XÂY DỰNG BỘ NGỮ LIỆU CHO ĐỘ ĐO PHONG CÁCH VĂN BẢN TIẾNG VIỆT

Đỗ Trần Anh Đức[1], Lương An Vinh[2], Nguyễn Tuyết Nhung[3], Nguyễn Thị Như Điệp[4]
[1]Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự Nhiên, ĐHQG-HCM
[2]Khoa Công Nghệ Thông tin, Đại học Công nghệ Sài Gòn
[3]Bộ môn Ngoại Ngữ, Trường Đại học An ninh Nhân dân
[4]Ban Khoa học Cơ bản, Đại học Công nghệ Sài Gòn
1512122@student.hcmus.edu.vn, vinh.luongan@stu.edu.vn,
velvetsnow.nguyen@gmail.com, nhudiep2004@gmail.com

Tóm tắt

Độ đo phong cách là phương pháp sử dụng định lượng để xác định các đặc trưng phong cách của người viết trong một văn bản. Những ứng dụng của độ đo phong cách được sử dụng phổ biến trong nhiều lĩnh vực có thể kể tới như: xác định nguồn gốc tác giả trong các văn bản vô danh, phân loại các văn bản có phong cách của cùng một tác giả, đưa ra các thông tin tác giả trong một văn bản nhằm hỗ trợ pháp lý,… Tuy nhiên, việc tiến hành các nghiên cứu độ đo phong cách cho văn bản vẫn còn rất hạn chế ở tiếng Việt vì sự thiếu sót các bộ ngữ liệu để thực nghiệm. Vì thế, trong bài báo này, chúng tôi trình bày về quá trình xây dựng một bộ ngữ liệu về độ đo phong cách cho tiếng Việt bằng cách thu thập tự động từ các trang báo nổi tiếng trên mạng (ví dụ: Tuổi Trẻ, VnExpress, …) và đồng thời đưa ra các phân tích đánh giá từ bộ ngữ liệu này dựa trên 3 tiêu chí, bao gồm: tính đại diện, tính cân bằng và lấy mẫu.

Từ khóa: xây dựng ngữ liệu, độ đo phong cách, tiếng Việt.

BUILDING A CORPUS FOR VIETNAMESE TEXT STYLOMETRY

Duc Do Tran Anh[1] , Vinh Luong An[2], Nhung Nguyen Tuyet[3], Diep Nguyen Thi Nhu[4]
[1]Faculty of Information Technology, University of Science, VNU-HCM
[2]Faculty of Information Technology, Saigon Technology University
[3]Department of Foreign Languages, University of People's Security
[4]Basic Sciences Department, Saigon Technology University
1512122@student.hcmus.edu.vn, vinh.luongan@stu.edu.vn,
velvetsnow.nguyen@gmail.com, nhudiep2004@gmail.com

Abstract

Stylometry is a method using to identify 'style' features of an author in a document quantitatively. Stylometry applications are widely used in various fields, such as: identifying authorship in disputed texts, classifying documents with the 'style' features of an author, providing an author information in any documents for legal aid. However, conducting stylometry experiments is still not well-concerned in Vietnamese language because of lacking of a corpus for these ones. Therefore, in this article, we present our procedure to build a corpus for stylometry in Vietnamese language by collecting automatically famous news papers in the Internet (such as: The Youth, VnExpress, …) as well as give the analysis from this corpus based on three standarts: representativeness, balance and sampling.

Key words: building corpus, stylometry, Vietnamese language.