TITLE=readme-j.pdf

DATE=2001-03-01

# 情報検索システム評価用テストコレクション 2 (NTCIR-2) README

## 1. ファイル構成

**この CD-ROM には、以下のファイルがあります。**

| | |
|---|---|
| readme-j.pdf | **このファイル** |
| readme-j.txt | **このファイルのテキストファイル版（EUC）** |
| readme-e.txt | **このファイルの英語版** |
| agreem2-j.pdf | **使用許諾に関する覚書（日本語）** |
| agreem2-e.pdf | **使用許諾に関する覚書（英語）** |
| manual-j.pdf | **使用説明書（日本語）** |
| manual-e.pdf | **使用説明書（英語）** |
| j-docs.tgz | **J コレクション（日本語文書）** |
| e-docs.tgz | **E コレクション（英語文書）** |
| topics.tgz | **検索課題（日本語、英語）** |
| rels.tgz | **日本語文書および英語文書に対する正解判定** |
| scripts.tgz | **文書番号（ACCN）の変換スクリプト** |

*.pdf **は、**Adobe AcrobatReader **が必要です。**

*.tgz **は、**tar **をして、**gzip **してあります。**gzip -dc <**ファイル名**> ｜ tar xvf - **として復元してください。**

*.tgz **の内容と、復元したときのファイルサイズは下記のとおりです。なお、ファイルサイズの** 1MB＝$1024^2$byte **です。**

(1) j-docs.tgz　　**Jコレクション（日本語文書）**

**復元すると** j-docs/**の下に以下のファイルがあります。**

| | |
|---|---|
| ntc2-j1g (118.2MB) | **文書セット（J コレクション） 日本語「学会発表データベース」から抽出** |
| ntc2-j1k (481.3MB) | **文書セット（J コレクション） 日本語「科学研究費補助金研究成果概要データベース」から抽出** |

(2) e-docs.tgz　　Eコレクション（英語文書）

復元すると e-docs/の下に以下のファイルがあります。

| | |
|---|---|
| ntc2-e1g (89.7MB) | 文書セット（E コレクション） 英語「学会発表データベース」から抽出 |
| ntc2-e1k (110.7MB) | 文書セット（E コレクション）英語「科学研究費補助金研究成果概要データベース」から抽出 |

(3) topics.tgz　　検索課題 （日本語、英語）

復元すると topics/の下に以下のファイルがあります。いずれも第２回 NTCIR ワークショップの評価用検索課題です。

| | |
|---|---|
| topic-j0101-0149 | 日本語検索課題 0101～0149。 |
| topic-e0101-0149 | 英語検索課題 0101～0149。 |

(4) rels.tgz　　正解判定

復元すると rels/の下に以下のファイルがあります。

・J コレクション（日本語文書）に対する正解判定

| | |
|---|---|
| rel1_ntc2-j2_0101-0149 | ntc1-j1.mod, ntc2-j1g, ntc2-j1k に対する検索課題 0101～0149 の正解判定（Level 1。正解ファイル。S 判定と A 判定を正解とした） |
| rel2_ntc2-j2_0101-0149 | ntc1-j1.mod, ntc2-j1g, ntc2-j1k に対する検索課題 0101～0149 の正解判定（Level 2。部分正解ファイル。S、 A、B 判定を正解とした） |
| rel*_ntc2-j2_0101-0149.nc | コメント部分を除いた正解判定（'*'は'1'または'2'） |

・E コレクション（英語文書）に対する正解判定

| | |
|---|---|
| rel1_ntc2-e2_0101-0149 | ntc1-e1.mod, ntc2-e1g, ntc2-e1k に対する検索課題 0101～0149 の正解判定（Level 1。正解ファイル。S 判定と A 判定を正解とした） |
| rel2_ntc2-e2_0101-0149 | ntc1-e1.mod, ntc2-e1g と ntc2-e1k に対する検索課題 0101～0149 の正解判定（ Level 2。部分正解ファイル。S、A、B 判定を正解とした） |
| rel*_ntc2-e2_0101-0149.nc | コメント部分を除いた正解判定（'*'は'1'または'2'） |

・J コレクションおよび E コレクションに対する正解判定

| | |
|---|---|
| rel1_ntc2-je2_0101-0149 | ntc1-j1.mod, ntc1-e1.mod, ntc2-j1g, ntc2-e1g, ntc2-j1k, ntc2-e1k に対する検索課題 0101～0149 の正解判定（Level 1。正解ファイル。S 判定と A 判定を正解とした） |
| rel2_ntc2-je2_0101-0149 | ntc1-j1,mod, ntc1-e1.mod, ntc2-j1g, ntc2-e1g, ntc2-j1k, ntc2-e1k に対する検索課題 0101～0149 の正解判定（Level 2。部分正解ファイル。S、A、B 判定を正解 |

とした）

rel*_ntc2-je2_0101-0149.nc　コメント部分を除いた正解判定（'*'は'1'または'2'）

（5）scripts.tgz　　文書番号（ACCN）の変換スクリプト

復元すると scripts/の下に以下のファイルとディレクトリがあります。

| | |
|---|---|
| readme-script-j.txt | 英語文書の ACCN 変換スクリプトについての説明の日本語版（EUC） |
| readme-script-e.txt | 英語文書の ACCN 変換スクリプトについての説明の英語版 |
| accn-tr.tar | 英語文書の ACCN 変換スクリプトをまとめたファイル |
| ntc1accn | NTCIR-1 の文書データの ACCN を NTCIR-2 の文書データと同じ形式に変換するスクリプト ACCN-j.pl と ACCN-e.pl、および、その説明のファイルのあるディレクトリ（詳細はディレクトリ下の説明ファイルを参照して下さい） |

## 2.データの形式と使用法

・テキストファイルの文字コードは EUC です。
・各ファイルの形式、使用法については、使用説明書（manual-e.pdf、manual-j.pdf）を参照してください。
・タスク、文書、検索課題番号によって、対応する正解判定ファイルが異なります。組み合わせを間違えないようにご留意ください。詳しくは図 1、および使用説明書の 5.2 節と図 5-2 を参照してください。

| タスク | 文書ファイル※1※2 | 検索課題（検索課題数） | 正解判定※3 |
|---|---|---|---|
| **単言語検索タスク** | | | |
| J-J タスク | j-docs/ntc2-j1*<br>ntc1-j1.mod | topics/topic-j0101-0149 (49) | rels/rel*_ntc2-j2_0101-0149 |
| E-E タスク | e-docs/ntc2-e1*<br>ntc1-e1.mod | topics/topic-e0101-0149 (49) | rels/rel*_ntc2-e2_0101-0149 |
| **言語横断検索タスク** | | | |
| J-E タスク | j-docs/ntc2-e1*<br>ntc1-e1.mod | topics/topic-j0101-0149 (49) | rels/rel*_ntc2-e2_0101-0149 |
| E-J タスク | e-docs/ntc2-j1*<br>ntc1-j1.mod | topics/topic-e0101-0149 (49) | rels/rel*_ntc2-j2_0101-0149 |

| | | | |
|---|---|---|---|
| J-J,E タスク | j-docs/ntc2-j1* | topics/topic-j0101-0149 (49) | rels/rel*_ntc2-je2_0101-0149 |
| | e-docs/ntc2-e1* | | |
| | ntc1-j1.mod | | |
| | ntc1-e1.mod | | |
| E-J,E タスク | j-docs/ntc2-j1* | topics/topic-e0101-0149 (49) | rels/rel*_ntc2-je2_0101-0149 |
| | e-docs/ntc2-e1* | | |
| | ntc1-j1.mod | | |
| | ntc1-e1.mod | | |

※1: 文書ファイルの ntc2-j1*および ntc2-e1*の'*'は、'g'および'k'です。'g'は「学会発表データベース」から抽出した文書、'k'は「科学研究費補助金研究成果概要データベース」から抽出した文書です。

※2: ntc1-*1.mod はNTCIR-1 の文書データ ntc1-*1 の ACCN を変換スクリプト ACCN-*.pl で変換したデータです。（ntc1-*, ntc1-*.mod, ACCN-*.pl の'*'は、'j'または'e'です。）

※3: 正解判定ファイルの rel*の'*'は、'1'または'2'です。'rel1_'で始まるファイル名は、正解レベル「Level 1」（S 判定と A 判定のみを適合とした）の「正解ファイル」、'rel2_'で始まるファイル名は、「Level 2」）S 判定、A 判定、および B 判定を適合とした）の「部分正解ファイル」です。

**図 1 検索タスク、文書、検索課題、正解判定の対応関係（ファイル名は、direcotory/filename）**

・テストコレクション２（NTCIR-2）の利用は、テストレコレクション２の使用許諾に関する覚え書きの範囲でのみ可能です。

## 3.文書についてのご注意

このコレクションの元になった「学会発表データベース」と「科学研究費補助金研究成果概要データベース」のレコードは、集められたまま、編集者や抄録作成者による編集や修正をしないで、使用しています。研究者によって書かれた著者抄録または研究成果概要を使用しており、抄録作成の専門家によって作成された抄録とは異なる内容構成のものもあります。また、データは可能な限りオリジナルに近い形を保つという基本方針のため、また、事実上、すべてのデータを手作業でチェックするのは不可能でもあるため、文書データには、「エラー」が含まれていることをご了承ください。「エラー」の中には、元のデータに含まれていたもの、入力作業時に発生したもの、国立情報学研究所でフォーマットを整える際に生じたもの、テストコレクション用にデータを抽出する際に生じたものなどが含まれている可能性があります。NTCIR プロジェクト事務局でのエラーのチェックは、内容の修正ではなく、開始タグと終了タグの対応、ACCN、タイトルなどの必須項目が含まれているかなどの形式面に重点をおいています。

また、NTCIR-2 中の文書データは、情報検索や関連研究の研究目的使用のために、「学会発表データベース」および「科学研究費補助金研究成果概要データベース」からその一部を抽出したものであり、網羅性に欠けるため、情報を得るという目的で使用することはでき

ません。

NTCIR-2 を使用して生じたいかなる損失にも、NTCIR プロジェクト事務局および国立情報学研究所は責任を負いません。あらかじめご了承ください。

## 4.問い合わせ先

NTCIR-2 に関するお問い合わせは、下記にお願いいたします。

(1)CD-ROM の入手方法や CD-ROM 利用者の発表論文の書誌事項の連絡など事務手続に関して

国立情報学研究所 NTCIR プロジェクト事務局
Email: ntc-secretariat@nii.ac.jp
〒101-8430 東京都千代田区一ツ橋 2-1-2
Phone 03-4212-2750（直通）
Fax 03-3556-1916

(2)CD-ROM に含まれるデータの形式や利用方法など技術的なことに関して

国立情報学研究所 NTCIR プロジェクトグループ
Email: ntcadm@nii.ac.jp
〒101-8430 東京都千代田区一ツ橋 2-1-2
Phone 03-4212-2529（直通）
Fax 03-3556-1916
担当：神門典子