Harry Potter ve Felsefe Taşı Kitabının İngilizce ve Türkçe Versiyonları Üzerinde Zipf Yasasının Uygulanması

Nisanur Bulut ICEANS 2022

Zipf Yasası

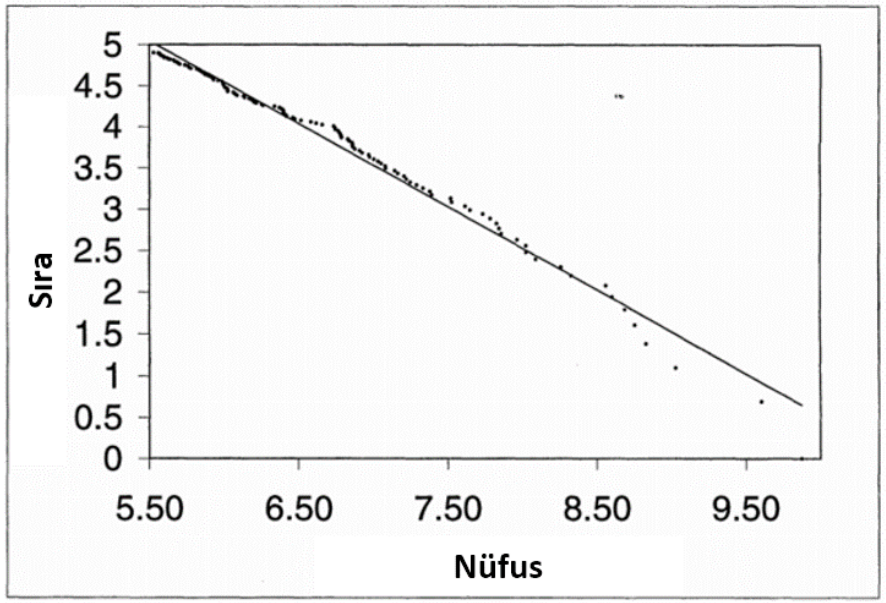
Zipf’e göre, içeriği ve dili fark etmeksizin, bir metindeki kelimeleri kullanım sıklığına göre azalan şekilde sıraladığımızda, belli bir örüntüye sahip liste elde edilir. Listedeki bir kelimenin kullanım sırasının sıklık ile çarpımı sabit bir sayıyı verir.

Bu yasa ilk olarak ilk olarak James Joyce'un Ulysses adlı eserinde tatbik edilmiştir, şöyledir;

- en çok kullanılan sözcük metinde,, 2653 kez kullanılırken
- en çok kullanılan,, 265 kez
- en çok kullanılan,, 133 kez kullanılmıştır.

Sosyal Yaşamda Zipf Yasası Örneği

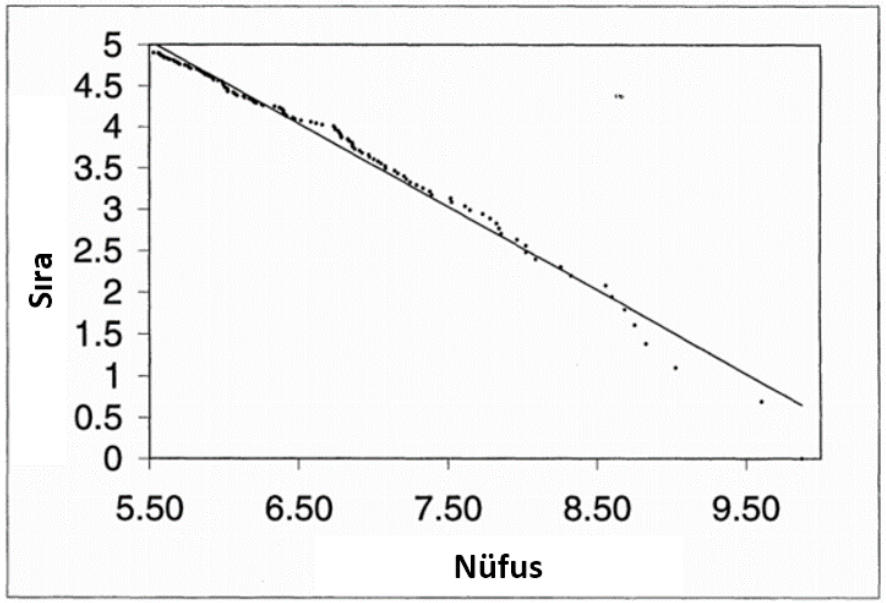
- Zipf'in bulgularına göre sözcüklerin büyük çoğunluğu çok nadiren kullanılırken az sayıda birtakım sözcük her zaman kullanılır.
- Yasa yalnızca bir insan diline ait değildir şimdiye değin içeriği pek çok farklı konuda olan pek çok farklı dildeki metinlerde denenmiştir.
- Sonrasında fark edilmiştir ki, Zipf yasası insanların gelir dağılımlarına ya da şehir nüfus oranlarına uygulandığında da aynı sonucu vermiştir. Örneğin, Bir ülkedeki en çok nüfusa sahip şehir genellikle bir sonrakinin iki katı kadardır



Şekil 1
1991'de ABD Metropoliten Alanlarının en büyük 135'inin Büyüklük-Sıra Grafiği
Kaynak: Statistical Abstract of the United States [1993]

Sosyal Yaşamda Zipf Yasası Örneği

- Birleşik Devletler'deki nüfusa göre en üst sıralarda yer alan şehirlere şöyle bir bakıldığında 2010 nüfus sayımında ABD'deki en büyük şehir olan New York'un nüfusu 8.175.133 iken ikinci sıradaki Los Angeles'ın nüfusu 3.792.621 ve sonraki üç sırada Chicago, Houston ve Philadelphia yer aldığı görülür. Şehirlerin nüfus sayıları sırasıyla 2.695.598, 2.100.263 ve 1.526.006'dır. Sayıların tam olmadığını açıkça görülmektedir fakat istatistiksel olarak bakıldığında Zipf'in öngörülerini kayda değer biçimde tutar. Yasanın sağlandığını görmek amacıyla, logaritmik fonksiyonda çizilebilir.



Şekil 1
1991'de ABD Metropoliten Alanlarının en büyük 135'inin Büyüklük-Sıra Grafiği
Kaynak: Statistical Abstract of the United States [1993]

Sıklık Kuralı

Bir metindeki benzeri olmayan kelimeler, yani bir kelimeyi birden fazla kez saymamak şartıyla, metinde kaç kez geçtiği sayılır. Freq(word)=Metinde bir kelimenin kaç kez kullanıldığını gösterir.

Olasılık Kuralı

Bu sıklık listesine göre, her kelime azalan sırada derecelendirilir. Ençok kullanılan kelimenin derecesi(rank) 1 iken ondan sonra gelen en çok kullanılan kelimenin derecesi(rank) 2 olur. Kelimenin anlamıyla ilgilenmeyiz yalnızca sıklık ve derece değerleriyle ilgileniriz.

r=derece(rank)

N=Metinde geçen kelime sayısıdır. Benzersiz kelime sayısı değildir yani bir kelime metin içinde bir den fazla geçebilir ve sayıma dahil edilir.

Prob(r) = freq(r) / N Zipf yasasına göre; r * Prob(r) = A

Bu formüle göre, A sabit bir sayıdır ve çoğunlukla 0.01 değerini alır. Zipf yasası istatistiksel bir çıkarımdır her zaman aynı değeri verir şeklinde kesin bir çıkarımı yoktur ancak her zaman yakınsal sonuçlar vermesi beklenir.

Prob(r)=freq(r)/N formülünü dikkate alır ve
Zipf yasasını yeniden yazarsak ;

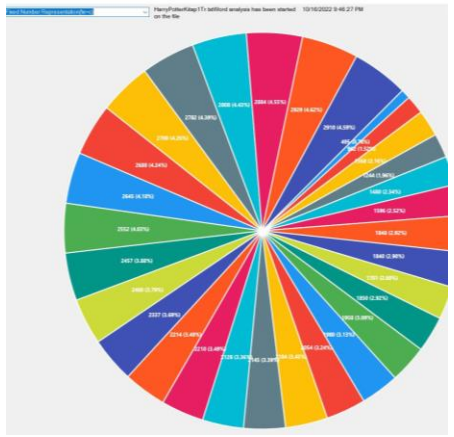
r * freq(r) = A * N formülü elde edilir.

Zipf Yasası Genel Uygulama Yaklaşımı

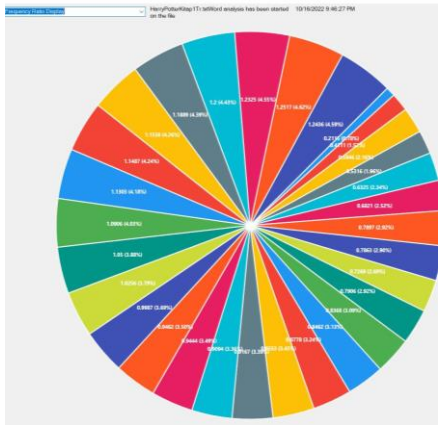
Zipf yasasını uygulayacak olursak, elde etmemiz gereken ilk şey freq(r) değeridir. Bu değeri derece yani r ile çarptığımızda elde edeceğimiz sonuçlar yaklaşık olarak aynıdır. Bunun anlamı, her kelime için r*freq(r) değerinin aynı olacağı değildir. Ama sonuçların birbirine yakın olması beklenir.

•Sıklık değeri en çok ya da en az olan kelimelere ait sonuçlara bakmak aldatıcıdır. Bunlara dair sonuçlar genelde en fazla hataya sahip sonuçlardır. İdeal gözlem için, kelimelerin derecesi 1,10,100,200… şeklinde olacak şekilde, Zipf yasasına olan uygunluklarına bakılmalıdır. Zipf yasası için ne uygun gösterim nokta grafikleridir. Sonuçlar logaritması alındıktan sonra görüntülendiğinde daha net görülecektir.[5]

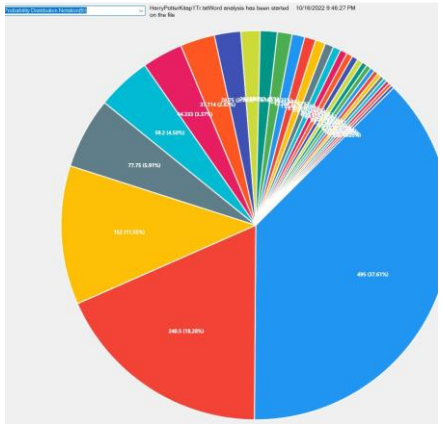
Ek ve Kelime Dağılım Göstergeleri (Türkçe)



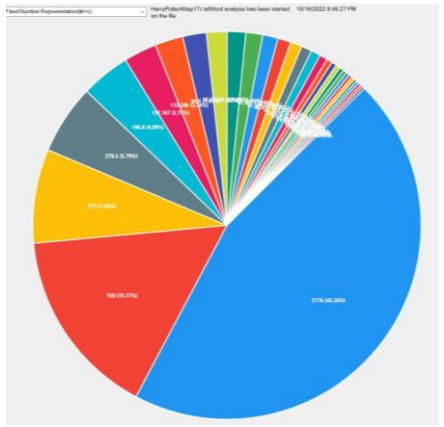
Fixed Number Representation (fxr=c)



Frequency Ratio Display

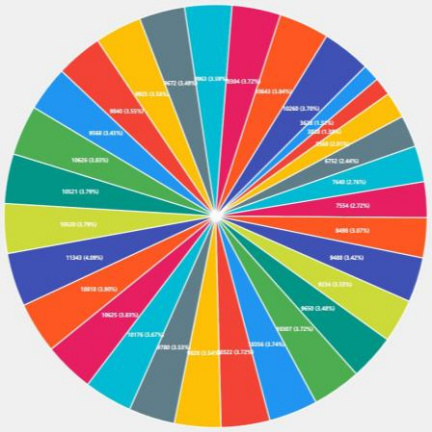


Probability Distribution Notation(f/r)

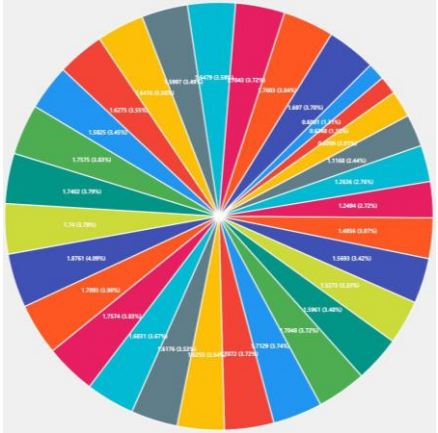


Türkçe Metin Üzerinde Ek Dağılımlarının f/r Gösterimi

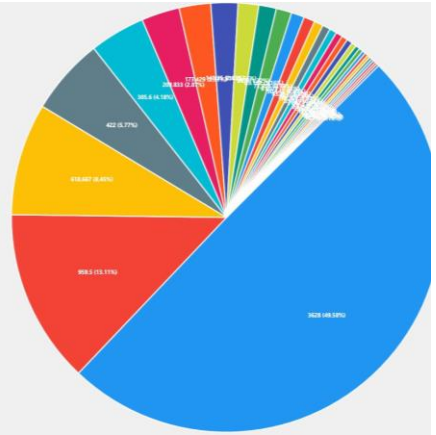
Kelime Dağılım Göstergeleri (İngilizce)



Fixed Number Representation (fxr=c)



Frequency Ratio Display



Probability Distribution Notation(f/r)

Ekosisteme baktığımızda kendi içinde istikrar taşıyan pek çok yasa vardır örneğin altın oran, pi sayısı, Euler sabiti. Bu yasalar, bir düzenle ortaya konan yasalar değil keşfedilen yasalardır. Tabiatta, hayatımızda mahiyetini kavramakta zorlandığımız acayip bir düzen var. Yüksek zekâlar bu sırları peşinde tecessüs ile koşmaya devam ediyorlar.

Zipf yasasının şehirlere uygulanmasındaki sonuçların uygunluğu ya da altın oranın varlığını algılamak zor olmasa gerek çünkü doğal bir işleyiş, canlılara ait bir yaşam parçasından bahsediliyor. Ancak metin analizlerine bakıldığında dil ve içerik fark etmeksizin yasasının ürettiği sabit sonuçların neden elde edildiği gizemi hala korunmaktadır.

Bu sıra sıklık kuralının sebebi olarak genlerimizde gizlenmiş olan seçim eğilimleri gösterilebilir. Ama burada da bir çıkmaza girilmektedir çünkü bu seçim eğilimlerinin de neden ve nasıl hareket geçtiği hala bir gizem konusudur. Belki de zipf yasası bir nedene bağlı sonuç olmaktan ziyade, toplumsal davranışlarımız, iletişimlerimiz, ticari meselelerimizin sonucunu gösteren bir yöndür.

Yalnız şu bir gerçektir ki zipf yasasının işleme mekanizmasının çözülmesi, insan eğilimlerinin ne olduğundan çok daha fazlasını ön görebilecektir.

Teşekkürler

“Tabiat kanunları tanrının eliyle matematik dilinde yazılmışlardır." "Tabiatı anlamak için onun yazıldığı dili anlamanız gerekir ve o dil de matematiktir.”

Galileo Galilei