情報システム論実習0630課題

課題１

※作成したプログラムは

【データセットの概要】

The Reuters Corpus

合計130万語の10,788のニュースドキュメント。90 のトピックに分類される。本実験では先頭の500ドキュメントを使用した。

【利用した文書類似度、距離尺度】

ベクトルベース手法のTF-IDFを使用した。

【実行結果（得られたクラスタの内容、クラスタの質について定性的評価を行う）】

Kはクラスタ数を表す

**K=8の時（実行結果一部抜粋）**

**新聞, テキスト, 建物, フロント が含まれている画像

自動的に生成された説明**

K0は砂糖(sugar)が含まれるものが属している傾向がある

K1は利得(profit)、損益が含まれるものが属している傾向がある

K2は国名,地域名(Japan,China...etc)が入っているものが属している傾向がある

K3は商売や経済に関して書かれたものが属している

K4は会社(coop)が含まれるものが属している傾向がある

K5は国際的なこと(international)が含まれるものが属している傾向がある

K6はお金(money)が含まれるものが属している傾向がある

K7は業務成績について書かれたものが属している傾向がある

クラスタの質に関しては,K0やK2などは区別がしやすかったが、K1,K3,K4,K6などは区別が難しく、判断することが困難だった。クラスタ数が少し大きいために人が区別することは困難になっている可能性がある。

**K=4の時（実行結果一部抜粋）**

**新聞, テキスト が含まれている画像

自動的に生成された説明**

K0は物質(sugar, tin...etc)について書かれたものが属している傾向がある

K1は会社について書かれたものが属している傾向がある

K2は経済について書かれたものが属している傾向がある

K3は国名、地域名が含まれるものが属している傾向がある。

クラスタの質に関しては、各クラスタの分別が容易だったためある程度良かったと考えらえる。人が認識できる程度の分類が可能

**K=2の時（実行結果一部抜粋）**

**新聞, スクリーンショット が含まれている画像

自動的に生成された説明**

K0は文全体の内数値が占める割合が小さいものが属している傾向がある

K1は文全体の内数値が占める割合が大きいものが属している傾向がある

クラスタの質に関しては、みた限り分類はできるが、意味のあるものをクラスタリングできているとは感じなかった（新たな発見がない）。クラスタ数が少なすぎると意味のある分別ができない可能性があることがわかった。（もちろん、事前にデータを二分化したいといった意図があってクラスタ数を２にするのは問題ない）

【感想】

本来は９０のトピックが存在しており、（もちろん１００個に限定しているためトピック数はかなり絞られるが）そのトピックを分類できることが一番良いが、その各文書のトピックが分からなくてもk-meansを用いて大きな括りで分類することができた。文書数を最大にしてクラスタ数を90に設定した場合、どんな結果が出るのか興味が湧いたため、課題を終わらせた後に時間があればやってみようと思った。