Neo4j Veritabanı'nda Sosyal Medya Verisi İçin Benzerlik Metriklerinin Geliştirilmesi

BIL 496

Yunus ÇEVİK

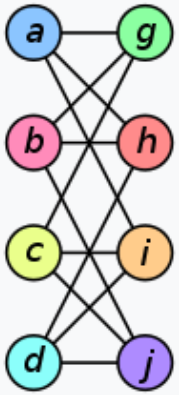
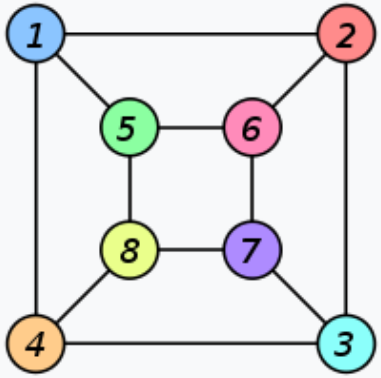
Proje Danışmanı: Dr. Öğr. Üyesi Burcu YILMAZ

Mayıs 2019

İçerik

- Projenin Şeması ve Tanımı
- Proje Tasarım Planı
- Projede Yapılanlar
- Başarı Kriterleri
- Kaynaklar

Proje Şeması ve Tanımı

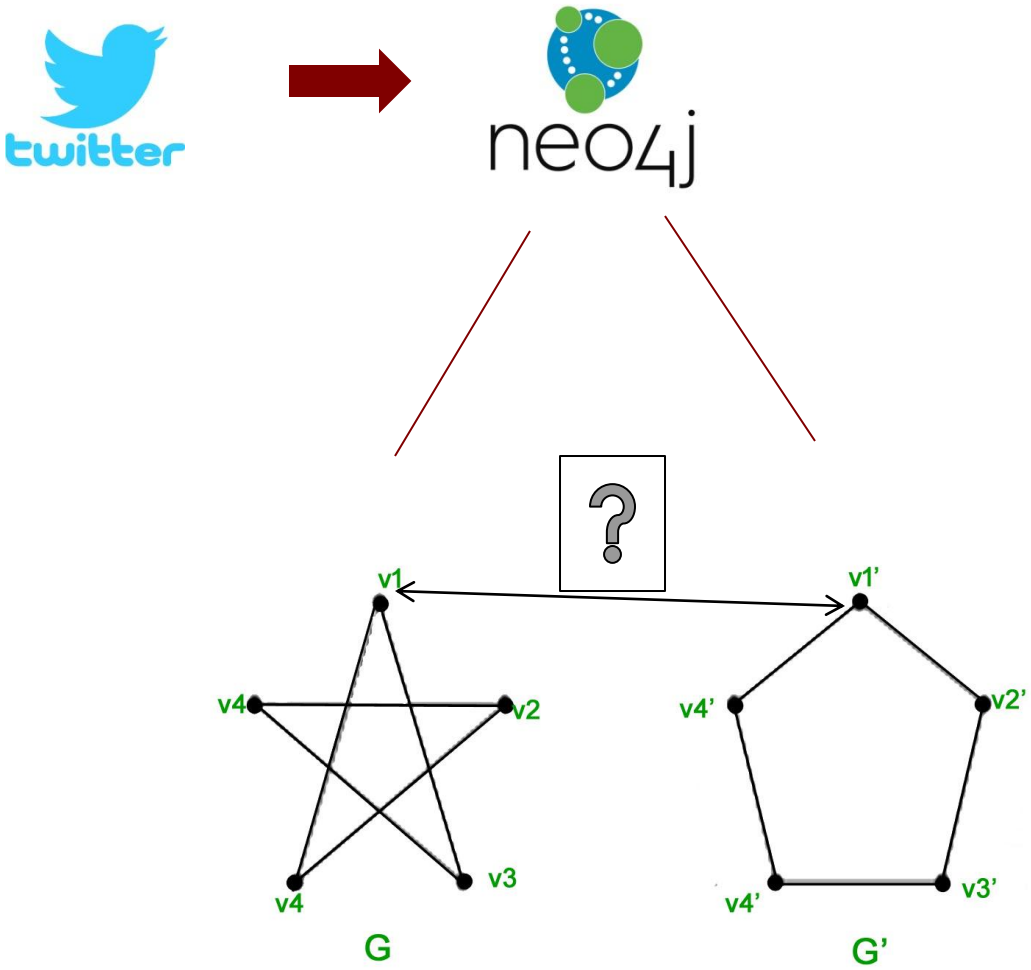
Graph G	Graph H	An isomorphism between G and H
		$f(a) = 1$ $f(b) = 6$ $f(c) = 8$ $f(d) = 3$ $f(g) = 5$ $f(h) = 2$ $f(i) = 4$ $f(j) = 7$

Şekil 1: İki Graph' ın İzomorfizmini Gösteren Resim

- Proje nedir?
  - İnsanlar, sosyal medya yardımı ile birbirlerine onlarca hatta yüzlerce mesaj gönderebiliyor.
  - Sosyal medya aracılığı ile yapılan mesajlaşmaların ve mesajlaşan kişilerin bilgileri belirli bir çizge(graph) yapısı içerisinde veritabanlarında bulundurulur.

✓ Yapacak olduğum projenin amacı, Neo4j veritabanında tutulan sosyal medya verilerinin, benzerlik metrikleri geliştirilerek, çeşitli makine öğrenmesi algoritmaları ile karşılaştırılmasıdır.

Proje Tasarım Planı

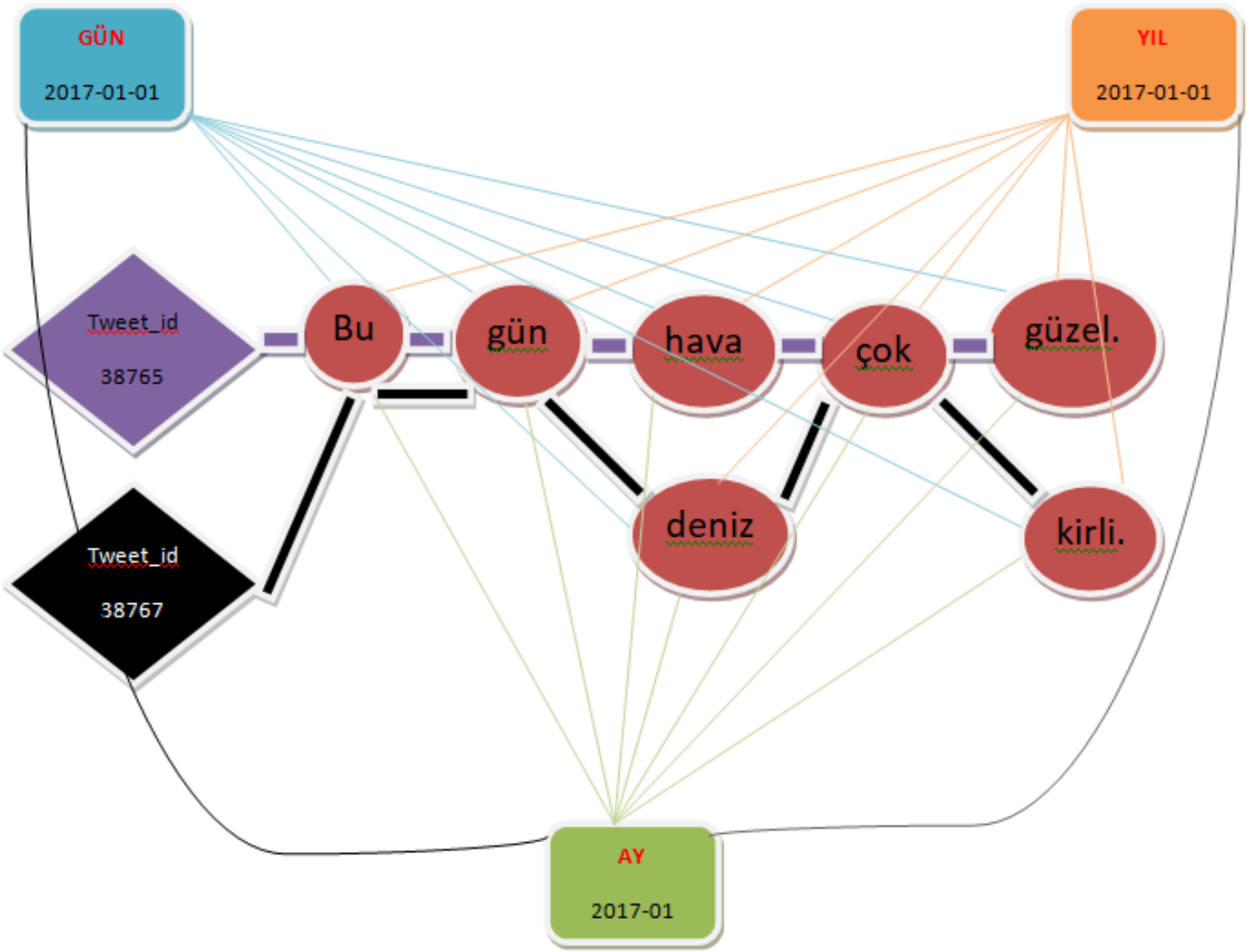


Projede Yapılanlar (MySQL' den veri alımı)

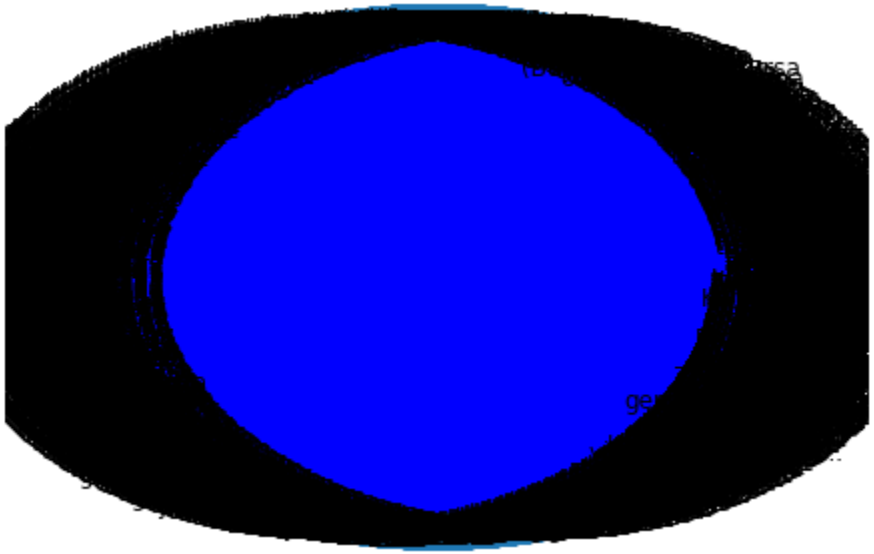
In [4]:
```
for val in read_data_from_database():
    twitId, date, title, metaData, url, content, urlInfo, section, firstParagraph = val
    print(twitId, date, metaData)
    print("\n")
```

```
343945 2017-01-01 Terör saldırısında 39 kişi hayatını kaybetti, 4'ü ağır 65 kişi yaralandı.


343946 2017-01-01 Ortaköy'de gerçekleştirilen silahlı saldırı sonrası Cumhurbaşkanı Erdoğan, telefonla bilgi aldı.


343947 2017-01-01 RTÜK, Ortaköy'deki terör saldırısı sonrası geçici yayın kısıtlamasına gitti.


343949 2017-01-01 Diyaneş İşleri Başkanı Görmez, Ortaköy'deki saldırının amacının yaşam biçimlerine göre toplumu bölmek ve ka
rşı karşıya getirmek olduğunu belirterek, Bu...


343950 2017-01-01 Ortaköy'deki saldırıya dünyadan tepki yağarken; ABD Başkanı Obama, gerekli destek ve yardımın yapılması hus
usunda ekibine talimat verdi


343951 2017-01-01 Ortaköy'deki saldırı sonrası açıklama yapan Bakan Bozdağ, Hiçbir terör saldırısı birliğimizi bozamayacak, k
ardeşliğimizi yok edemeyecek...
```

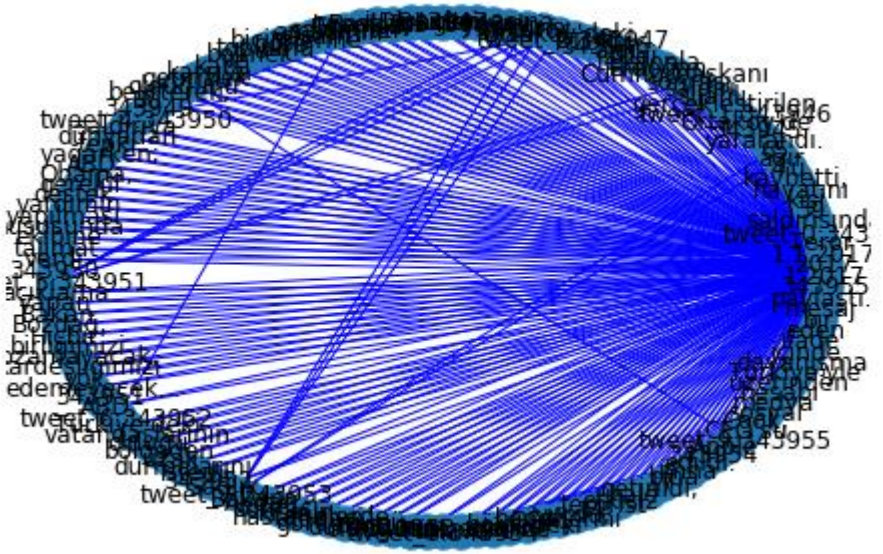
Çizge Modeli



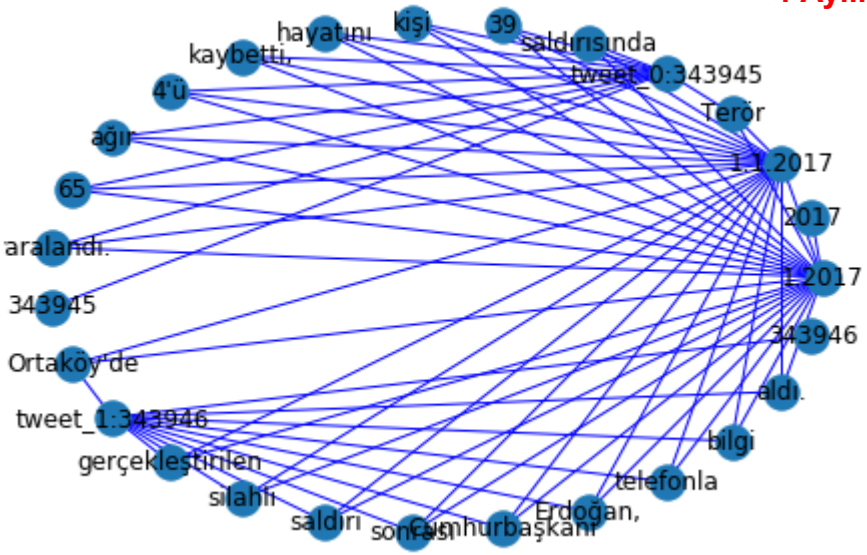
MySQL' deki 1 aylık verilerin networkx' de çizge olarak gösterimi



1 aylık tweet çizgesi

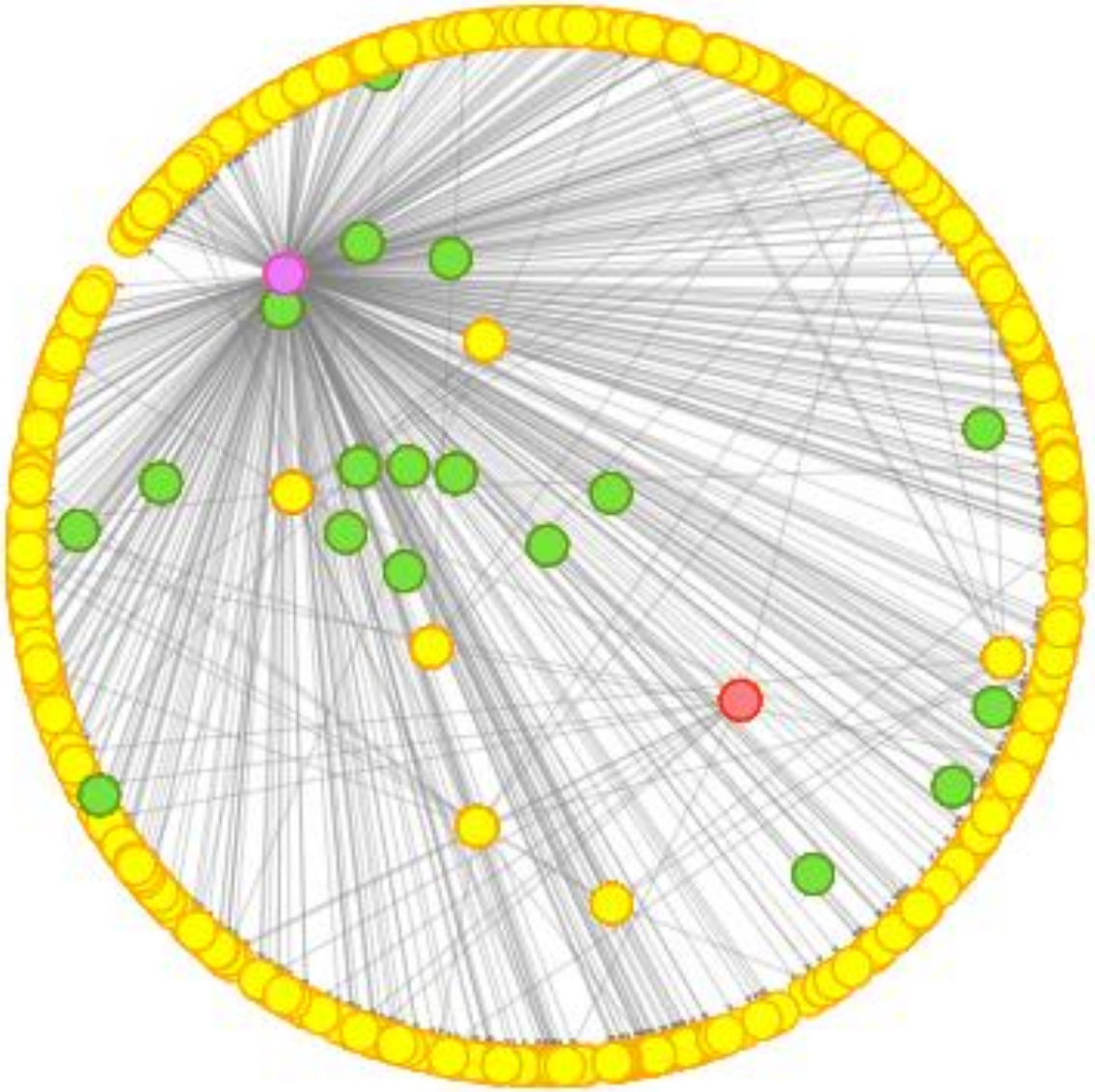


1 Aylık verilerin 10 tane tweeti



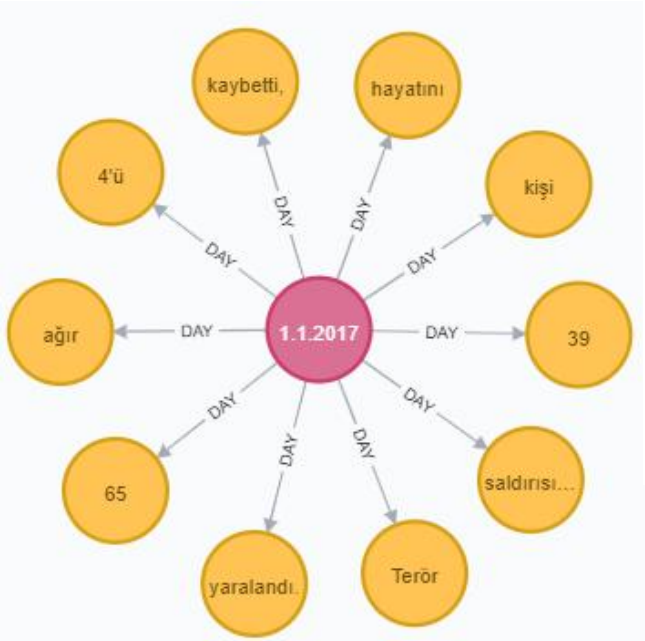
1 Aylık verilerin 2 tane tweeti

Jupyter Notebook' da Neo4j'den alınan verilerin gösterimi

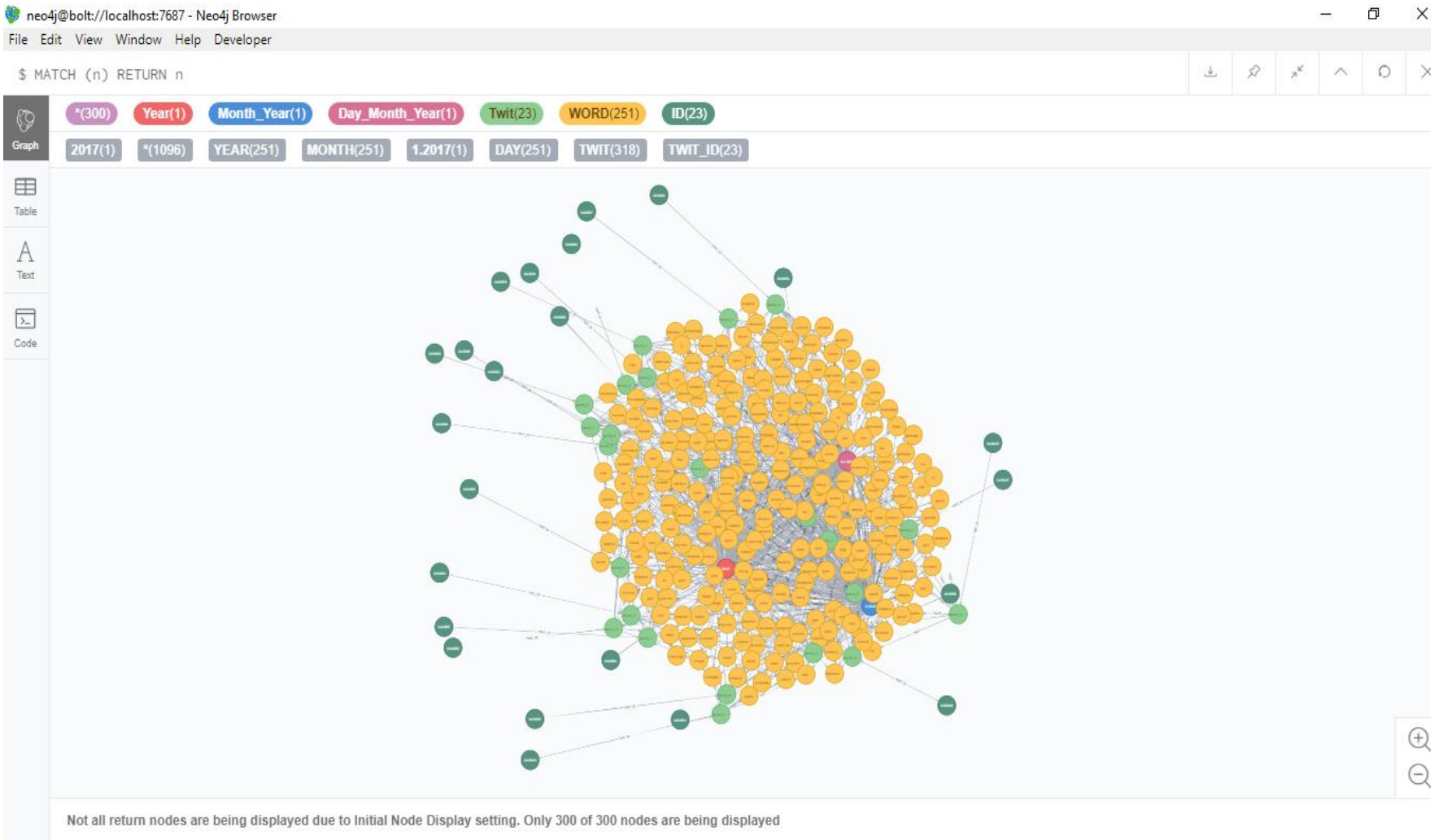


- Tweet_No: Tweet_ID
- Tweet_ID
- Tweet_Node
- Tweet_Tarihi (Gün – Ay – Yıl)

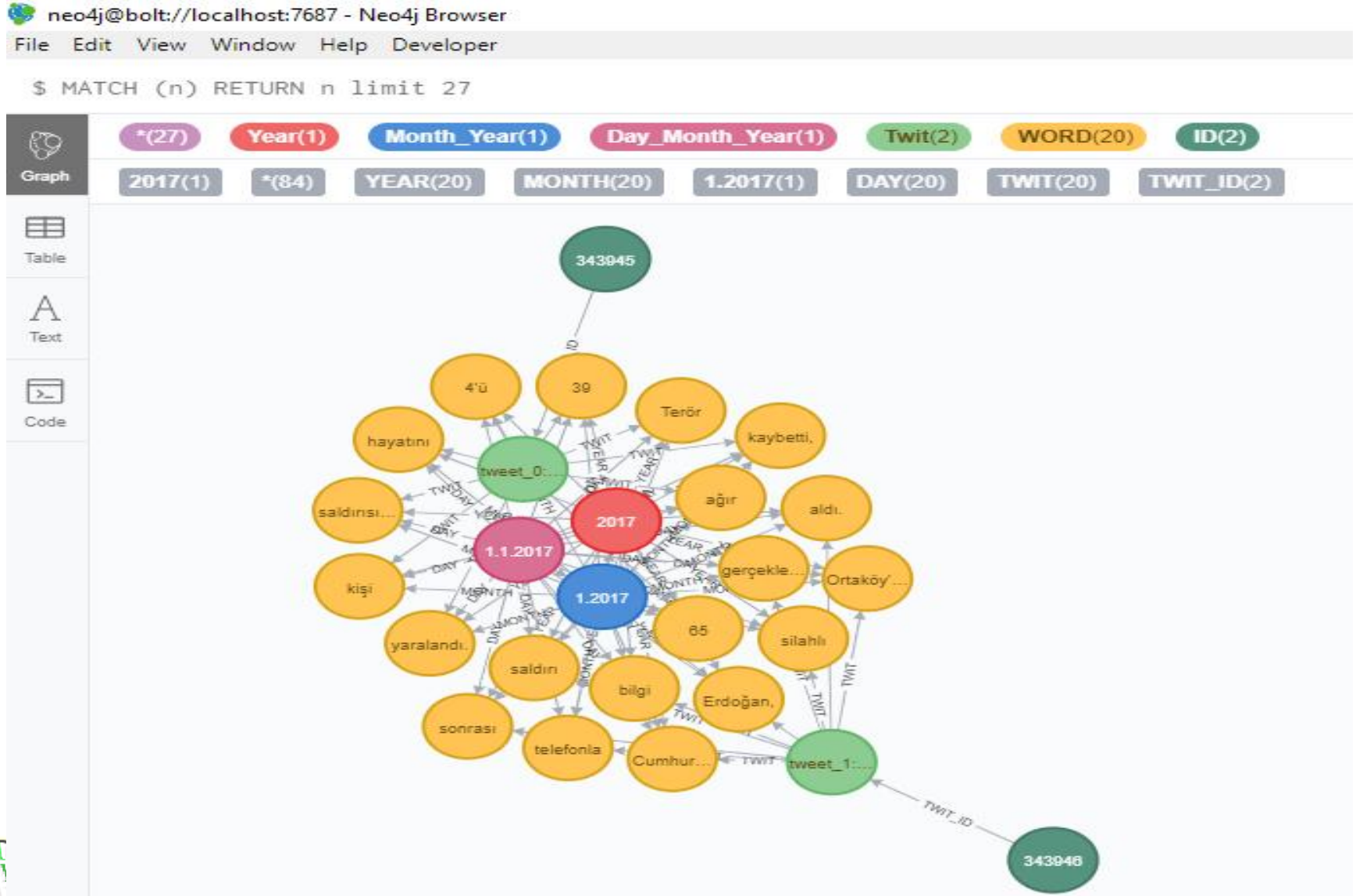
Gün – Ay – Yıl ‘ da olan tweet bilgileri



MySQL' deki verilerin Neo4j' e kaydedilmesi ve gösterimi



Neo4j' de 2 tweetin gösterimi



1 yılda kullanılan tweetlerin kelime – kelime ilişkilerinin cypher sorgusu ile gösterimi

```
1  %load_ext cypher
```

```
1  %%cypher
2  http://neo4j:123456@localhost:7474/db/data
3  MATCH (year:Year)-[:YEAR]->(word:WORD)
4  RETURN word.name, year.name
```

363 rows affected.

word.name	year.name
Muallim	2017
4,6	2017
olma	2017
260	2017
Hiçbir	2017
ilgili	2017
görevimiz	2017
Başkanı	2017
silahlı	2017
İstanbul	2017

Çizge Benzerliği Metrikleri

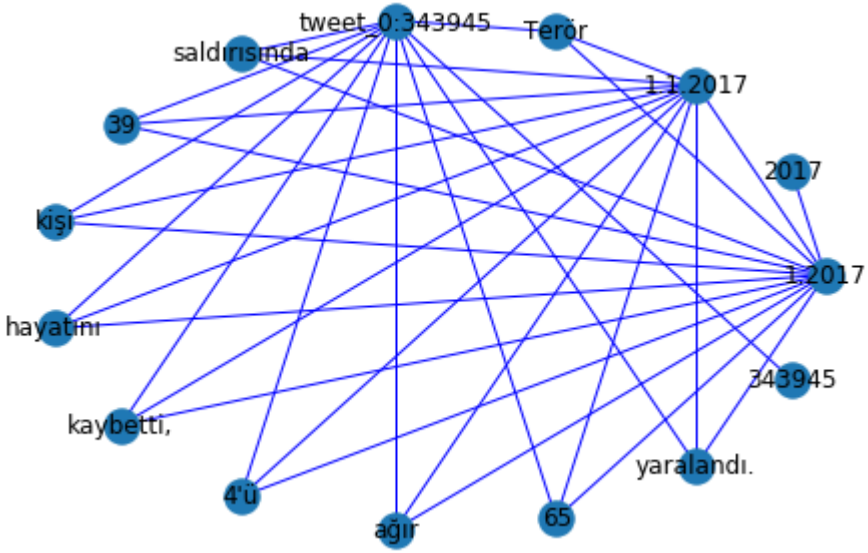
Çizge Benzerliği:
Çizge benzerliği, sosyal ağlar, görüntü işleme, ve bilgisayar görmesi gibi çeşitli uygulama alanlarına sahiptir ve bu nedenle birçok algoritma ve benzerlik ölçütü önerilmiştir. Önerilen teknikler üç ana kategoriye ayrılır: mesafenin düzenlenmesi/çizge izomorfizmi(edit distance/graph isomorphism), özellik çıkarımı (feature extraction) ve yinelemeli yöntemlerdir(iterative methods).

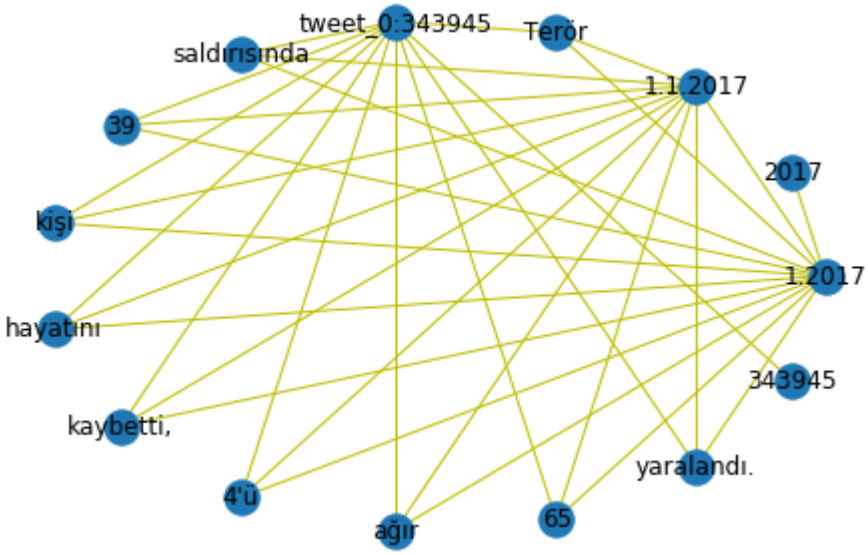
Edit distance/graph isomorphism:
Çizge benzerliğini değerlendirmede ki bir yaklaşım çizge izomorfizmidir. Eğer iki çizge izomorfikse ya da biri diğerinin alt-çizgesi (sub-graph) ile izomorfikse ya da bu iki çizge izomorfik alt-çizgelere sahipse onlar benzerdir.

Özellik çıkarımı(Feature extraction):
Bu yöntemlerin arkasındaki ana fikir, benzer çizgelerin muhtemelen derece dağılımı, çap, özdeğerler(eigenvalues) gibi belirli özellikleri paylaşmasıdır. Bu özelliklerin çıkarılmasından sonra, toplam istatistikler arasındaki benzerliği ve çizgeler arasındaki benzerliği değerlendirmek için bir benzerlik ölçüsü uygulanır.

Yinelemeli yöntemler(Iterative methods):
Yinelemeli yöntemlerin arkasındaki felsefe, “eğer iki düğümün komşuları benzerse o iki düğümde benzerdir”. Her yinelemede, düğümler benzerlik skorları değişiminde bulunurlar ve bu işlem yakınsama sağlandığında sona erer.

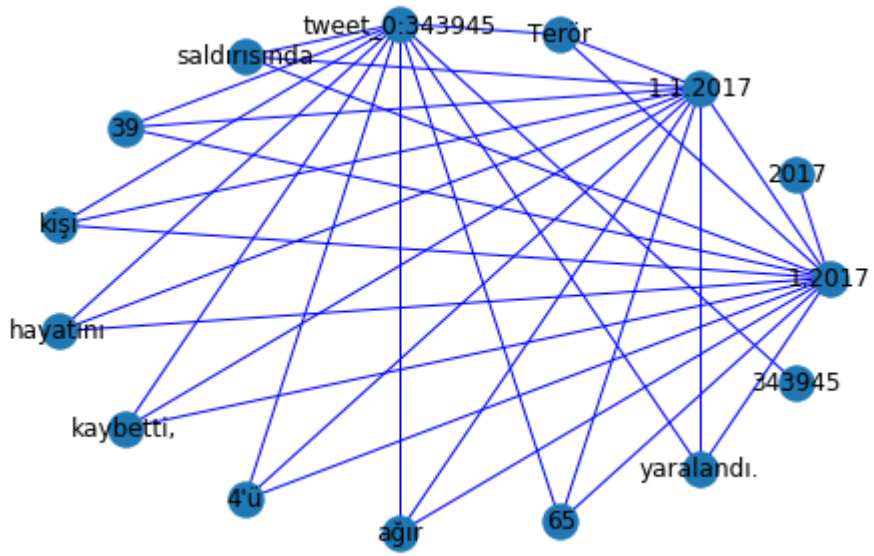
İki tweet'in benzerliklerinin karşılaştırılması

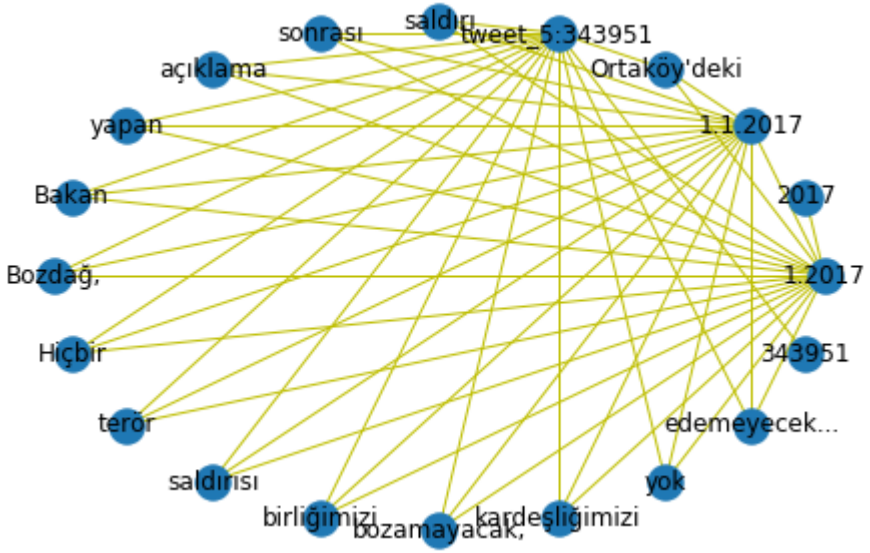





```
['1.2017', 2017, '1.1.2017', 'Terör', 'tweet_0:343945', 'saldırısında', '39', 'kişi', 'hayatını', 'kaybetti,', "4'ü", 'ağır',
'65', 'yaralandı.', '343945']   ==> Graph1
['1.2017', 2017, '1.1.2017', 'Terör', 'tweet_0:343945', 'saldırısında', '39', 'kişi', 'hayatını', 'kaybetti,', "4'ü", 'ağır',
'65', 'yaralandı.', '343945']   ==> Graph2
                Cosine Similarity between two days =>    1.0
                Cosine Similarity between two months =>  1.0
                Cosine Similarity between two years =>   1.0
                Cosine Similarity between two users =>   1.0000001
                Cosine Similarity between two texts  =>  1.0

                Weigthed Average Score of Cosine Similarities =>  1.000000023841858
                ------------------------------------------------------------------
                Graph-isomorphism between two graphs =>  True
                Graph-edit-distance between two graphs =>  0.0
```

İki tweet'in benzerliklerinin karşılaştırılması






```
['1.2017', 2017, '1.1.2017', 'Terör', 'tweet_0:343945', 'saldırısında', '39', 'kişi', 'hayatını', 'kaybetti,', "4'ü", 'ağır',
'65', 'yaralandı.', '343945']   ==> Graph1
['1.2017', 2017, '1.1.2017', "Ortaköy'deki", 'tweet_5:343951', 'saldırı', 'sonrası', 'açıklama', 'yapan', 'Bakan', 'Bozdağ,',
'Hiçbir', 'terör', 'saldırısı', 'birliğimizi', 'bozamayacak,', 'kardeşliğimizi', 'yok', 'edemeyecek...', '343951']  ==> Graph2
```

```
Cosine Similarity between two days =>    1.0
Cosine Similarity between two months =>  1.0
Cosine Similarity between two years =>   1.0
Cosine Similarity between two users =>   -0.13250376
Cosine Similarity between two texts  =>  0.5744927774241578

Weigthed Average Score of Cosine Similarities =>  0.6883978029165867
-----------------------------------------------------------------------
Graph-isomorphism between two graphs =>  False
Graph-edit-distance between two graphs =>  20.0
```

Başarı Kriterleri

- Geliştirilen çizge benzerlik metriğinin klasik benzerlik metriklerine göre %10 daha başarılı sonuçlar vermesi.
- Çizgelerin kenar, düğüm ve metinsel olarak en az üç açıdan benzerliğinin ölçülmesi.
- Çizgelerin, Neo4j' de en az 500 mb datayı 10 saniyede oluşturması.
- Neo4j' deki sorguların 15 saniyede sonuçlanması.

Kaynaklar

1. https://javacii.wordpress.com/2017/03/10/neo4j-nedir-nerelerde-kullanilir/ **(Neo4j nedir ve nerelerde kullanılır?)**
2. https://medium.com/5bayt/neo4j-nedir-e7160602211e **(Neo4j nedir?)**
3. https://netvent.com/yepyeni-bir-veritabani-sistemi-neo4j/ **(Neo4j nedir?)**
4. https://prezi.com/kngegnwvp2yp/cizge-veri-tabanlarnn-incelenmesi-ve-cizge-veri-taban-yar/ **(çizge veritabanlarının incelenmesi ve yaratılması)**
5. https://towardsdatascience.com/graph-embeddings-the-summary-cc6075aba007 **(graph emmeding)**
6. http://bilgisayarkavramlari.sadievrenseker.com/2009/06/18/denksekillilik-isomorphism/ **(isomorphism)**
7. https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781785282287/8/ch08lvl1sec117/creating-a-document-graph-with-cosine-similarity **(creating a document graph with cosine similarity)**