Sosyal Medya Verilerden Çizge Vektörlerinin Çıkarılması

BIL 496
2. İzleme

Hasan MEN

Proje Danışmanı: Yrd. Doç. Dr. Burcu YILMAZ

MAYIS 2018
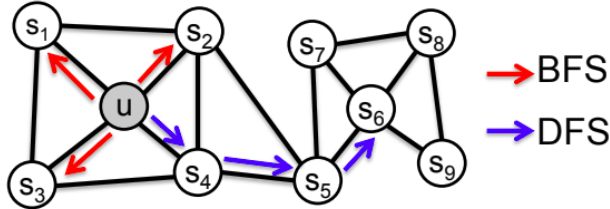
İçerik

- Projenin Şeması ve Tanımı

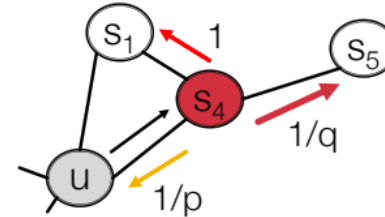- Neler Yapıldı ve Sonuçlar?

- Başarı Kriterleri

- Kaynaklar

- **Proje Nedir?**
  - Sosyal medya verilerinden çizge vektörlerinin çıkarılması ve düğümler arasındaki yakınlıkların node2vec vektörleriyle gösterilmesi
  -

- **Node2Vec**
  - Random Walk Based(Unsupervised) Feature Learning
  - Node -> word, path -> sentence

$N_{BFS}(u) = \{ s_1, s_2, s_3 \}$ Local microscopic view

$N_{DFS}(u) = \{ s_4, s_5, s_6 \}$ Global macroscopic view
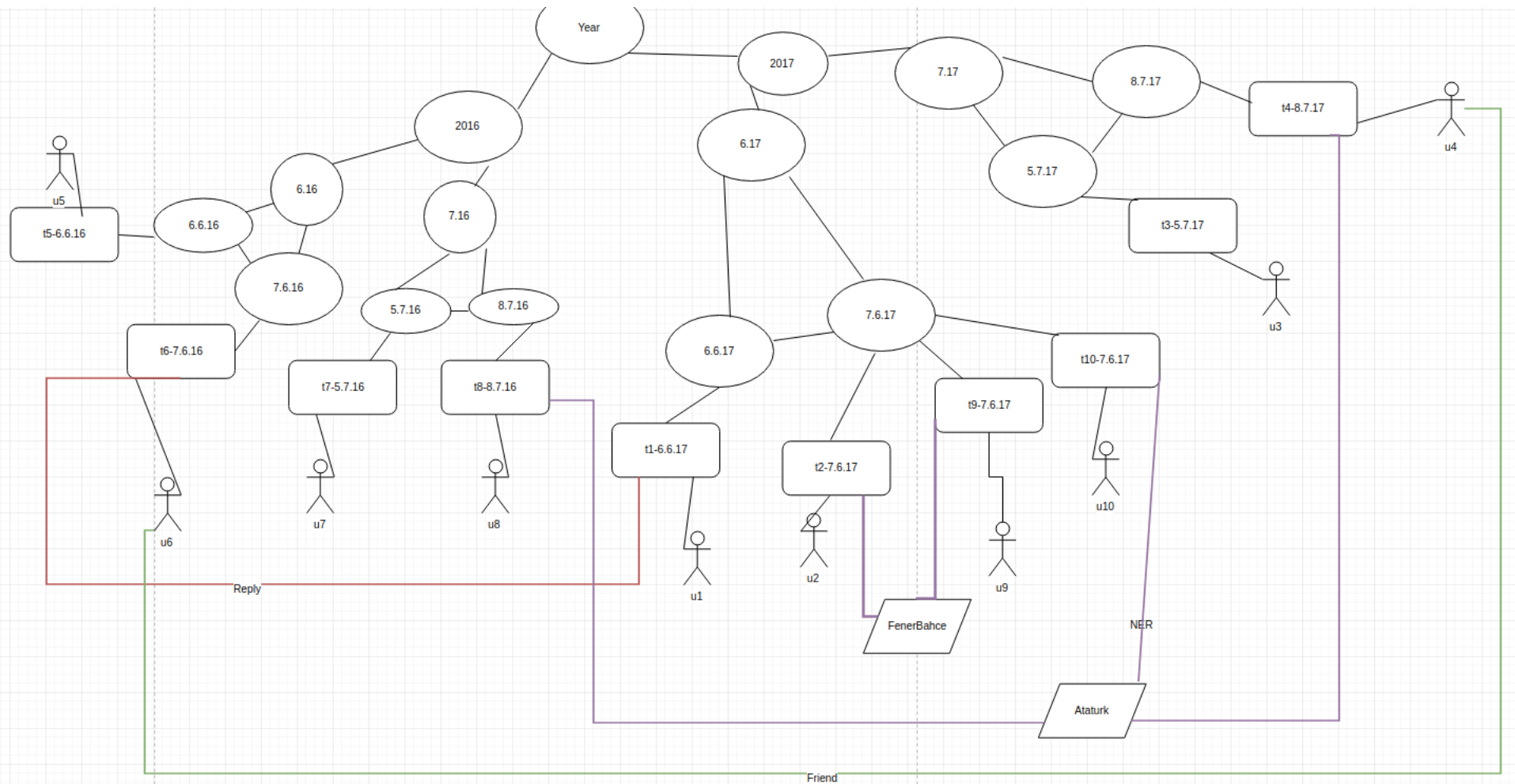
$u \rightarrow s_4 \rightarrow ?$

p-> Bir önceki düğüme geri dön
q-> BFS, DFS

- Örnek tweet veritabanı oluşturulup benzerliklere bakıldı.

- Tweetler arası Cosine Similarity

	t1-6.6.17	t2-7.6.17	t9-7.6.17	t10-7.6.17	t3-5.7.17	t4-8.7.17	t5-6.6.16	t6-7.6.16	t7-5.7.16	t8-8.7.16	Ataturk	Fenerbahce
t1-6.6.17	1.000000	0.572257	0.559279	0.441329	0.311849	0.472269	0.442075	0.631455	0.173818	0.270076	0.421397	0.549774
t2-7.6.17	0.572257	1.000000	0.886658	0.642856	0.343982	0.468388	0.356305	0.474113	0.316525	0.429324	0.566828	0.839289
t9-7.6.17	0.559279	0.886658	1.000000	0.642328	0.342840	0.465838	0.344921	0.481054	0.347912	0.416909	0.574576	0.845471
t10-7.6.17	0.441329	0.642856	0.642328	1.000000	0.334221	0.546497	0.197289	0.389078	0.325023	0.566032	0.674901	0.618860
t3-5.7.17	0.311849	0.343982	0.342840	0.334221	1.000000	0.562075	0.161957	0.317815	0.158607	0.255837	0.391825	0.319875
t4-8.7.17	0.472269	0.468388	0.465838	0.546497	0.562075	1.000000	0.288689	0.579483	0.189491	0.440060	0.523337	0.419330
t5-6.6.16	0.442075	0.356305	0.344921	0.197289	0.161957	0.288689	1.000000	0.515816	0.124489	0.133310	0.217849	0.331898
t6-7.6.16	0.631455	0.474113	0.481054	0.389078	0.317815	0.579483	0.515816	1.000000	0.143896	0.284175	0.435890	0.452710
t7-5.7.16	0.173818	0.316525	0.347912	0.325023	0.158607	0.189491	0.124489	0.143896	1.000000	0.555741	0.407248	0.348478
t8-8.7.16	0.270076	0.429324	0.416909	0.566032	0.255837	0.440060	0.133310	0.284175	0.555741	1.000000	0.630416	0.435569
Ataturk	0.421397	0.566828	0.574576	0.674901	0.391825	0.523337	0.217849	0.435890	0.407248	0.630416	1.000000	0.539073
Fenerbahce	0.549774	0.839289	0.845471	0.618860	0.319875	0.419330	0.331898	0.452710	0.348478	0.435569	0.539073	1.000000

- Günler arası Cosine Similarity

	2016	2017	6.16	6.6.16	7.6.16	7.16	5.7.16	8.7.16	6.17	6.6.17	7.6.17	7.17	5.7.17	8.7.17
2016	1.000000	0.598326	0.647383	0.630970	0.604029	0.641924	0.670506	0.627349	0.592653	0.482654	0.379833	0.558716	0.413130	0.489229
2017	0.598326	1.000000	0.571504	0.456125	0.473436	0.596240	0.527112	0.493643	0.521053	0.475840	0.421288	0.564449	0.591205	0.596420
6.16	0.647383	0.571504	1.000000	0.785970	0.716352	0.682797	0.601981	0.547013	0.588789	0.504605	0.341317	0.532749	0.394160	0.486608
6.6.16	0.630970	0.456125	0.785970	1.000000	0.768501	0.562804	0.454280	0.432787	0.457399	0.457325	0.283527	0.392932	0.346725	0.415442
7.6.16	0.604029	0.473436	0.716352	0.768501	1.000000	0.559089	0.475362	0.445300	0.490915	0.575356	0.270918	0.465337	0.389163	0.504084
7.16	0.641924	0.596240	0.682797	0.562804	0.559089	1.000000	0.777051	0.724714	0.604624	0.418651	0.383337	0.501499	0.384700	0.451763
5.7.16	0.670506	0.527112	0.601981	0.454280	0.475362	0.777051	1.000000	0.789299	0.482408	0.371788	0.398187	0.431579	0.332632	0.415712
8.7.16	0.627349	0.493643	0.547013	0.432787	0.445300	0.724714	0.789299	1.000000	0.468031	0.347931	0.429005	0.447705	0.376765	0.472852
6.17	0.592653	0.521053	0.588789	0.457399	0.490915	0.604624	0.482408	0.468031	1.000000	0.504262	0.425077	0.586516	0.515765	0.492454
6.6.17	0.482654	0.475840	0.504605	0.457325	0.575356	0.418651	0.371788	0.347931	0.504262	1.000000	0.452786	0.411083	0.314536	0.390264
7.6.17	0.379833	0.421288	0.341317	0.283527	0.270918	0.383337	0.398187	0.429005	0.425077	0.452786	1.000000	0.309105	0.235599	0.312527
7.17	0.558716	0.564449	0.532749	0.392932	0.465337	0.501499	0.431579	0.447705	0.586516	0.411083	0.309105	1.000000	0.737889	0.660702
5.7.17	0.413130	0.591205	0.394160	0.346725	0.389163	0.384700	0.332632	0.376765	0.515765	0.314536	0.235599	0.737889	1.000000	0.742958
8.7.17	0.489229	0.596420	0.486608	0.415442	0.504084	0.451763	0.415712	0.472852	0.492454	0.390264	0.312527	0.660702	0.742958	1.000000

- Kişi vektörleri için Cosine Similarity

	u1	u2	u3	u4	u5	u6	u7	u8	u9	u10
u1	1.000000	0.220591	0.268791	0.473416	0.397479	0.500663	0.320796	0.269443	0.200918	0.231947
u2	0.220591	1.000000	0.083940	0.175423	0.166321	0.167029	0.276929	0.292574	0.713989	0.358272
u3	0.268791	0.083940	1.000000	0.439557	0.284998	0.398915	0.291745	0.326546	0.094297	0.255834
u4	0.473416	0.175423	0.439557	1.000000	0.443557	0.514040	0.388228	0.440042	0.194506	0.393321
u5	0.397479	0.166321	0.284998	0.443557	1.000000	0.478535	0.409996	0.356223	0.134811	0.211692
u6	0.500663	0.167029	0.398915	0.514040	0.478535	1.000000	0.355785	0.418397	0.181256	0.367694
u7	0.320796	0.276929	0.291745	0.388228	0.409996	0.355785	1.000000	0.653034	0.274495	0.387762
u8	0.269443	0.292574	0.326546	0.440042	0.356223	0.418397	0.653034	1.000000	0.275342	0.523879
u9	0.200918	0.713989	0.094297	0.194506	0.134811	0.181256	0.274495	0.275342	1.000000	0.387041
u10	0.231947	0.358272	0.255834	0.393321	0.211692	0.367694	0.387762	0.523879	0.387041	1.000000

* u1-u6 Reply
* u2-u9 Same NER
* u6-u4 Friend

Neler Yapıldı? - 2-2

- 300 gerçek tweet ile denemeler yapıldı.

```
############ Tweet: ############
SendByScreenName:demirmadeni
UserID:u1658471437
TweetID:t765466344707989504
InReplyToStatusID:t-1
InReplyToUserID:u-1
Date:2016-08-16 11:32:52
Text: Galatasaray basketbol şubesinden son haberler!
https://t.co/WQTFKf7X3D https://t.co/T7Ex0UiHgn

############ Tweet: ############
SendByScreenName:demarkesports
UserID:u1499266621
TweetID:t765582802343657472
InReplyToStatusID:t-1
InReplyToUserID:u-1
Date:2016-08-16 19:15:38
Text:#Rio2016 Kadınlar Basketbol çeyrek final maçında Sırbistan, Avustralya'yı 73-71 mağlup etti ve yarı finale yük
seldi. https://t.co/akxCoNG6nJ
```

```
model.docvecs.similarity('t765466344707989504','t765582802343657472')
0.7528703228033607
```

	t765562268767678464	t763141405766250496	basketbol	t765466344707989504	t765582802343657472	t765648654883512321
t765562268767678464	1.000000	0.168840	0.235586	0.336817	0.355017	0.443149
t763141405766250496	0.168840	1.000000	0.244844	0.123604	0.303882	0.155905
basketbol	0.235586	0.244844	1.000000	0.644623	0.664798	0.359757
t765466344707989504	0.336817	0.123604	0.644623	1.000000	0.627713	0.477728
t765582802343657472	0.355017	0.303882	0.664798	0.627713	1.000000	0.434969
t765648654883512321	0.443149	0.155905	0.359757	0.477728	0.434969	1.000000

- Günler arası

	2016	AGU.2016	10.8.2016	16.8.2016	EKI.2016	29.10.2016	30.10.2016	2017	OCA.2017	12.1.2017
2016	1.000000	0.642372	0.492105	0.153328	0.742586	0.347860	0.208554	0.766375	0.569516	0.181680
AGU.2016	0.642372	1.000000	0.753962	0.335239	0.587421	0.292050	0.230589	0.427036	0.319298	0.145191
10.8.2016	0.492105	0.753962	1.000000	0.275777	0.464674	0.205210	0.197475	0.354981	0.344252	0.261786
16.8.2016	0.153328	0.335239	0.275777	1.000000	0.136094	0.137177	0.195119	0.102880	0.107122	0.181643
EKI.2016	0.742586	0.587421	0.464674	0.136094	1.000000	0.575391	0.449526	0.484274	0.336304	0.125472
29.10.2016	0.347860	0.292050	0.205210	0.137177	0.575391	1.000000	0.398337	0.230422	0.192817	0.193661
30.10.2016	0.208554	0.230589	0.197475	0.195119	0.449526	0.398337	1.000000	0.150760	0.185048	0.224175
2017	0.766375	0.427036	0.354981	0.102880	0.484274	0.230422	0.150760	1.000000	0.835190	0.381484
OCA.2017	0.569516	0.319298	0.344252	0.107122	0.336304	0.192817	0.185048	0.835190	1.000000	0.631471
12.1.2017	0.181680	0.145191	0.261786	0.181643	0.125472	0.193661	0.224175	0.381484	0.631471	1.000000

Neler Yapılacak?

- Test sonuçları arttırılıp bir bildiri haline getirilecek.

- Modül dökümanı tamamlanacak.

Başarı Kriterleri

- Ardışık gün vektörlerinin diğer günlere göre daha yakın olması

- Arkadaş kişi vektörlerinin diğer kişilere göre daha yakın olması

- Benzer konu&lokasyon(NER) vektörlerinin daha yakın olması

Kaynaklar

1. Node2Vec: http://snap.stanford.edu/node2vec/ [Ziyaret Tarihi: 15/03/18]
2. Bag of Words: https://www.wikiwand.com/en/Bag-of-words_model [Ziyaret Tarihi: 15/03/18]
3. Networkx: https://networkx.github.io/ [Ziyaret Tarihi: 15/03/18]

4. Paragraph Vectors: https://github.com/inejc/paragraph-vectors [Ziyaret Tarihi: 15/03/18]
5.