**TRUY XUẤT THÔNG TIN BẰNG MÔ HÌNH XÁC SUẤT BM25**

1. **TF \* IDF là gì?**

TF \* IDF là một cách gần đúng để ước tính cách người dùng đánh giá mức độ liên quan của một kết hợp văn bản. Trực giác cơ bản TF \* IDF khá rõ ràng và dựa trên hai yếu tố chính được nhúng trong tên của công thức tính điểm có xu hướng tương ứng với cách trí óc con người có xu hướng đánh giá mức độ liên quan của tìm kiếm:

* Tần suất thuật ngữ hay còn gọi là tf: “dog” xuất hiện trong bài viết thường xuyên như thế nào? 3 lần? 10 lần?
* Tần suất tài liệu nghịch đảo hay còn gọi là idf: Tần suất tài liệu đo lường số tài liệu mà một thuật ngữ xuất hiện. Tần suất tài liệu nghịch đảo (1/df) sau đó đo lường mức độ đặc biệt của thuật ngữ. Có phải thuật ngữ “chó” rất hiếm (chỉ xuất hiện trong một tài liệu) không? Hoặc tương đối phổ biến (xảy ra trong gần như tất cả các tài liệu?).

Nói cách khác, TF \* IDF đo lường mức độ tập trung tương đối của thuật ngữ trong một đoạn văn bản nhất định. Nếu “chó” phổ biến trong bài viết này, nhưng tương đối hiếm ở những nơi khác, thì điểm TF \* IDF sẽ cao. Bài viết này nên được coi là rất liên quan đến cụm từ tìm kiếm “chó”. Nếu “con chó” xuất hiện một lần ở đây, nhưng rất nổi bật trong nhiều tài liệu khác, điểm của nó sẽ tương đối thấp.

Một thước đô bổ sung là độ dài của văn bản. “Con chó” xuất hiện hai lần trong một cuốn sách 500pg hầu như không nói gì về mức độ mà cuốn sách đó viết về “con chó”. Tuy nhiên, “Dog” xuất hiện hai lần trong một tweet ngắn có nghĩa là tweet đó nói rất nhiều về “dog”! Do đó, sự thiên vị bổ sung được giới thiệu được gọi là “fieldNorms”. Sự đóng hóp này mang lại sự thiên vị đáng kể đối với việc khớp các tài liệu ngắn hơn so với các tài liệu dài hơn. Các thuật ngữ được “tập trung” nhiều hơn trong tài liệu ngắn hơn. Do đó, tài liệu ngắn hơn có nhiều khả năng là về cụm từ được tìm kiếm hơn và do đó sẽ được cho điểm cao hơn.

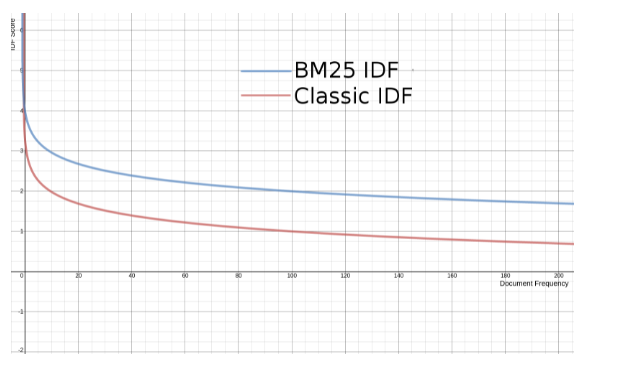
1. **Giới thiệu**

Trong tìm kiếm thông tin, Okapi BM25 là hàm tính thứ hạng được các công cụ tìm kiếm sử dụng để xếp hạng các văn bản theo độ phù hợp với truy vấn nhất định. Hàm xếp hạng này dựa trên mô hình xác suất, được phát minh ra vào những năm 1970-1980. Phương pháp có tên BM25 (BM – best match), nhưng người ta thường gọi “Okapi BM25”, vì lần đầu tiên công thức được sử dụng trong hệ thống tìm kiếm Okapi, được sáng lập tại trường đại học London những năm 1980 và 1990.

BM25 cải thiện dựa trên TF \* IDF. BM25 là viết tắt của “Best Match 25”. Được phát hành vào năm 1994, đây là lần lặp lại thứ 25 của việc điều chỉnh tính toán mức độ liên quan. BM25 có nguồn gốc từ việc truy xuất thông tin xác suất. Truy xuất thông tin xác suất là một lĩnh vực hấp dẫn đối với chính nó. Về cơ bản, nó coi mức độ liên quan như một bài toán xác suất. Điểm liên quan, theo khả năng truy xuất thông tin xác suất, phải phản ánh xác suất mà người dùng sẽ coi kết quả phù hợp.

1. **IDF**

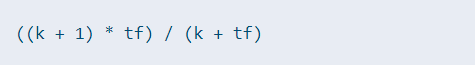
Trên biểu đồ, IDF của BM25 trông rất giống với IDF Lucence cổ điển. Lý do duy nhất cho sự khác biệt ở đây là nó xuất phát từ việc truy xuất thông tin xác suất. Lucence thực hiện một thay đổi đối với IDF thông thường của BM25. IDF của BM25 có khả năng cho điểm tiêu cực đối với các điều khoản có tần suất tài liệu rất cao. Vì vậy, IDF trong BM25 của Lucene thực hiện một thủ thuật tuyệt vời này để giải quyết vấn đề này. Họ thêm 1 vào giá trị, trước khi lấy điều này khiến không thể tính giá trị âm. Kết quả cuối cùng là một IDF trông cực kỳ giống vói đường cong IDF hiện tại của Lucene, như thê hiện ở biểu đồ sau.



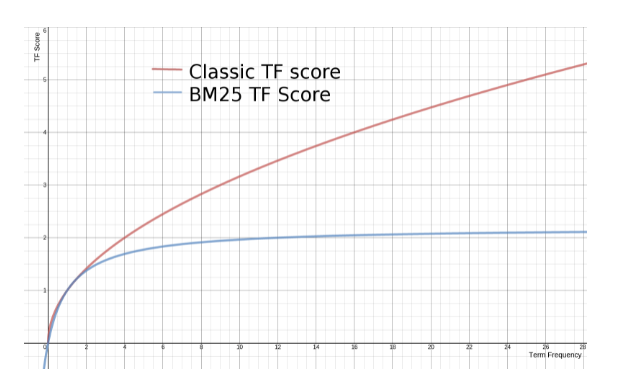
1. **TF**

Tần suất kỳ hạn trong BM25 làm giảm tác động của tuần suất kỳ hạn hơn nữa so với TF\*IDF truyền thống. Tác động của tần suất thuật ngữ luôn tăng, nhưng tiệm cận một giá trị.

Không tính đến độ dài của tài liệu, tần suất thuật ngữ tính theo công thức



như được vẽ dưới đây:



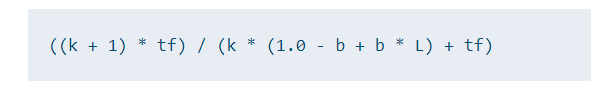
Như bạn thấy, đường cong này tiếp cận (k + 1) tiệm cận (ở đây là k = 1,2). Nó có một hiệu ứng thực sự thú vị. Nhiều hơn tf luôn có nghĩa là nhiều liên quan hơn. Tuy nhiên, bạn nhanh chóng đạt được lợi nhuận giảm dần. Bạn không bao giờ vượt qua k, nhưng bạn luôn tiếp cận nó! Mặt khác, tf Lucene cổ điển liên tục tăng và không bao giờ đạt đến điểm bão hòa.

Giá trị này là bao nhiêu? Đối với BM25, k thường được đặt thành 1,2. Đa số để lại k yên. Thay đổi k có thể là một cách tiếp cận điều chỉnh hữu ích để sửa đổi tác động của TF. Điều chỉnh k rõ ràng làm cho đường tiệm cận di chuyển. Tuy nhiên, điều quan trọng hơn là k cao hơn khiến TF mất nhiều thời gian hơn để đạt được độ bão hòa. Bằng cách kéo dài điểm bão hòa, bạn kéo dài sự khác biệt về mức độ liên quan giữa các tài liệu có tần suất thuật ngữ cao hơn và thấp hơn!

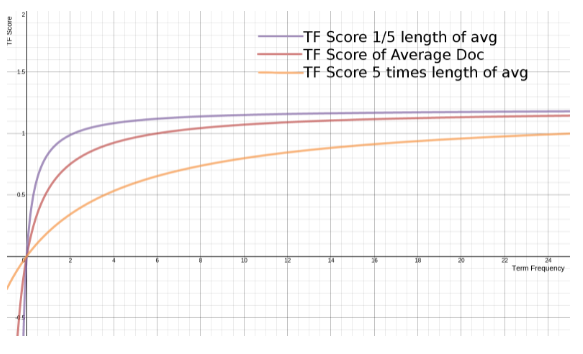
1. **BM25 sử dụng độ dài tài liệu như thế nào?**

Điểm TF ở trên còn bị ảnh hưởng bởi việc tài liệu cao hơn hay thấp hơn độ dài trung bình của tài liệu trong kho ngữ liệu.

Điều này như thế nào? Chúng ta hãy xây dựng dựa trên công thức TF từ trước, giới thiệu hai biến: một hằng số b và một giá trị độ dài L. Lấy công thức trên và cộng (1.0 – b + b \* L) dưới dạng bội số của k ở mẫu số.



Đây L là độ dài của một tài liệu so với độ dài trung bình của tài liệu. L là 2 nếu tài liệu được cho điểm gấp đôi độ dài tài liệu trung bình của kho tài liệu. L là 0,1 nếu tài liệu được cho điểm bằng một phần mười độ dài tài liệu trung bình. L do đó thực sự được trình bày dưới dạng | d | / avgDl – độ dài tài liệu này chia cho độ dài tài liệu trung bình.



Như ta có thể thấy trong biểu đồ, kết quả cuối cùng cho các giá trị khác nhau L là các tài liệu ngắn hơn đạt được tiệm cận nhanh hơn nhiều. Họ gần như bão hòa ngay lập tức để đạt được điểm TF tốt nhất có thể. Điều này có lý, các tài liệu ngắn có ít thuật ngữ hơn. Càng nhiều kết quả trùng khớp trong các tài liệu ngắn này, ta càng cảm thấy chắc chắn về mức độ liên quan. Vì vậy, số lượng tăng nhanh hơn. Mặt khác, một cuốn sách dài lê thê, cần phải có nhiều kết quả phù hợp nữa để đi đến điểm mà chúng ta có thể cảm thấy tự tin. Vì vậy, việc đạt được “mức độ liên quan tối đa” mất nhiều thời gian hơn.

Hằng số b sẽ cho phép chúng ta tinh chỉnh mức độ ảnh hưởng của L giá trị của chũng ta đối với việc tính score. Lưu ý trong công thức trên, a b của 0 loại bỏ hoàn toàn ảnh hưởng của L, trở lại công thức của phần trước. b Độ dài tài liệu cao hơn sẽ ảnh hưởng nhiều hơn đến score. Nói cách khác, trong cổ điển TF\*IDF ta luôn tắt định mức trên một trường để loại bỏ ảnh hưởng của độ dài trường. Ở đây, ta có thể chỉ cần đặt b thành 0 trên độ tương tự để loại bỏ tác động của độ dài trường.

Khi BM25 được kết hợp với nhau:

IDF \* ((k + 1) \* tf) / (k\* (1.0 - b + b \* (|d| / avgDl)) + tf)