Bangla Word Embedding (Vector Space Model)

R & D Team, Pipilika.

Word Embedding(Word2vec) - Overview

- Word Embedding represents words in a continuous vector space where semantically similar words are mapped to nearby points.

- Word Embedding depend on the Distributional Hypothesis, which states that words that appear in the same contexts share semantic meaning.

- Word2vec “vectorizes” words, and by doing so, it makes natural language computer-readable.

Word Embedding(Word2vec)

- Word2vec is a particularly computationally-efficient predictive model for learning word embedding's from raw text.

- Word2vec trains words against other words that neighbor them in the input corpus.

- Word2vec is a two-layer neural net that processes text.

Word2Vec Publication

Distributed Representations of Words and Phrases and their Compositionality (2013)

[Published at arXiv (Cornell University), Submitted on 16 Oct 2013]

Tomas Mikolov (Google Inc.),
Ilya Sutskever (Google Inc.),
Kai Chen (Google Inc.),
Greg Corrado(Google Inc.),
Jeffrey Dean(Google Inc.)

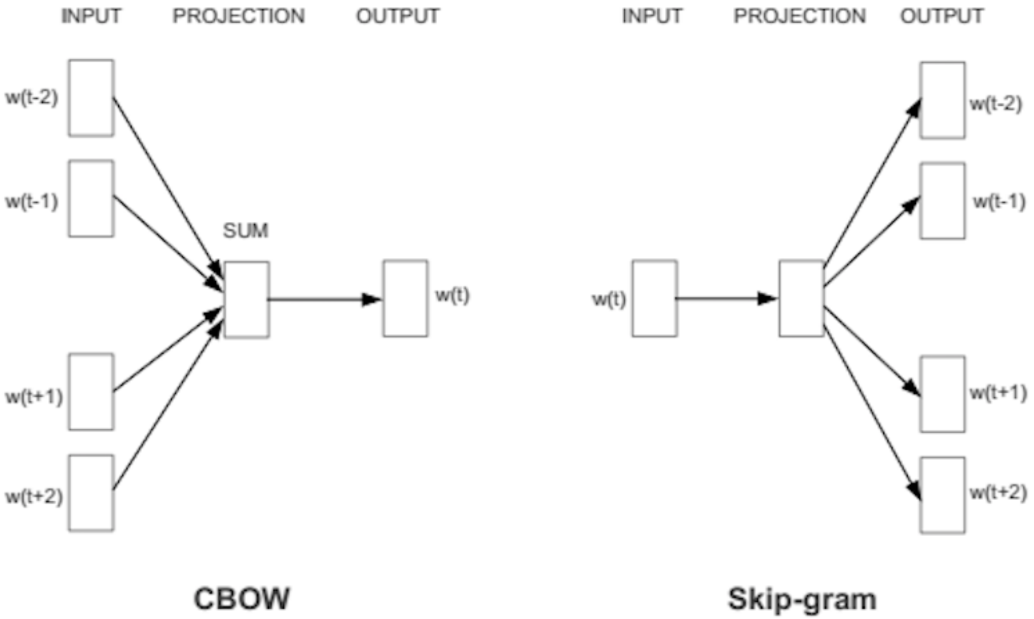
Word Embedding(Word2vec)

- Word2vec is a particularly computationally-efficient predictive model for learning word embedding's from raw text.

- Word2vec trains words against other words that neighbor them in the input corpus.

- Word2vec is a two-layer neural net that processes text.

Word2vec - How to compute



- Using context to predict a target word
- Using a word to predict a target context

□ We used skip-gram method, as its computationally faster and performs well for large dataset.

Word2vec - How to compute

- All vectors are initialized as random points in space.

- The entries in the vectors are treated as parameters to be learned.

- Word2vec use stochastic gradient based training method over SGNS (negative sampling) to reduce cost function.

- The negative sampling objective tries to maximize P(D = 1|w, c) for observed (w, c) pairs while maximizing P(D = 0|w, c) for randomly sampled “negative” examples. (w=word, c = context)

- Optimizing cost makes observed word-context pairs have similar embeddings, while scattering unobserved pairs.

Word2vec - Uses

- Word’s association with other words (e.g. “man” is to “boy” what “woman” is to “girl”)
- Cluster documents and classify them by topic.
- Named Entity Recognition (NER)
- Parts of Speech tagging (POS)
- Machine translation (MT)
- Sentiment analysis (SA)
- Search.
- Recommendations in such diverse fields as scientific research, legal discovery, e-commerce and customer relationship management.

Word2vec - How to compute

- All vectors are initialized as random points in space.

- The entries in the vectors are treated as parameters to be learned.

- Word2vec use stochastic gradient based training method over SGNS (negative sampling) to reduce cost function.

- The negative sampling objective tries to maximize P(D = 1|w, c) for observed (w, c) pairs while maximizing P(D = 0|w, c) for randomly sampled "negative" examples. (w=word, c = context)

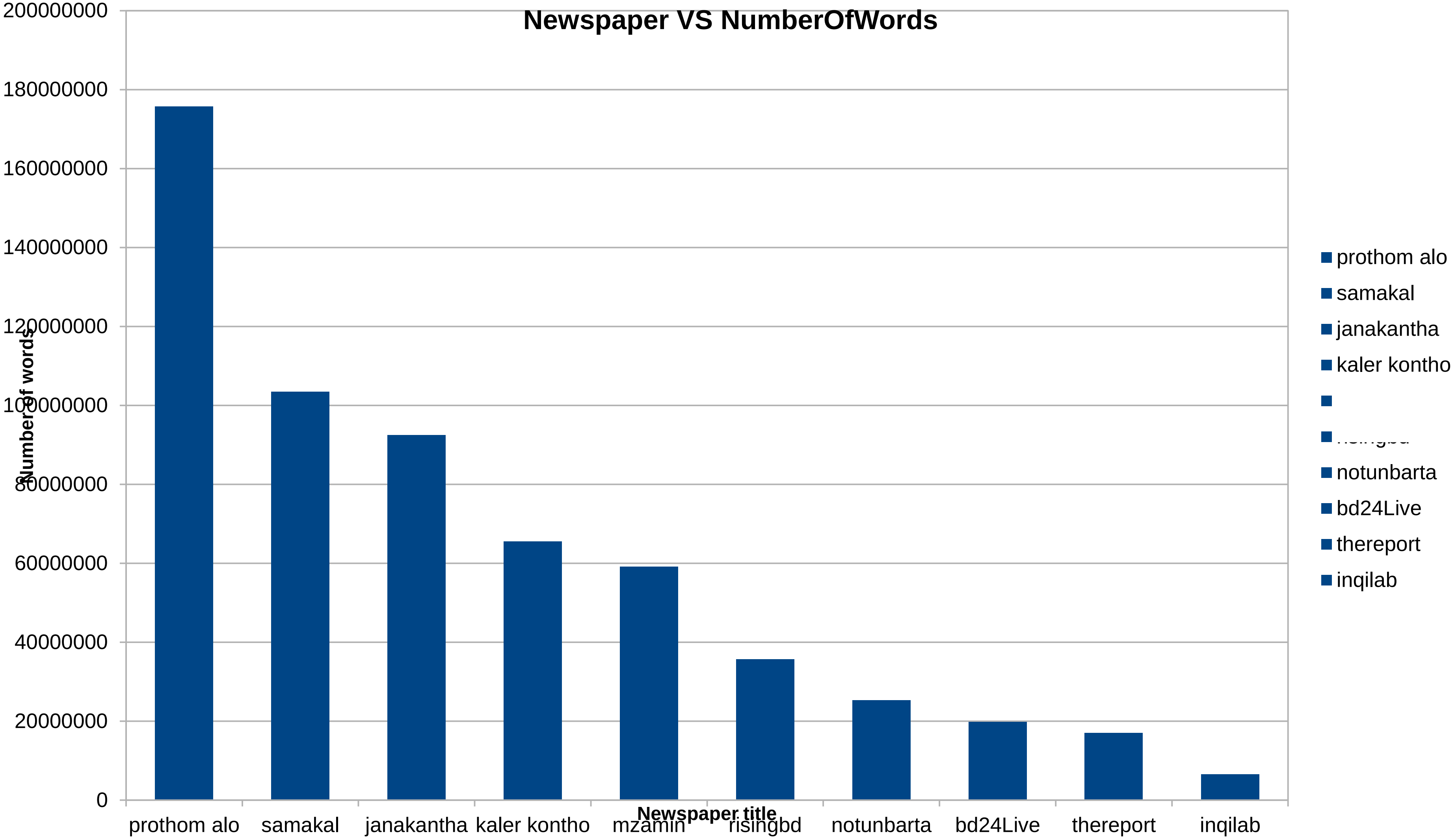
- Optimizing cost makes observed word-context pairs have similar embeddings, while scattering unobserved pairs.

Bangla Word Embedding

Bangla Word Embedding

- Collect online newspaper data and parse articles.
- Refine text data and remove noise.
- Token sentences.
- Train word2vec model with Neural Network.
- Evaluate model.

Bangla Word Embedding - Dataset



Bangla Word Embedding - Process

- Generated Embedding model for every newspaper separately.

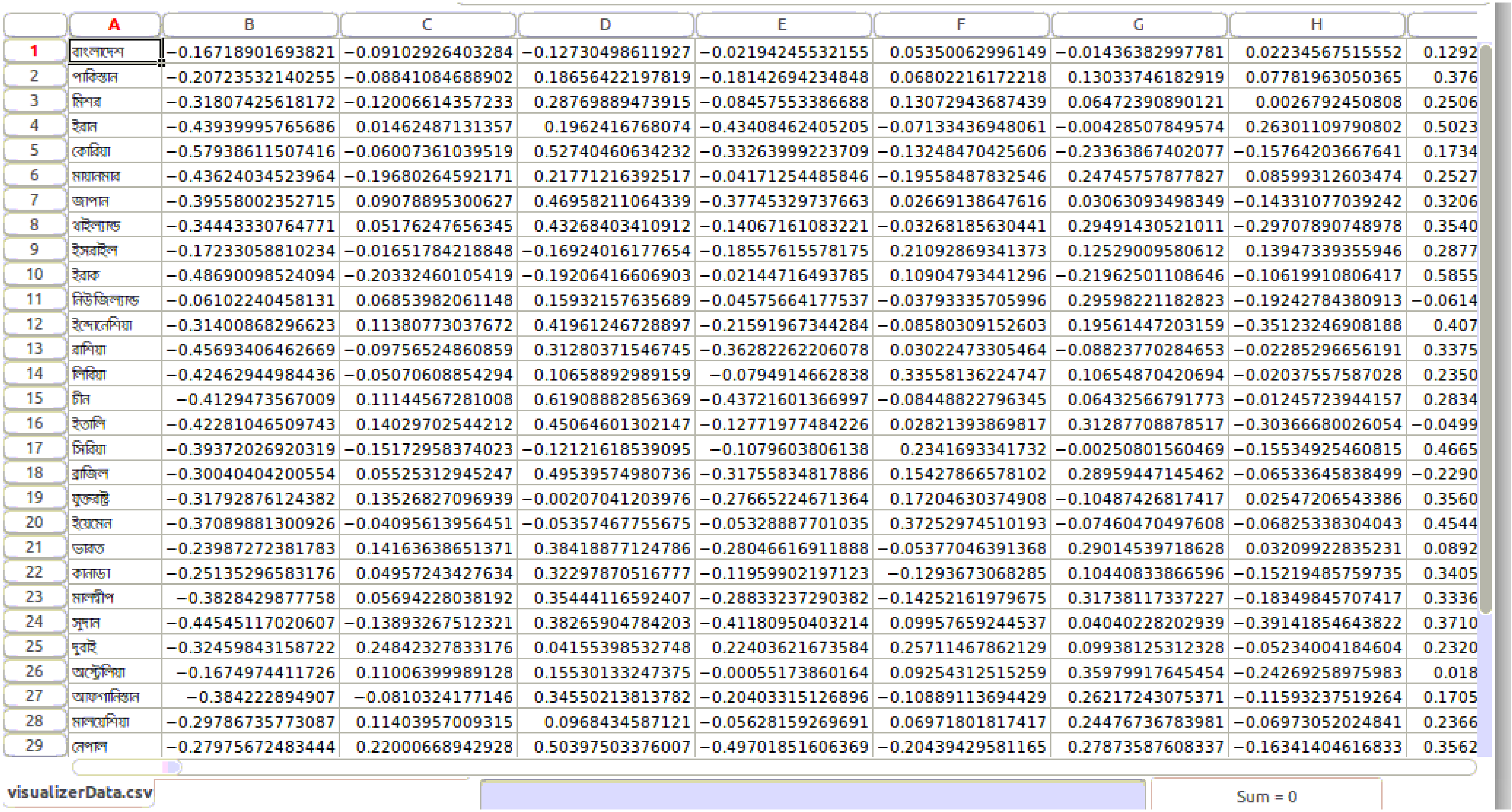
- Generated Embedding model using the total content.

- We used vector size 100, window size 5, min occurrence 5 and a two layer Neural Network.

- Vector size of final model is 663843 (Unique words)

- Embeddings was generated using DeepLearning4J's word2vec implementation (open source java library).

Word embedding – Vector representation



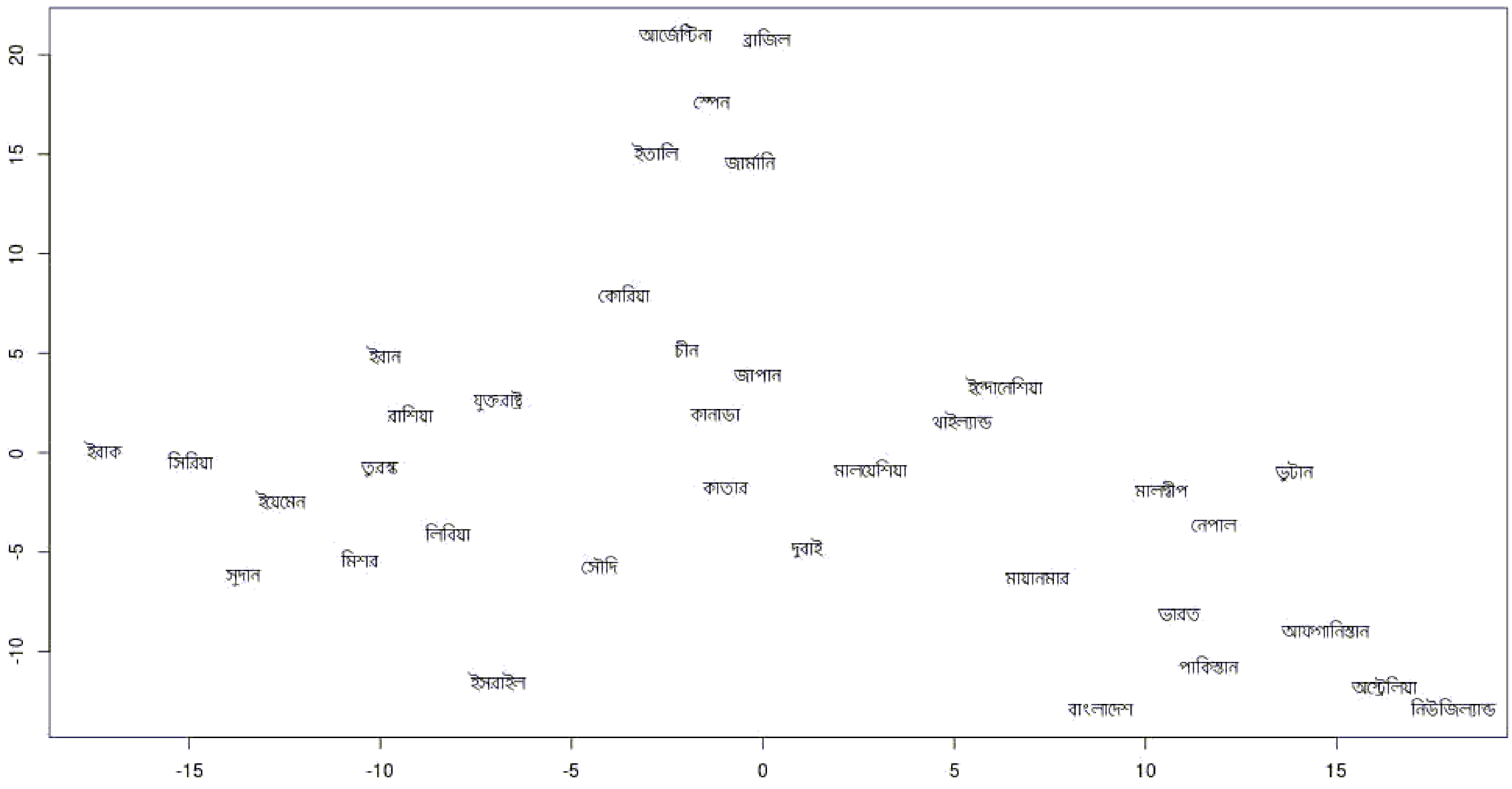
	A	B	C	D	E	F	G	H	
1	বাংলাদেশ	−0.16718901693821	−0.09102926403284	−0.12730498611927	−0.02194245532155	0.05350062996149	−0.01436382997781	0.02234567515552	0.1292
2	পাকিস্তান	−0.20723532140255	−0.08841084688902	0.18656422197819	−0.18142694234848	0.06802216172218	0.13033746182919	0.07781963050365	0.376
3	মিশর	−0.31807425618172	−0.12006614357233	0.28769889473915	−0.08457553386688	0.13072943687439	0.06472390890121	0.0026792450808	0.2506
4	ইরান	−0.43939995765686	0.01462487131357	0.1962416768074	−0.43408462405205	−0.07133436948061	−0.00428507849574	0.26301109790802	0.5023
5	কোরিয়া	−0.57938611507416	−0.06007361039519	0.52740460634232	−0.33263999223709	−0.13248470425606	−0.23363867402077	−0.15764203667641	0.1734
6	মায়ানমার	−0.43624034523964	−0.19680264592171	0.21771216392517	−0.04171254485846	−0.19558487832546	0.24745757877827	0.08599312603474	0.2527
7	জাপান	−0.39558002352715	0.09078895300627	0.46958211064339	−0.37745329737663	0.02669138647616	0.03063093498349	−0.14331077039242	0.3206
8	থাইল্যান্ড	−0.34443330764771	0.05176247656345	0.43268403410912	−0.14067161083221	−0.03268185630441	0.29491430521011	−0.29707890748978	0.3540
9	ইসরাইল	−0.17233058810234	−0.01651784218848	−0.16924016177654	−0.18557615578175	0.21092869341373	0.12529009580612	0.13947339355946	0.2877
10	ইরাক	−0.48690098524094	−0.20332460105419	−0.19206416606903	−0.02144716493785	0.10904793441296	−0.21962501108646	−0.10619910806417	0.5855
11	নিউজিল্যান্ড	−0.06102240458131	0.06853982061148	0.15932157635689	−0.04575664177537	−0.03793335705996	0.29598221182823	−0.19242784380913	−0.0614
12	ইন্দোনেশিয়া	−0.31400868296623	0.11380773037672	0.41961246728897	−0.21591967344284	−0.08580309152603	0.19561447203159	−0.35123246908188	0.407
13	রাশিয়া	−0.45693406462669	−0.09756524860859	0.31280371546745	−0.36282262206078	0.03022473305464	−0.08823770284653	−0.02285296656191	0.3375
14	লিবিয়া	−0.42462944984436	−0.05070608854294	0.10658892989159	−0.0794914662838	0.33558136224747	0.10654870420694	−0.02037557587028	0.2350
15	চীন	−0.4129473567009	0.11144567281008	0.61908882856369	−0.43721601366997	−0.08448822796345	0.06432566791773	−0.01245723944157	0.2834
16	ইতালি	−0.42281046509743	0.14029702544212	0.45064601302147	−0.12771977484226	0.02821393869817	0.31287708878517	−0.30366680026054	−0.0499
17	সিরিয়া	−0.39372026920319	−0.15172958374023	−0.12121618539095	−0.1079603806138	0.2341693341732	−0.00250801560469	−0.15534925460815	0.4665
18	ব্রাজিল	−0.30040404200554	0.05525312945247	0.49539574980736	−0.31755834817886	0.15427866578102	0.28959447145462	−0.06533645838499	−0.2290
19	যুক্তরাষ্ট্র	−0.31792876124382	0.13526827096939	−0.00207041203976	−0.27665224671364	0.17204630374908	−0.10487426817417	0.02547206543386	0.3560
20	ইয়েমেন	−0.37089881300926	−0.04095613956451	−0.05357467755675	−0.05328887701035	0.37252974510193	−0.07460470497608	−0.06825338304043	0.4544
21	ভারত	−0.23987272381783	0.14163638651371	0.38418877124786	−0.28046616911888	−0.05377046391368	0.29014539718628	0.03209922835231	0.0892
22	কানাডা	−0.25135296583176	0.04957243427634	0.32297870516777	−0.11959902197123	−0.1293673068285	0.10440833866596	−0.15219485759735	0.3405
23	মালদ্বীপ	−0.3828429877758	0.05694228038192	0.35444116592407	−0.28833237290382	−0.14252161979675	0.31738117337227	−0.18349845707417	0.3336
24	সুদান	−0.44545117020607	−0.13893267512321	0.38265904784203	−0.41180950403214	0.09957659244537	0.04040228202939	−0.39141854643822	0.3710
25	দুবাই	−0.32459843158722	0.24842327833176	0.04155398532748	0.22403621673584	0.25711467862129	0.09938125312328	−0.05234004184604	0.2320
26	অস্ট্রেলিয়া	−0.1674974411726	0.11006399989128	0.15530133247375	−0.00055173860164	0.09254312515259	0.35979917645454	−0.24269258975983	0.018
27	আফগানিস্তান	−0.384222894907	−0.0810324177146	0.34550213813782	−0.20403315126896	−0.10889113694429	0.26217243075371	−0.11593237519264	0.1705
28	মালয়েশিয়া	−0.29786735773087	0.11403957009315	0.0968434587121	−0.05628159269691	0.06971801817417	0.24476736783981	−0.06973052024841	0.2366
29	নেপাল	−0.27975672483444	0.22000668942928	0.50397503376007	−0.49701851606369	−0.20439429581165	0.27873587608337	−0.16341404616833	0.3562

If we want to plot this data, we need to apply dimension reduction first.

Clustering Using Word embedding

Word embedding – Plotting Vectors

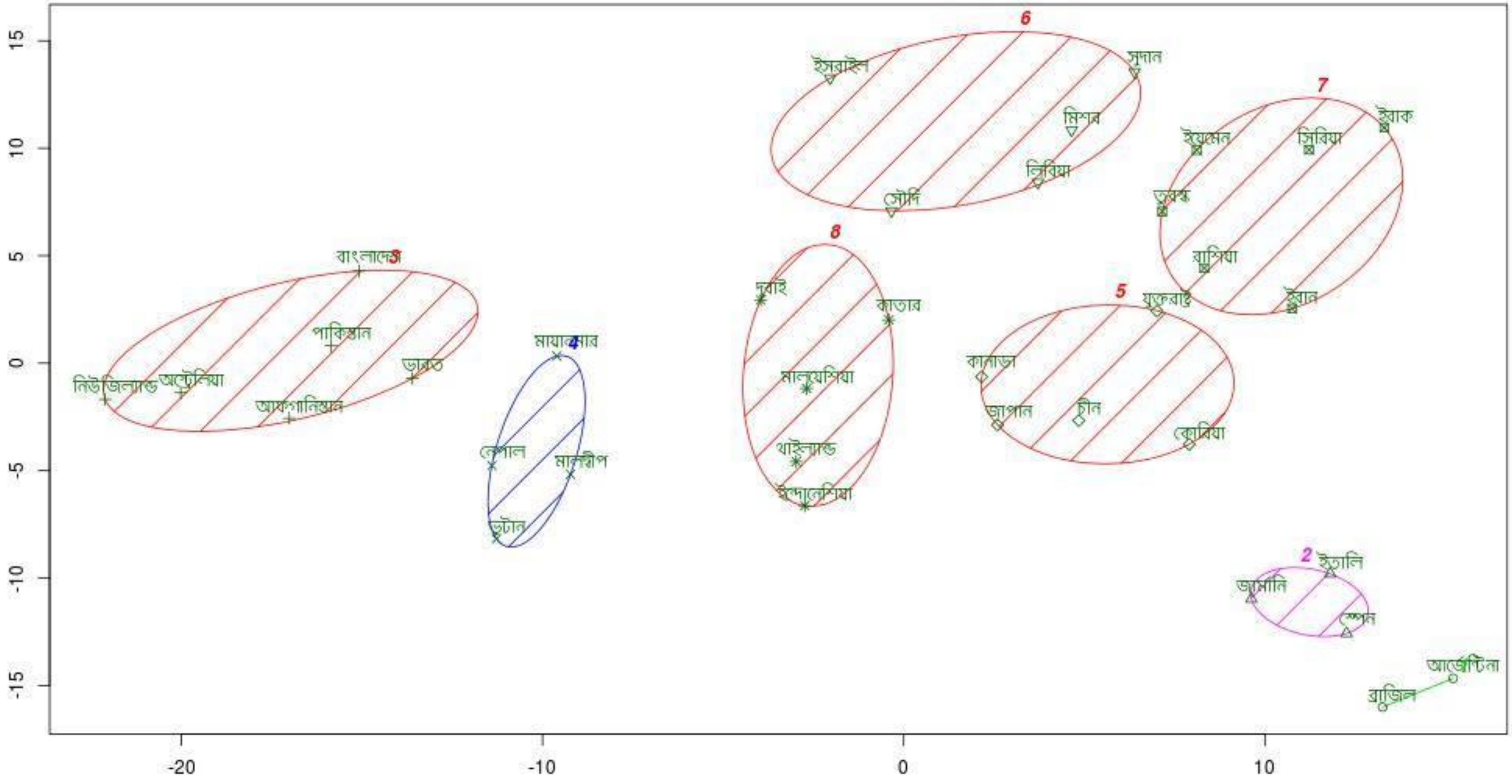
Words in two dimensions



After reducing dimension size to 2 from 100 using **t-sne** algorithm

Word embedding - Clustering countries

Clusters of Countries, K-means (k = 8)

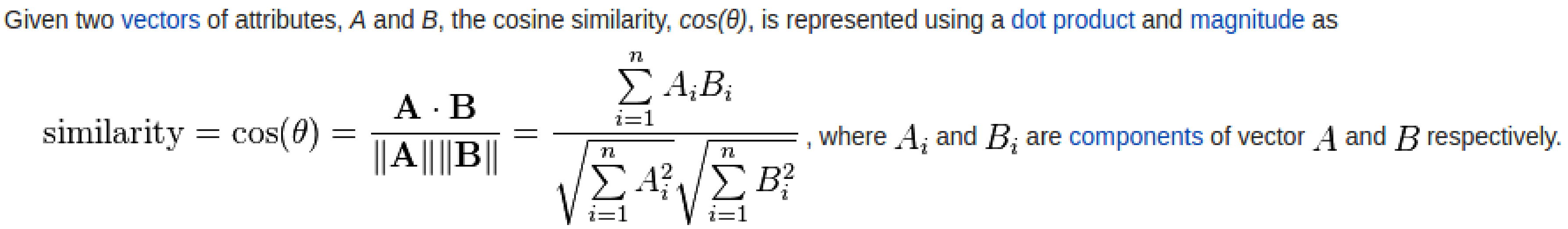


Interesting: Countries with common affairs tend to stay in same cluster.

Cosine Similarity between words

Cosine Similarity

Formal Definition :



Given two vectors of attributes, *A* and *B*, the cosine similarity, *cos(θ)*, is represented using a dot product and magnitude as

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

, where A_i and B_i are components of vector A and B respectively.

Step 1 : Take a subset of words from embedding model ***S.***

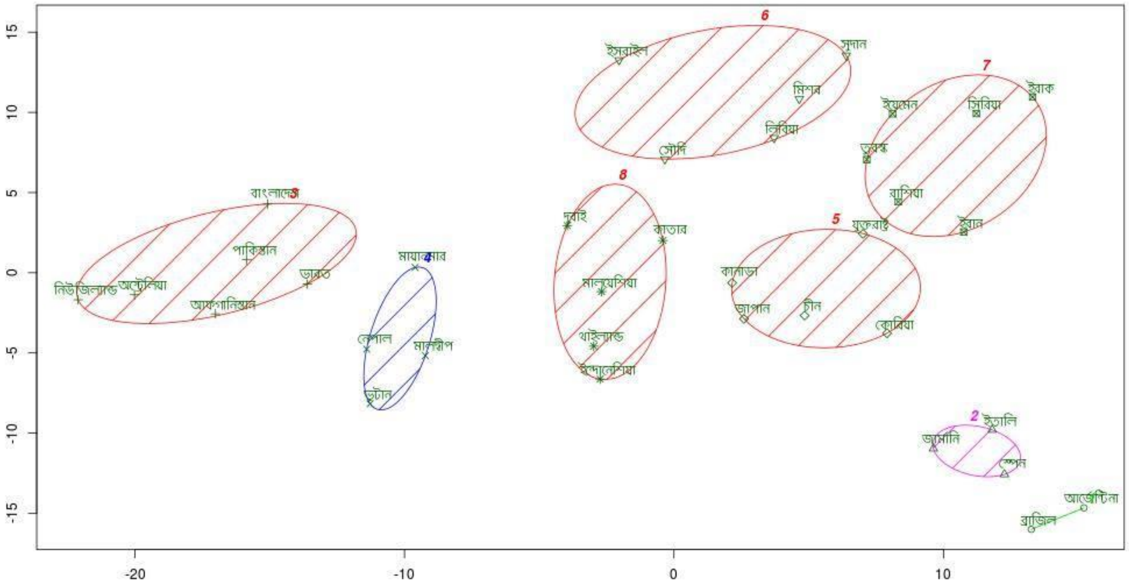
Step 2 : Take a word ***A.***

Step 3 : Calculate cosine similarity of each word in set ***S*** with word ***A.***

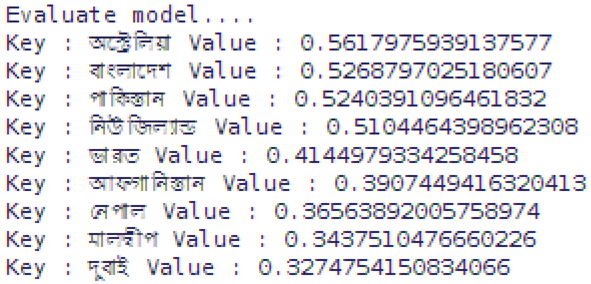
Step 4 : Sort the values of ***S*** according to score,
top elements are most similar to ***A.***

Cosine similarity

(Interesting properties of word embedding)

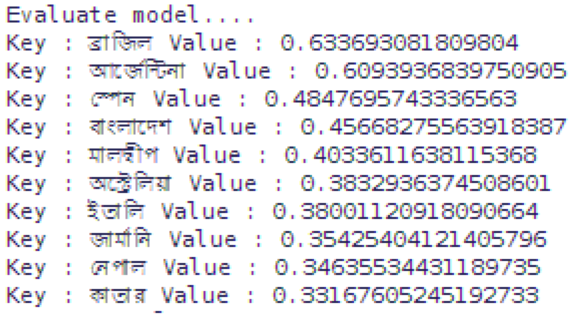


Cosine similarity with : ক্রিকেট



```
Evaluate model....
Key : অস্ট্রেলিয়া Value : 0.5617975939137577
Key : বাংলাদেশ Value : 0.5268797025180607
Key : পাকিস্তান Value : 0.5240391096461832
Key : নিউজিল্যান্ড Value : 0.5104464398962308
Key : ভারত Value : 0.4144979334258458
Key : আফগানিস্তান Value : 0.3907449416320413
Key : নেপাল Value : 0.36563892005758974
Key : মালদ্বীপ Value : 0.3437510476660226
Key : দুবাই Value : 0.3274754150834066
```

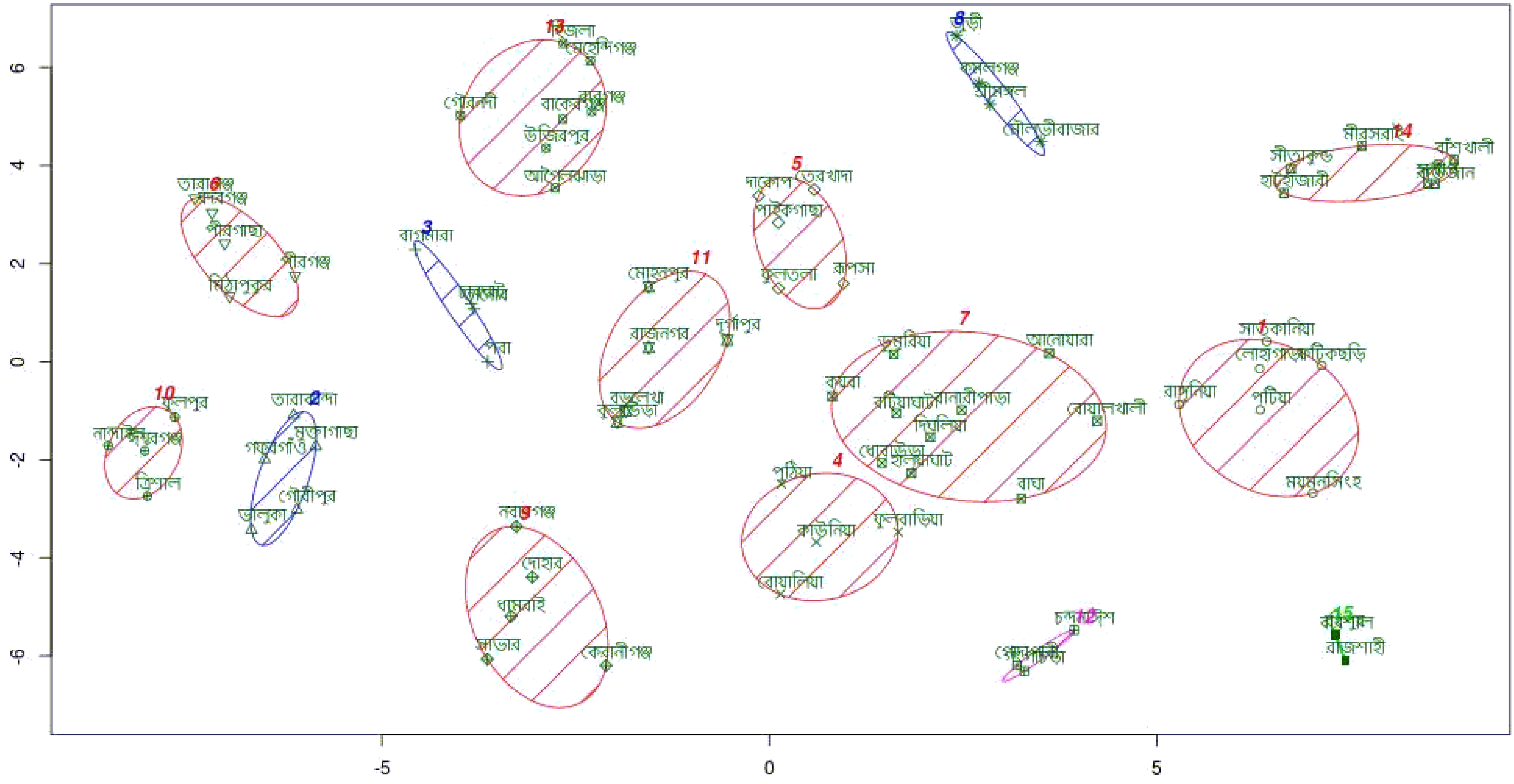
Cosine similarity with : ফুটবল



```
Evaluate model....
Key : ব্রাজিল Value : 0.633693081809804
Key : আর্জেন্টিনা Value : 0.6093936839750905
Key : স্পেন Value : 0.4847695743336563
Key : বাংলাদেশ Value : 0.45668275563918387
Key : মালদ্বীপ Value : 0.4033611638115368
Key : অস্ট্রেলিয়া Value : 0.3832936374508601
Key : ইতালি Value : 0.38001120918090664
Key : জার্মানি Value : 0.35425404121405796
Key : নেপাল Value : 0.34635534431189735
Key : কাতার Value : 0.33167605245192733
```

Word embedding - Clustering sub-districts

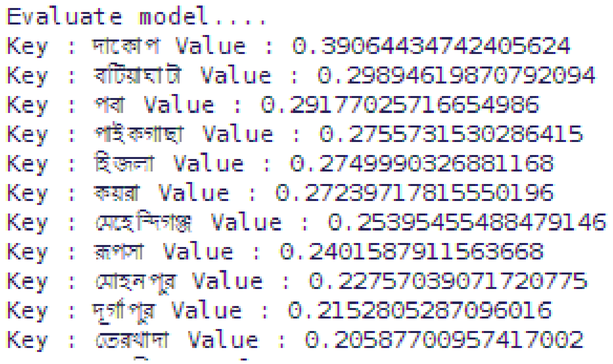
Clusters of sub-districts, K-means (k = 15)



Interesting: Sub-districts of a particular district tend to stay in same cluster.

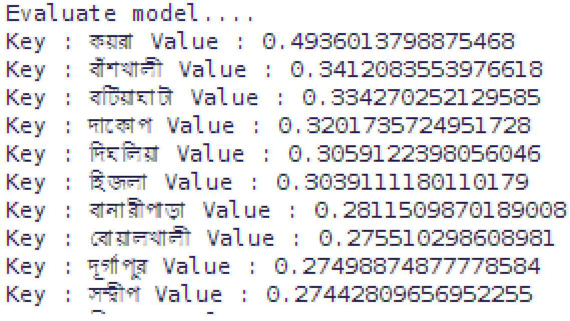
Cosine similarity
(Interesting properties of word embedding)

Cosine similarity with : নদীভাঙ্গন



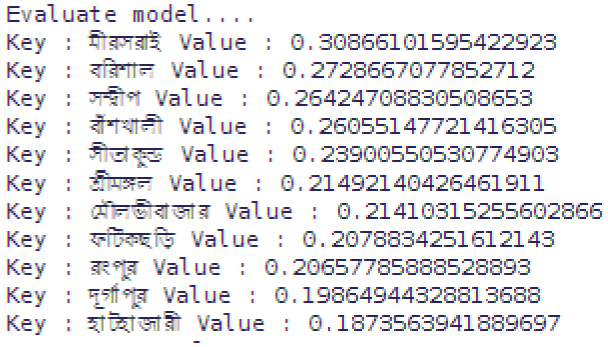
```
Evaluate model....
Key : দাকোপ Value : 0.39064434742405624
Key : বটিয়াঘাটা Value : 0.29894619870792094
Key : পবা Value : 0.29177025716654986
Key : পাইকগাছা Value : 0.2755731530286415
Key : হিজলা Value : 0.2749990326881168
Key : কয়রা Value : 0.27239717815550196
Key : মেহেন্দিগঞ্জ Value : 0.25395455488479146
Key : রূপসা Value : 0.2401587911563668
Key : মোহনপুর Value : 0.22757039071720775
Key : দুর্গাপুর Value : 0.2152805287096016
Key : তেরখাদা Value : 0.20587700957417002
```

Cosine similarity with : জলদস্যু



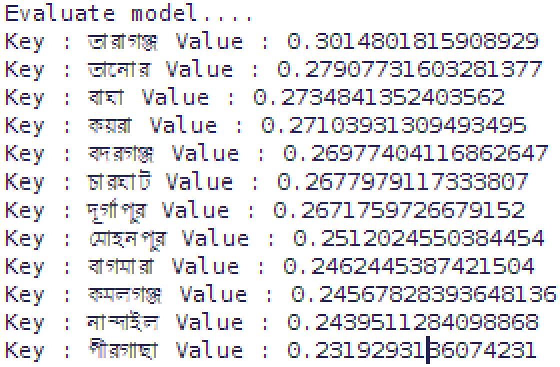
```
Evaluate model....
Key : কয়রা Value : 0.4936013798875468
Key : বাঁশখালী Value : 0.3412083553976618
Key : বটিয়াঘাটা Value : 0.334270252129585
Key : দাকোপ Value : 0.3201735724951728
Key : দিঘলিয়া Value : 0.3059122398056046
Key : হিজলা Value : 0.3039111180110179
Key : বানারীপাড়া Value : 0.2811509870189008
Key : বোয়ালখালী Value : 0.275510298608981
Key : দুর্গাপুর Value : 0.27498874877778584
Key : সন্দ্বীপ Value : 0.27442809656952255
```

Cosine similarity with : পাহাড়



```
Evaluate model....
Key : মীরসরাই Value : 0.30866101595422923
Key : বরিশাল Value : 0.2728667077852712
Key : সন্দ্বীপ Value : 0.26424708830508653
Key : বাঁশখালী Value : 0.26055147721416305
Key : সীতাকুন্ড Value : 0.23900550530774903
Key : শ্রীমঙ্গল Value : 0.21492140426461911
Key : মৌলভীবাজার Value : 0.21410315255602866
Key : ফটিকছড়ি Value : 0.2078834251612143
Key : রংপুর Value : 0.20657785888528893
Key : দুর্গাপুর Value : 0.19864944328813688
Key : হাটহাজারী Value : 0.1873563941889697
```

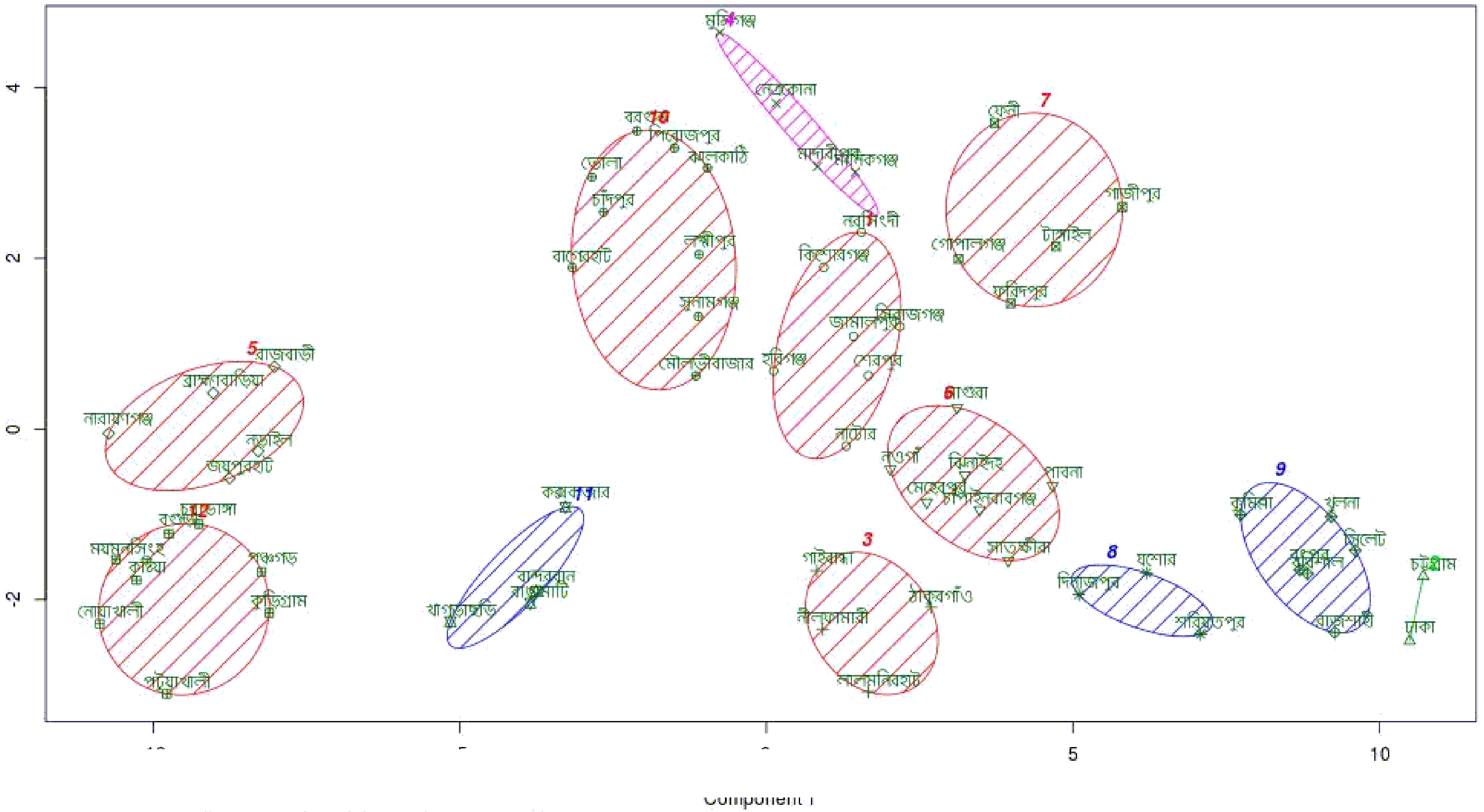
Cosine similarity with : লিচু



```
Evaluate model....
Key : তারাগঞ্জ Value : 0.3014801815908929
Key : তানোর Value : 0.27907731603281377
Key : বাঘা Value : 0.2734841352403562
Key : কয়রা Value : 0.27103931309493495
Key : বদরগঞ্জ Value : 0.26977404116862647
Key : চারঘাট Value : 0.2677979117333807
Key : দুর্গাপুর Value : 0.2671759726679152
Key : মোহনপুর Value : 0.2512024550384454
Key : বাগমারা Value : 0.2462445387421504
Key : কমলগঞ্জ Value : 0.24567828393648136
Key : নান্দাইল Value : 0.2439511284098868
Key : পীরগাছা Value : 0.2319293136074231
```

Word embedding - Clustering districts

Clusters of districts, K-means (k = 12)

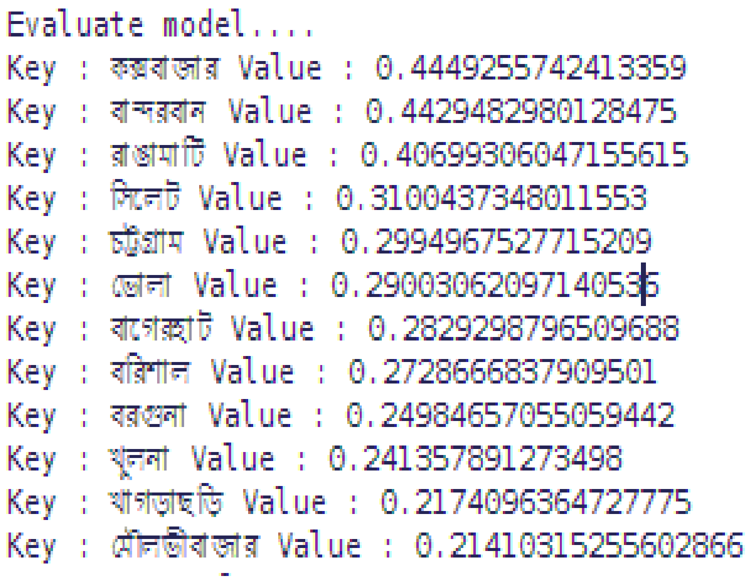


These two components explain 100 % of the point variability.

Interesting: Dhaka and Ctg. are unlike any other districts.

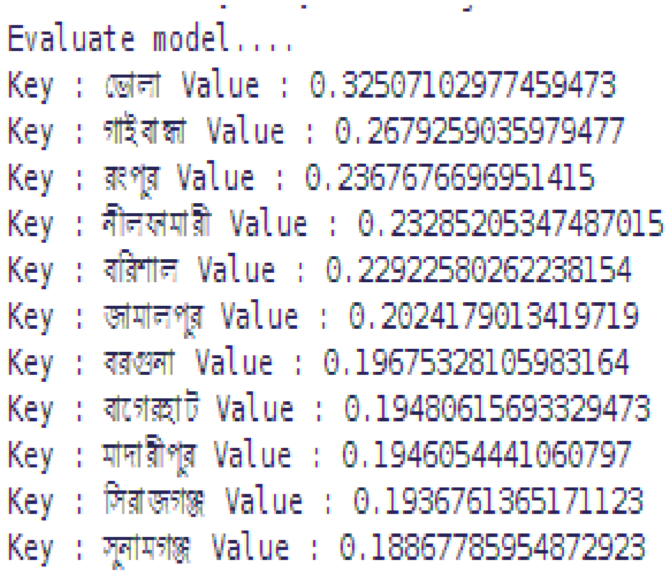
Cosine similarity
(Interesting properties of word embedding)

Cosine similarity with : পাহাড়



```
Evaluate model....
Key : কক্সবাজার Value : 0.4449255742413359
Key : বান্দরবান Value : 0.4429482980128475
Key : রাঙামাটি Value : 0.4069930604715615
Key : সিলেট Value : 0.3100437348011553
Key : চট্টগ্রাম Value : 0.2994967527715209
Key : ভোলা Value : 0.29003062097140535
Key : বাগেরহাট Value : 0.2829298796509688
Key : বরিশাল Value : 0.2728666837909501
Key : বরগুনা Value : 0.24984657055059442
Key : খুলনা Value : 0.241357891273498
Key : খাগড়াছড়ি Value : 0.2174096364727775
Key : মৌলভীবাজার Value : 0.21410315255602866
```

Cosine similarity with : বজ্রপাত



```
Evaluate model....
Key : ভোলা Value : 0.3250710297745947З
Key : গাইবান্ধা Value : 0.2679259035979477
Key : রংপুর Value : 0.236767669695141５
Key : নীলফামারী Value : 0.23285205347487015
Key : বরিশাল Value : 0.2292258026223815４
Key : জামালপুর Value : 0.2024179013419719
Key : বরগুনা Value : 0.1967532810598316４
Key : বাগেরহাট Value : 0.1948061569332947З
Key : মাদারীপুর Value : 0.194605444106079７
Key : সিরাজগঞ্জ Value : 0.1936761365171123
Key : সুনামগঞ্জ Value : 0.18867785954872923
```

Thank You!