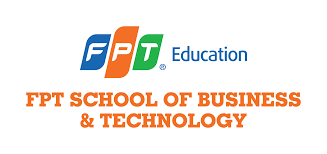
VIỆN QUẢN TRỊ & CÔNG NGHỆ FSB



SO SÁNH PHƯƠNG PHÁP kNN, NAIVE BAYES, SVM VÀ NEURAL NETWORK TRONG BÀI TOÁN PHÂN LOẠI VĂN BẢN

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

Thành viên: Nguyễn Hoài Phương

Lớp: MSE09

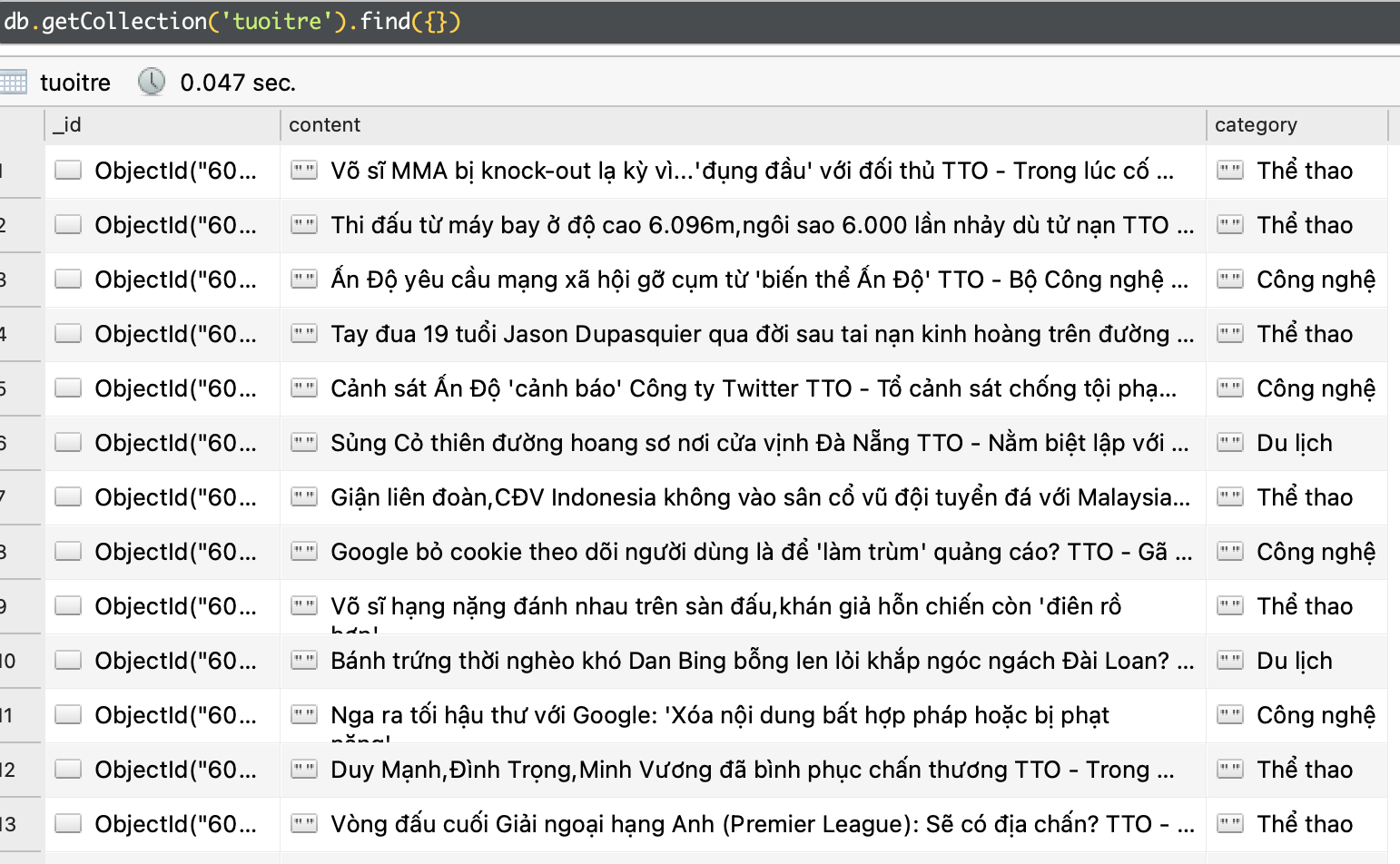
# Dataset

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 1579 bài báo về “Công nghệ”, 2535 bài báo về “Du lịch” và 3058 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 5737 bài báo được xài cho quá trình training và 1435 bài báo được xài cho quá trình test.

### Training data

Có tổng cộng 1244 bài báo liên quan tới công nghệ, 2020 bài báo liên quan tới du lịch và 2471 bài báo liên quan tới thể thao.

### Test data

Có tổng cộng 335 bài báo liên quan tới công nghệ, 513 bài báo liên quan tới du lịch và 587 bài báo liên quan tới thể thao.

|  |  |  |
| --- | --- | --- |
|  | Training | Testing |
| Công nghệ | 1244 | 335 |
| Du lịch | 2020 | 513 |
| Thể thao | 2471 | 587 |
| Tổng cộng | 5737 | 1435 |

# Kết quả sử dụng k-NN

Accuracy : 0.8592

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 265 | 58 | 12 |
| Du lịch | 55 | 437 | 21 |
| Thể thao | 19 | 37 | 531 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.78171091 | 0.82142857 | 0.94148936 |
| recall | 0.79104478 | 0.85185185 | 0.90459966 |
| f-measure | 0.78635015 | 0.83636364 | 0.92267593 |

# Kết quả sử dụng Bayes (không balancing)

Accuracy : 0.8523

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 197 | 132 | 6 |
| Du lịch | 5 | 491 | 17 |
| Thể thao | 2 | 50 | 535 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.96568627 | 0.72956909 | 0.95878136 |
| recall | 0.5880597 | 0.95711501 | 0.91141397 |
| f-measure | 0.7309833 | 0.82799325 | 0.93449782 |

# Kết quả sử dụng Bayes (có dùng balancing)

Accuracy : 0.8746

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 285 | 46 | 4 |
| Du lịch | 58 | 446 | 9 |
| Thể thao | 20 | 43 | 524 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.78512397 | 0.83364486 | 0.97579143 |
| recall | 0.85074627 | 0.86939571 | 0.89267462 |
| f-measure | 0.81661891 | 0.85114504 | 0.93238434 |

# Kết quả sử dụng SVM (default kernel)

Accuracy : 0.8878

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 258 | 72 | 5 |
| Du lịch | 21 | 480 | 12 |
| Thể thao | 6 | 45 | 536 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.90526316 | 0.8040201 | 0.96925859 |
| recall | 0.77014925 | 0.93567251 | 0.91311755 |
| f-measure | 0.83225806 | 0.86486486 | 0.94035088 |

# Kết quả sử dụng SVM (linear kernel)

Accuracy : 0.8829

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 260 | 69 | 6 |
| Du lịch | 28 | 473 | 12 |
| Thể thao | 9 | 44 | 534 |

array([0.87542088, 0.80716724, 0.9673913 ]),

array([0.7761194 , 0.92202729, 0.90971039]),

array([0.82278481, 0.86078253, 0.93766462]),

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.87542088 | 0.80716724 | 0.9673913 |
| recall | 0.7761194 | 0.92202729 | 0.90971039 |
| f-measure | 0.82278481 | 0.86078253 | 0.93766462 |

# Kết quả sử dụng SVM (poly kernel)

Accuracy : 0.8272

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 160 | 163 | 12 |
| Du lịch | 9 | 489 | 15 |
| Thể thao | 3 | 46 | 538 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.93023256 | 0.70057307 | 0.95221239 |
| recall | 0.47761194 | 0.95321637 | 0.9165247 |
| f-measure | 0.63116371 | 0.80759703 | 0.93402778 |

# Kết quả sử dụng Neural Network 2 layers (1)

layer 1: 128 neural, activation = ‘linear’

layer 2: 3 neural, activation = ‘softmax’

epochs = 100

batch\_size = 128

Accuracy : 0.8488

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 244 | 83 | 8 |
| Du lịch | 40 | 445 | 28 |
| Thể thao | 9 | 49 | 529 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0.83276451 | 0.7712305 | 0.93628319 |
| recall | 0.72835821 | 0.86744639 | 0.9011925 |
| f-measure | 0.77707006 | 0.81651376 | 0.91840278 |

# Kết quả sử dụng Neural Network 2 layers (2)

layer 1: 1024 neural, activation = ‘linear’

layer 2: 3 neural, activation = ‘softmax’

epochs = 100

batch\_size = 128

Accuracy : 0.8467

Confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| Công nghệ | 236 | 91 | 8 |
| Du lịch | 33 | 457 | 23 |
| Thể thao | 11 | 54 | 522 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Công nghệ | Du lịch | Thể thao |
| precision | 0. 84285714 | 0.75913621 | 0.94394213 |
| recall | 0.70447761 | 0.89083821 | 0.88926746 |
| f-measure | 0.76747967 | 0.81973094 | 0.91578947 |

# 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).