive track data and relevant, available literature on consistency of relevance judgements in an attempt to answer the following initial questions:

1. How much inconsistency exists in the TREC-5 interactive track assessments and of what type(s) is it?
2. Are the levels of consistency or lack of it within "normal" bounds?
3. For any abnormal levels, are there any likely explanations?

5.1 Amount and Types of Inconsistency

5.1.1 The Data

Figure 4 shows the number of judgements by type: total agreements, total disagreements, and then the two possible types of disagreements:

1. "Document is relevant, but does not cover any aspects"
2. "Document is not relevant, but does cover one or more aspects"

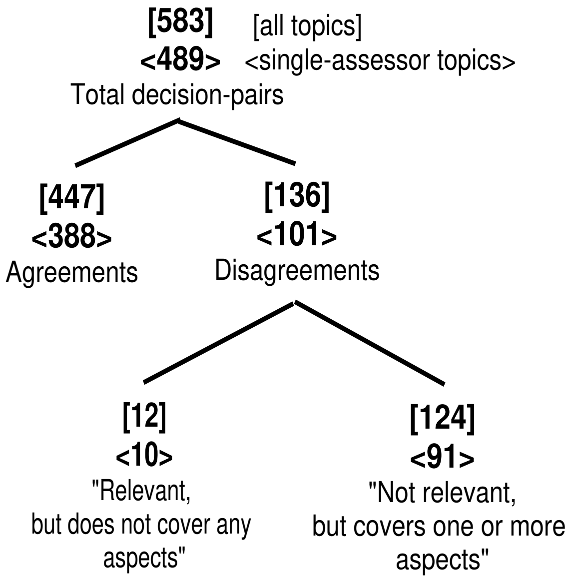


Figure 4: Categorization of assessments.

Looking at all topics, relevance and aspect assessments are in agreement 77% of the time (80% of the time in the case of the single-assessor topics). The disagreements are unequally distributed with about 90% being of the type "Document is not relevant, but does cover one or more aspects" - another unexpected result.

Figure 3 shows more detail, for each topic: the number of assessor agreements and disagreements by type, the percentage of agreements, and the relevance and aspect assessors. The topics are arranged from left to right in order of decreasing percentage of agreements. For example, the sixth column from the left represents the data on topic 299. Assessor H made both the relevance and the aspectual coverage assessments, which agreed 85% of the time, i.e., for 33 of 39 documents the judgements were: "Relevant and covers one or more aspects" or "Not relevant and covers no aspect." There was one document judged to be relevant but not to cover any aspect, and there were five documents judged to be not relevant, but found to cover one or more aspects.

5.1.2 Two Errors in Results Processing

In the course of reviewing the documents on which NIST assessors disagreed, two errors made in the assessment recording were uncovered. The information presented here on the review of consistency reflects the corrected results, but because the effect would be small and in order to maintain consistency with the data on which the site reports were based, the summary evaluation presented here has not been corrected.

For topic 254 the document WSJ920225-0057 was assessed as covering aspect 1 but did not. The relevance assessment as "not relevant" stands. The aspectual recall of the Rutgers University search 254i-005-2 might have been affected, but was not because another document saved by the searcher also covered aspect 1. For topic 284 the document AP880310-0205 was incorrectly recorded as covering aspect 1. The relevance assessment as "not relevant" stands. Correcting the error would reduce by 4% the aspectual recall of the following searches:

- City University 284_and
- City University 284_kar
- Rutgers University 284i-003-1
- Rutgers University 284i-011-1

5.2 What's normal consistency?

Are the levels of consistency or lack of it within "normal" parameters? The published literature on relevance assessments contains mostly studies which measure differences between different individuals' assessments of the same material, not between the same individual's. The heterogeneity of the between-assessor studies with regard to assessor characteristics, type/length of material judged, type of judgement, similarity measure, etc., makes comparison very difficult, but, with that caveat, one might minimally expect the TREC-5 within-assessor data to exhibit greater consistency than the between-assessor data, and for the most part the TREC-5 data do.

By any of the measures in Figure 5, averaged over all the single-assessor topics, the TREC-5 relevance and aspect judgements agree to a greater extent than most of those in the published studies comparing different individuals' assessments of the same topic-document pairs. (Note: As can be seen from Figure 5 the 2 two-assessor topics have little effect on the average across all topics but to simplify comparison with other work, references here will be to the single-assessor topics only, unless explicitly noted.)

For example, with regard to the ability of one judge to duplicate the judgement of another as measured by "effectiveness" (where 0 is no agreement and 1 is perfect agreement) (Goffman & Newill, 1966) the following averages have been reported or can be calculated from reported data:

- 0.35 (Barhydt, 1967)
- 0.30 (Janes & McKinney, 1992)
- 0.495 (Janes, 1994)
- 0.58 TREC-5 Interactive

Similarly, with regard to agreement on relevant documents as defined by Lesk and Salton (1969):

- 0.31 (Lesk & Salton, 1969)
- 0.52 (Janes & McKinney, 1992)
- 0.36 - 0.64 (Burgin, 1992)
- 0.70 TREC-5 Interactive

With regard to direct comparisons:

- 57.2% (Figueiredo, 1978) as reported by Janes (1994)
- 66% and 70% (Janes & McKinney, 1992)
- 80% TREC-5 Interactive

With regard to correlation:

- 0.55 - 0.75 for subject experts (Saracevic, 1975)
- 0.45 - 0.60 for "information providers" (Saracevic, 1975)

Topic Number	Decision-pair types: Relevance - Aspects R = Relevant(Covers aspect) N = Not Relevant(No aspect)				Measures of Agreement						
	RR	RN	NR	NN	% Decision Pairs that agree 100*(RR+NN)/All	Pearson's r	Agreement on Relevant Docs RR / (RR+RN+NR)	Agreement on Non-Relevant Docs NN / (NN+NR+RN)	Effectiveness of Aspect Assessment in Approximating Relevance Assessment: Sensitivity RR / (RR+RN)	Specificity NN / (NN+NR)	Effectiveness Se+Sp-1
264	53	0	0	7	100	1.00	1.00	1.00	1.00	1.00	1.00
274	45	0	4	5	93	0.71	0.92	0.56	1.00	0.56	0.56
293	6	0	2	23	93	0.83	0.75	0.92	1.00	0.92	0.92
260	6	1	1	15	91	0.79	0.75	0.88	0.86	0.94	0.79
258	27	1	4	25	91	0.83	0.84	0.83	0.96	0.86	0.83
299	11	1	5	22	85	0.69	0.65	0.79	0.92	0.81	0.73
255	13	0	6	17	83	0.71	0.68	0.74	1.00	0.74	0.74
286	32	4	17	19	71	0.45	0.60	0.47	0.89	0.53	0.42
256	5	1	11	22	69	0.37	0.29	0.65	0.83	0.67	0.50
254	22	2	13	5	67	0.31	0.59	0.25	0.92	0.28	0.19
292	15	1	32	23	54	0.31	0.31	0.41	0.94	0.42	0.36
284	23	1	29	6	51	0.27	0.43	0.17	0.96	0.17	0.13
All topics — Totals ⇨	258	12	124	189	Avgs. ⇨ (weighted) 77	0.58	0.65	0.58	0.95	0.60	0.55
1-assessor topics	237	10	91	151	80	0.61	0.70	0.60	0.96	0.62	0.58

Figure 5: Various measures of agreement between the TREC-5 relevance and aspect assessments.

TREC conference	Assessors per document	Average percent of agreement	Type of agreement	Number of documents	Judgment type(s) per document
TREC-4	3	81%	2-way	14968	relevance relevance relevance
TREC-5	1	80%	2-way	489	relevance aspect
TREC-2	2	79%	2-way	6867	relevance relevance
TREC-5	1 or 2	77%	2-way	583	relevance aspect

Figure 6: Summary of TREC assessment consistency data.

- 0.61 TREC-5 Interactive

Exceptions include Shaw, Wood, Wood, and Tibbo (1991), in which direct comparison shows agreement of 98% or better by experts who were allowed to confer and resolve disagreements; Cleverdon (1970), who cites rank correlations "never below 0.92" on full Cranfield documents; the direct comparison, which can be calculated from raw data reported by Lesk and Salton: 98.4% overall agreement on abstracts versus TREC-5's average for single-assessor topics of 70%; and Rees (1967) who reports a correlation (r) of 0.72 on abstracts by judges with varying degrees of medical knowledge.

An example of within-assessor consistency in relevance judgements can be found in a follow-on study performed by Rees et al. (1967). In a follow-up to their larger study Rees et al. had 21 medical students rate abstracts of 12 documents with regard to 5 "aspects" on 11-point scales in two sessions 2 weeks apart. The average correlation per student calculated from the published raw data on the first "aspect" (overall relevance) is 0.72. For the TREC-5 data on relevance/aspectual coverage of full documents, the average correlation for single-assessor topics is is 0.61.

TREC-5's within-assessor agreement of 80% lies surprisingly close to the between-assessor agreement for TREC-2 (79%) and TREC-4 (81% for decision pairs). During the assessment of TREC-2 results, 6867 documents were judged for relevance by a second assessor. For TREC-4 14,968 documents were judged for relevance by a second and a third assessor. Figure 6 summarizes the existing TREC data on consistency of assessments. One might have expected greater agreement where only one assessor was involved but less to the extent that the tasks - judging relevance versus judging aspectual content - were perceived as different.

5.3 Likely Explanations?

In an informal attempt to understand some of the variation in consistency by topic, the author read all of the documents on which the TREC-5 interactive track assessors disagreed and reviewed the relevance and aspect assessments. The author disagreed to varying degrees both with the assessors' relevance judgements and with their aspectual decisions, but no pattern was detectable and so no results of this examination are reported except for the two topics showing the highest levels of disagreement: 284 and 292. The following were examined for these two topics:

1. topic text
2. documents judged against the topic for TREC-5 interactive
3. assessor assignments
4. relevance assessor comments
5. aspect assessor comments

Topic 284 - International drug enforcement cooperation Of 29 disagreements all are of the type "Not relevant, but covers aspects." At the risk of just adding one more set of judgements, it is very difficult after reviewing the documents to understand how 26 of them could have been judged as "not relevant" unless different interpretations of the topic were used in relevance and aspect assessment. However, the relevance assessor also performed the aspect assessment. The assessor provided comments after making the relevance judgements and later after making the aspectual assessments. The comments provide no evidence of any change in interpretation between relevance assessment and aspect assessment. See Appendix F for assessor comments and extracts from documents, which would seem to make them relevant.

Topic 292 - Worldwide welfare The topic description deals in rather fuzzy terms e.g., "social programs" and "poor people." The narrative lists a number of characteristics a relevant document "should" or "would" have - raising the possibility that judgements could differ due to different assumptions about the degree to which each of these characteristics (at least the "would's") is required.

Of 33 disagreements all but one were of the type "Not relevant, but covers aspects" The aspect assessor bracketed each aspect identified in the document being judged. Few if any of the bracketed answers/aspects meet the majority of the criteria in the narrative, much less all.

The relevance assessor, who was different from the aspect assessor, made comments which indicate the criteria in the topic narrative were taken as a list of requirements, although exceptions were made. In contrast, the aspect assessor's comments indicate no such strict interpretation and the bracketed aspects suggest a consistently looser view of the topic narrative. It should be noted that the other topic (260) for which aspect assessment was done by a different person than did relevance assessment also has a narrative which lists several criteria for relevance, but this topic had a high level of between-assessor agreement.

5.4 Summary with respect to assessor disagreements

The levels of agreement between relevance and aspect judgements for the TREC-5 interactive assessments are in rough agreement with the reported data for repeated assessments by a single assessor. The notion that the lack of even greater agreement is due largely to differences in how the assessors' perceived the two tasks - relevance and aspect assessment - requires further testing in a better designed, controlled setting. The fact that disagreements were overwhelmingly of the sort "Not relevant, but does cover aspects"seems like additional evidence that the concepts of being relevant and covering aspects were understood as different and that relevance taken more narrowly than aspectual coverage, but there is no independent evidence for this assumption. Neither can we move beyond speculation as to why topics 284 and 292 generated the most disagreement. Finally, if the TREC-5 interactive task specification is reused in TREC-6, it would seem advisable to define precision in terms of the aspect assessment rather than in terms of relevance assessment. Figure 7 shows the results that would have been obtained if precision had been based on aspect assessment rather than the standard relevance assessments.

References

Barhydt, G. C. (1967). The effectiveness of non-user relevance assessments. *The Library Quarterly, 23*(2), 146–149.

Burgin, R. (1992). Variations in relevance judgements and the evaluation of retrieval performance. *Information Processing and Management, 28*(5), 619–627.

Cleverdon, C. W. (1970). *The effect of variations in relevance assessments in comparative experimental tests of index languages* (Cranfield Library Report No. 3). Cranfield, UK: Cranfield Institute of Technology.

Figueiredo, R. C. (1978). [Estudo comparativo de julgamentos de relevância do usuário e não-usuário de servi cos de D. S. I.]. *Ciéncia da Informa cao, 7*, 69–78.

Goffman, W. and Newill, V. A. (1966). A methodology for test and evaluation of information retrieval systems. *Information Storage and Retrieval, 3*, 19–25.

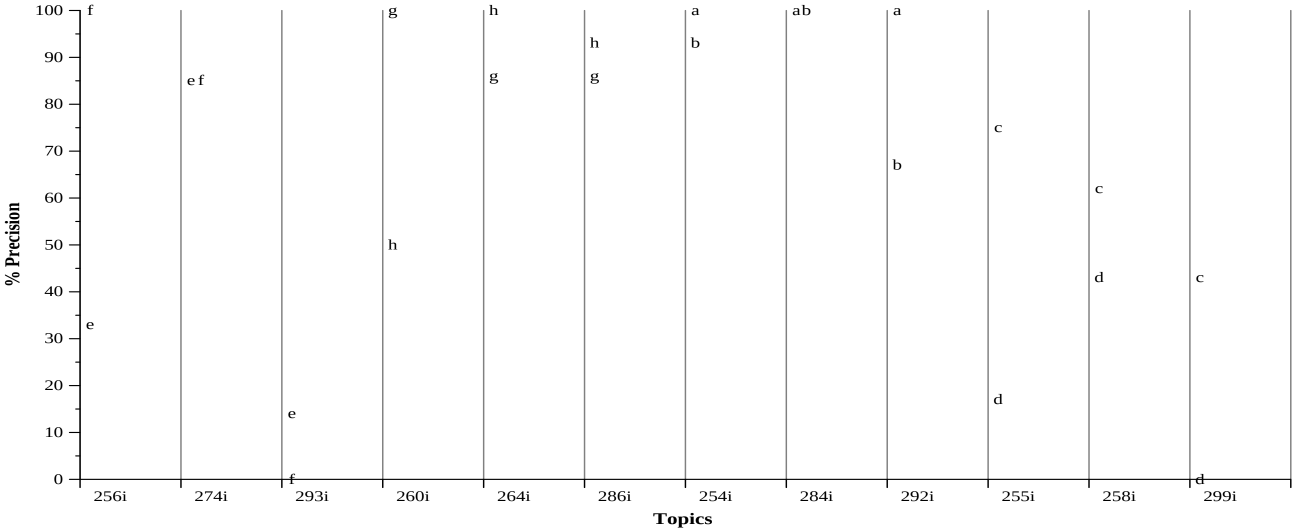
Janes, J. W. (1994). Other people's judgements: a comparison of users' and others' judgements of document relevance, topicality, and utility. *Journal of the American Society for Information Science, 45*(3), 160–171.

Janes, J. W. and McKinney, R. (1992). Relevance judgements of actual users and secondary judges: a comparative study. *The Library Quarterly, 62*(2), 150–168.

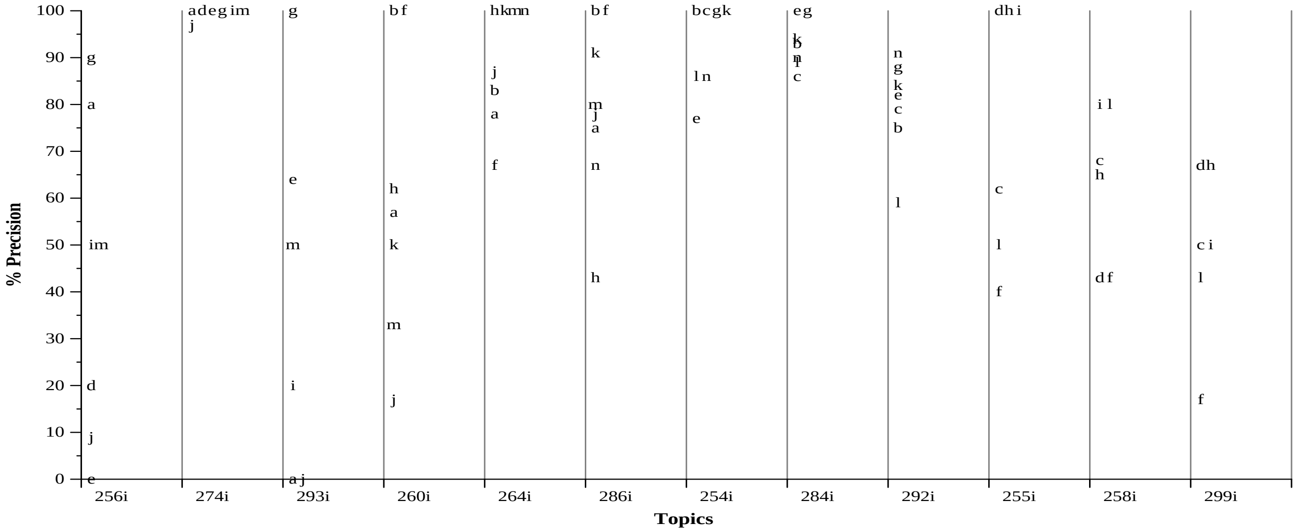
Lesk, M. E. and Salton, G. (1969). Relevance assessment and retrieval system evaluation. *Information Storage and Retrieval, 4*, 343–359.

Rees, A. M. (1967). Evaluation of information systems and services. In C. A. Cuadra (Ed.), *Annual review of information science and technology* (Vol. 2, pp. 343–359). John Wiley & Sons.

Rees, A. M., Schultz, D. G., Baumanis, G., Marcus, S., Rothenberg, L., Saracevic, T., Stern,



Precision based on aspect assessments for all searches by topic - City University, London
Searchers: a=and b=kar c=bri d=col e=cla f=lis g=ala h=sar



Precision based on aspect assessments of all searches by topic - Rutgers University
Searchers: a=s001 b=s002 c=s003 d=s004 e=s005 f=s006 g=s007 h=s008 i=s009 j=s010 k=s011 l=s012 m=s013 n=s014

Figure 7: Revised Precision Results.

M., and Zull, C. (1967). A field experimental approach to the study of relevance assessments in relation to document searching. In (Final Report to the National Science Foundation). Cleveland, OH: Case Western Reserve University. (NSF Contract No. C-423)

Saracevic, T. (1975). Relevance: a review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, *26*, 321–343.

Shaw, W. M. J., Wood, J. B., Wood, R. E., and Tibbo, H. R. (1991). The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research*, *13*, 347–366.

6 Appendix A - Specification of the TREC-5 interactive task

The TREC interactive task will use a single set of topics selected for this purpose. The topics are given below. They are taken from the new topics designed for the main TREC adhoc experiment; however, the task given to the searchers is not the same as that required for the main experiment, and a few of the topics have been slightly modified to suit the interactive task set. Searches will be made on the same set that is being used for the main adhoc experiment.

Each searcher will search some topics on the local system under investigation, and some on the NIST ZPRISE system. ZPRISE is a basic query-formulation-and-ranked-retrieval system, and should be set up in a standard configuration, specifically without the relevance feedback facility available in the latest version, but with a facility for recording user selection of items. It may be set up locally or used remotely at one of the other sites involved in the experiment.

6.1 Experimental design

Topics come in four blocks (B1-B4) of three topics each, that is 12 topics altogether. A minimum of four searchers is required (S1-S4), each searching one block of three topics on each of two systems: ZPRISE as control (C) and the local experimental system (E). Each searcher therefore does six topics in total. The design is:

	B1	B2	B3	B4
S1	E	C		
S2		E	C	
S3			E	C
S4	C			E

6.2 Topic category and task

The topics which have been chosen require the discovery of a range of different aspects for satisfactory resolution, and the task given to searchers is defined in these terms.

The task is then to conduct an interactive search with the following specification: “Your job is to identify as many aspects of each of the following topics as you can, in twenty minutes. You should find and save at least one document for as many different aspects as possible. ‘Aspects’ can mean, for instance, different developments in a field, or different methods, or different countries, or different opinions. To identify precisely the type of aspect, you should read carefully the ‘Description’ of the topic.”

Participants may need to refine the specification given, and/or provide additional guidance, in their instructions to their searchers. Any such refinement/guidance should form part of the report.

6.3 Data to be recorded

It will be necessary for the system and/or the searcher to record and report on the progress and outcome of the search, in various ways. There follows a series of notes on specific items that need to be recorded; below is a partial specification of the reporting format.

Documents selected: the searcher’s selection (choice) of items for the final output list must be identified. This selection and the time taken form the basis for the summary evaluation (see below).

Time taken: the elapsed (clock) time taken for the search, from the time the searcher first sees the topic until s/he declares the search to be finished, should be recorded. It is assumed that the interactive search takes place in one uninterrupted session. If a session is unavoidably interrupted, it is recommended that it be abandoned and the topic given to another searcher.

Sequence of events: all significant events in the course of the interaction should be recorded. The events listed below are those that seem to be fairly generally applicable to different systems and interactive environments; however, the list may need extending or modifying for specific systems.

- Timing of events: it may be necessary to record the times of individual events in the interaction (see below).
- Intermediate search formulations: if appropriate to the system, these should be recorded.
- Documents viewed: "viewing" is taken to mean the searcher seeing a title or some other brief information about a document; these events should be recorded.
- Documents seen: "seeing" is taken to mean the searcher seeing the text of a document, or a substantial section of text; these events should be recorded.
- Terms entered by the searcher: if appropriate to the system, these should be recorded.
- Terms seen (offered by the system): if appropriate to the system, these should be recorded.
- Selection/rejection: documents or terms selected by the user for any further stage of the search (in addition to the final selection of documents).

6.4 Reporting formats

The evaluation (see below) is separated into two stages: a summary evaluation, involving a small number of measures defined here, and an analytical evaluation, which may involve a number of different measures not all of which have yet been defined. The "sparse" reporting format defined here is intended to provide the minimum data required for the summary evaluation; the "rich" format is to provide the additional information needed for the analytical evaluation.

"Sparse" format: a list of the identifiers of the selected documents for each topic, together with the elapsed (clock) time of the search.

"Rich" format: for each topic, the sequence of events as indicated above, and perhaps the times of events. The specification of this rich format will be based on that discussed for TREC-4; it is likely to require further interaction among the groups taking part, to ensure that all groups can comply.

Two further items should be reported. A full narrative description should be given of the interactive session for one specified topic (274) As indicated above, any further guidance and/or refinement of the task specification given to the searchers should also be reported.

6.5 Summary evaluation

The summary evaluation measures will be aspect-recall (that is, proportion of the aspects of the topic covered), precision (according to the usual relevance judgements), both measured for the selected set, and elapsed search time, as defined above. These constitute a triplet of measures, to be taken together. Some account will need to be taken of the results on the control set of searches on ZPRISE.

6.6 Analytical evaluation

Participants are invited to consider ways of measuring performance, and/or of diagnosing the effects of system features or other variables.

6.7 Topic blocks

```
B1:  256i, 274i, 293i
B2:  260i, 264i, 286i
B3:  254i, 284i, 292i
B4:  255i, 258i, 299i
```

The order of topics within a block is not significant.

7 Appendix B

7.1 General instructions

These are the general instructions to be given to each searcher at the beginning of the interaction with both the control and the test system.

> Imagine that you have just returned from a visit to your doctor during which it was discovered that you are suffering from high blood pressure. The doctor suggests that you take a new experimental drug, but you wonder what alternative treatments are currently available. You decide to investigate the literature on your own to learn what different alternatives are available to you for high blood pressure treatment. You are only interested in finding one document for each of the different treatments for high blood pressure. You find and save a single document that lists four treatment drugs. Then you find and save separate documents that discuss the use of calcium, regular exercise, biofeedback, and the snakeroot plant as possible alternative treatments. In all, you have identified eight different aspects for this topic in five documents.
>
> Now we would like you to identify as many aspects as possible for each topic that will be presented to you. You will be given 20 minutes to search for each topic's aspects. Please save one document for each of the aspects that you identify. If you save one document that contains many aspects, try not to save additional documents that contain only those aspects, unless a document contains additional aspects as well.
>
> Carefully read each description and narrative for each topic because the interpretation of "aspects" changes from topic to topic. For example, aspects can refer to different developments in a field, to different instances in which an event can occur, or to different kinds of treatments – as it did in our example above.
>
> Do you have any questions about
>
> - what we mean by aspects
> - the way in which you are save non-redundant documents for each aspect?

7.2 Specific instructions

These are the specific instructions for each individual topic. The searcher should be given the text of each topic exactly as follows, each topic text at the beginning of each search. Timing of the search begins with the searcher's receipt of the topic.

7.3 Topic text

7.3.1 Block B1

Number: 256i

Topic: Negative Reactions to Reduced Requirements for College Undergraduate Core Studies

Description: Colleges for a long time have been reducing their requirements in such core subjects as history, literature, philosophy, and science. Criticism of this trend has occurred.

Please save at least one document that identifies EACH DIFFERENT criticism of this trend. If one document discusses several criticisms, then you need not save other documents that repeat those aspects, since your goal is to identify the different criticisms that have been made.

Narrative To be relevant, a document will provide negative opinions/facts concerning the fact that colleges have reduced their basic requirements for the granting of degrees to undergraduates.

Number: 274i

Topic: Electric Automobiles

Description: What are the latest developments in the production of electric automobiles?

Please save at least one document that identifies EACH DIFFERENT recent development in the field of electric automobiles. If one document discusses several developments, then you need not save other documents that repeat those developments, since your goal is to identify the different ones that have been discussed.

Narrative: The economic feasibility of electric automobiles appears to be limited by a number of factors, including the limited range of operation between recharges of batteries. What progress has been made in addressing these factors?

Number: 293i

Topic: Evacuation of U.S. Citizens by U.S. Military

Description: Evacuation of U.S. Citizens by U.S. Military since the year 1900.
Please save at least one document that identifies EACH DIFFERENT instance of evacuation in this century. If one document discusses several historical instances, then you need not save other documents that mention those evacuations, since your goal is to identify the different evacuations that have occurred.

Narrative: The U.S. Military has many times been called upon to evacuate non-combatant U.S. personnel from Foreign Countries. A relevant document will identify instances, since 1900, where U.S. military personnel have been called on to evacuate non-military U.S. citizens from possible chaotic and dangerous situations around the world. Evacuation of military personnel should not be considered relevant unless accomplished in concert with the evacuation of civilians.

7.3.2 Block B2

Number: 260i

Topic: Evidence of human life

Description: What evidence is there of the existence of human life in the New World 10,000 and more years ago?
Please save at least one document that identifies EACH DIFFERENT source or type of evidence of human life in the New World 10,000 and more years ago. If one document discusses several cases of such evidence, then you need not save other documents that mention those same types or sources, since your goal is to identify the different types and sources of evidence that have been found and discussed.

Narrative: To be relevant a document would show evidence that humans existed in the New World 10,000 and more years ago. The evidence could range from human hairs, tools, fossils, etc. to bones of animals indicating human presence. Relevant documents would have to specify the evidence and the age suspected, as well as the basis for any theory presented. Information on the Clovis Culture would be relevant. Archaeology digs would probably be related in some way to each relevant document.

Number: 264i

Topic: U.S. Citizens in Foreign Jails

Description: Identify instances where U.S. citizens have been or are being held in foreign jails since the year 1900.
Please save at least one document that identifies EACH DIFFERENT case of a U.S. citizen being held in a foreign jail. If one document discusses several such cases, then you need not save other documents that mention those cases, since your goal is to identify all of the different cases of U.S. citizens being held in foreign jails.

Narrative: With International travel becoming more commonplace, many travelers, whether due to ignorance of the law or with criminal intent, are ending up in foreign jails. Relevant documents will discuss the country involved, the reason for the seizure and jailing, and the sentence, if convicted. U.S. intervention, if any, would also be of interest.

Number: 286i

Topic: Paper Cost

Description: Identify the factors which have led to the cost of paper rising.

Please save at least one document that identifies EACH DIFFERENT factor that has led to increase in the cost of paper. If one document discusses several such factors, then you need not save other documents that mention those same explanations, since your goal is to identify all of the different factors which have been discussed or suggested.

Narrative: In the last year or so the publishing industry is said to have faced a 40% increase in the cost of paper. What factors have led to this price rise? For example, Is there a shortage of materials? Has the cost of processing risen? Have higher taxes been imposed?

7.3.3 Block B3

Number: 254i

Topic: Non-invasive procedures for persons with heart ailments

Description: What kinds of medications or procedures other than heart surgery have been used to treat heart ailments?

Please save at least one document that identifies EACH DIFFERENT medication or procedure which has been used. If one document discusses several such medications or procedures, then you need not save other documents which discuss these individual medications or procedures, since your goal is to identify the different ones that have been used.

Narrative: A relevant document will report/discuss those cases in which persons diagnosed with heart ailments were treated with medications and/or techniques such as angioplasty, stents, lasers, arthrectomy (roto router) etc., in place of surgery. Also advantages of non-invasive procedures over surgery and comparative studies which show any disparity in longevity when either procedure is used.

Number: 284i

Topic: International Drug Enforcement Cooperation

Description: Identify instances where it is shown that international cooperation is taking place in an effort to combat the worldwide drug problem.

Please save at least one document that identifies EACH DIFFERENT instance. If one document discusses several cases of international cooperation, then you need not save other documents which mention those same instances, since your goal is to identify the different cases of international cooperation.

Narrative: Drugs are a critical problem being experienced throughout the world. International Cooperation will be required if this problem is to be alleviated. Examples of relevant documents would be those related to shared information regarding shipments across international borders as well as sharing of information with various prosecutors and other law enforcement personnel.

Number: 292i

Topic: Worldwide Welfare

Description: Identify social programs for poor people in countries other than the U.S.

Please save at least one document that identifies EACH DIFFERENT social program in EACH DIFFERENT COUNTRY. If one document discusses several programs in different countries, then you need not save other documents which mention those same programs, since your goal is to identify the different such programs which exist.

Narrative: To be relevant a document would identify a welfare program in a foreign country and explain how it works to aid citizens who have little or no income. It would include those who can't work because of a disability and people who have the extra burden of small children. The document should indicate how these people are supported or not supported. A relevant document should identify the source of the monies used to support such welfare programs.

7.3.4 Block B4

Number: 255i

Topic: Environmental Protection

Description: Name countries that do not practice or ignore environmental protective measures.

Please save at least one document that identifies EACH DIFFERENT COUNTRY without environmental controls. If one document discusses several such countries, then you need not save other documents that also mention those nations, since your goal is to identify all of the different countries which do not have environmental controls.

Narrative: Nations that do not practice or ignore environmental protective controls degrade the progress other nations have made in this vital area. There are international efforts to protect the environment. The actions of some countries are of some concern, however, because they may be ignoring efforts to conserve and protect the world's resources. The objective of this topic is to identify countries that do not have environmental controls.

Number: 258i

Topic: Computer Security

Description: Identify instances of illegal entry into sensitive computer networks by non-authorized personnel.

Please save at least one document that identifies EACH DIFFERENT case of illegal entry into a computer network. If one document discusses several cases or instances, then you need not save other documents which mention the same ones, since your goal is to identify the different instances of illegal entry.

Narrative: Illegal entry into sensitive computer networks is a serious and potentially menacing problem. Both 'hackers' and foreign agents have been known to acquire unauthorized entry into various networks. Items relevant to this subject would include but not be limited to instances of illegally entering networks containing information of a sensitive nature to specific countries, such as defense or technology information, international banking, etc. Items of a personal nature (e.g. credit card fraud, changing of college test scores) should not be considered relevant.

Number: 299i

Topic: Impact on local economies of military downsizing

Description: What kinds of impact on local economies have been caused by U.S. military downsizing and base closure at U.S. and foreign soil locations?

Please save at least one document that identifies EACH DIFFERENT KIND of impact on local economies. If one document discusses several impacts, then you need not save other documents which discuss the same consequences, since your goal is to identify the different types of impacts due to military base closure or military downsizing.

Narrative: Relevant documents would contain information on the specific type of impact on the local economy of one or more U.S. military base closures.

8 Appendix C - Reporting of results to NIST

1. Sparse format - TWO files from each site

    (a) Search file

One line for EACH SEARCH, each line containing the following blank-delimited items from left to right:

1. Unique site ID

2. Search ID

Participant's choice but ids for searches using PRISE must begin with the string "PRISE."

3. Searcher ID - participant's choice

4. Block number (e.g., "B1")

5. Block sequence number for this searcher ("1" or "2")

6. TREC topic number

7. Elapsed time - number of secs., fractions truncated

Clock time from the moment the searcher sees the topic until the moment the searcher indicates the search is complete.

(b) Documents file

One line for each document in a given search result, each line containing the following blank-delimited items from left to right:

1. Chronological sequence number (e.g., "1", "2") Use number of last time saved if saved multiple times.

2. Search ID (from search file)

3. TREC document identifier (DOCNO)

NOTE: Reported data items listed within each line must NOT contain whitespace.

2. Rich format

   a) ASCII file - no wider than can be printed in landscape mode using 10pt. font on 8.5x11 inch paper.

3. Full narrative description for topic 274: Electric Automobiles

   a) ASCII file - no wider than can be printed in portrait mode using 10pt. font on 8.5x11 inch paper.

4. Description of any further guidance/refinement of the task specification given to the searchers

   a) ASCII file - no wider than can be printed in portrait mode using 10pt. font on 8.5x11 inch paper.

9 Appendix D - Instructions for NIST aspect assessors

You will be presented with a series of topics. Each is a description of an information need - roughly, a "question" to which there may be multiple "answers". For example, if the topic is:

Topic 1

Description: There has been a great deal of effort expended in recent years in the search for the ways to treat high blood pressure. These have included investigations of a variety of means, not all of which involve new drugs.

Answers: each different treatment for high blood pressure available today.

Narrative: Relevant documents will discuss the actual use of various means, traditional or non-traditional, for the treatment of hypertension in humans today.

then answers might be the following:

- use of calcium
- regular exercise
- biofeedback
- the Mediterranean diet
- etc.

You will also be presented with a set of documents, each of which should be relevant to the topic because it discusses one or more of the different possible answers to the question posed by the topic.

You are to do the following separately for each topic:

1. read the topic carefully.
2. read each of the documents for the topic and gradually:

3. create a list of the possible answers found somewhere in the documents. You should select a short phrase to describe each answer.
4. determine which documents contain which answers.

So, if you were given the example topic on treatments for hypertension and a set of 8 documents (call them: D1, D2,... D8,) you might end up with the following:

Topic: 1 - Answers

1. use of calcium
2. regular exercise
3. biofeedback
4. the Mediterranean diat

Answers contained by each document

- D1: 1,2
- D2: 1,3
- D3: 1
- D4: 3
- D5: 1,4
- D6: 2
- D7: 4
- D8: none

In this example we are acting as though Document D1 discusses the use of calcium and regular exercise to treat hypertension but none of the other "answers". Document D4 mentions the use of biofeedback as a treatment. Document D8 contains NO discussion of any answer to the question posed by the topic - this is possible but unlikely.

Feel free to read the documents as many times as needed and to revise your list of answers as well as the list of which answers are provided by which documents.

Do you have any questions about what you are supposed to do ?

10 Appendix E - Summary Evaluation

10.1 City University, London

Summary Results for Each Search

```
 Search id                          NOTE:  See the associated system/site
|        Searcher id                       report for information about what
|        |  Block number                   portion of the interactive task
|        |  |    Block sequence number     was executed to produce these data.
|        |  |    |   Topic
|        |  |    |   |    Elapsed time (secs)
|        |  |    |   |    |       Num docs saved
|        |  |    |   |    |       |   Precision(%)
|        |  |    |   |    |       |   |    Aspectual Recall(%)
|        |  |    |   |    |       |   |    |  Summary vector of aspects covered
|        |  |    |   |    |       |   |    |  | (1=covered; leftmost = #1)
|        |  |    |   |    |       |   |    |  |
```

254_and	a	B3	1	254	1246	7	71	78	111001111
254_kar	b	B3	1	254	1206	14	79	33	110000001
255_bri	c	B4	1	255	1308	4	75	23	10111100100000000000000000
255_col	d	B4	1	255	1207	6	0	4	00000000000000010000000000
256_cla	e	B1	1	256	1291	6	0	14	0000001
256_lis	f	B1	1	256	1213	1	0	14	0000001
258_bri	c	B4	1	258	1302	8	62	33	100011001001101001000000
258_col	d	B4	1	258	835	7	29	17	100001001000000001000000
260_ala	g	B2	1	260	1215	3	67	67	101011
260_sar	h	B2	1	260	1294	2	50	17	000010
264_ala	g	B2	1	264	1269	7	86	29	10010001000000110
264_sar	h	B2	1	264	1278	7	100	24	00010100100000100
274_cla	e	B1	1	274	1249	13	69	91	11111111110
274_lis	f	B1	1	274	1222	13	85	73	11111110100
284_and	a	B3	1	284	1303	6	67	24	1001101000000100000100000
284_kar	b	B3	1	284	1223	9	44	36	1001101010110010001000000
286_ala	g	B2	1	286	1278	7	86	44	110100001
286_sar	h	B2	1	286	1249	14	86	78	111100111
292_and	a	B3	1	292	1253	2	50	6	00110000000000000000000000000000
292_kar	b	B3	1	292	1375	6	33	16	00000001000100001100100000000000
293_cla	e	B1	1	293	1216	7	14	17	100000
293_lis	f	B1	1	293	1236	4	0	0	000000
299_bri	c	B4	1	299	1040	7	29	47	001100011100101
299_col	d	B4	1	299	1129	3	0	0	000000000000000

10.2 Rutgers University

```
Summary Results for Each Search

Search id                             NOTE:   See the associated system/site
|           Searcher id                       report for information about what
|           |  Block number                   portion of the interactive task
|           |  |    Block sequence number     was executed to produce these data.
|           |  |    |   Topic
|           |  |    |   |    Elapsed time (secs)
|           |  |    |   |    |       Num docs saved
|           |  |    |   |    |       |  Precision(%)
|           |  |    |   |    |       |  |     Aspectual Recall(%)
|           |  |    |   |    |       |  |     |  Summary vector of aspects covered
|           |  |    |   |    |       |  |     |  | (1=covered; leftmost = #1)
|           |  |    |   |    |       |  |     |  |
254i-002-2  b  B3   2   254  1103    7  100   33 111000000
254i-003-1  c  B3   1   254  1095   12   75   78 111110101
254i-005-2  e  B3   2   254  1260   13   15   44 101001100
254i-007-1  g  B3   1   254  1075    3   67   33 101000001
254i-011-1  k  B3   1   254  1108    9   78   44 101010001
254i-012-2  l  B3   2   254  1246    7   71   44 111000001
254i-014-2  n  B3   2   254  1043    7   71   44 101010100
255i-003-2  c  B4   2   255  1141   16   38   81 11011111011111111111111000
255i-004-1  d  B4   1   255  1280    1  100   15 00111100000000000000000000
255i-006-2  f  B4   2   255  1214    5   20   19 10000000000000000100000111
255i-008-1  h  B4   1   255  1242    2  100   23 11111100000000000000000000
255i-009-2  i  B4   2   255  1330    2  100   19 00111100000010000000000000
255i-012-1  l  B4   1   255  1289    6   50   15 11010000000010000000000000
256i-001-1  a  B1   1   256  1315    5   20   29 1000001
256i-004-2  d  B1   2   256  1222    5    0   14 0000100
256i-005-1  e  B1   1   256  1132    3    0    0 0000000
256i-007-2  g  B1   2   256  1013   10   40   71 1111010
256i-009-1  i  B1   1   256  1163    2   50   29 1001000
256i-010-2  j  B1   2   256   617   11    0   14 1000000
256i-013-1  m  B1   1   256  1092    2   50   43 1011000
258i-003-2  c  B4   2   258  1182   28   64   79 111111111111111101011000
258i-004-1  d  B4   1   258  1138    7   29   38 111011110000100000100000
258i-006-2  f  B4   2   258  1225    7   43   38 111011110000100100000000
258i-008-1  h  B4   1   258  1094   17   53   67 111101101001100111001111
258i-009-2  i  B4   2   258  1231    5   80   54 110111100000110010001111
258i-012-1  l  B4   1   258  1214   10   90   67 111111110001110010001111
260i-001-2  a  B2   2   260  1149    7   57  100 111111
260i-002-1  b  B2   1   260  1061    2  100   67 101011
260i-006-1  f  B2   1   260  1333    1  100   50 001011
```

260i-008-2	h	B2	2	260	1211	8	62	83	111011
260i-010-1	j	B2	1	260	1195	12	17	50	001011
260i-011-2	k	B2	2	260	1178	2	50	50	001011
260i-013-2	m	B2	2	260	751	3	33	33	100010
260i-014-1	n	B2	1	260	1304	0	0	0	
264i-001-2	a	B2	2	264	1147	9	78	47	10111101100100000
264i-002-1	b	B2	1	264	1131	6	83	35	10010101100010000
264i-006-1	f	B2	1	264	1217	3	67	6	10000000000000000
264i-008-2	h	B2	2	264	1182	11	100	29	10010101000000010
264i-010-1	j	B2	1	264	1344	15	87	47	00110101101010100
264i-011-2	k	B2	2	264	1127	13	100	47	01010101101010100
264i-013-2	m	B2	2	264	830	3	100	18	00010100100000000
264i-014-1	n	B2	1	264	1298	15	100	71	01110111110011101
274i-001-1	a	B1	1	274	1095	10	100	73	11111110100
274i-004-2	d	B1	2	274	1047	8	100	64	11111100100
274i-005-1	e	B1	1	274	1385	6	83	73	01111110110
274i-007-2	g	B1	2	274	712	14	100	73	11111110100
274i-009-1	i	B1	1	274	1247	5	100	73	11111110100
274i-010-2	j	B1	2	274	1243	36	94	100	11111111111
274i-013-1	m	B1	1	274	1191	6	100	64	11111100010
284i-002-2	b	B3	2	284	1157	15	60	48	1100011000010110001111001
284i-003-1	c	B3	1	284	1188	22	50	52	1101101001010110001110100
284i-005-2	e	B3	2	284	1157	13	69	44	1101111000000100011101000
284i-007-1	g	B3	1	284	1015	9	44	24	1000100000000110010010000
284i-011-1	k	B3	1	284	1018	17	41	44	1101101000011101011000000
284i-012-2	l	B3	2	284	1201	18	67	56	0111101110010111101000010
284i-014-2	n	B3	2	284	818	10	80	36	0001101000010111011000000
286i-001-2	a	B2	2	286	1152	28	64	89	110111111
286i-002-1	b	B2	1	286	939	6	100	44	110100001
286i-006-1	f	B2	1	286	1226	4	50	44	011110000
286i-008-2	h	B2	2	286	1121	14	7	44	110110000
286i-010-1	j	B2	1	286	1013	9	44	56	111100010
286i-011-2	k	B2	2	286	1251	11	82	56	110110001
286i-013-2	m	B2	2	286	590	5	60	44	110101000
286i-014-1	n	B2	1	286	1087	15	73	56	110110001
292i-002-2	b	B3	2	292	1110	4	50	19	0000000100011000000000100001100000
292i-003-1	c	B3	1	292	1067	19	21	56	10111110110111000000101111001100
292i-005-2	e	B3	2	292	1209	22	36	41	00010001001100110011010000111001
292i-007-1	g	B3	1	292	1254	16	31	47	00010001001100111111110000111000
292i-011-1	k	B3	1	292	1213	19	47	47	00010001001100110111110000111010
292i-012-2	l	B3	2	292	1212	22	32	44	10010001001100110011110000111000
292i-014-2	n	B3	2	292	945	11	45	38	00010001001100110011010000111000
293i-001-1	a	B1	1	293	1206	5	0	0	000000
293i-004-2	d	B1	2	293	1357	0	0	0	
293i-005-1	e	B1	1	293	1261	11	45	100	111111
293i-007-2	g	B1	2	293	1337	1	100	17	000100

293i-009-1	i	B1	1	293	1136	5	20	17	000100
293i-010-2	j	B1	2	293	964	6	0	0	000000
293i-013-1	m	B1	1	293	1088	2	50	17	100000
299i-003-2	c	B4	2	299	1195	18	39	73	111111111100010
299i-004-1	d	B4	1	299	1251	3	33	47	101111101000000
299i-006-2	f	B4	2	299	1337	6	0	7	000000000000100
299i-008-1	h	B4	1	299	1140	9	44	87	101111111111011
299i-009-2	i	B4	2	299	1236	2	50	40	101111100000000
299i-012-1	l	B4	1	299	1157	7	57	60	111111110000010

11 Appendix F

11.1 Comments of the relevance assessor on topic 284

International Drug Enforcement Cooperation

Scan terms: World-wide problem international cooperation law enforcement personnel prosecutors borders shipments

Comments: The drug culture has had a profound impact upon persons, families, the country, education and the world economy.

International cooperation regarding a coherent policy to confront, interdict and neutralize drug traffickers has been hampered by a combination of political rivalries, jealousies, official corruption and inefficiency. In contrast, organized international criminals have become sophisticated and adept at extending its drug operations across countries and continents and of manipulating the world economic system.

The United States is the major country targeted by the drug traffickers because of the demand and personal financial resources. Other countries, however, particularly in Europe, are now feeling the effects of the drugs within their cultures. If the supply and availability of drugs is to be curtailed or limited, it will require the cooperation of all countries – growers, manufacturers and users alike. Only through a concerted international effort can drugs be identified, interdicted and destroyed.

Those items considered to be relevant were those in which there was cooperation, collaboration and collusion between the U.S. and another country(ies) to locate, interdict and destroy international shipments of drugs and incarcerate those persons involved in the trafficking.

Items included: cooperation among countries to combat international drug trafficking; UN and drug companies exchanging information on illegal drug marketing; agreement among Thailand, Burma, Malaysia, Pakistan and China to work with the United States in cutting opium production; Hong Kong asking for and receiving assistance from the DEA (in the form of helicopters), to identify, locate and dismantle labs making cocaine; U.S. giving speed boats to Santo Domingo to interdict the flow of drugs; Bolivia and Brazil agreeing to join forces to fight the drug cartel; and five countries and two continents cooperating to crack a drug smuggling ring in Milan, Italy.

Not considered relevant were those items which reported on internal actions to interdict, subvert, destroy or apprehend drug dealings within a country, regardless of the origin of the drugs or the nationality of the traffickers.

11.2 Comments of the aspect assessor on topic 284

Criteria employed in identifying an article included statements regarding: - international meetings at which plans were discussed to stop drug trafficking - information regarding drug shipments - cooperation between two or more countries to interdict drug shipments - training of police/agents in foreign countries to target drug maintenance - providing (by U.S.) of vehicles, helicopters, materials, herbicides to destroy crops, warehouses storing the drugs

Specific actions by an individual country to interdict a drug shipment, apprehend drug smugglers within the confines of a country: i.e. unilateral action, was not considered relevant

11.3 Author's review of document judgements for topic 284

11.3.1 AP880213-0180 Relevance error?

"While the two leaders (Reagan and de la Madrid) met, senior officials from both countries held talks of their own, focussing on foreign policy, trade, drugs and law enforcement."

11.3.2 AP880222-0072 Relevance error?

The document mentions meeting between Reagan and de la Madrid and cites statement by de la Madrid about the need to attach production, distribution, and consumption - no specific mention of cooperation but the two leaders were meeting and considering the issue so it seems relevant.

11.3.3 AP880401-0280 Relevance error?

"We're talking the same language," said white House official Donald Ian MacDonald of increasing U.S. - Soviet cooperation on drugs. ... regional organizations in several parts of the world have been formed to tackle the problem.

Two clear references to international cooperation.

11.3.4 AP880406-0185 Relevance error?

The document describes cooperation between Honduras, the Dominican Republic, and the United States in the arrest of a drug dealer.

11.3.5 AP880409-0127 Relevance error?

The document U.S. Attorney General's tour of Latin America to assess anti-drug trafficking efforts and in particular meetings with Peruvian officials involved in the anti-drug fight.

11.3.6 AP880411-0168 Relevance error?

The document quotes Edwin Meese as saying: "Peru has been one of the leaders in the use of herbicides for eradicating coca plants. We have been cooperating with Peru in these experimental studies."

11.3.7 AP880415-0060 Relevance error?

The document quotes Mexico's federal attorney general as saying: "Mexico is undertaking a plain effort of international cooperation in the campaign against drug trafficking." There are also references to financial assistance from the United States to Mexico for anti-drug efforts.

11.3.8 AP880415-0165 Relevance error?

"It [the Mexican government] also cooperates with the U.S. Drug Enforcement Administration, which maintains an office and agents here [Mexico]."

11.3.9 AP880519-0173 Relevance error?

The document reports on the arrest and conviction of Carlos Lehder Rivas, a major Columbian drug dealer. Attorney General Meese quoted calling the case "an excellent example of international cooperation."

11.3.10 AP880519-0201 Relevance error?

Essentially the same information as AP880519-0173

11.3.11 AP880520-0105 Relevance error?

Essentially the same information as AP880519-0173

11.3.12 AP880617-0255 Relevance error?

The document reports statements about goals and expectations of discussions at an economic summit about drug trafficking, drug-money laundering, "a general discussion of the whole drug problem..."

11.3.13 AP880811-0234 Relevance error?

"Soviet customs agents helped with the best tip of the year. They told the Mounties that a drug ring planned to ship 11,277 pounds of hashish from Kabul, Afghanistan, through Leningrad to Montreal. The hash was seized on arrival and three Canadians are now behind bars."

11.3.14 AP880927-0103 Relevance error?

"Treaties with six countries calling for cooperation in narcotics prosecutions and other criminal matters were approved by the Senate Foreign Relations Committee on Tuesday...."

11.3.15 AP881011-0041 Relevance error?

The document reports on the Inter-regional Narcotics Eradication Air Wing. "[U.S.] civilian-piloted gunships will join Peruvian narcotics police in raids to destroy cocaine laboratories, warehouses and airstrips."

"The State Department began its international anti-drug campaigns in 1978 when Congress offered Mexico four helicopters to spray marijuana and opium poppies with the herbicide paraquat. The State Department now maintains about 120 aircraft involved in drug efforts in Mexico, Columbia and Burma."

11.3.16 AP881112-0073 Relevance error?

The document mentions the existence of U.S. State Department anti-drug programs in Bolivia and Columbia without spelling out the details. It also mentions a $10 million [U.S.] narcotics program program in Columbia.

11.3.17 FT911-2671 Relevance error?

The document describes three international agencies to be merged into one U.N. organization and says these agencies "implement and monitor treaties on legal drug production and run intervention programmes to try to change the economic and cultural bases of Third World countries producing opium and coca leaves."

11.3.18 FT921-6988 Relevance error?

The document talks about a U.S - Latin America drug summit and mentions U.S. aid and military presence in drug-producing states.

11.3.19 FT922-9654 Relevance error?

The document describes a "debt-for-drugs" scheme being investigated by the U.N. pursuant to a motion passed by the U.N. Commision on Narcotics Drugs. According to the document, most of the nations which would be asked to forgive debt in exchange for drug-fighting efforts Third World countries say they can't otherwise afford, are opposed to the idea. So this may not be any cooperation here.

11.3.20 FT924-10092 Relevance error?

The document mentions a U.N. pilot project in Laos which required a change in the U.N.'s approach to include "planned, co-ordinated management of U.N. agencies, funding bodies and international aid from developed countries..."

11.3.21 FT934-10491 Relevance error?

"The new [U.S.] efforts, say officials, would build on improved co-operation already evident, particularly with Bolivia and Columbia in the Andes."

11.3.22 FT942-10977 Relevance error?

The document refers to "Umopar, a para-military jungle unit trained by the DEA, the U.S. Drug Enforcement Agency, to eliminate the Bolivian drug trade."

11.3.23 WSJ910529-0062 Relevance error?

"The Bush administration pursues an integrated policy that attacks illegal drugs and money laundering on the international front. Its anti- money-laundering strategy seeks to establish a network of countries joined together .... These programs will facilitate international cooperation...."

11.3.24 WSJ920113-0038 Relevance error?

The document mentions Hong Kong's participation in the Financial Action Task Force, "a 26 nation group.... The FATF last year issued recommendations aimed at destabilizing trafficking organizations ..."

11.3.25 WSJ920204-0078 Relevance error?

The documents cites several instances in which Panama under General Noriega assisted the United States in the war on drugs.

11.3.26 WSJ920214-0116 Relevance error?

"President Fujimori's commitment. U.S. support for the decriminalization of coca farmers, reports of President Bush's personal backing, and similar support from some European governments and the United Nations triggered ... the drafting of initial plans for private investment" in Peru to allow coca farmers to switch to other crops.

11.4 Comments of the relevance assessor on topic 292

> Many documents mentioned the fact that a country has a social program for the poor and unemployed but gave no insight as to how the program worked. These were not considered relevant. Some non-relevant documents supplied information on how poor a country was but did not indicate that they had any welfare system. Other documents identified the source of income for their welfare program but did not include sufficient information on how this money supported the recipients. Generally when I found information on a country including how much funding was spent for a specific type of recipient, I made it relevant. Some documents included enough information on its unemployment problem to make it relevant. If a document stated that a welfare recipient lost their benefits because they attained a specific amount of savings or income, I made it relevant. A document was relevant when it stated how a country supported its welfare program, i.e. taxes. While FT data appeared to be the best source for relevant documents, the CR and FR data had none.

11.5 Comments of the aspect assessor on topic 292

1. If the document treats government programs (pensions, health, housing, et al) as a unit, I listed them as ONE answer.
2. Of the first 11 documents 7 were on the U.S.; therefore did not provide answers since the topic defined as "in countries other than the U.S." (I checked to make certain there was not incidental or comparative reference to other countries.
3. Ref 1 above: Since several docs deal with reductions in funding of government social programs, they may mention specific programs but by lack of description they tend to lump the programs all together. In these cases I have listed a single "answer", e.g. "South Africa - government programs"
4. Doc 18 (FT931-10916) is essentially a think piece with only fleeting references to the British welfare system and how it does (or doesn't) work.
5. Doc 34 (FT934-12800) Demographic data on countries was not considered an "answer" unless it was related directly to a welfare benefit in a particular country.
6. Doc 62 (FT944-6725) The article does not really say that the Social Action Program has accomplished anything, but refers to it as "launched", "embraced", and "extended". I counted it as an answer.