# Sözlükte Bulunmayan Kelimelere Merkür Gerilemesi Döneminde Amazon Taramalarında Daha Fazla Rastlanıyor

*Renay Oshop*
*Çeviren: Gamze Tetik*

## 1. Giriş

Benim kültürümde Merkür gerilemesi büyük endişe duyulan bir zamandır.

Astronomik gerilemelerin kişilerin iletişimi, seyahatleri, elektronik eşya kullanımı, düşünce şekilleri üzerinde etkili olduğuna inanılır.

Merkür gerilemesinin iletişim üzerindeki etkisini test etmek için 1 Haziran 2019 ve 23 Temmuz 2014 tarihleri arasında 69.242,585 görüntülemenin rasgele %5'ini inceledim. (Amazon.com: Elektronik, giyim, bilgisayar, Kitap, DVD & daha fazlası için internet alışveriş verisi)

Her arama girişindeki bilgisayarlı sözlükte olmayan kelimeleri saydım ki indirim rakamları, internet uzantıları, emojiler, üç veya daha az karakterden oluşan kelimeler de buna dahil. Ayrıca İngilizce olmayan arama girişleri de dahil edilmedi.

Ayrıca, bağlam hataları, gramer, noktalama ve büyük küçük harf hataları göz önüne alınmadı. Daha çok Merkür gerilemesi ve Merkür gerilemesi olmayan dönemlerdeki sözlükte bulunmayan kelimelerin kullanımını karşılaştırmak istedim.

## 2. Kaynaklar

Milyonlarca Amazon görüntülemesinin kaynağı olarak Stanford Üniversitesi'nin büyük verisi kullanıldı. (Jure Leskovec and Andrej Krevl 2014) Amazon verisi, araştırmam için kibarca kullanımıma açan Julian McAuley tarafından yönetildi. (J. McAuley 2015)

Tamamen tekilleştirilmiş versiyonu olan, hiç çoğaltılmamış bir veri seti seçtim. Bu veri seti 23 Temmuz 2014'e kadar olan 82.84 milyon Amazon ürün görüntülemesini içeriyor.

## 3. Metodlar

Yazılım olarak *Mathematica,* standard veri bilimi ve sinyal işleme tekniklerini uygulamak için kullanıldı. (Wolfram Research, Inc. 2015) Algoritmalar herkes tarafından görüntülenebilir. (Oshop, Zenodo 2015)

Öncelikle, 82,469,759 girdinin %16'sı veritabanındaki standarda uygun olmadığı ve Mathematica ile değerlendirilemeyeceği için temizlendi. Bu kadar büyük bir verinin otomatik analizi başarının birincil anahtarı olacağından, kaybedilen yüzde konusunda rahattım. Ancak, Amazon ürün departmanının gönderim zamanı gibi kayıp girişlerle ilgili önemli bir tema olup olmadığını bilemiyorum. Bu kadar geniş veri setlerini (burada 57 GB birleştirilmemiş) kullanmak, özellikle daha eski bir veri kümesi için en iyi teknolojilerin en iyi operatörleri tarafından bile hala sanat biçimidir.

Bu yüzden, Unix formatında orijinal girişin zamanı ve düzenlemeler dahil gözden geçirme metni olmak üzere her satırda iki giriş olacak şekilde 69,242,585 satırlık bir tablo oluşturulmuştur. Bu veri tabanında gözden geçirme düzenlemelerinin zamanlamasına dair kayıt bulunmamaktadır.

Ardından, tüm girişlerin rastgele seçilmiş %5'ini analiz etmek için rastgele sayı üreteci ve sözlük kullanan bir değerlendirme programladım. (DictionaryLookup Kaynak Bilgileri- Wolfram Language Documentation n.d.)

Her girdinin incelemesi metinsel sözcük alt dizilerine bölündü. Emojileri ve bağlantıları bulmak için kelime dizilerinin karakter karakter analiz edilmesi gerekiyordu.

Ardından, eğer sözlük kelime dizisiyle eşleşme bulamazsa, o giriş 'yazım hatası' olarak sayıldı. Bu giriş yanlış yazımlarının toplamı, girişteki toplam metinsel kelimelerin sayısına bölündü ve sonuç, orijinal gönderim günü ile birlikte kaydedildi.

Yalnızca %5'lik rastgele örneklemeyle bile, bu programı masaüstü bilgisayarımda çalıştırmak, saniyede 1.000 girişlik bir analiz hızında yaklaşık 31 saat sürdü.

Bu tekniğin bir ürünün olağandışı bir marka adını, örneğin herhangi bir genel eğilime bir tür ses olarak tanıtılacak bir yanlış yazım olarak kabul edeceğini unutmayın.

Ortaya çıkan girişler tarihe göre sınıflandırıldı. Bir günün grup ortalaması o tarihle kaydedildi ve saklandı. 1 Ocak 2009 ve 23 Temmuz 2014 arasındaki tüm tarihler için bu değerler bir zaman serisinde oluşturuldu. Zaman serisine bir ortalama filtre uygulandı ve sonuçlar onunla birikte çizildi.

Her gün için gerçek ortalama değer ile filtreden gelen eğilim arasındaki fark, çıkarma yolula bulundu ve çizildi.

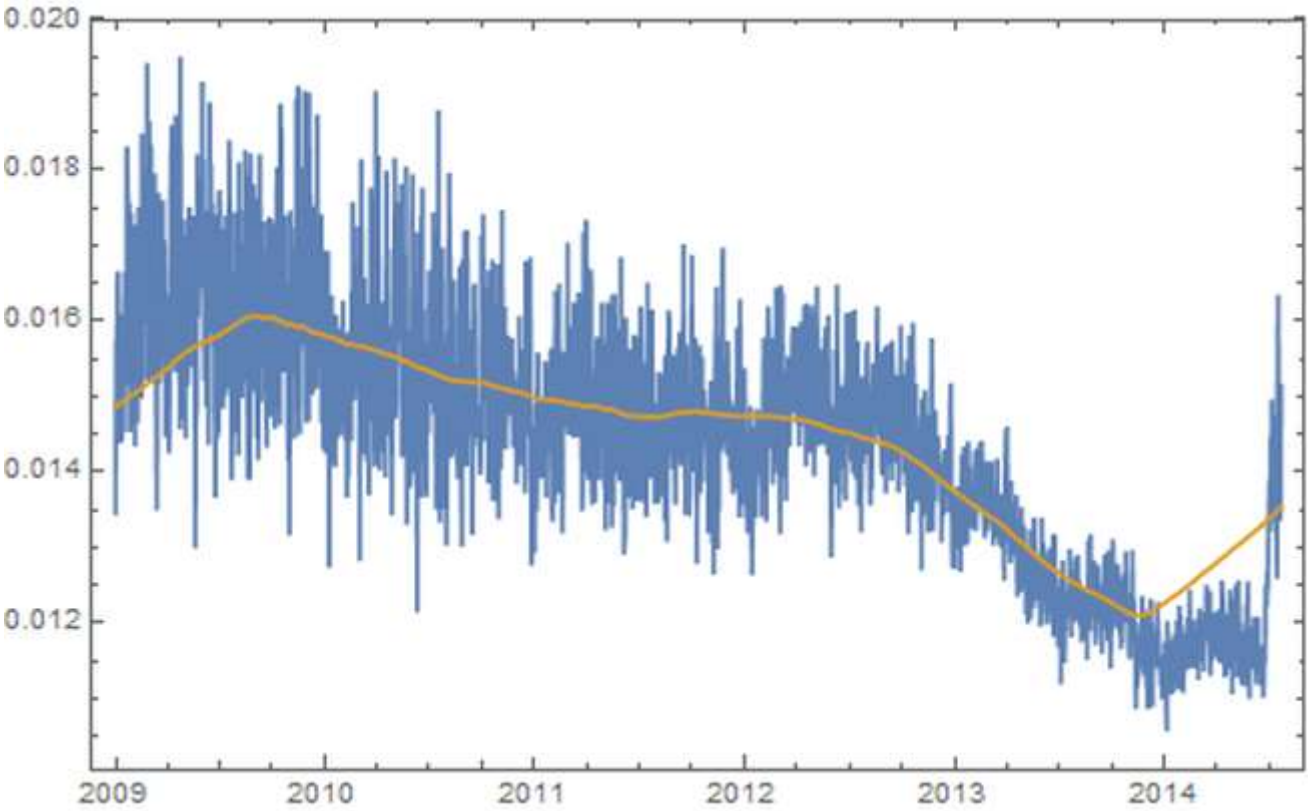
Çeşitli dönemler ortaya çıktı. Bu çeşitlilik, ayrı bir Fourier dönüşümü ile analiz edildi. En yüksek iki değerin katkısından gelen ana yapının grafiği çizildi ve Merkür retrosunun düzenliliğine iyi bir şekilde karşılık geldiği bulundu.

Aşağıdaki resimde, ince dikey şeritler, o zaman diliminde Merkür gerilemesinin gerçek sınırlarını göstermektedir.

# 4. Sonuçlar

## A. Günlük Ortalamaların Hareketi

Mavi çizgi, gerçek günlük ortalamaların yukarı aşağı hareketini tanımlar. Daha açık renk olan altın sarısı çizgi genel eğilimi gösterir. Bu genel eğilim, mavi veriler üzerinde bir ortalama filtre ile elde edildi.

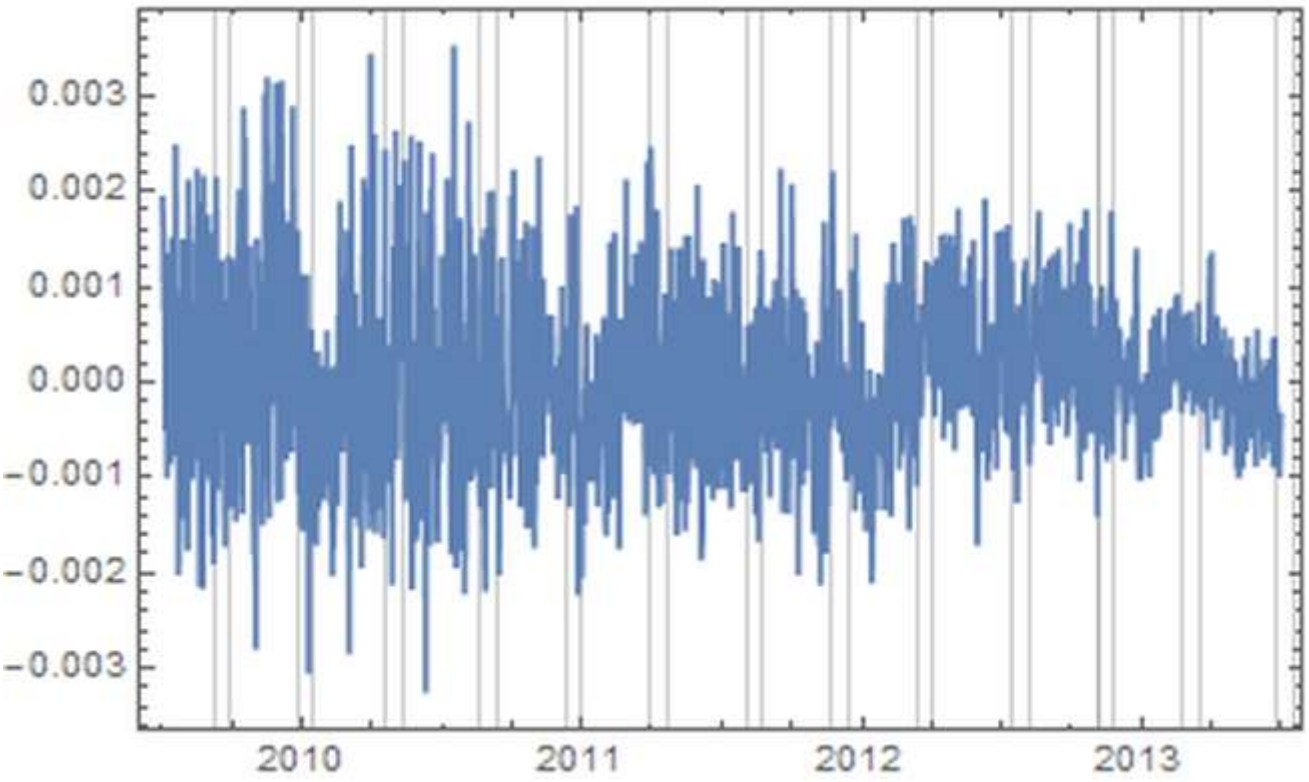



Şekil 1: Zaman içerisinde hata oranındaki ortalamaların trend grafiği (altın rengi) ve gerçek ortalama günlük hata oranları (mavi renkte)

Ortalama filtrenin doğası gereği sağ ve sol uç bölgeler yukarıdaki grafikte görebileceğiniz kadar doğru değildir ve değerlendime dışı bırakılmıştır. Buna göre aşağıdaki analizler yalnızca 1 Temmuz 2009 ve 1 Temmuz 2013 arasındaki tarih değerlerini içerir, yine de dört yıllık oldukça uzun bir süre olduğugöz önüne alınmalıdır.

B. Günlük Ortalamaların Genel Trendden Farkları

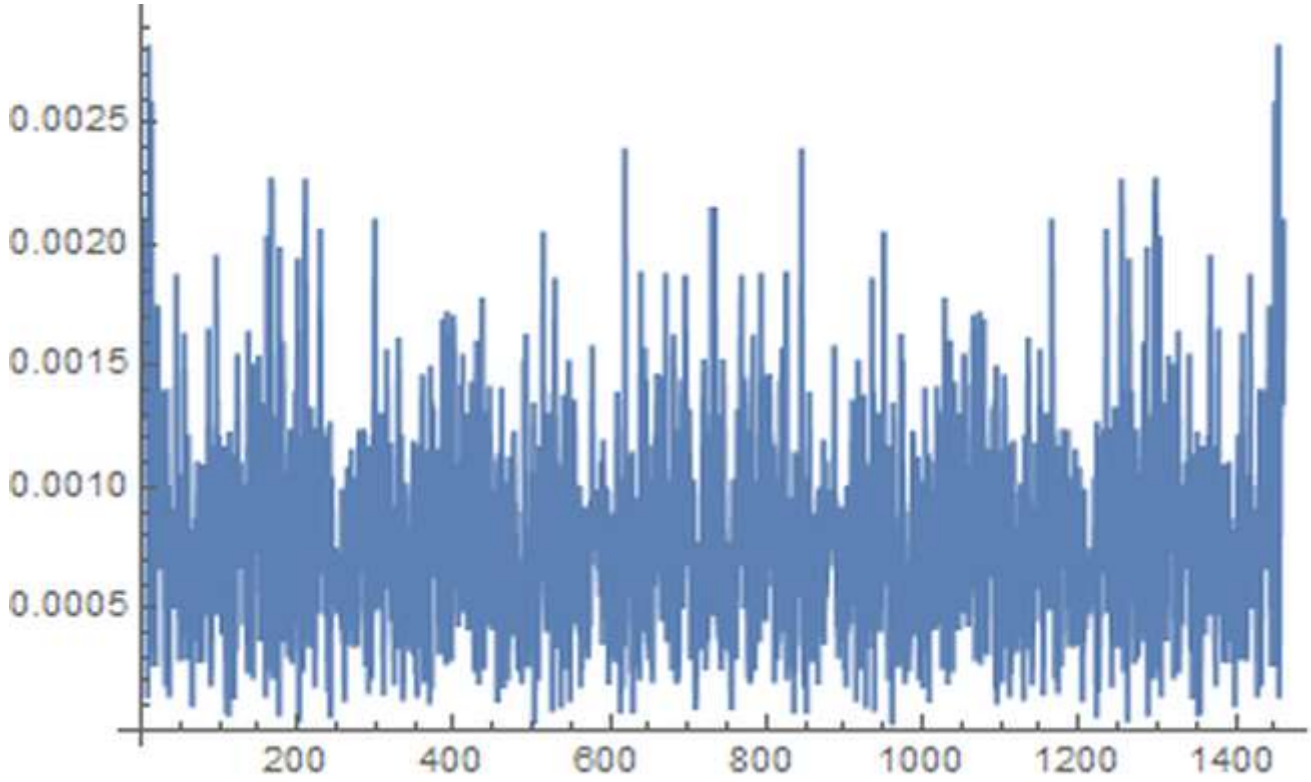
Trend filtre değeri (Şekil 1'deki altın rengi) her gün için ortalama filtre değerinden (mavi renkli) çıkarıldı ve sonuç çizildi.




Şekil 2: 1 Temmuz 2009 ile 1 Temmuz 2013 arasındaki tarihler için gerçek günlük ortalama değerlerden (Şekil 1'de mavi) trendin (Şekil 1'de altın) çıkarılması

Gerçek günlük değer ile genel eğilim arasındaki bu farklılıklar, ayrık Fourier dönüşümüne tabi tutuldu. Spektograf aşağıdadır.

1. Farkların Ayrık Fourier Dönüşümü Grafiği

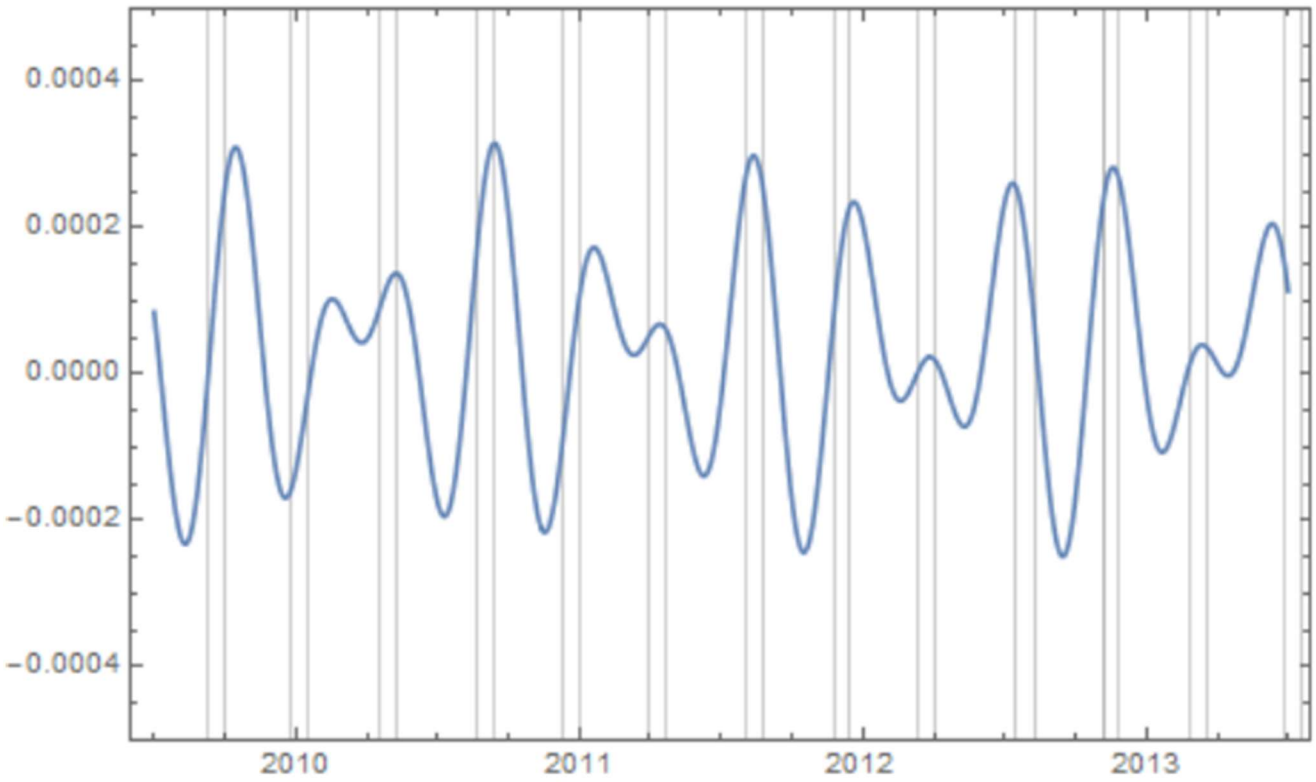


Şekil 3: Farkların Ayrık Fourier Dönüşümü Grafiği – Şekil 2'de görüldüğü gibi

Temel ilk çizgi, yatay eksen boyunca 9'da en yüksek ve ikinci ana tepe 13'te. Bu sayılarla, yükseklikleriyle birlikte, verilerin ana seyrini, yani Amazon incelemelerinde günlük yazım hataları oranlarının yükselmesine ve düşmesine en çok katkıda bulunan dalgayı oluşturabiliriz.

## 2. Temel ve Birinci Büyük Harmonik Toplamının Yapısı

Gördüğünüz gibi, her Merkür gerilemesi öncesi sözlük dışı kullanım oranının yerel ve küresel anlamda minimumda ve Merkür gerilemesi döneminde ise yerel ve küresel olarak maksimumda olduğu görülür.



Şekil 4: Şekil 3'teki Fourier dönüşümünün ilk büyük harmoniği ile temel grafik

# 5. Tartışma

Yazılımın insan sesinin, ses klibi grafiğine çok benzeyen Şekil 2, ilk bakışta büyüleyici görünebilir ancak yorumlanması zordur.

Ses grafiği söz konusu olduğunda, temsil ettiği sesin tamamı oradadır, fakat görsel resimden tekrar sese nasıl geri dönersiniz? Sonuçta sesin kaydedilmesi, sıkıştırılması, iletilmesi ve yeniden oluşturulması bu tür yazılımlardan gerçekleşir.

Matematiksel bir yol olmalı ve var. Son yüzyılın en önemli bilim araçlarından birisi Fourier dönüşümüdür.

Birisinin kusursuz bir piyanoda saf C notasını çaldığını düşünün, Fourier dönüşümü bu notayı Şekil 3'e çok benzeyen bir grafiğe böler, sadece solda tek bir çizgi olurdu (aynası sağda olacak şekilde) ve temel olarak da C'nin frekansıyla eşleşen olarak adlandırılır.

Doğada işler daha karmaşık olma eğilimindedir. Gerçek bir piyano, büyük ihtimalle büyük olanı oluşturan ve değiştiren çok küçük alt harmoniklere sahip olacaktır. Büyük olan hala mükemmel C'nin sinüzoidal dalgasının tam şekline sahip olacaktır, fakat bu küçük frekanslar bu ilk baskın eğriyi, saf bir C notasından daha karmaşık, ancak buna benzeyen bir son dalga formuna ayarlayacaktır.

Bu daha küçük notalar seti, Fourier dönüşüm grafiğinde de görülür, ancak daha kısa

çizgiler olarak ve yatay eksen boyunca frekanslarının her birine karşılık gelen yerleşimlerde, yükseklikler birinci notanın değişiklik sırasını verir. Tipik olarak tüm dalga sadece birkaç satırda özetlenebilir. Sesteki gürültünün sıkıştırılması ve giderilmesi bu şekilde gerçekleşir.

İnsan sesine dönersek, zengin çeşitlilikteki alt-harmoniklerin karmaşık yığını, Fourier dönüşümünün harikası tarafından hâlâ çözülebilir, ancak düzinelerce satıra bölünebilir.

Benzer şekilde, Şekil 3'teki dalga formuna bakıldığında, çekilebilecek baskın “notalar” vardır ve hepsinden en temel, baskın olanı, diğerlerinin değiştirdiği, yaklaşık olarak her birinde bir meydana gelen bir frekans bandındadır. -2.90338 zaman penceresi için bir faz kayması ile 142 ila 162 gün. Bu, 9. Konumdaki en yüksek banttır. Büyük ilk harmonik sonraki yükseklikte (ve etkinin gücünde) 13 konumunda görünür ve bu, 1.05873 zaman penceresi için bir faz kayması ile yaklaşık olarak her 111 ila 115 günde bir meydana gelen bir frekans bandını temsil eder, tam frekans ve faz Merkür'ün kayması, düşünülen zaman diliminde kendi kendine gerilediği farkediliyor.

Temel ve ilk harmoniği bir araya getirerek bir dalga elde ederiz ve Şekil 4'teki grafiğinde görüldüğü gibi, dalganın en düşük çukurlarının tümü, Merkür gerilemesi olarak da bilinen zaman periyotları arasında gerçekleşir ve en yüksek tepe noktaları Merkür Gerilemesindedir.

# 6. Sonuçlar

Bu çalışmanın sonuçları, bu çalışma ile önemli bir ilk adım atılmış olsa da, Merkür gerilemesinin yazım hatalarına neden olduğunu kanıtlamaktan hala çok uzak. Burada, Merkür gerilemesi zamanında veri setinde sözlük dışı kelime kullanımına maksimum periyodik katkının tek tip bir zamanı olduğunu ve %2,19'a kadar çıkabilen bir artış olduğu gösterilebilir.

Eğer tek bir hareketteki bu küçük ama önemli hata oranı artışı, bir dizi eylemde geçerliyse, daha sonra her gün yaptığımız binlerce küçük şeyle birleştirilirse (küçük adımlardan inşa edilmiş bir şey), 21-23 gün süren Merkür gerileme dönemi boyunca insanın büyük bir hata yapması neredeyse kesin olur.

Dikkat çekici bir not, istatistiksel olarak anlamlı %2 artışın, tamamen farklı bir dönemde tamamen farklı bir dönemde tamamen alakasız bir veritabanının sözlük dışı kullanım oranı için Merkür gerilemesi sırasında bulunanla aynı olmasıdır. (Oshop, Across Millions of Entries, Reddit is More Likely to Show Misspellings During Mercury Retrograde in May 2015, 2015)

Hala Merkür gerilemesi ile tam olarak aynı periyoda ve başlangıç noktasına sahip olan bilinmeyen bir pazar veya kültürel faktör olabilir. Eğer öyleyse, bu yine de Merkür gerileme mevsimlerinde sözlük dışı kelimelerin artan kullanımına yönelik bir eğilimi, aslında mutlak bir korelasyonu düşündürür. Bu süre zarfında hata riskinin arttığına inanan insanlar yine de haklı olacaktır.

Kürelerin armoniklerinin şiirsel ifadesi sadece bir metafor olmayabilir.

# Kaynakça

*Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more.* n.d. http://www.amazon.com (erişim 10 24, 2015).

"DictionaryLookup Source Information—Wolfram Language Documentation." *Wolfram Language & System Documentation Center.* n.d. http://reference.wolfram.com/language/note/DictionaryLookupSourceInformation.html(erişim 10 24, 2015).

*Discrete Fourier Transforms—Wolfram Language Documentation.* n.d. http://reference.wolfram.com/language/tutorial/FourierTransforms.html (erişim 10 25, 2015).

J. McAuley, R. Pandey, J. Leskovec. "Inferring networks of substitutable and complementary products." *Knowledge Discovery and Data Mining*, 2015.

Jure Leskovec and Andrej Krevl. *{SNAP Datasets}: {Stanford} Large Network Dataset Collection.* June 2014. http://snap.stanford.edu/data (erişim 10 24, 2015).

Oshop, Renay:
- *Across Millions of Entries, Reddit is More Likely to Show Misspellings During Mercury Retrograde in May 2015.* 10 10, 2015. http://www.ayurastro.com/articles/across-millions-of-entries-reddit-is-more-likely-to-show-misspellings-during-mercury-retrograde-in-may-2015 (erişim 11 2, 2015).
— (2015). Non-Dictionary Words Occur More Often in Amazon Reviews During Mercury Retrograde. Zenodo. DOI: 10.5281/zenodo.32847

Wolfram Research, Inc. *Mathematica.* Version 10.2. Champaign, IL: Wolfram Research, Inc., 2015.

*Orijinal Metin: “Non-Dictionary Words in Amazon Reviews Occur More During Mercury Retrograde”, Renay Oshop*