データサイエンス特論　授業課題　第六回分（特徴量抽出）

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。

例：

	特徴量	データ型
1	Congratulations が含まれるかどうか	0 か 1 のブーレアン型
2	メールの総文字数	整数型

『ヒント』:
事前に処理された迷惑メール判定用の訓練データ（ならびにテストデータ）の例が、UCI Machine Learning Archive にある。
https://archive.ics.uci.edu/ml/datasets/spambase
こちらから使えそうな特徴量（属性）を適宜、選択してよい。