Habib University
CS 416 - Algorithms for Machine Learning
Fall 2018

Emad Bin Abid
*ea*02893

Assignment 03
Procedures
Submitted: November 08th, 2018

Scraping:

An scraping is a process generally known to extract something of interest from any source or origin. In our case, scraping is described as a process to extract data from any data source. ***Web***, probably, is the largest source of data which currently exists. Hence, web scraping is the process which suits us the most. *Web scraping* is described as the process of extracting the data from anywhere out of the whole web.

URLs:

The URLs used for scraping were the following:
https://jang.com.pk/roman/news/1,
https://jang.com.pk/roman/news/2,
...,
...,
...,
https://jang.com.pk/roman/news/8000

Corpus:

The main corpus consists of 12324 total words. Out of the total number of words, 3001 words are unique.

The main corpus can be found inside the **data** folder as **corpus-1-reduced.txt**.

Visualizing Embedding using t-SNE:

The model was trained using multiple values of *context size* and *embedding dimension*. The results of t-SNE were visualized and compared manually with the original corpus after 10 epochs. Some of the results are given below. Some words (out of many) checked for synonyms, analogies and misspellings are given below:

• Synonyms: mauqa (mawaqay), gayi (gaya), gaya (gaye), gaye (gayi), ilzamaat (ilzaam), ...

• Analogies: ilaaj (hospital), khitaab (conference), riyasat (hukoomat), tafteesh (police), ...

• Misspellings: shareef (sharif), karen (keren), unhein (unhen), rozgari (rozgaar), ne (nay), par (per), girftar (giraftari), nay (ny), ne (ny), se (say), bhi (bi), kiya (kya), tak (taq), ne (naay), ny (naay), nay (naay), unhon (unho), ke (kay), inho (inhon), ...

i. $\underline{context\_size = 5}$ **and** $\underline{embedding\_dimension = 100}$

Table 1: Synonyms

Word	**Embedding Vector**	**Synonym**	**Embedding Vector**
mauqa	[-47.659054 -52.449436]	mawaqay	[26.422606 117.54547]
gaya	[-124.96698 -26.481392]	gayi	[23.78688 -67.207245]
gaya	[-124.96698 -26.481392]	gaye	[159.87271 -47.362774]
gayi	[23.78688 -67.207245]	gaye	[159.87271 -47.362774]
ilzamaat	[104.44866 4.5253577]	ilzaam	[-42.44178 21.932611]

Table 2: Analogies

Word	**Embedding Vector**	**Analogy**	**Embedding Vector**
ilaaj	[-120.40903 -85.55526]	hospital	[123.35004 119.61952]
khitaab	[-0.15578835 45.074234]	conference	[118.7522 -54.207916]
riyasat	[32.268997 28.609953]	hukoomat	[40.57867 164.86522]
tafteesh	[-4.4760947 13.971443]	police	[-15.499614 -73.69939]

Table 3: Misspellings

Word	**Embedding Vector**	**Misspelling**	**Embedding Vector**
shareef	[62.620174 0.8829548]	sharif	[153.77322 73.19686]
karen	[-60.286724 -95.09334]	keren	[-82.29357 28.164558]
unhein	[78.26031 75.88128]	unhen	[20.93452 -152.23032]
rozgari	[113.64553 44.28771]	rozgaar	[-71.11416 -17.093113]
ne	[-26.179857 -15.459935]	nay	[-137.08281 32.882595]
par	[-135.95943 69.16075]	per	[71.12073 120.3799]
tak	[41.668976 -106.20394]	tak	[50.57748 -41.83501]

ii. *context_size* = 10 **and** *embedding_dimension* = 200

Table 4: Synonyms

Word	**Embedding Vector**	**Synonym**	**Embedding Vector**
mauqa	[-20.300875 68.08492]	mawaqay	[95.557594 11.881878]
gaya	[-25.760405 31.146553]	gayi	[47.922714 -34.287678]
gaya	[-25.760405 31.146553]	gaye	[-88.59407 7.45824]
gayi	[47.922714 -34.287678]	gaye	[-88.59407 7.45824]
ilzamaat	[5.5550995 10.515322]	ilzaam	[-9.729152 48.697796]

Table 5: Analogies

Word	**Embedding Vector**	**Analogy**	**Embedding Vector**
ilaaj	[-23.083122 -49.57438]	hospital	[-16.20287 6.802215]
khitaab	[63.034203 -7.6697407]	conference	[23.584167 19.641176]
riyasat	[-13.1628 -25.791626]	hukoomat	[13.114212 -26.074453]
tafteesh	[-78.29451 -27.71728]	police	[45.376167 10.575589]

Table 6: Misspellings

Word	**Embedding Vector**	**Misspelling**	**Embedding Vector**
shareef	[43.86594 36.88414]	sharif	[25.46588 87.632195]
karen	[47.522118 75.09346]	keren	[-26.250793 -10.243973]
unhein	[75.261604 21.61777]	unhen	[8.512618 -68.270195]
rozgari	[-66.71932 68.05706]	rozgaar	[38.444836 -11.457395]
ne	[-52.245945 37.065567]	nay	[-89.320816 45.2088]
par	[-39.275383 56.77006]	per	[2.0172753 74.43931]
tak	[-37.319244 11.421612]	tak	[1.2284299 -44.874336]

iii. $\underline{context\_size = 15}$ **and** $\underline{embedding\_dimension = 300}$

Table 7: Synonyms

Word	**Embedding Vector**	**Synonym**	**Embedding Vector**
mauqa	[17.225733 -67.149124]	mawaqay	[-50.913193 -4.3995857]
gaya	[50.08408 82.03919]	gayi	[-3.3924904 -4.467893]
gaya	[50.08408 82.03919]	gaye	[109.81827 14.929681]
gayi	[-3.3924904 -4.467893]	gaye	[109.81827 14.929681]
ilzamaat	[18.571644 -12.561453]	ilzaam	[-4.8602138 -28.036772]

Table 8: Analogies

Word	**Embedding Vector**	**Analogy**	**Embedding Vector**
ilaaj	[-54.297276 62.82052]	hospital	[-39.080215 96.573875]
khitaab	[-25.701286 -12.245479]	conference	[-9.744786 -59.419174]
riyasat	[6.481866 48.694958]	hukoomat	[-3.9154415 22.09118]
tafteesh	[42.266827 -26.547346]	police	[76.66628 -3.7746143]

Table 9: Misspellings

Word	**Embedding Vector**	**Misspelling**	**Embedding Vector**
shareef	[86.35 -67.80652]	sharif	[27.279932 32.07133]
karen	[-22.321964 45.310703]	keren	[33.4251 56.86645]
unhein	[-42.134888 31.809668]	unhen	[73.961716 22.961979]
rozgari	[110.23308 -28.32726]	rozgaar	[5.591937 108.079735]
ne	[-83.7667 -0.6159746]	nay	[19.529768 -101.332306]
par	[-63.935364 24.294868]	per	[-28.00743 10.901405]
tak	[47.886497 21.39272]	tak	[55.12442 -86.533966]

iv. $\underline{context\_size = 20}$ **and** $\underline{embedding\_dimension = 300}$

Table 10: Synonyms

Word	**Embedding Vector**	**Synonym**	**Embedding Vector**
mauqa	[98.48787 -54.281155]	mawaqay	[67.391495 62.35655]
gaya	[-49.16667 111.76317]	gayi	[-1.1245097 117.9099]
gaya	[-49.16667 111.76317]	gaye	[-82.86818 -80.45088]
gayi	[-1.1245097 117.9099]	gaye	[-82.86818 -80.45088]
ilzamaat	[123.82037 -14.104588]	ilzaam	[-9.375072 46.343353]

Table 11: Analogies

Word	**Embedding Vector**	**Analogy**	**Embedding Vector**
ilaaj	[-0.86358505 -120.63874]	hospital	[38.280777 -98.26652]
khitaab	[48.66667 -7.755692]	conference	[-35.38935 30.37804]
riyasat	[72.65573 108.0003]	hukoomat	[14.115551 38.360466]
tafteesh	[25.842173 -63.78161]	police	[24.431042 104.01102]

Table 12: Misspellings

Word	**Embedding Vector**	**Misspelling**	**Embedding Vector**
shareef	[102.56679 43.573265]	sharif	[25.714264 69.540016]
karen	[-93.443985 -31.35119]	keren	[67.0984 21.011854]
unhein	[-9.555634 81.860855]	unhen	[-18.67653 10.61822]
rozgari	[3.8241947 -42.107246]	rozgaar	[-40.73494 -2.5979357]
ne	[93.82841 14.390238]	nay	[54.42057 -56.610504]
par	[75.49697 -86.37513]	per	[2.3979635 -92.409454]
tak	[21.623108 -12.251095]	tak	[8.414068 11.07134]