データサイエンス特論　プログラム課題 第 2 回(多値ラベル分類)

以下の課題を実施し、実行したプログラムと実行結果(プログラムは Python, Jupyter notebook(拡張子. ipynb, .html)等でログをまとめるのも可、実行結果を MS Word(PDF)等でまとめることも可) を、ZIP でまとめ、MoodleLMS にアップロードすること。期限は【8 月 8 日（月）】の深夜 23:59 分までとする。（プログラム課題はこれで終わりです）

UCI（カリフォルニア大学アーバイン校）の機械学習用データレポジトリにある有名なロイターのニュース記事（Reuters-21578） https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection は、英語で書かれた1980 年当時のニュース記事の 集合である。このデータは、機械学習の練習にしばしば利用されてきた。ニュース記事には カテゴリー情報が 1 種類以上付与（マルチラベルが付与）されており、このカテゴリーラベルを予測するテキスト分類問題を考える。ただし、元のロイターニュース記事には、各種のノイズ文書がある。そこで、55 カテゴリー、10,700 文書（7713 訓練文書, 2987 テスト文書）に絞り込んだ 部分集合記事データを作成した。以下のサイトからダウンロードできる。 https://www.kde.cs.tut.ac.jp/~aono/data/ma_reuters.zip においてある。以降、こ のデータをMA-Reuters と呼ぶこととする。

作業手順を述べる。

(Step 1) MA-Reuters を各自の作業できるPC 等にダウンロードせよ。

(Step 2) MA-Reuters はPython のNLTK パッケージのコーパスに従って作成してある。そこで、各自のPython の環境にNLTK がまだインストールされていない場合、NLTK をインストールせよ。同時に https://www.nltk.org/data.html を参照してデータのダウンロードを最初に一回行い、適当な NLTK 用のデータフォルダーを作成しておくこと。たとえば Windows 環境下で、C:\nltk_data に作成したとすると、コーパス フォルダが、C:\nltk_data\corpora 以下にできるので、(1)で入手したma_reuters.zip を corpora フォルダにコピーしておくこと。

(Step 3) NLTK と同時にscikit-learn（サイキットラーン）パッケージも利用することになるので、こちらのインストールがまだの場合しておくこと。

以下が課題である。本課題で Python のパッケージのインストールやニューラルネットワーク（NN）の環境設定などのヘルプはできないことを注記する。

[1] https://www.kde.cs.tut.ac.jp/~aono/2022/DataScience/Reuters-Multi-Label-SVM-Example.html にある Jupyter notebook での Python コードを順番に実行し、多値クラス多値ラベル問題を解決できることを確認せよ。ただし、このコード内でオレンジカテゴリー内の(最初の)文書のプリントは、違うカテゴリー（オレンジカテゴリー以外）の文書に変更して実行すること。

[2] 5 5 個のカテゴリーのなかで、Jaccard 係数（資料第 6 回参照）が最も高いカテゴリーと最も低いカテゴリーが何であったかのべよ。

[3] ここでの NLTK の TF-IDF モデル＋SVM 以外の組み合わせで本マルチラベル問題に対応できる手法を適宜ためし、実行結果をのべよ。たとえば、テキスト表現に授業の資料やビデオで述べた Word2Vec や BERT 等を用いてもよい。NLTK の BoW(TF-IDF)モデルのまま場合、少なくとも分類器は変更すること。たとえば、ニューラルネット(NN)を用いてもよい。NN の場合、評価手法はバイナリークロスエントロピー等の誤差がマルチラベル・マルチクラス分類ではよく用いられる。