Final NLP Course Project Report

Telegram Channels Analysis

Mohammad yarmoghadam

Supervise by

Dr Sauleh Etemadi

June 2021

Department of Computer Engineering

Iran University of Science and Technology

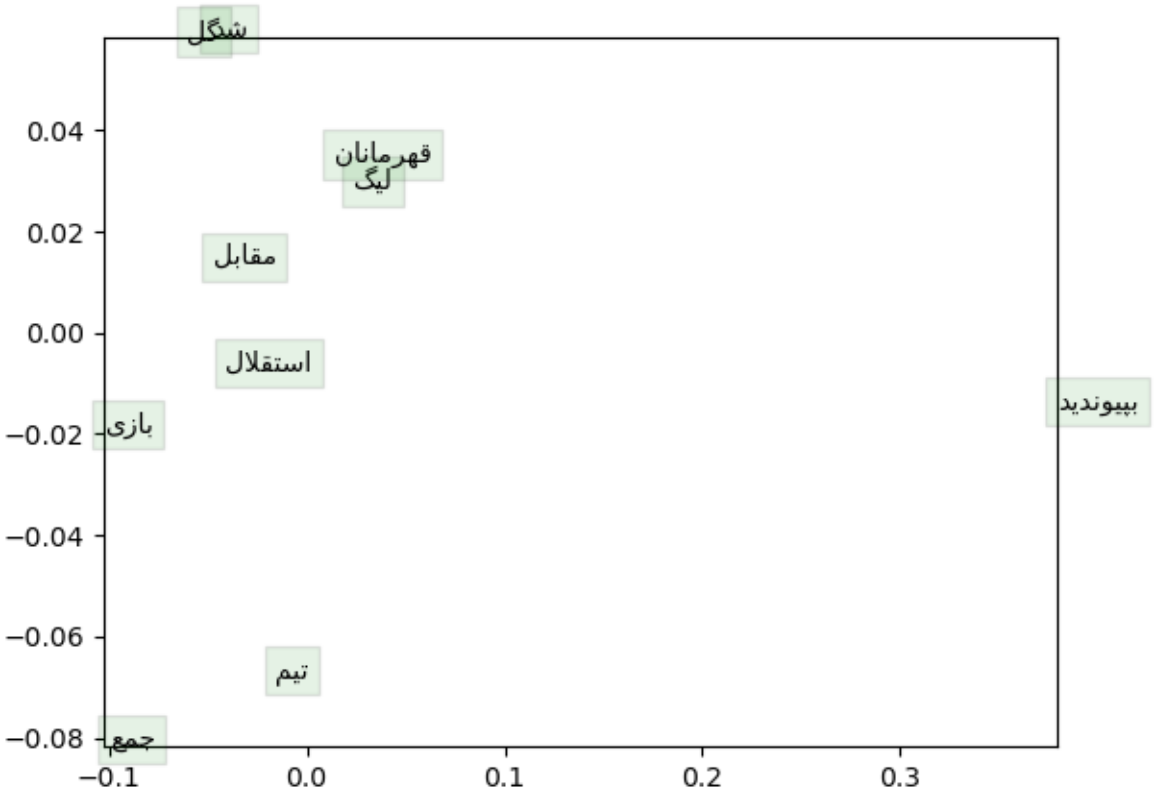
Contents

Word2Vec Feature

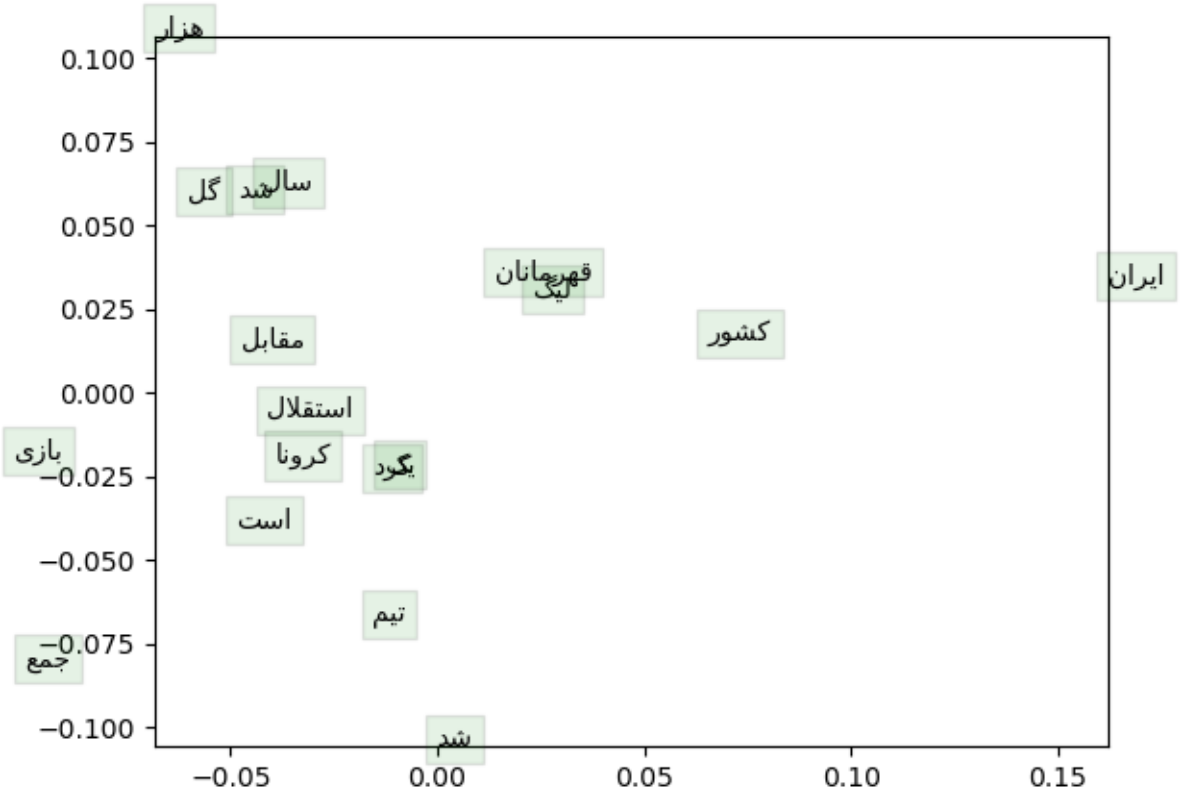
In this part I used my base code of assignment2 from cs224n course and add some changes to it.

The model trained for 40L iteration of dataset. The result of this model is saved on “models/word2vec” directory. From each label I set that Most 30 repeated word to be chosen.

By finish this part and training word2vec model I reach to these shapes that you can see below.



1.1 training Word2vec for sport label



1.2 training word2vec with news label and show result of both labels in same shape

Tokenization Feature

For this part I used from assignment 4 of cs224n which is about SentencePiece that is a text tokenizer.

In this part I divide each corpus to 5 segment to test and train model on each model as the doc wanted.

This help us to have a better evaluation on each text.

The size of vocab that is defined as follow is between 1500 till 6200.

Number of unknown token <unk> is represented on tokenization file which is located in reports Directory.

The percentage of <unk> token also existed in this file for each input file.

As a final result I understand that the number of <unk> token has increased as the size of vocab grown.

Number of <unk> token ↑ ➔ loss of data ↓

Parsing

Trained the parser model, to identify the dependency parsing of each sentence.

The result of model is saved in parsing directory which is located inside of models directory.

The result of this model is located on model.wights file.

The UAS parameter is 85 at the end which is good.

Language model

For this part used a LSTM model with 200 hidden layer to generate text for each label.

I suggest some words in the begging of the text and model is predict the rest to it according to the Training.

In below I write some of examples that model is generated for each label:

Sport:

چهار میلیون تومان تثبیت شد بانک

گروه اسرائیل با موفقیت داشته است

کمک میدهند یکی جهات از ی

سالیوان qarne khabarfoori com **ملاقات آلتا**

چند parsinehnews **ژانر دادگاهی نمیبره سخنرانی**

News:

چهار میلیون تومان تثبیت شد بانک

گروه اسرائیل با موفقیت داشته است

کمک میدهند یکی جهات از ی

سالیوان khabarfoori com **ملاقات آلنا** qarne

چند parsinehnews **ژانر دادگاهی نمیبره سخنرانی**

As we have some English word on dataset and can't remove them from data, generated data have some English word which is not good but can in improved in next versions.

Results can be accessed in language_model directory which is in reports folder.

Fine Tuning

As we saw in last part, language model do not have good result at all.

So it's better to used some better models like BERT and so on.

I write the code of this part but for timing problems and nearing to deadline of project, I can't run it Completely and saving the result.

But you can run it and see the result like other parts.

Note:

For log files that is requested by doc I add all of the log file that is needed but I got an error throught it and I should to comment all of this files.

You can see them in all files.

The error is : " ModuleNotFoundError: No module named 'src' "