**データサイエンス特論　授業課題　第六回分（特徴量抽出）**

**豊橋技術科学大学大学院　情報知能工学専攻**

**音声言語処理研究室**

**203319 M1木内貴浩**

電子メールが迷惑メール(spam)かそうでないか(ham)のフィルターを、第六回で紹介した伝統的な機械学習法（最近傍法、ナイーブベイズ法、サポートベクトルマシン法など）で判定したい。電子メールの中身は、英語であること仮定してよいとする（例 Enron Spam dataset : <https://www.kaggle.com/datasets/wanderfj/enron-spam>/）。メールがspamかそうでないかのラベルはあると仮定して、生の英語の電子メールテキストデータから、判定に使いたい特徴量の抽出（判定に使ってみたい特徴量とそのデータ型の概要のリスト作成）を行い、リストを作成せよ。10個以上抽出すること。ただし、ここでは深層学習的特徴量や埋込み特徴量（Word2vec, GloVe, fastTextなど）は使わないとする。以下の例を含めてかまわない。

|  |  |  |
| --- | --- | --- |
|  | 特徴量 | データ型 |
| 1 | Congratulationsが含まれるかどうか | 0か1のboolean型 |
| 2 | メールの総文字数 | 整数型 |
| 3 | 特殊文字の割合(特殊文字/文字数) | 実数型 |
| 4 | Word cloudのtop10の文字が含まれるかどうか | 0か1のboolean型 |
| 5 | 単語の出現レベル(100\*メールの単語出現回数/全体の単語出現回数) | 実数型 |
| 6 | 文字の出現レベル(100\*メールのchar出現回数/全体のchar出現回数) | 実数型 |
| 7 | 大文字の連続した平均長さ | 実数型 |
| 8 | 最も長い大文字の列の長さ。 | 整数型 |
| 9 | 途切れない大文字の列の長さの総和 | 整数型 |
| 10 | 電子メールに含まれる大文字の総数 | 整数型 |