Reuters-21578(路透社文档)

数据摘要：

This is a very often used test set for text categorisation tasks.

中文关键词：

数据挖掘,路透社,文本归类,文本分类,

英文关键词：

Data mining,Reuters,Text categorization,Text Classification,

数据格式：

TEXT

数据用途：

The data can be used to data mining and analysis.

数据详细介绍：

The Reuters-21578 text dataset

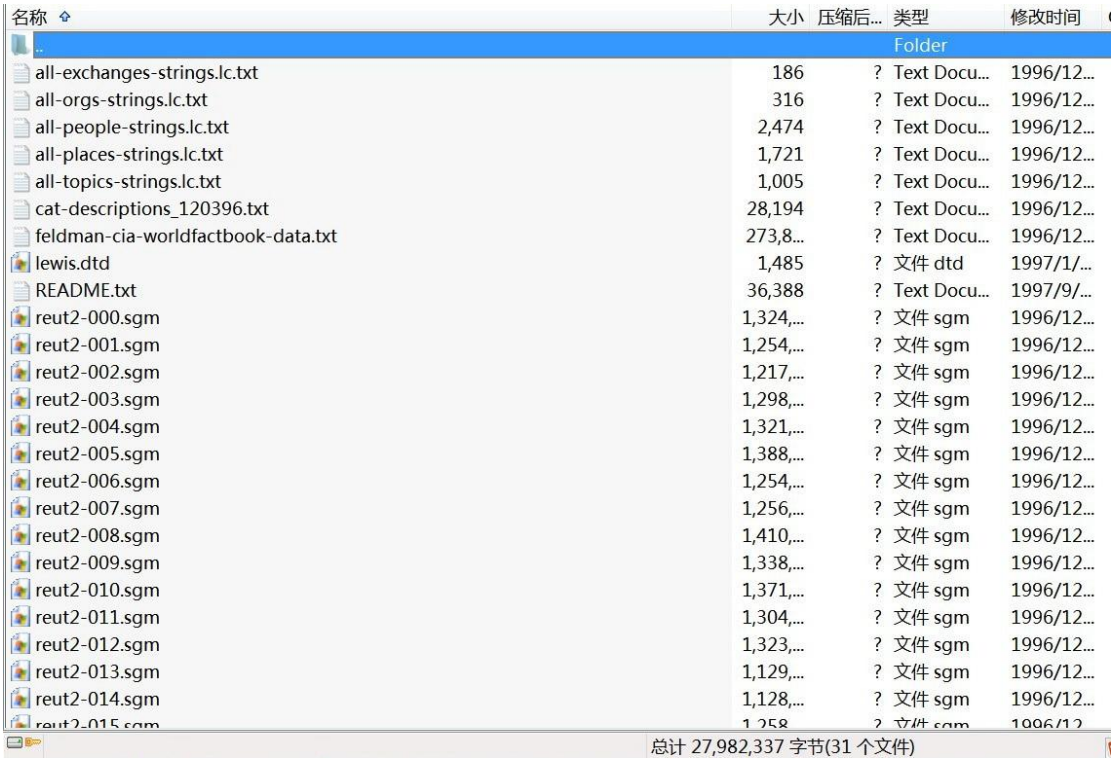
- This is a very often used test set for text categorisation tasks. It contains 21578 Reuters news documents from 1987. They were labeled manually by Reuters personnel. Labels belong to 5 different category classes, such as 'people', 'places' and 'topics'. The total number of categories is 672, but many of them occur only very rarely. Some documents belong to many

different categories, others to only one, and some have no category. Over the past decade, there have been many efforts to clean the database up, and improve it for use in scientific research. The present format is divided in 22 files of 1000 documents delimited by SGML tags (here is as an example one of these files). Extensive information on the structure and the contents of the dataset can be found in the README file. In the past, this dataset has been split up into training and test data in many different ways. You should use the 'Modified Apte' split as described in the README file.

- **Size**:
  - 21578 documents; according to the 'ModApte' split: 9603 training docs, 3299 test docs and 8676 unused docs.
  - 27 MB
- **References**: This is a popular dataset for text mining experiments. The aim is usually to predict to which categories of the 'topics' category class a text belongs. Different splits into training ,test and unused data have been considered. Previous use of the Reuters dataset includes:
  - *Towards Language Independent Automated Learning of Text Categorization Models (1994)* by C. Apte, F. Damerau and S. M. Weiss: This paper tests a rule induction method on the Reuters data. This is where the 'Apte' split of the data was introduced.
  - *An Evaluation of Statistical Approaches to Text Categorization (1997)* by Y. Yang: This paper contains a comparison of 14 different classification methods on 6 different datasets (or at least 6 different splits over 2 datasets).
  - *Inductive learning algorithms and representations for text categorization (1998)* by S. T. Dumais, J. Platt, D. Heckerman and M. Sahami: 5 different learning algorithms for text categorisation are compared. The dataset they use is the 'Modified Apte' split which you will also use.

Carnegie Group, Inc. and Reuters, Ltd.

数据预览：



名称	大小	压缩后...	类型	修改时间
..			Folder	
all-exchanges-strings.lc.txt	186	?	Text Docu...	1996/12...
all-orgs-strings.lc.txt	316	?	Text Docu...	1996/12...
all-people-strings.lc.txt	2,474	?	Text Docu...	1996/12...
all-places-strings.lc.txt	1,721	?	Text Docu...	1996/12...
all-topics-strings.lc.txt	1,005	?	Text Docu...	1996/12...
cat-descriptions_120396.txt	28,194	?	Text Docu...	1996/12...
feldman-cia-worldfactbook-data.txt	273,8...	?	Text Docu...	1996/12...
lewis.dtd	1,485	?	文件 dtd	1997/1/...
README.txt	36,388	?	Text Docu...	1997/9/...
reut2-000.sgm	1,324,...	?	文件 sgm	1996/12...
reut2-001.sgm	1,254,...	?	文件 sgm	1996/12...
reut2-002.sgm	1,217,...	?	文件 sgm	1996/12...
reut2-003.sgm	1,298,...	?	文件 sgm	1996/12...
reut2-004.sgm	1,321,...	?	文件 sgm	1996/12...
reut2-005.sgm	1,388,...	?	文件 sgm	1996/12...
reut2-006.sgm	1,254,...	?	文件 sgm	1996/12...
reut2-007.sgm	1,256,...	?	文件 sgm	1996/12...
reut2-008.sgm	1,410,...	?	文件 sgm	1996/12...
reut2-009.sgm	1,338,...	?	文件 sgm	1996/12...
reut2-010.sgm	1,371,...	?	文件 sgm	1996/12...
reut2-011.sgm	1,304,...	?	文件 sgm	1996/12...
reut2-012.sgm	1,323,...	?	文件 sgm	1996/12...
reut2-013.sgm	1,129,...	?	文件 sgm	1996/12...
reut2-014.sgm	1,128,...	?	文件 sgm	1996/12...
reut2-015.sgm	1,258	?	文件 sgm	1996/12

总计 27,982,337 字节(31 个文件)

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

Some notes on the Reuters CategoriesDavid D. Lewis3-Dec-96 The letter W. Bruce Croft received from Phil Hayes (March 9, 1990)gave a list of 135 TOPICS categories, which were used in HAYES89. Ireproduce this list below. The email message David D. Lewis received (Nov. 26, 1990) from PhilHayes gave lists (which I reproduce below) for categories in thesefields: number stated number actually at top of list on list ORGS 56 56 EXCHANGES 39 39 PLACES (i.e. COUNTRIES) 176 175 PEOPLE 269 267 Total 540 537 Total stated in IEEE paper: 539 The total 674 categories mentioned in HAYES90b is the sum of 135 and539. So the published numbers are a little off. For the purposes of Reuters-21578 we have taken as ground truth the actual lists ofcategory names, and the number of items on those lists. **********Subject Codes (135) Money/Foreign Exchange (MONEY-FX)Shipping (SHIP)Interest Rates (INTEREST)**Economic Indicator Codes (16)Balance of Payments (BOP)Trade (TRADE)Consumer Price Index (CPI)Wholesale Price Index (WPI) Unemployment (JOBS)Industrial Production Index (IPI)Capacity Utilisation (CPU)Gross National/Domestic Product (GNP)Money Supply (MONEY-SUPPLY)Reserves (RESERVES)Leading Economic Indicators (LEI)Housing Starts (HOUSING)Personal Income (INCOME)Inventories (INVENTORIES)Instalment Debt/Consumer Credit (INSTAL-DEBT)Retail Sales (RETAIL)**Currency Codes (27)U.S. Dollar (DLR)Australian Dollar (AUSTDLR)Hong Kong Dollar (HK)Singapore Dollar (SINGDLR)New Zealand Dollar (NZDLR)Canadian Dollar (CAN)Sterling (STG) D-Mark (DMK)Japanese Yen (YEN)Swiss Franc (SFR)French Franc (FFR)Belgian Franc (BFR)Netherlands Guilder/Florin (DFL)Italian Lira (LIT)Danish Krone/Crown (DKR)Norwegian Krone/Crown (NKR)Swedish Krona/Crown (SKR)Mexican Peso (MEXPESO)Brazilian Cruzado (CRUZADO)Argentine Austral (AUSTRAL)Saudi Arabian Riyal (SAUDRIYAL)South African Rand (RAND)Indonesian Rupiah (RUPIAH)Malaysian Ringitt (RINGGIT) Portuguese Escudo (ESCUDO)Spanish Peseta (PESETA)Greek Drachma (DRACHMA)**Corporate Codes (2) Mergers/Acquisitions (ACQ)Earnings and Earnings Forecasts (EARN)**Commodity Codes (78)ALUMBARLEYCARCASS CASTOR-MEALCASTOR-OILCASTORSEEDCITRUSPULPCOCOACOCONUT-OILCOCONUTCOFFEECOPPERCOPRA-CAKECORN-OILCORN CORNGLUTENFEEDCOTTON COTTON-MEALCOTTON-OILCOTTONSEEDF-CATTLEFISHMEALFLAXSEEDGOLDGRAINGROUNDNUT GROUNDNUT-MEALGROUNDNUT-OILIRON-STEELLEADLIN-MEALLIN-OILLINSEEDLIVESTOCKL-CATTLEHOGLUMBERLUPINMEAL-FEED NICKELOATOILSEEDORANGEPALLADIUMPALM-MEALPALM-OILPALMKERNELPLATINUMPLYWOODPORK-BELLYPOTATORAPE-MEALRAPE-OILRAPESEEDRED-BEANRICERUBBERRYESILKSILVERSORGHUMSOY-MEALSOY-OILSOYBEANSTRATEGIC-METALSUGARSUN-MEALSUN-OILSUNSEEDTAPIOCATEATINTUNG-OILTUNGVEG-OILWHEATWOOLZINC**Energy Codes (9)Crude Oil (CRUDE)Heating Oil/Gas Oil (HEAT)Fuel Oil (FUEL)Gasoline (GAS)Natural Gas (NAT-GAS)Petro-Chemicals (PET-CHEM)Propane (PROPANE)Jet and Kerosene (JET)Naphtha (NAPHTHA)

_____@heading[Organization Codes (56)]@begin[format]African Development Bank (ADB-AFRICA)@*Agency for International Development (AID)@*Asian Development Bank (ADB-ASIA)@*Association of International Bond Dealers (AIBD)@*Association of Natural Rubber Producing Countries (ANRPC)@*Association of South East Asian Nations (ASEAN)@*Association of Tin

[点此下载完整数据集]