Poetry Generator

URDU LANGUAGE

(NATURAL LANGUAGE PROCESSING)

ABDULLAH NAVEED

National University of Computer & Emerging Sciences, Islamabad

GITHUB

WEBSITE

PROBLEM STATEMENT

The goal of this assignment was to implement n-gram language modeling to generate poetry in Urdu language using spaCy library.

PROBLEM ANALYSIS

N-GRAM Modeling is the process of calculating probabilistic language model from a corpus for predicting the next item in a given sequence of words.

In this assignment, the required n-gram models are:

- Unigram Model.
- Forward Bigram Model.
- Backward Bigram Model.
- Bi Directional Bigram Model.
- Trigram Model.

SOLUTION DESIGN

My technique to solve this assignment in a step-wise manner is as following:

- First objective was to load the corpus and tokenize it into words. For this purpose, **spaCy** library was used which has built in function to tokenize Urdu language words.

- Second step was the actual Language Modeling.
    - Unigram model was calculated using the tokenized words
    - Bigram models and Trigram model were calculated using corpus and tokenized words.

- The next objective was to generate poetry from the computed n-gram models. For which the following steps were taken:
    - Inside a loop for each stanza, 4 verses were initialized
    - A starting word for each verse was picked from the starting words list that was generated from corpus.
    - Then for each starting word, next word of the sequence was calculated by taking the **argmax** of candidate words that were provided by the language models.
    - Finally for each poem generated: 3 stanzas with 4 verses with a space between each stanza was printed.

EVALUATION

FORWARD BIGRAM MODEL:

The forward bigram model was not up to the mark as it was just looking at one word to generate the probability of next word also in case of Urdu it was working in the opposite direction.

BACKWARD BIGRAM MODEL:

The backward bigram model was giving far better results than the forward bigram model as it was working in the same direction in which Urdu language is written.

BIDIRECTIONAL BIGRAM MODEL:

The bi-directional bigram model was taking both forward and backward bigram models into account for choosing the candidate words in sequence and then picking the word with highest probability from both models. It's results were tilted more towards the backward model as it was going in natural direction of Urdu language.

TRIGRAM MODEL:

The trigram model was giving the perfect results in comparison to bigram models because it was considering two words to generate the probability of next word hence providing some sense to overall sequence of the verse.

TIME COMPLEXITIES:

Let:

- "C" be the size of corpus
- "W" be the size of tokenized words list.
- "S" be the size of starting words list.

1. **Unigram Model:**

   O(W^2)

2. **Forward Bigram Model:**

   O(W*(C+W))

3. **Backward Bigram Model:**

   O(W*(C+W))

4. **Trigram Model:**

   O(W*(C+W))

SPACE COMPLEXITIES:

1. **Corpus:**

   O(C)

2. **Tokenized Words:**

   O(W)

RESULTS

With Different Inputs:

میرے دل کا نہ ہو گئی ہے کہ ’

کھینچنے والوں سےگلا نہ ہو گئی ہے کہ

ہے کہ ’ یوں ہو گئی ہے کہ ’

نہ ہو گئی ہے کہ ’ یوں ہو گئی ہے

میرے دل کا نہ ہو گئی ہے کہ ’ یوں ہو

’ ایک ہی کیوں نہ ہو گئی ہے کہ

اقبال کے لیے ہم نے کی ہے کہ ’ یوں ہو

سبزۂ خط سے گزر میں ہے کہ ’ یوں

کام بند غم سے گزر میں ہے کہ

جو تو کیا ہے کہ ’ یوں ہو گئی ہے کہ

کہ ’ یوں ہو گئی ہے کہ ’ یوں ہو

بیٹھنا اُس کو دیکھ لیا دانتوں میں ہے

FORWARD BIGRAM MODEL

گزر سے تجھ ، ہے کیا گے جائیں مر کہ

ہنوز ہے کیا گے جائیں مر کہ ہے کیا گے

کیا گے جائیں مر کہ ہے کیا گے

بخش کو مجھ ڈبویا ہوتا نہ کیوں ہی

ہیں دیکھتے کو مجھ ڈبویا ہوتا نہ کیوں ہی

بارہا نے ہم ہے کیا گے جائیں مر کہ ہے

آیا پسند مشکل سخت پسند مشکل سخت پسند

پایا مزا کا دل سے تجھ ، ہے کیا

یہ ہے کیا گے جائیں مر کہ ہے

ہدیہ جگر تا دل سے تجھ ، ہے کیا

آیا پسند مشکل سخت پسند مشکل سخت پسند مشکل

تھا مرد آزاد عجب کرے مغفرت حق ہے کیا

BACKWARD BIGRAM MODEL

(Both Providing Satisfactory Results with BIGRAM MODEL on different inputs)

With Same Inputs:

تا جگر کہ ’ تو کیا ہے کہ
ہو گئیں خاک نہیں آتی ہے کہ ’ تو کیا ہے
گدائے کوچۂ میخانہ نامراد نہیں آتی ہے کہ ’ تو کیا
شورِ پندِ ناصح نے کیا ہے کہ ’ تو کیا

کوئی کشمکش اندوہِ عشق کی ہے کہ ’ تو
فلک الافلاک اے خدا کی ہے کہ ’ تو
عشق کی ہے کہ ’ تو کیا ہے
عریاں نکلا دل کا نہ ہوا ہے کہ

قفس میں ’ تو کیا ہے کہ ’ تو کیا
کہو کہ ’ تو کیا ہے کہ ’
کاوش کا نہ ہوا ہے کہ ’ تو کیا ہے
صحن گلشن میں ’ تو کیا ہے کہ ’

FORWARD BIGRAM MODEL

تا بلب زنجیری دود سپند آیا فضائے خندۂ گل صدائے مرغ
ہو مومنوں کے پیتے تھے پیوست گلو بنا سکتے
گدائے کوچۂ میخانہ نامراد نہیں طولانی لطف خلش پیکاں آسودگیٔ فتراک
شورِ پندِ ناصح گر لکھنے بیٹھوں تو زنداني

کوئی ہدف تیرے زمان و معاني میں رہتا ہوں بجھا
فلک الافلاک اے فقرِ غیٔور کھا گئی روحِ فرنگی
عشق بتاں سے باندھا گیا کیونکر میسر میر سپاہ ناسزا
عریاں نکلا قیس تصویر بہتر ہے کھرا ہے کھرا

سمجھتا ہوں بجھا چاہتا ہوں بجھا چاہتا ہوں بجھا
قفس اداس بیٹھا کہتا ہوں بجھا چاہتا ہوں بجھا
کہو اچھا یوں بوسہ ، شاہیں کا پیغام ہے کھرا ہے
کاوش ہائے خانقہی فقیہ شہر قاروں ہے کھرا

BACKWARD BIGRAM MODEL

تا بلب زنجیری دود سپند آیا فضائے خندۂ گل صدائے مرغ
ہو مومنوں کے پیتے تھے پیوست گلو بنا ہے کھرا
گدائے کوچۂ میخانہ نامراد نہیں طولانی لطف خلش
شورِ پندِ ناصح نے روما کی بستی دکاں نہیں طولانی

کوئی ہدف تیرے زمان و معاني میں رہتا ہوں بجھا چاہتا
فلک الافلاک اے فقرِ غیٔور کھا گئی روحِ فرنگی کو
عشق بتاں سے باندھا گیا کیونکر میسر میر
عریاں نکلا قیس تصویر بہتر ہے کھرا ہے کھرا

سراپا رہن عشق بتاں سے باندھا گیا کیونکر میسر
قفس میں الجھ کر تلف کھول کے پیتے تھے
کہو کہ چھپ کے پیتے تھے پیوست گلو بنا ہے کھرا
کاوش کا پیغام ہے کھرا ہے کھرا ہے

BIDIRECTIONAL BIGRAM MODEL

(Here BIDIRECTIONAL Model is tilted more towards backward model because of its similar direction with Urdu Language)

With Same Inputs:

دل کا نہ ہوا ہے کہ ، تو کیا ہے کہ

آہ و دل کا نہ ہوا ہے کہ ، تو کیا

روانی باۓ موج ، تو کیا ہے کہ ، تو کیا

لیکن اب اس کی ہے کہ ، تو کیا ہے کہ

نہ ہوا ہے کہ ، تو کیا ہے کہ ، تو

دل کا نہ ہوا ہے کہ ، تو کیا ہے کہ

تم نے کیا ہے کہ ، تو کیا ہے کہ ،

پندار کا نہ ہوا ہے کہ ، تو کیا ہے کہ

کہ ، تو کیا ہے کہ ، تو کیا ہے کہ

کاش اس کی ہے کہ ، تو کیا ہے کہ ،

پیماں سے ، تو کیا ہے کہ ، تو کیا ہے

توڑا جو تو کیا ہے کہ ، تو کیا ہے کہ

FORWARD BIGRAM MODEL

دل جلوں میں الجھ گیا کیونکر میسر میر سپاہ ناسزا لشکریاں

آہ سحرگہی مجھ کو،کہ جہاں جادہ غیر از نمود کچھ کھٹکتا

روانی باۓ صحبت مخالف ہے کھرا ہے کھرا ہے کھرا ہے

لیکن عبث کہ چھپ کے پیتے تھے پیوست گلو بنا سکتے

نہ بیدار وہ سرود کیا دبدبۂ نادر کیا دبدبۂ نادر کیا

دل جلوں میں الجھ گیا کیونکر میسر میر سپاہ ناسزا لشکریاں

تم سبھی دوست ہونہیں سکتا غریب الدیار ہوں بجھا چاہتا ہوں

پندار کا پیغام ہے کھرا ہے کھرا ہے کھرا

کہ چھپ کے پیتے تھے پیوست گلو بنا سکتے غلاموں کی

کاش رِضواں ہی خودکشی کرے اخذِ فیضِ جاں ستاں ناوک خیز

پیماں سے باندھا گیا کیونکر میسر میر سپاہ ناسزا لشکریاں شکستہ

توڑا جوشِ قدح پہ معشوق فریبی عنواں اٹھایئے گر لکھنے بیٹھوں

BACKWARD BIGRAM MODEL

دل جلوں میں الجھ کر تلف کھول کے پیتے تھے پیوست

آہ سحرگہی مجھ کو،کہ جہاں جادہ غیر از نمود کچھ کھٹکتا

روانی باۓ موج لرزاں ہے کھرا ہے کھرا ہے کھرا ہے

لیکن عبث کہ چھپ کے پیتے تھے پیوست گلو بنا ہے

نہ بیدار وہ سرود کیا دبدبۂ نادر کیا دبدبۂ نادر کیا

دل جلوں میں الجھ کر تلف کھول کے پیتے تھے پیوست

تم سبھی کچھ کھٹکتا تھا شغف فقیری ملی انکو سرفرازی نازاں

پندار کا پیغام ہے کھرا ہے کھرا ہے کھرا

کہ چھپ کے پیتے تھے پیوست گلو بنا ہے کھرا ہے

کاش رِضواں ہی خودکشی کرے اخذِ فیضِ جاں ستاں ناوک خیز

پیماں سے باندھا گیا کیونکر میسر میر سپاہ ناسزا لشکریاں شکستہ

توڑا جو چرا کر تلف کھول کے پیتے تھے پیوست گلو

BIDIRECTIONAL BIGRAM MODEL

(Again, BIDIRECTIONAL Model is tilted more towards backward model because of its similar direction with Urdu Language)

With Different Inputs:

جسے تم سمجھ رہے ہو وہ اب زر کم عیار ہوگا

عروج آدم خاکی کے منتظر ہیں تمام یہ کہکشاں

مدعا تیری زندگی کا تو اور سوئے غیر نظر ہائے

مرے جرم خانہ خراب کو ترے عفو بندہ نواز میں

ٹھہر سکا نہ ہوائے چمن میں لالہ دکھاتا پھرتا

میں کہ شاہیں بناتا نہیں آشیانہ زندگی کچھ اور

بیٹھے گا محشر میں جنوں میرا یا اپنا گریباں چاک

یا اپنا گریباں چاک یا دامن یزداں چاک یہ پیران کلیسا

بدگمانی بلکہ میری سخت جانی سے نگاہ بے حجاب ناز

ستارے یہ نیلگوں افلاک یہی زمانۂ حاضر کی کائنات ہے کیا

چاک یا دامن یزداں چاک یہ پیران کلیسا

کرتے ہیں میری آواز گر نہیں آتی کوئی

TRIGRAM MODEL

ابھی عشق کے انداز چھٹ جائیں گے کیا ؟

فقط آواز ہے طاؤس فقط رنگ یہ دیر کہن کیا ہے

یہ کہا کہ وہ اثر کہن نہ تری حکایت سوز میں

فراغت گاہ آغوش وداع دل پسند آیا عدم

کبھی اے حقیقت منتظر نظر آ لباس مجاز میں کہ شاہیں

کیا کہ چھپا ہوا ہو سکوت پردۂ ساز میں تو

کہہ دی بڑا بے ادب ہوں سزا چاہتا ہوں مری

گو فقر بھی رکھتا ہے انداز ملوکانہ ناپختہ ہے

ناتوانی سے حریفِ دمِ افعی نہ ہوا میں

قید ہے خرم کیا چھینے گا غنچے سے کوئی

لے گا قافلہ مور ناتواں کا ہزار موجوں کی ہو کشاکش

کیوں ہے ؟ ذوق حسن و زیبائی سے

TRIGRAM MODEL

(TRIGRAM model providing more accurate results in comparison to other models)

SOME CHALLENGES FACED

- Corpus having garbage symbols and values which effected the model.
- Implementation of all the models was challenging yet fun to see it work at the end.

CONCLUSION AND FUTURE WORK

URDU Poetry Generation with n-gram modeling was first step of making the machine learn from text and produce an output. It was a challenging yet fun to implement assignment

---

ABDULLAH NAVEED