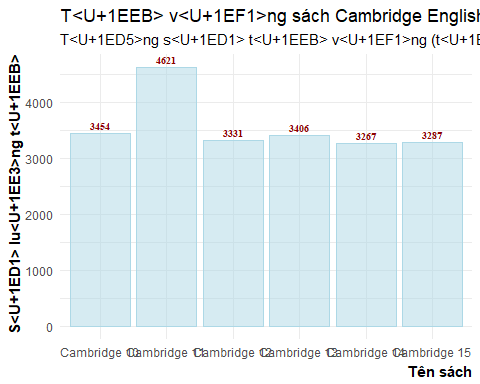
Một số insight về từ vựng sách Cambridge English IELTS

Phạm Phương Thảo

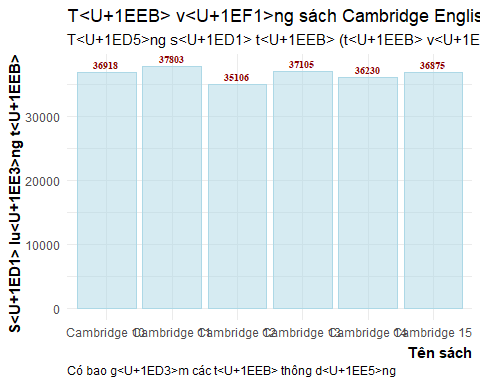
8/30/2021

(Ngắn gọn cho bạn nào lười đọc): File Excel từ vựng tiếng Anh xuất hiện trong sách Cambridge English IELTS từ cuốn 10 - 15. Xem từng ảnh để biết thêm về các tiêu chí lọc từ của mình và cách sử dụng file từ vựng.

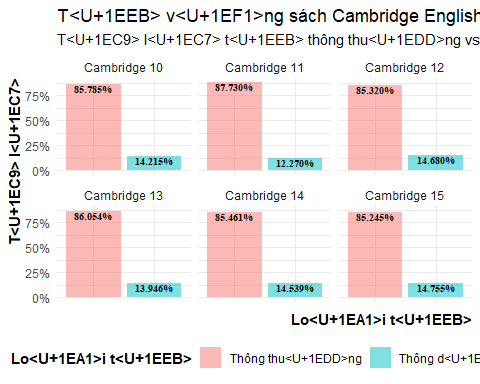


Nói một chút về nguồn sách mình sử dụng. Sách Cambridge English IELTS là sách có bản quyền nên nguồn sách có trên mạng tất cả đều là sách lậu. Sách này được chia sẻ ở hai dạng: một là PDF chữ và hai là PDF dạng scan hình ảnh. Trong số 6 cuốn sách mình sử dụng, chỉ duy nhất cuốn Cambridge 11 là ở dạng PDF chữ. Quá trình chuyển từ PDF chữ thành text để khai thác dữ liệu khá đơn giản và chính xác. Các cuốn sách còn lại đều phải đưa vào R để thực hiện việc chuyển file hình ảnh thành text, độ chính xác thấp hơn do chất lượng hình ảnh nhiều chỗ không tốt.

Biểu đồ ở ảnh bên cho thấy, số lượng từ vựng xuất hiện trong sách Cambdrige qua các năm xấp xỉ bằng nhau, riêng chỉ cuốn Cambridge 11 có lượng từ lớn hơn rất nhiều. Có hai lí do để giải thích. Một là như mình đã đề cập, cuốn 11 được chuyên từ PDF chữ nên độ chính xác cao hơn. Hoặc một lí do khác đó là cuốn 11 thực sự có lượng từ lớn hơn các cuốn khác.

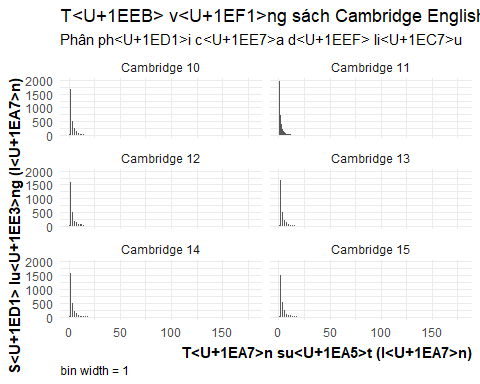


Tuy nhiên, khi khai thác ở một khía cạnh khác là tổng số từ (tức các từ vựng có lặp lại) thì lượng từ ở tất cả các cuốn sách đều ngang bằng nhau. Điều này giúp mình tin rằng cuốn 11 thực sự có lượng từ vựng nhiều hơn các cuốn khác, chứ không phải vì lí do R đã thiếu chính xác trong việc xử lí file hình ảnh thành text.



R có tích hợp một tập dữ liệu các từ thông dụng trong tiếng Anh (gọi là stopwords), trong đó bao gồm 1149 từ. Nếu mình không có chú thích thì tất cả các đồ thị mà mình thực hiện đều đã loại bỏ các từ thông dụng này bởi đây là những từ mà bạn chắc chắn phải học. Biểu đồ cột bên cho ta thấy, số lượng từ vựng thuộc nhóm từ thông dụng chỉ chiếm một phần nhỏ trong bộ từ vựng sử dụng cho bài thi IELTS (khoảng 14%). Có 533 từ vựng xuất hiện đồng thời trong cả 6 cuốn Cambridge IELTS (chiếm khoảng 15% tổng số từ). Nếu cộng cả hai loại trên thì cũng chỉ mới được khoảng 29% số lượng từ vựng của IELTS, một con số rất thấp.

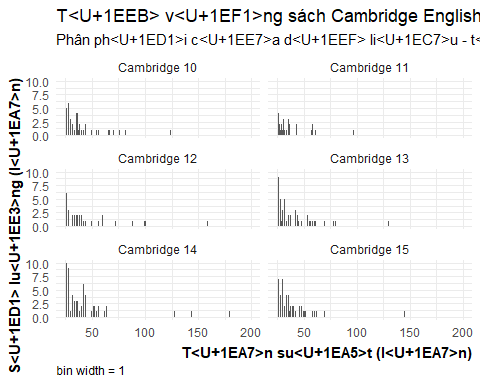
Nếu bạn có một vốn từ vựng ít thì mình nghĩ con số 29% này chính là điểm bạn nên bắt đầu vì đây là những từ rất rất thông dụng.



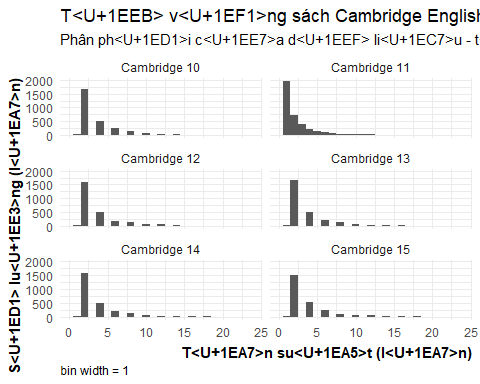
Mục đích của mình khi thực hiện việc khai thác dữ liệu chữ từ sách chính là có thể học số lượng tối thiểu từ vựng có khả năng xuất hiện tối đa trong bài thi IELTS. Hay nói ngắn gọn chính là học những từ có khả năng xuất hiện cao nhất.

Trước tiên, mình vẽ các histogram để xem xét sự phân bố của dữ liệu. Ở đây, mình muốn biết phân phối tần suất xuất hiện của từ vựng trong từng cuốn sách. Khái niệm tần suất và số lượng mình chú thích trên biểu đồ sẽ rất dễ gây nhầm lẫn nên mình muốn giải thích rõ hơn. Về giá trị tần suất, ví dụ, nhìn trên biểu đồ, bạn sẽ thấy có một cột rất cao. Cột đó có giá trị tần suất là 2. Nghĩa là, trong mỗi cuốn Cambridge, số lượng từ vựng xuất hiện 2 lần trong sách là rất nhiều (có trên 1500 từ như vậy). Có rất ít từ xuất hiện trên 25 và vì số lượng rất ít nên bạn gần như không thể thấy nó được biểu diễn trên biểu đồ. Vậy tại sao mình lại biết là có từ xuất hiện trên 25 lần mặc dù nó không được thể hiện trên biểu đồ? Đó là vì, ở trục x, giá trị tần suất được thể hiện đến con số 150. Tức là phải có một từ nào đó xuất hiện lớn hơn 150 lần nhưng vì số lượng rất ít nên gần như là không thấy.

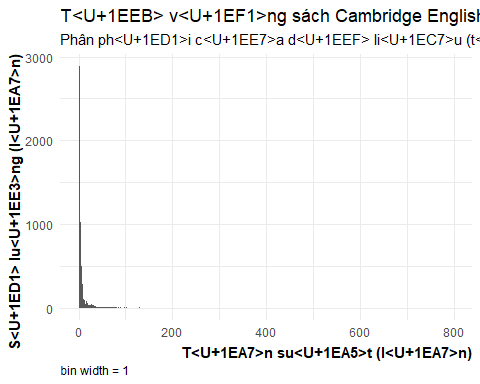
Để thể hiện rõ hơn, mình sẽ chia biểu đồ và phóng to vào từng vùng của biều đồ để nhìn số liệu được chính xác hơn. Có thể thấy từ biểu đồ chung thì hầu hết các từ vựng đều nằm ở vùng có tần suất nhỏ hơn 25. Do vậy, hai vùng được chia bao gồm: từ 0 - 25 và lớn hơn 25.



Trước tiên, chúng ta sẽ xem xét những từ vựng có tuần suất xuất hiện trên 25 lần trong một cuốn sách Cambridge. Cần biết là mỗi cuốn Cambridge bao gồm 4 test. Xuất hiện 25 lần tức là trung bình mỗi bài test từ sẽ xuất hiện 6 lần. Số lượng từ nằm trong nhóm này dù không nhiều nhưng là những từ nhất định phải biết bởi chúng rất thường xuất hiện trong bài thi IELTS. Cần phải nói thêm, trong tập dữ liệu sử dụng cho biểu đồ này, mình đã loại bỏ một số từ thường xuyên xuất hiện trong câu hỏi của đề thi (ví dụ từ questions xuất hiện đến 205 lần). Bạn cũng có thể thấy là số lượng từ và phân phối của từ giữa các cuốn sách rất tương đồng nhau (ở cả biểu đồ này và những biểu đồ trước đó). Điều này cho thấy, sách thực sự được làm rất chất lượng.

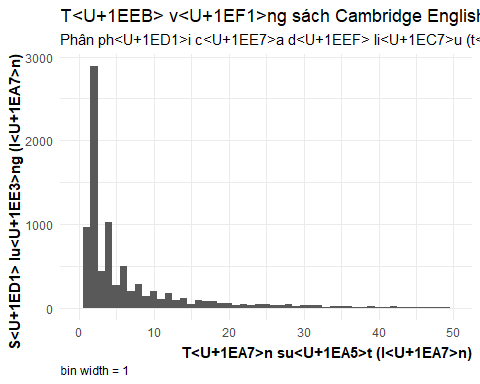


Ở đây, các biều đồ phân phối khá tương đồng nhau, chỉ riêng cuốn Cambridge 11 là có chút khác biệt. Mục đích của mình khi vẽ biểu đồ phân phối dữ liệu cho từng cuốn riêng biệt là nhằm xác định liệu có sự khác nhau hay không. Nếu có khác, mình sẽ phải thực hiện việc lọc dữ liệu cho từng cuốn riêng biệt. Nhưng vì sự tương đồng là rất lớn nên mình sẽ gộp chung toàn bộ từ vựng của 6 cuốn và thực hiện việc lọc dựa trên một tập dữ liệu lớn.



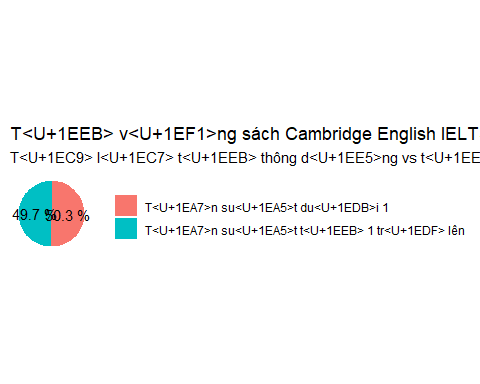
Giải thích thêm về biểu đồ. Mình đã gộp chung từ vựng của tất cả 6 cuốn Cambridge làm một. Giá trị tần suất các từ trùng nhau của các cuốn sách sẽ được cộng gộp. Ví dụ, từ ‘fun’ xuất hiện 1, 4, 7, 3, 8, 2 lần trong lần lượt các cuốn sách từ Cambridge 10 - 15. Vậy từ ‘fun’ đã xuất hiện tổng cộng 25 lần. Đó cũng là giá trị tần suất sẽ được ghi nhận trên biểu đồ.

Nhìn qua biểu đồ bạn có thể thấy hầu hết từ vựng nằm tập trung ở khoảng tần suất từ 0 đến 50. Do đó, mình sẽ phóng to vào khoảng này để tìm hiểu thêm.



Đến đây thì toang hẳn rồi :v. Nếu quay lại các biểu đồ trước, khi phân tích riêng từng cuốn sách, bạn sẽ thấy là tần suất xuất hiện nhiều nhất là 2 lần. Mình đã hi vọng rằng dù tần suất xuất hiện là 2 nhưng nếu các từ trùng lặp từ cuốn này sang cuốn khác thì vẫn được xem là từ thông dụng. Nhưng không, bọn nó chỉ xuất hiện đúng 2 lần thật. Nếu chỉ nhìn bằng mắt thường, chỉ riêng ở giá trị tần suất 1 và 2 đã chiếm tới gần 4000, tức chiếm tới gần 45% tổng số từ vựng (từ vựng trong 6 cuốn sách là 8703 từ, không bao gồm từ thông dụng). Và nếu giả sử, một từ được coi là thông dụng khi nó xuất hiện ít nhất một lần trong một bài test, tức là 24 lần trong 6 cuốn sách thì lượng từ vựng có tần suất nhỏ hơn 24 chiếm khoảng 90% tổng lượng từ :((

Tóm lại là cố gắng trau dồi từ vựng đi các bạn ạ. Không ăn gian được đâu :))



Tin tốt lành cuối cùng. Mặc dù chỉ chiếm khoảng 10% nhưng những từ vựng thông dụng chiếm tới 50% tổng lượng từ xuất hiện.

Mặc dù kết quả không như mong đợi nhưng dù gì cũng đã chạy đông chạy tây được đến đây nên mình vẫn tạo một bảng từ vựng các từ xuất hiện trong sách Cambridge English IELTS 10 - 15. Hi vọng nó sẽ có ích nếu bạn cần một bảng từ vựng để sử dụng cho Anki chẳng hạn.

File từ vựng sẽ bao gồm:

* stopwords: 1149 từ vựng thông dụng trong tiếng Anh
* greater than 1: các từ vựng có tần suất xuất hiện trung bình từ 1 lần trở lên trong mỗi cuốn sách
* less than 1: các từ vựng có tần suất xuất hiện trung bình dưới 1 lần trong mỗi cuốn sách