Doğruluk Problemi İçin Veri Kümesi Hazırlanması

Arif Kürşat Karabayır, Ozan Onur Tek, Özgür Fırat Çınar, ve Selma Tekir
arifkarabayir@gmail.com, ozanonurtek@gmail.com,
ozgurfiratcinar@gmail.com, selmatekir@iyte.edu.tr

İzmir Yüksek Teknoloji Enstitüsü Bilgisayar Mühendisliği Bölümü 35430 Urla/İzmir

Özet. İnternet günümüzde en önemli bilgi kaynaklarından biri haline gelmiştir. İnternet ile birlikte, bilgiye ulaşımın ve paylaşımın kolaylaşması, çelişkili bilgilerin açığa çıkmasına sebep olmuştur. Çelişkili bilgilerin artmasıyla, bunlar arasında doğru olanı bulmak da her geçen gün zorlaşmaktadır. Bu sorun literatürde doğruluk (veracity) problemi olarak tanımlanmıştır. Bu alanda geliştirilen algoritmalar girdi olarak yapısal veriyi kabul etmektedir. Bu algoritmaların internet üzerinde kullanılabilmesi için internetteki yapısal olmayan verinin yapısal forma dönüştürülmesi gerekmektedir. İnternet'teki verinin çeşitliliği düşünüldüğünde bu işin konudan bağımsız, otomatik olarak gerçekleştirilmesi zordur.
Bu çalışmada doğruluk problemi üzerine geliştirilen algoritmaların sınanabilmesi için internetteki yapısal olmayan verilerin yapısal bir veri kümesine dönüştürülmesinde gerekli aşamalar belirlenip otomatize edilmesine katkı sağlanmıştır. Bu aşamalar kullanılarak örnek bir özdeyiş veri kümesi oluşturulmuş ve belirlenen bir doğruluk sınama algoritması bu veri kümesinde uygulanarak elde edilen sonuçlar yorumlanmıştır.

Anahtar Kelimeler: Doğruluk Problemi · Veri Kümesi · Web

Construction of a Dataset for the Veracity Problem

Arif Kürşat Karabayır, Ozan Onur Tek, Özgür Fırat Çınar, ve Selma Tekir
arifkarabayir@gmail.com, ozanonurtek@gmail.com,
ozgurfiratcinar@gmail.com, selmatekir@iyte.edu.tr

İzmir Institute of Technology Department of Computer Engineering 35430 Urla/İzmir

Özet. Internet has become one of the most important information sources. With the advent of Internet, the ease of access and sharing of information have caused the emergence of conflicting information. The increase in conflicting information makes it a challenge to find the truth out of it. This problem is named as the veracity problem. The algorithms that were developed in response to this problem accept structured data as input. Thus, to be able to use these algorithms on Internet, there is a need to transform the unstructured data on the Internet into a structured form. This need is hard to fulfill in a domain-independent and automatic way considering the variety on Internet.
In this work; structured data preparation to test the effectiveness of the truth-finder algorithms is experienced. The process of transforming the unstructured data on the Internet into a structured form is described in steps to contribute its generalization in a domain-independent way. As a result of this process, a new quotes data set is constructed and a truth-finder algorithm is tested on this dataset by giving some comments on it.

Anahtar Kelimeler: Veracity Problem · Dataset · Web

1 Giriş

İnternet günümüzde hayatımızın önemli bir parçası haline gelmiştir. İnsanlar gündelik hayatında internette önemli bir süre geçirmektedir. Bu durum birçok insanın bu platformu önemli bir bilgi kaynağı olarak görmesine neden olmaktadır. İnternetteki bilgilerin ve bilgi sağlayan kaynakların sayısı da gitgide artmaktadır. Fakat sunulan bilgilerin güvenilir olduğunun bir garantisi yoktur. Popüler kaynakların güvenilir bilgi sağlayacağı düşünülebilir fakat yapılan araştırmalar bu varsayımın her zaman doğru olmadığını göstermektedir [1].

İnternetteki çelişkili bilgilerden doğru olanı bulup, kaynakların güvenilirliğini saptamak için 2008 yılında doğruluk problemi [1] tanımlanmış ve bu problemin çözümü için bir model oluşturulmuştur. Takip eden yıllarda bu problem üzerine çalışmalar sürdürülmüştür [2,3,4,5]. Bu alanda geliştirilen algoritmalar girdi olarak yapısal veriyi kabul etmektedir. Ancak internetteki verilerin büyük kısmı yapısal olmayan formdadır. Bu yüzden, bu konudaki algoritmaların doğrudan kullanımı

mümkün değildir. Algoritmaların sınanması için önceden hazırlanmış birkaç yapısal veri kümesi kullanılmaktadır.

Bu çalışmada, bir alan belirlenerek bu alan hakkında internette sunulan yapısal olmayan veri yapısal bir forma dönüştürülerek doğruluk bulma algoritmalarının sınanabileceği bir veri kümesi oluşturulmuş ve izlenen adımlar kayıt altına alınarak sürecin otomatize edilmesine yönelik bir katkıda bulunulmuştur.

Bildirinin geri kalan bölümünde ilk olarak doğruluk problemi tanıtılmakta ve ilgili literatür verilmektedir. Metodoloji bölümünde doğruluk bulma algoritmalarını sınamak üzere özdeyiş veri kümesi oluşturma süreci anlatılmaktadır. Bildirinin sonuç bölümünde ise bulgular ortaya konmakta ve gelecek çalışmalar üzerine düşünceler dile getirilmektedir.

2 Doğruluk Problemi ve İlgili Literatür

Doğruluk problemi ilk defa 2008 yılında Yin vd. [1] tarafından ortaya konmuştur. Doğruluk, birden fazla kaynağın, (örneğin web sitesi) bir veya birden fazla nesne (örneğin kitap) üzerinde iddia ettiği bilgilerin (web sitesinin kitap hakkında iddia ettiği yazar ismi) doğruluğunun çıkarımı üzerine kurgulanmış bir problemdir. Bu problem kapsamındaki temel terimler aşağıda verilmektedir:

- **Kaynak**, bilginin sağlandığı yer.
- **İddia**, bir kaynağın nesne hakkında sağladığı veri.
- **Gerçek**, en yüksek güvenilirlik puanına sahip iddia.
- **Kaynak Güvenilirliği**, bir kaynağın doğru bilgiyi verme olasılığı.
- **İddia Güvenilirliği**, bir iddianın gerçek olma olasılığı.

Bugüne kadar geliştirilen doğruluk bulma algoritmaları iki farklı veri tipi üzerine yoğunlaşmıştır. Bunlar sayısal ve kategorik verilerdir.

Sayısal Veriler genellikle bir ölçümü veya miktarı ifade eder. Bu tipteki veriler için kesin doğru veya kesin yanlış demek doğru değildir. Bunun yerine bu verilerin doğruluğu, kesinlik oranı ile ifade edilir. Örnek olarak, bir nesnenin bir özelliği hakkında 25 ve 90 olarak iki farklı iddia varsa ve gerçek değer 100 ise, 90 doğruya yakın sayılıp 25'e göre iddia güvenilirlik puanı yüksektir. **Tablo 1**'de sayısal veri tipinde bir veri kümesi örneği verilmektedir.

Kategorik Veriler ise daha çok nesnelerin karakteristiğini ifade ederken kullanılır. Bu tipteki veriler ya doğru ya da yanlış olarak sınıflandırılır. Kategorik veri tipi ile ilgili örnek veri kümesi **Tablo 2**'de sunulmaktadır.

Doğruluk probleminin çözümü için tasarlanan algoritmalar bu kavramlar üzerinden modellerini oluşturmuştur. Her bir çalışma konuya farklı bir yönden yaklaşmakta olup aralarında bazı temel farklılıklar bulunmaktadır. Bu çalışmalar ve öne çıkan farklılıkları takip eden bölümde açıklanmaktadır.

Tablo 1. Sayısal Veri Kümesi Örneği - Nüfus Veri Kümesi

Nesne	Kaynak	İddia
Abu Dhabi	Contributor#1513217: Mohammed	1850230.0
Amsterdam	Contributor#141597: Ilse@	741329.0
Amsterdam	Contributor #1300620: Krator	742884.0
Adelaide	Contributor #3922171: Pirate05	1124315.0
Athens	Contributor #1876487: El Greco	4200000.0
Athens	Contributor #3007532: Theiasofia	4242000.0
Athens	Contributor #1876487: El Greco	745514.0
Athens	Contributor #0 (87.203.23.2)	745514.0
Athens	Contributor #1876487: El Greco	745514.0
Navalcán	Contributor#363486: Emijrp	2238.0
Navalmoralejo	Contributor #363486: Emijrp	63.0
Los Navalmorales	Contributor#363486: Emijrp	2636.0
Los Navalucillos	Contributor#363486: Emijrp	2636.0

Tablo 2. Kategorik Veri Kümesi Örneği - Özdeyiş Veri Kümesi

İddia	Nesne	Kaynak
George Bernard Shaw	Beauty is a short-lived tyranny	famous-quotes
Socrates	Beauty is a short-lived tyranny	brainyquote
Albert Einstein	It is strange to be known so universally and yet to be so lonely.	famous-quotes
Napoleon Bonaparte	A soldier will fight long and hard for a bit of colored ribbon.	quotables

Yin vd. [1] tarafından geliştirilen model temel "TruthFinder" modelidir ve kaynak-iddia ilişkilendirmesi üzerinde çalışarak her kaynak ve her iddia için belirli bir güvenilirlik puanı ataması yapar ve bu puanları yinelemeli bir şekilde günceller. Bu kapsamda, kaynak güvenilirliği ($t(\omega)$) değeri, $F(\omega)$, ω kaynağının sağladığı iddiaların güvenilirlik $s(f)$ toplamının; $F(\omega)$, ω kaynağının sağladığı iddiaların toplam sayısına oranıdır:

$$t(\omega) = \frac{\sum_{f \in F(\omega))} s(f)}{|F(\omega)|} \tag{1}$$

Daha sonra, bulunan kaynak güvenilirliği $t(\omega)$, iddia güvenilirliğini hesaplamak için kullanılır.

$$s(f) = \sum_{w \in W(f))} t(\omega) \tag{2}$$

Bu modelin en önemli noktası, bir nesne hakkındaki farklı iddiaların birbirinin güvenilirlik puanına etki etmesidir. Bulunan iddia güvenilirlik puanları bu etkiler göz önünde bulundurularak güncellenir.

Zhao ve Han [2] doğruluk problemini çözmek için "GTM" adında bayesçi olasılıksal bir model oluşturmuştur. Geliştirilen model, kategorik veri üzerine yoğunlaşan önceki çalışmaların aksine özellikle sayısal verilerde daha başarılı sonuçlar vermektedir.

Doğruluk problemi için hazırlanan veri kümelerinde, kaynaklardan olabildiğince çok veri sağlanması temel bir gereksinimdir. Ancak pratikte her kaynak yeteri kadar veri sağlamayabilir. Bu durum doğruluk bulma algoritmaları için çözülmesi gereken bir problemdir. Li vd. [3] çalışmasında, az sayıda iddiaya sahip kaynakların da bulunduğu bir veri kümesinde doğruluk bulma başarımını artıracak bir model öne sürülmüştür. Geliştirilen model hem kategorik hem de sayısal veri kümelerinde çalışabilmektedir.

Xin vd. [4] kaynakların verileri birbirinden kopyalama durumu üzerine odaklanmıştır. İnsanlar sezgisel olarak bir iddiayı ne kadar çok kaynakta görürse o iddianın doğruluğundan o kadar emin olurlar. Ancak kaynakların verileri birbirlerinden kopyalamaları durumunda yanlış bilgiler hızla yayılabilir ve doğru bilgiyi saptamak bu durumda oldukça zorlaşabilir. Çalışmada, iddiaları birbirinden kopyalayan kaynakların tespit edilmesi ile gerçek iddiaların saptanmasını kolaylaştıran bir model geliştirilmiştir.

Qi vd. [5] doğruluk bulma algoritmalarında sayısal ve kategorik verileri farklı şekillerde ele almıştır. Literatürde her iki veri kümesi için de tasarlanan çalışmalar mevcut olmasına rağmen ilk defa bu çalışmada veri kümesinin heterojen olduğu bir başka ifade ile her iki veri tipini de kapsadığı durumda çalışabilecek bir model geliştirilmiştir. Modelin heterojen veri kümesinde çalışabilmesi, farklı veri tipleri için farklı uzaklık fonksiyonlarının tanımlanabilmesi ile sağlanmıştır.

3 Metodoloji

Doğruluk bulma algoritmalarının çalışması için yapısal bir veri kümesi gerekmektedir. Ancak internetteki verilerin çoğu yapısal olmayan formdadır. Bu kısımda, internetteki yapısal olmayan verilerin, doğruluk bulma algoritmalarının sınanabileceği yapısal bir forma dönüştürülmesi süreci tanımlanmaktadır.

3.1 Tanım Kümesinin ve Kaynakların Belirlenmesi

İdealde internetteki hemen her konunun tanım kümesi olarak seçilebileceği düşünülebilir. Bununla beraber, gerçekleştirdiğimiz denemelerde doğruluk bulma algoritmalarının mevcut çalışma yapıları gereğince, tanım kümeleri için belirli şartların sağlanması durumunda daha sağlıklı sonuçlar verdiği gözlemlenmiştir. Bu denemeler sonucunda tespit edilen gerekli şartlar aşağıdaki gibi belirlenmiştir:

- Nesneler hakkındaki iddialar değişiklik gösterebilmelidir.
- Tanım kümesi ile ilgili birçok kaynaktan veri elde edilebilmelidir.
- Her nesne hakkında tek bir kesin doğru veri bulunmalıdır.

Bu şartlar göz önünde bulundurularak tanım kümesi olarak "Özdeyişler" belirlenmiştir.

Tanım kümesi belirlendikten sonraki aşama, verilerin çekileceği kaynakların seçilmesidir. Tanım kümesi kapsamında birçok kaynak bulunabilir ancak bu

kaynakların hepsi veri kümesi oluşturulması için uygun olmayabilir. Özdeyiş tanım kümesi için bu çalışmada, internetteki popüler özdeyiş paylaşım siteleri tespit edilmiştir. Ayrıca sosyal medyada paylaşılan verilerin ne kadar güvenilir olduğunu görebilmek adına özdeyiş paylaşan belirli Twitter hesapları da kaynaklara dahil edilmiştir. Bu kaynakların belirlenmesinde;

- Kaynakların sözdizimi ve yazım kurallarını doğru kullanması,
- Verileri sağlarken tamamen ikinci bir kaynağa bağımlı kalmaması,
- Yeterli sayıda veri sağlamış olması,
- Verilerin sağlandığı formatın tutarlı olması

gibi kriterler dikkate alınmıştır.

3.2 Verilerin İşlenmesi

Kaynakların belirlenmesinin ardından kaynaklara ait programlama arayüzleri tespit edilmiştir. Bu programlama arayüzleri kullanılarak veriye ulaşılmış ve veri düzenli bir formatta kayıt altına alınmıştır.

Bölüm 3.1'de de belirtildiği üzere, kaynaklar belirli kriterler çerçevesinde seçilmiş olmasına rağmen, kayıt altına alınan veri incelendiğinde formata uygun olmayan veya konu dışı içerikler, özetle kirlilik tespit edilmiştir. Formata uygun olmayan içerikleri temizlemek için düzenli ifadelerden yararlanılmıştır. Konu dışı içerikler için, önce bu içeriklerin ortak noktaları tespit edilmiş ve yine düzenli ifadeler ve Jaro-Winkler Distance algoritması [6] yardımıyla veri kümesi bu içeriklerden temizlenmiştir. Jaro-Winkler Distance temelde Jaro-Winkler Benzerliği üzerine oturtulmuş iki karakter dizisinin benzerliğini tespit etme algoritmasıdır. Jaro-Winkler Benzerliği ise temelde Jaro Benzerliği'ni kullanır. Bu benzerlik matematiksel olarak aşağıdaki formül ile hesaplanır;

$$sim_j = \begin{cases} 0 & \textit{if } m{=}0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \textit{otherwise} \end{cases} \tag{3}$$

s_i karakter dizisinin uzunluğunu,
m eşleşen karakter sayısını,
t karakterler için yer değişikliğini ifade eder.

İki dizinin birbiriyle eşleşir olarak sayılması için;

$$\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1 \tag{4}$$

sağlaması gerekir. Jaro-Winkler Benzerliği ise bu noktadan yola çıkılarak;

$$sim_w = sim_j + (\ell p(1 - sim_j)), \tag{5}$$

eşitliği ile verilir. Bu eşitlikte;
sim_j Jaro benzerliğini,

l ortak önek karakterlerin uzunluğunu,
p sabit ölçekleme faktörünü ifade eder.

Son olarak Jaro-Winkler Distance değeri ise;

$$d_w = 1 - sim_w \tag{6}$$

şeklinde ifade edilir. Jaro-Winkler algoritması ayrıca, N-gram benzerlik algoritmasıyla beraber kullanılarak, her bir nesne birbiri ile karşılaştırılmış ve belirli bir eşik değerinin üzerindeki eşleşmelerde nesneler özdeş sayılmıştır. Bu sayede aynı nesnenin küçük noktalama farklılıkları, yazım yanlışları ve farklı kelime kullanımı ile oluşan farklılıkların önüne geçilmeye çalışılmıştır. N-gram benzerlik algoritması [7], bir kaynak metinde, n uzunluğundaki tüm komşu kelime veya karakter gruplarının kombinasyonlarını baz alan yöntem olarak ifade edilebilir. Sonraki adımda ise bir kaynağın aynı iddiayı birden fazla kez barındırma sorunuyla karşılaşılmıştır. Bu sorun N-gram benzerlik algoritmasıyla, benzerlik ilişkisi sayısal olarak birbirine çok yakın ve aynı kaynaktan iddia edilmiş verilerin temizlenmesiyle çözümlenmiştir. Son aşamada ise veri kümesi özdeyiş-söyleyen ilişkisine göre gruplandırılmış ve TruthFinder algoritması için uygun bir hale getirilmiştir.

Bu işlemler sonucunda 17 farklı kaynaktan yaklaşık 70000 iddia içeren özdeyiş veri kümesi oluşturulmuştur. Bu veri kümesi TruthFinder algoritmasında sınanmış ve bulgular bir sonraki kısımda anlatılmıştır.

4 Sonuç

4.1 Bulgular

Yapılan bu çalışmanın temel amacı, doğruluk problemi ve ilgili algoritmaları destekleyici bir alt yapı oluşturmak ve söz konusu alana katkı sağlamaktır.

Bahsedilen temel kazanımların sağlanması için öncelikli olarak alan üzerindeki benzer çalışmalar incelenmiş ve varolan yaklaşımlar farklı veri kümelerinde sınanmıştır. Kitap-yazar, şehir-nüfus gibi eşleşmeler içeren kategorik ve sayısal veri kümeleri incelenmiştir.

Çalışma kapsamında oluşturulan özdeyişler içeren yeni bir veri kümesi ile TruthFinder algoritması çalıştırılmıştır. TruthFinder algoritması nesne-gerçek çiftlerine ihtiyaç duyduğu için kişilerin nesne, özdeyişlerin ise gerçek olarak kullanılması planlanmıştır fakat, bu durum TruthFinder algoritmasının genel sezgisel yaklaşımlarından en temeli olan "Her nesne yalnızca bir tane gerçek doğruya sahiptir." önermesine ters düştüğü için bu fikir terkedilmiş ve tam tersi bir analoji ile devam edilmiştir.

TruthFinder algoritması çeşitli kaynaklardan alınan, özdeyişlerden oluşan veri kümesinde çalıştırılmıştır. Elde edilen, "kaynakların güvenilirlik skorları" **Tablo 3**'te verilmektedir. Tablodaki skorlar incelendiğinde kaynak güvenilirliği puanı 1'e yakın olan kaynakların sağladığı iddiaların gerçek olma olasılığı diğer kaynaklara göre daha yüksektir. Diğer bir değişle daha yüksek puanlı kaynaklar

Tablo 3. TruthFinder algoritmasının özdeyiş veri kümesi üzerindeki sonuçları

Kaynak	Kaynak Güvenilirliği
AboutCheGuevara	0.680000
FranzKafkaQtss	0.680430
SFreudSayings	0.692487
mahatmaa_gandhi	0.695539
GreatestQuotes	0.691252
QuoteDaily	0.711092
NietzscheQuotes	0.718113
einsteinquotes	0.719569
PureNietzsche	0.725397
NietzscheQ	0.738335
gandhhii	0.767866
quotables	0.827372
famous-quotes	0.832783
brainyquote	0.884118
Einsteiin_Quote	0.980225
GandhiiQuotes	0.981335
einssttein	0.983000

daha güvenilir iddialar sağlamaktadır.

Tablo 4'te özdeyiş veri kümesinden bir örnek sunulmaktadır. Sunulan örneğe bakıldığında, "Beauty is a short-lived tyranny" ve "Emancipate yourselves from mental slavery, none but ourselves can free our minds!" nesneleri hakkında *brainyquote* kaynağının gerçek iddiayı sağladığı ve gerçek olmayan iddia sağlayan diğer iki kaynaktan daha yüksek güvenilirlik puanına sahip olduğu görülmektedir.

Tablo 4. Özdeyiş Veri Kümesi

İddia	Nesne	Kaynak	Gerçek
George Bernard Shaw	Beauty is a short-lived tyranny	famous-quotes	Socrates
Socrates	Beauty is a short-lived tyranny	brainyquote	Socrates
Marcus Garvey	Emancipate yourselves from mental slavery, none but ourselves can free our minds!	brainyquote	Marcus Garvey
Bob Marley	Emancipate yourselves from mental slavery, none but ourselves can free our minds!	quotables	Marcus Garvey

Özdeyiş veri kümesi üzerinde algoritma çıktısındaki iddia güvenilirliği puanları kullanılarak yapılan hesaplamada TruthFinder algoritmasının F1 puanı 0.903 olarak hesaplanmıştır. Bu sonucun, algoritmanın ilk olarak sınandığı yazar-kitap veri kümesindeki 0.87 F1 puanına yakın olması bu çalışmada oluşturulan veri kümesinin doğruluk problemi algoritmalarının sınanabilmesi için uygun olduğunu desteklemektedir.

4.2 Gelecek Çalışmalar

Özellikle son yıllarda -internetteki bilgi kirliliğinin de artmasıyla doğruluk problemi önem kazanmıştır ve bu önemini koruyacağı öngörülmektedir. Bu konuda yapılan hem akademik hem de kurumsal çalışmaların sayısı her geçen gün artmaktadır.

Haberin doğruluğunu editörler aracılığı ile kontrol eden, bunu kamuoyu ile paylaşan, çalışmalarına Türkiye'de başlamış, Uluslararası Doğruluk Kontrolü Ağı tarafından yayınlanan İlkeler Kılavuzu'nu tanıyarak imzalayan teyit.org Facebook'un üçüncü taraf doğrulama programının Türkiye uygulayıcısı olarak duyurulmuştur [8].

Çalışmanın devamında, yukarıda verilen örnek gibi; haber kaynakları tarafından iletilen bilgilerin doğrulanmasında ya da ansiklopedik veri kaynaklarının doğruluğunun denetlenmesinde kullanılabilecek otomatize araçlar geliştirilebilir.

Sistem, gerçek zamanlı veri üzerinde ve/veya farklı veri tipleri üzerinde çalışabilir hale getirilerek (sıralanmış veri, sayısal-kategorilenmiş veri, grafiksel veri vs.) otomatize çalışan bir çok platform için bir iskelet oluşturabilir.

Kaynaklar

1. X. Yin., J. Han., P. S. Yu.: Truth discovery with multiple conflicting information providers on the Web. Trans. Knowl. Data Eng. (TKDE) 20(6), 796-808 , (2008).
2. Zhao, B., Han, J.: A Probabilistic Model for Estimating Real-valued Truth from Conficting Sources. In: QDB, (2012).
3. Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., Fan, W., Han, J.: A Confidence-aware Approach for Truth Discovery on Long-tail Data. In: Proceedings of the VLDB Endowment, vol.8, pp. 425-436. (2014).
4. Dong, Q.L., Berti-Equille, L., Srivastava, D.: Integrating Conflicting Data: The Role of Source Dependence. Proc. VLDB Endow. 2(1), 550-561. (2009).
5. Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., Han, J.; Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, pp. 1187-1198. (2014).
6. Jaro, M.: Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, 84(406), 414-420, (1989).
7. Kondrak, G.: N-gram Similarity and Distance. In: Proceedings of the 12th International Conference on String Processing and Information Retrieval (SPIRE'05), pp. 115-126. (2005).
8. https://teyit.org/facebookun-dogrulama-programi-turkiyede-teyit-org-is-birligiyle-hayata-geciyor/. Last accessed 31 May 2018