TITLE=README-j.pdf

DATE=1999-11-01

情報検索システム評価用テストコレクション１(NTCIR-1) 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
agreem-j.pdf	使用許諾に関する覚書(日本語)
agreem-e.pdf	使用許諾に関する覚書(英語)
tagree-j.pdf	タグ付きデータコレクションの使用許諾に関する覚書(日本語)
tagree-e.pdf	タグ付きデータコレクションの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)
adhoc.tgz	随時検索用の文書(日英混在)と正解判定
mlir.tgz	日本語単言語検索用の文書と正解判定
clir.tgz	言語横断検索用の文書（英語文書）と正解判定
topics.tgz	検索課題 (日本語)
tmrec.tgz	用語抽出研究用のタグ付きデータコレクション

*.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＝1024^2byte です。

(1) adhoc.tgz　　随時検索用の文書(日英混在)と正解判定

復元すると　adhoc/の下に以下のファイルがあります。

ntc1-je1 (576.6MB)	文書セット（JE コレクション）　日本語英語混在
rel1_ntc1-je1_0001-0030	ntc1-je1 に対する検索課題 0001～0030 の正解判定 (正解ファイル。A 判定のみを正解とした)
rel2_ntc1-je1_0001-0030	ntc1-je1 に対する検索課題 0001～0030 の正解判定 （部分正解ファイル。A 判定と B 判定を正解とした）
rel1_ntc1-je1_0031-0083	ntc1-je1 に対する検索課題 0031～0083 の正解判定 (正解ファイル。A 判定のみを正解とした)
rel2_ntc1-je1_0031-0083	ntc1-je1 に対する検索課題 0031～0083 の正解判定 （部分正解ファイル。A 判定と B 判定を正解とした）

(2) mlir.tgz　　日本語単言語検索用の文書(日本語)と正解判定

復元すると mlir/の下に以下のファイルがあります。

ntc1-j1 (311.5MB)	文書セット（J コレクション） 日本語
rel1_ntc1-j1_0001-0030	ntc1-j1 に対する検索課題 0001～0030 の正解判定 (正解ファイル。A 判定のみを正解とした)
rel2_ntc1-j1_0001-0030	ntc1-j1 に対する検索課題 0001～0030 の正解判定 （部分正解ファイル。A 判定と B 判定を正解とした）
rel1_ntc1-j1_0031-0083	ntc1-j1 に対する検索課題 0031～0083 の正解判定 (正解ファイル。A 判定のみを正解とした)
rel2_ntc1-j1_0031-0083	ntc1-j1 に対する検索課題 0031～0083 の正解判定 （部分正解ファイル。A 判定と B 判定を正解とした）

(3) clir.tgz　　言語横断検索用の文書（英語）と正解判定

復元すると clir/の下に以下のファイルがあります。

ntc1-e1 (217.5MB)	文書セット（E コレクション） 日本語英語混在
rel1_ntc1-e1_0001-0030	ntc1-e1 に対する検索課題 0001～0030 の正解判定 (正解ファイル。A 判定のみを正解とした)
rel2_ntc1-e1_0001-0030	ntc1-e1 に対する検索課題 0001～0030 の正解判定 （部分正解ファイル。A 判定と B 判定を正解とした）
rel1_ntc1-e1_0031-0083	ntc1-e1 に対する検索課題 0031～0083 の正解判定 (正解ファイル。A 判定のみを正解とした)
rel2_ntc1-e1_0031-0083	ntc1-e1 に対する検索課題 0031～0083 の正解判定 （部分正解ファイル。A 判定と B 判定を正解とした）

(4) topics.tgz　　検索課題 (日本語)

復元すると topics/の下に以下のファイルがあります。

topic0001-0030	検索課題 0001～0030。第 1 回 NTCIR ワークショップの訓練用
topic0031-0083	検索課題 0031～0083。第 1 回 NTCIR ワークショップの評価用

(5) tmrec.tgz　　用語抽出研究用のタグ付きデータコレクション

復元すると tmrec/の下に以下のファイルがあります。

README.j	このディレクトリ中のファイルの説明(日本語)
README	このディレクトリ中のファイルの説明(英語)
README.termtagj	用語候補の選択とタグ付けに関する説明書(日本語)
README.termtage	用語候補の選択とタグ付けに関する説明書(英語)
ntc1-tt0	言語タグ付きデータ
ntc1-tu0	言語タグなしデータ
ntc1-ttg	プレインテキストデータに、用語候補のタグを加えたもの
ntc1-tml	ntc1-ttg から、用語候補を抜き出し多少の正規化を加えたもの

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。詳しくは使用説明書の 5.2 節と図 5-2 を参照してください。
・テストコレクション 1 (NTCIR-1)の利用は、テストレコレクション 1 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」は、速報性を重視したデータベースで、 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。著者抄録を使用しており、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、学術情報センター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 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。なお、用語抽出研究等に使用するタグ付きデータコレクション ntc1-tt0 およびタグなしデータコレクション ntc1-tu0 の日本語部分については手作業でデータを修正しましたが、英語部分に付きましては時間的な問題および日本語をタスクの対象としたため、エラーの修正は形式的なチェックにとどめています。

また、NTCIR-1 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」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-1 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学術情報センター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-1 に関するお問い合わせは、下記にお願いいたします。

学術情報センター研究開発部 NTCIR プロジェクト
Email: ntcadm@rd.nacsis.ac.jp
〒112-8640 東京都文京区大塚 3-29-1
Phone 03-3942-6969（直通）
Fax 03-5395-7064
担当： 神門 典子（かんど のりこ）