XÂY DỰNG NGỮ LIỆU TIẾNG VIỆT DỰA TRÊN BÁO TUOITRE ONLINE

I. QUÁ TRÌNH XÂY DỰNG NGỮ LIỆU TIẾNG VIỆT:

- **Xác định khai thác 50 chủ đề bao gồm**: *Âm nhạc, Nghĩ, Xã hội, Tư vấn, Giới tính, Kiều bào, Góc học tập, Nhịp sống trẻ, TV Show, Xu hướng, Chuyện pháp đình, Mua sắm, Đầu tư, Kinh doanh, Nhân vật, Giả-thật, Sức khỏe, Đời sống, Pháp luật, Khám phá, Thư giãn, Hồ sơ, Phát minh, Giải trí, Bình luận, Văn hóa, Pháp lý, Điện ảnh, Phòng mạch, Thời trang, Yêu, Văn học - Sách, Việc làm, Mẹ & Bé, Học đường, Hậu trường, Thường thức, Giáo dục, Dinh dưỡng, Thế giới, Du học, Muôn màu, Câu chuyện giáo dục, Thời sự, Xe, Phóng sự, Biết để khỏe, Tài chính, Khoa học, Doanh nghiệp*.
- **Thời gian các bài báo: từ ngày 1/1/2008 - 9/5/2018.**

Giai đoạn thu thập và thống kê tất cả các đường dẫn tới các bài báo:

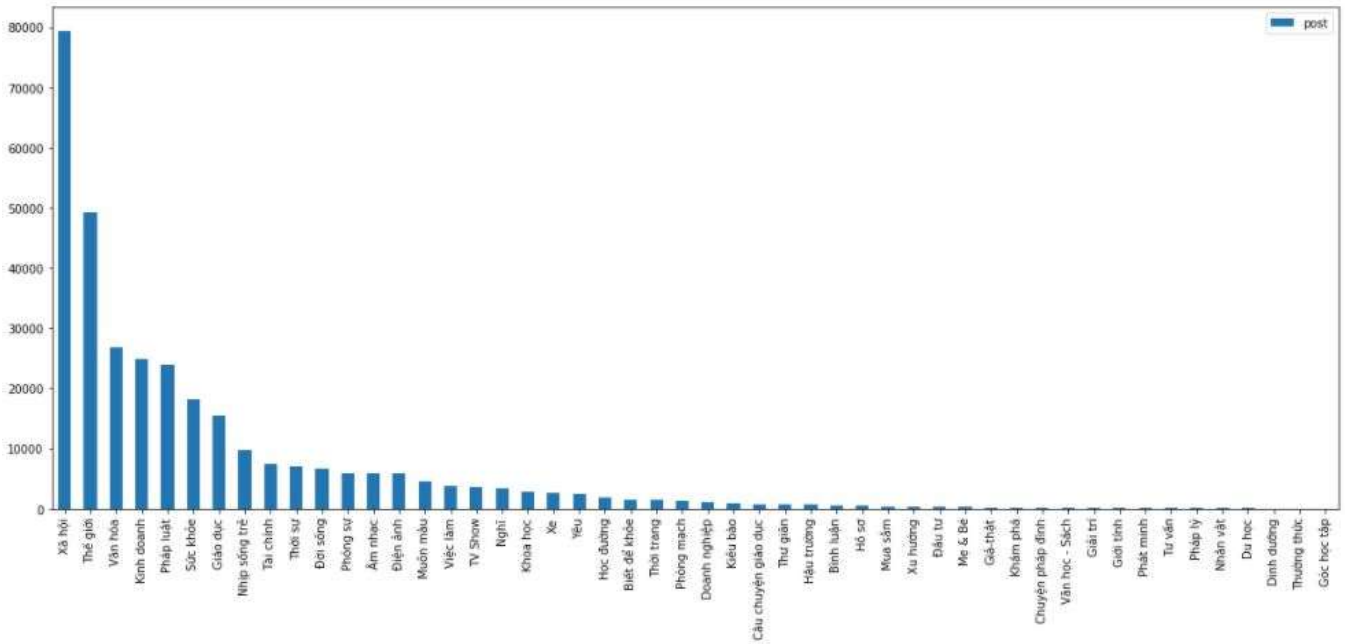
Thu thập tất cả các đường dẫn, chủ đề, năm đăng của các bài báo qua từng năm.

Lọc lại tất cả các url xem có bị trùng cùng 1 đường dẫn nhưng lại bị lặp lại hay không, xảy ra vấn đề này do 1 bài báo không chỉ duy nhất được gắn nhãn cho 1 chủ đề cố định mà có thể có rất nhiều chủ đề cùng 1 lúc.

Thống kê các bài theo chủ đề, theo năm, theo thể loại của từng năm và trung bình từng loại mỗi năm.

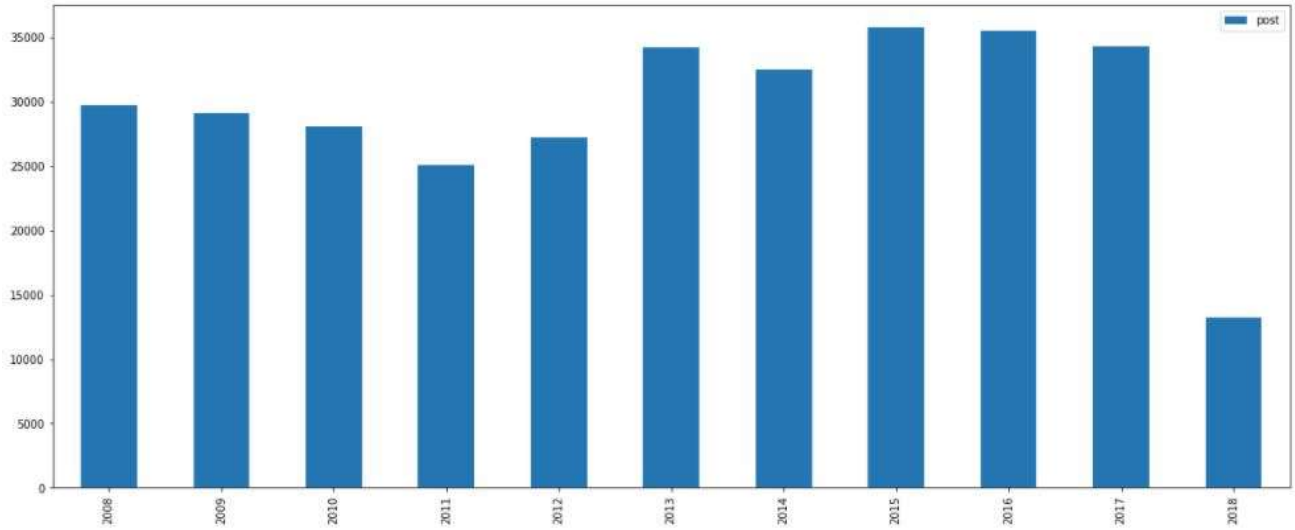
Visualize kết quả thống kê, đánh giá và cân bằng bộ ngữ liệu.

Theo chủ đề, ta nhận thấy có rất nhiều bài báo tập trung cho 1 vài chủ đề nổi bật (Xã hội, thế giới, Văn hoá, Kinh Doanh, Pháp luật), còn những chủ đề còn lại rất ít bài báo từ đó ta thấy độ lệch chuẩn quá lớn đối với các chủ đề.



Hình 0.1. biểu diễn các bài báo theo chủ đề

Theo từng năm, ta nhận thấy phần lớn các năm có các bài báo được đăng tải lên tương đối đều nhau



Hình 0.2. biểu diễn các bài báo theo năm

Thông qua kết quả visualize, ta chọn cụm thể loại có số lượng bài báo nhiều và tương đối nhiều để cân bằng cụm đó, còn lại ta vét cạn tất cả các bài.

Cách thức cân bằng

Kết quả cuối cùng ta thu được là khai thác 72.23% trên tổng số bài báo thuộc 50 chủ đề trong 10 năm qua.

Xem kết quả thống kê từng chủ đề theo từng năm ở đây.

Giai đoạn thu thập text:

Dựa trên các url sau khi đã cân bằng ở bước trên, ta khai thác ngữ liệu và định dạng theo chuẩn TEI[1]

Thoạt đầu, mình định khai thác gồm cả heading và contents của từng đoạn, thế nhưng mà do cách đăng bài của tuoitre không có định dạng cho việc đó, chỉ dựa trên style bold người đọc xác định tiêu đề từng đoạn, trong khi đó rất nhiều tác giả sử dụng in đậm như hình thức nhấn mạnh. Do các bài báo trên mạng có sự khác nhau về định dạng nên không thể xác định được heading và content của từng đoạn nên chỉ có cách lưu lại tất cả các đoạn riêng lẻ theo từng " ".

Trong quá trình xây dựng ngữ liệu, mình đã xoá đi các khoảng trắng dư ở các đầu, một vài ký tự ascii đặc biệt.

II. ĐỊNH DẠNG TEI:

Trong bài này mình sử dụng định dạng **TEI**.

Theo định dạng sau:

Corpus TEI:

```
<?xml version="1.0" encoding="UTF-8"?>
<teiCorpus version="3.3.0" xmlns="http://www.tei-c.org/ns/1.0">
        <teiHeader>
                <fileDesc>
                  <titleStmt>
                        <title>TÊN_CORPUS</title>
                  </titleStmt>
                  <publicationStmt>
                        <pubPlace>thành phố Hồ Chí Minh</pubPlace>
                        <publisher>tuổi trẻ online</publisher>
                  </publicationStmt>
            </fileDesc>
            <profileDesc>
                  <langUsage default="NO"></langUsage>
               <language id="vi">vietnamese</language>
            </profileDesc>
        </teiHeader>
        <TEI/>
        <TEI/>
        <TEI/> (xem phần dưới về định dạng từng text TEI)
</teiCorpus>
```

Text TEI:

```
<TEI lang="vi" id="MÃ_ID">
        <teiHeader>
                <fileDesc>
                  <titleStmt>
                        <title>TÊN_BÀI_VIẾT</title>
                  </titleStmt>
                  <publicationStmt>
                        <authority>TÁC_GIẢ</authority>
                     <date>NGÀY_THÁNG_NĂM_ĐĂNG</date>
                  </publicationStmt>
                  <profileDesc>
                                <textDesc>
                                        <domain type="CHỦ_ĐỀ"/>
                                </textDesc>
                        </profileDesc>
            </fileDesc>
        </teiHeader>
        <text>
                <front>ĐOẠN_MỞ_ĐẦU</front>
            <body>
                  <div1>
                     <p>ĐOẠN_1 </p>
                  </div1>
                  <div2>
                     <p> ĐOẠN_2</p>
                  </div2>
            </body>
        </text>
</TEI>
```

III. KẾT QUẢ:

Thu được 230,148 bài báo.

Tổng 114,581,753 tiếng.

Chiếm 98.12% số bài báo cần khai thác.

Chiếm 64.11% tổng số bài báo của vài chủ đề có trong gần 10 năm qua trên tuoitre.vn.

Chủ đề	Sô bài
Doanh nghiệp	1091
Kiều bào	820
Câu chuyện giáo dục	812
Bình luận	617
Hậu trường	611
Hồ sơ	604
Thư giãn	532
Mua sắm	351
Xu hướng	319
Đầu tư	290
Mẹ & Bé	282
Giả-thật	210
Chuyện pháp đình	164
Văn học - Sách	150
Khám phá	149
Giới tính	142
Tư vấn	129
Giải trí	128
Phát minh	126
Pháp lý	95
Nhân vật	82
Dinh dưỡng	56
Du học	51
Góc học tập	41
Thường thức	32

Chủ đề	Sô bài
Thế giới	24548
Xã hội	23096
Kinh doanh	22412
Văn hóa	22408
Pháp luật	21133
Sức khỏe	18004
Giáo dục	14254
Nhịp sống trẻ	9623
Tài chính	7478
Đời sống	6701
Thời sự	6031
Phóng sự	5903
Âm nhạc	5800
Điện ảnh	5757
Muôn màu	4543
Việc làm	3823
TV Show	3447
Nghĩ	3416
Khoa học	2750
Xe	2537
Yêu	2397
Học đường	1887
Biết để khỏe	1510
Thời trang	1419
Phòng mạch	1387