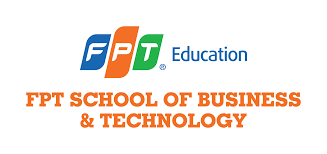
VIỆN QUẢN TRỊ & CÔNG NGHỆ FSB



BÁO CÁO ASSIGNMENT 1 – SO SÁNH PHƯƠNG PHÁP kNN, NAIVE BAYES VÀ SVM TRONG BÀI TOÁN PHÂN LOẠI VĂN BẢN

Giảng viên hướng dẫn: **TS. QUẢN THÀNH THƠ**

Thành viên: Nguyễn Hoài Phương

Lớp: MSE09

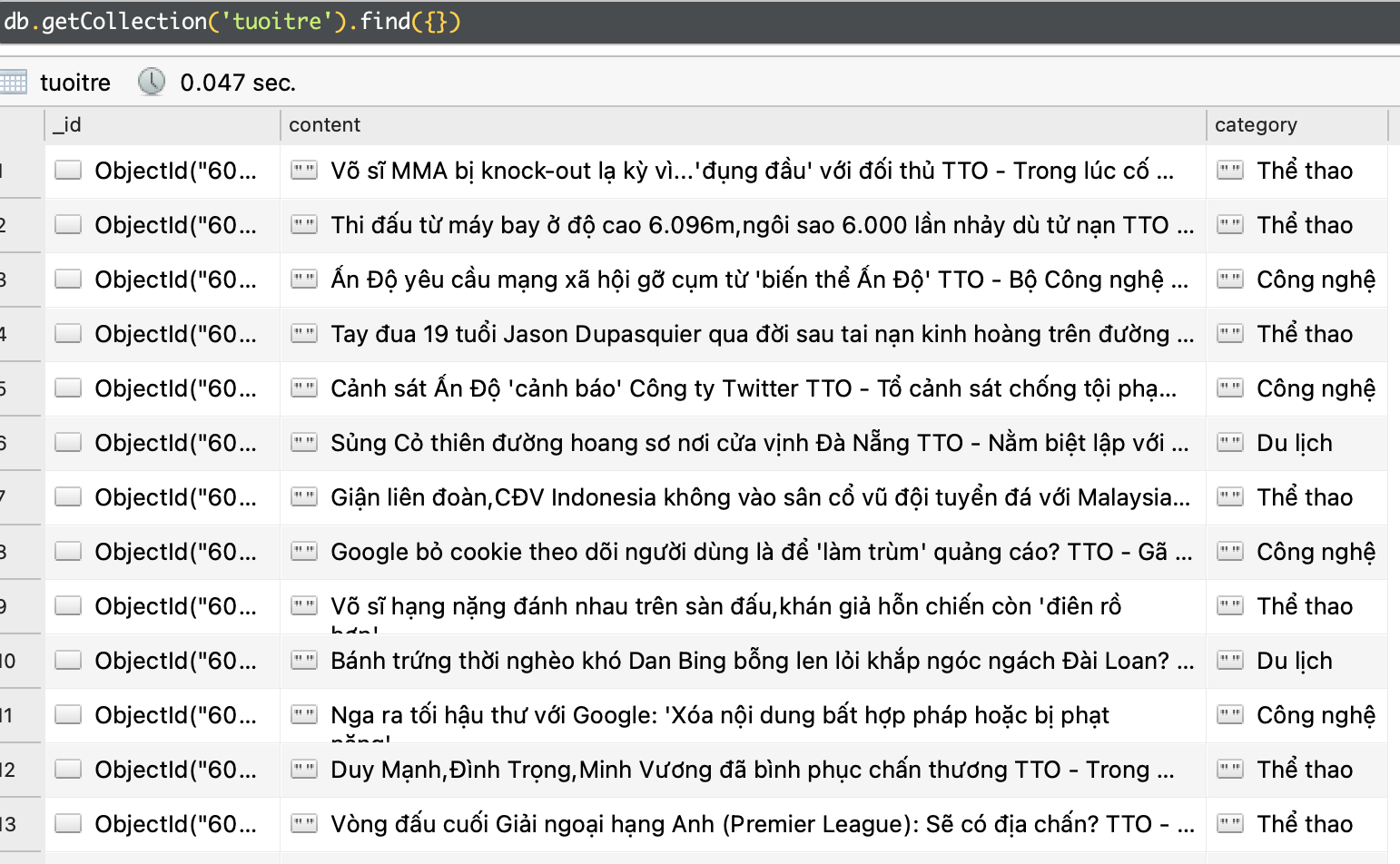
# Dataset

Tập training and test data được lấy từ trang web https://tuoitre.vn/. Tiêu đề và phân loại bài báo sẽ được lưu vào cột text và category tương ứng. Category có thể là “Công nghệ", “Thể thao" hoặc “Du lịch".

Framework scrapy được sử dụng Để “cào” được dữ liệu trên trang web tuổi trẻ online. Source code được lưu trữ ở git repo bên dưới:

https://github.com/bluesky2106/python-ai/tree/master/spyder

Sau khi dữ liệu được “cào” thì lưu trữ vào mongodb.

Có tổng cộng 474 bài báo về “Công nghệ”, 455 bài báo về “Du lịch” và 445 bài báo về “Thể thao” đã được cào. Từ dữ liệu thô trên, chúng ta phân chia thành 2 tập dữ liệu training và test data theo tỷ lệ 80 – 20. Như vậy có tổng cộng 1099 bài báo được xài cho quá trình training và 275 bài báo được xài cho quá trình test.

### Training data

Có tổng cộng 376 bài báo liên quan tới công nghệ, 369 bài báo liên quan tới du lịch và 354 bài báo liên quan tới thể thao.

### Test data

Có tổng cộng 98 bài báo liên quan tới công nghệ, 86 bài báo liên quan tới du lịch và 91 bài báo liên quan tới thể thao.

|  |  |  |
| --- | --- | --- |
|  | Training | Testing |
| Công nghệ | 376 | 98 |
| Du lịch | 369 | 86 |
| Thể thao | 354 | 91 |

# Kết quả sử dụng k-NN

Accuracy : 0.9055

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 86 | 10 | 2 |
| Du lịch | 6 | 79 | 1 |
| Thể thao | 2 | 5 | 84 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.9149 | 0.8404 | 0.9655 |
| recall | 0.8776 | 0.9186 | 0.9231 |
| f-measure | 0.8958 | 0.8778 | 0.9438 |

# Kết quả sử dụng Bayes (không balancing)

Accuracy : 0.9164

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 86 | 10 | 2 |
| Du lịch | 6 | 80 | 0 |
| Thể thao | 3 | 2 | 86 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.9053 | 0.8696 | 0.9773 |
| recall | 0.8776 | 0.9302 | 0.9451 |
| f-measure | 0.8912 | 0.8989 | 0.9609 |

# Kết quả sử dụng Bayes (có dùng balancing)

Accuracy : 0.9236

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 86 | 10 | 2 |
| Du lịch | 5 | 81 | 0 |
| Thể thao | 2 | 2 | 87 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.9247 | 0.871 | 0.9775 |
| recall | 0.8776 | 0.9419 | 0.956 |
| f-measure | 0.9005 | 0.905 | 0.9667 |

# Kết quả sử dụng SVM (default kernel)

Accuracy : 0.9236

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 89 | 7 | 2 |
| Du lịch | 7 | 79 | 0 |
| Thể thao | 2 | 3 | 86 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.9082 | 0.8876 | 0.9773 |
| recall | 0.9082 | 0.9186 | 0.9451 |
| f-measure | 0.9082 | 0.9029 | 0.9609 |

# Kết quả sử dụng SVM (linear kernel)

Accuracy : 0.9164

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 88 | 8 | 2 |
| Du lịch | 8 | 78 | 0 |
| Thể thao | 2 | 3 | 86 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.898 | 0.8764 | 0.9773 |
| recall | 0.898 | 0.907 | 0.9451 |
| f-measure | 0.898 | 0.8914 | 0.9609 |

# Kết quả sử dụng SVM (poly kernel)

Accuracy : 0.8764

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 93 | 3 | 2 |
| Du lịch | 15 | 71 | 0 |
| Thể thao | 12 | 2 | 77 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.775 | 0.9342 | 0.9747 |
| recall | 0.949 | 0.8256 | 0.8462 |
| f-measure | 0.8532 | 0.8765 | 0.9059 |

# Kết luận:

Xét trên tất cả các khía cạnh (accuracy, precision, recall, f-measure), phương pháp Naive Bayes có balance là cho ra kết quả tôt nhất ở mọi category (Công nghệ, Du lịch và Thể thao).