文章の類似検査

ここに幾つかの文章がテキストデータの形で存在しています。これらの文章が互いにどれだけ似通っているかを判定し 100 点満点で評価するプログラムを書け、と言われたらどうしますか。勿論プログラミング言語を知らなければ記述することはできませんが、どのような考えでどのように処理するか、その手順や方法を考えることは言語を知らなくても日本語で考えたり書いたりできます。

単語Ａ単語Ｂが連続して出現する統計的確率を求めることで２つの文章の類似検査をする簡単なプログラムを作ってみました。短いプログラムですがコピペをした文章や一部を改変した文章、前後左右を入れ替えて元の文章とは全く異なってしまった文章などを発見できます。

他人が書いた文章を許可を得ず勝手にコピーして利用することや一部をコピーし改変して利用することなどは著作権を侵害する行為です。このような文章を自動的に発見することができるかもしれません。

ビッグデータの解析研究等はまだ始まったばかりだと聞いています。高校生諸君の柔軟な発想で画期的な分析手法を考え出してもらいたいものです。

文章の類似検査プログラムは下記 URL よりダウンロードできます。
http://www.cfs.chiba-u.jp/koudai-renkei/information/files/ruiji.xlsm