Programming Assignment 2

Subrajit Makur (CS17M046) and Amar Vashishth (CS17M052)

March 6, 2018

Data

We have trained out models on a Data of 600 MB. It has been collected from many sources, mentioning few of them below, for complete list of sources please look into the links folder of this project.

- https://www.ebanglalibrary.com/shebaprokashoni/
- https://www.ebanglalibrary.com/purnendupatri/
- http://www.atnbangla.tv/
- http://www.rtnn.net/
- http://www.the-editor.net/
- http://www.al-ihsan.net/
- http://www.bdnews24.com/bangla
- http://www.dailysangram.com/
- http://www.newsnetbd.com/
- http://www.bartamanpatrika.com/archive/
- http://www.dhakapost.com/
- http://www.dw-world.de/dw/article
- http://www.bdnewseveryday.com/
- http://www.dw-world.de/
- http://bangla.irib.ir
- http://www.thedailysangbad.com
- http://www.jugantor.com/
- http://www.dainikdestiny.com/
- http://www.dainikazadi.com/
- http://www.borerkagoj.com/
- http://www.dailydeshbangla.com/

are some বাংলা sources.

Similar Words Prediction

Word	Similar Word
সিংহ	বাঘ
সাপ	ইঁদুর
বলা	আলাপ
কলম	বই
ভ্রমণ	পর্যটক
রেলপথ	কাজ

Word2vec

Input Word	Semantically Closest Words	Cosine Similarity
চাচা চাচী ছেলে	মেয়ে	0.581942
ভাই বোন দাদা	দিদি	0.689487
রাজকুমার রাজকুমারী ভাই	বোন	0.542658
চীন ভারত বাংলাদেশ	শ্রীলংকা	0.581942
মা বাবা শিক্ষক	শিক্ষিকা	0.560474

CBOW

Exp. #	Size	Window Size	Sample Size	Accuracy
1	100	8	16 Millions	40.73%
2	200	12	20 Millions	52.22%
3	250	15	30 Millions	55.62%
4	400	8	42 Millions	58.33%

Skip-gram

Exp. #	Size	Window Size	Sample Size	Accuracy
1	150	12	16 Millions	22.35%
2	250	16	20 Millions	19.54%
3	350	20	30 Millions	25.6%
4	400	8	42 Millions	30.28%

GloVe

Experi. No.	Window Size	Vector Size	Accuracy(in %)
1	15	50	62.09
2	20	100	65.49
3	25	150	68.13
4	30	200	73.06

Words Embeddings Visualizations

All data available in ”wordembd.tsv” file.

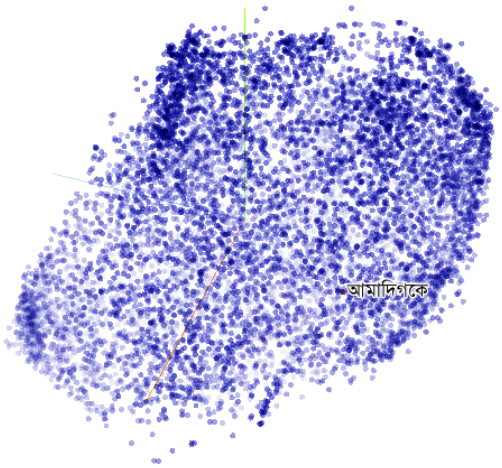

Figure 1: Complete Word Embedding Visualization using tSNE

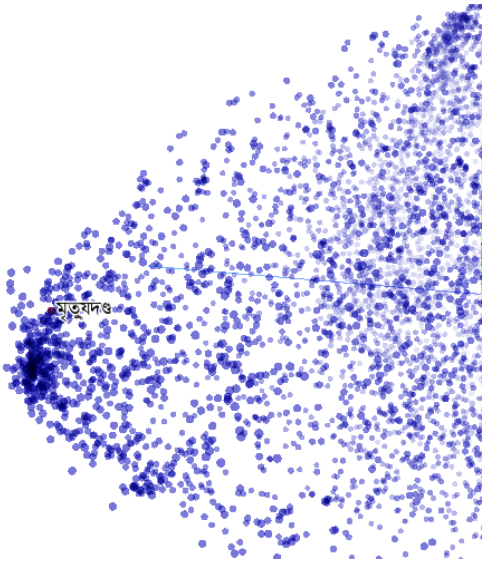

Figure 2: A cluster within embedding, representing a group of words sharing similar word sense(tSNE)

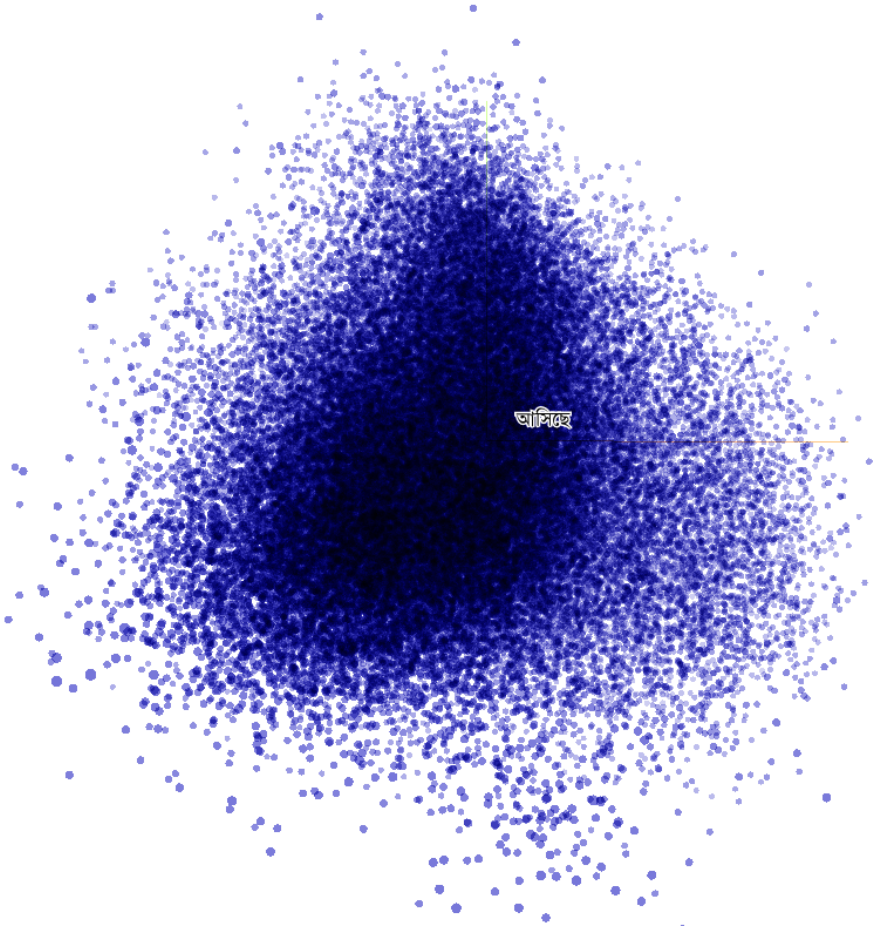

Figure 3: Complete Word Embeddings Visualization using PCA