情報メディア実験（MAST）/ 知能情報メディア実験（COINS）（イントロダクション）

乾孝司（システム情報系情報工学域）
inui@cs.tsukuba.ac.jp

1　はじめに

文書分類とは，文字通り，文書を分類することであり，文書が入力として与えれた時，入力文書を適切なクラスに割り振る（分類する）ことが仕事である（図 1）．文書分類の技術は，スパムフィルタや，Web ページのペアレンタル・コントロールなどにおける核となる要素技術であり，情報社会にとっての必須技術のひとつである．

本実験課題では，実際に文書分類をおこなうプログラムを作成することを通じて，言語データ（人間が日常生活で使っている日本語や英語等の言語が記述されたデータ）を計算機上で処理する過程を体験的に学習することを目的とする．具体的には，以下の 2 つの手法を取り上げ，文書自動分類の処理過程を学習する．

- k 近傍法
- ナイーブベイズ法

2　参加メンバ

- 初回に，学籍番号，氏名，メールアドレスを確認します．
- 後日，実験用のメーリングリストを作成します．

3　ティーチング・アシスタント（TA）

4　進め方

- 春日からの移動を考慮して，開始時刻は毎回 15 分遅れでおこなう（休憩時間を先に使う）．
  - 水曜は 12:30 開始，金曜は 15:30 開始．
- 週 2 回のうち，基本的に水曜にその週の演習内容の説明をおこなう．水曜の残り時間および金曜は，各自の演習時間に割り当てる．

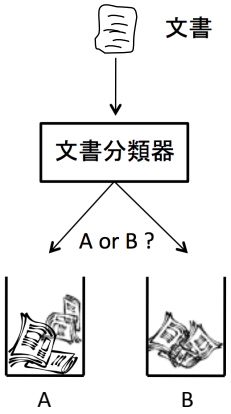

図 1: 文書分類

5 教材

- 教材として利用する資料やデータは，必要に応じて，適宜配布する．資料の pdf は以下の URL から入手できるようにする予定（パスワード認証付き）．

```
http://www.mibel.cs.tsukuba.ac.jp/~inui/kougi/jikken/index.html
```

- 日本語データを扱う場合は EUC を仮定する．
- サンプルプログラムは Ruby or Python で配布する．
- また，実験の実施は，Ruby or Python でおこなうこと．
- 上記プログラミング原語の経験がない人は実験開始までに参考書を読むなどしておいて下さい．
  - 参考書の貸し出し有

6 出欠管理

- 共通の出席管理表を毎回各自チェックすること．
- 個別の出欠管理はおこなわない．

7 評価

- 本実験では，実験時間中に文書分類のプログラムを作成してもらい，性能評価会（全２回）において，自作プログラムの分類性能を競ってもらう．
- すべての性能評価会に参加すれば合格圏内を保証する（もちろん，出席，最終レポート等の課題共通の単位習得条件はクリアしていることが必要である）．その他，提出された分類性能やプログラムの工夫点等に応じて，加点方式で評価する．
- 必須となる提出物は，競技会の結果レポートおよび最終レポートの２種類である．ただし，実験の課題ごとに（進捗状況を確認するため），作成したプログラムの動作を目視確認する．