Text Generation using Markov Chains

Rahul Sailesh Wadhwa
Saloni Kashiv
Steven Strickland

Text Prediction

Given an input sequence, try to predict the next word, based on the probability matrix created using the input corpus. For example, if the input sequence is-:

- “I am sick”, the auto-completed word will be one of - {“at”, ”of”}
- “I have seen”, the auto-completed word will be one of- {'NOTHING', 'IT', 'MAY', 'SEE', 'PLAY', ‘YOU', 'HOURES', 'HIM’, 'HER', 'TEMPESTS', 'TH’, 'MORE'}

(These are examples taken from our code)

Wide range of applications-

- Predicting text in apps like messages
- Autocomplete suggester where a bunch of words are predicted.

Why Markov chains?

Markov assumption-

The Markov assumption states that given a certain number of previous states, the predicted state is independent of all the other states that come before it.

$$\textbf{Markov Assumption:} \quad P(q_i = a | q_1 \ldots q_{i-1}) = P(q_i = a | q_{i-1})$$

These are some advantages of employing Markov Chains for text generation compared to other method:

- Simple and easy to implement
- Lower computation time

However the disadvantages of using Markov Chains to build text generator:

- The generated text is as good as the input corpus (garbage in garbage out)

Input Corpus

Gutenberg Corpus-

NLTK includes a small selection of texts from the Project Gutenberg electronic text archive, which contains some 25,000 free electronic books.

From this we have chosen 3 classic books by Shakespeare- Hamlet, Macbeth, and Julius Caesar.

How it works?

Use of dictionaries-

We made use of nested dictionaries in order to store the training data for the model and create the transition probability matrix.

Why dictionaries?

- Fast access time for different key values.
- Easy to create transition probability matrix for given n-gram model.

Order 1

```
['I']
['WILL']
['TELL']
['PALEHEARTED']
['FEARE']
['AND']
['HEALTH']
['GIUE']
['THEE']
['A']
['STRANGE']
['AND']
['FLIGHTS']
['OF']
['BEAUTIE']
['INTO']
['THE']
['DESART']
```

I WILL TELL PALEHEARTED FEARE AND HEALTH GIUE THEE A STRANGE AND FLIGHTS OF BEAUTIE INTO THE DESART WITH MY SELFE A GREENE A
ND LET ME MY SELFE I THINKE I TAKE MY CREDIT NOW HAMLET HAM THE RECOUERY OF FOOT AND HAUE LONGED LONG IF THEIR GRAND COMMISS
ION TO YOU LAY IT SO FROM FELICITIE AWHILE BUT WE OUR SCALE WEIGHING THE HONIE OF MARCH IS SOUTHERLY I DO NOT COME YOU KNOW
HIM PEACE LEADE ON BROOD AND VOLUMNIUS THOU DARST THOU SHALT THOU GUILDENSTERNE AND MAKES THEM DO MAKE MALICIOUS MOCKERY BAR
N WELL OF NORWAY GIUING MORE SWEET

Order 3

```
['I', 'AM', 'SURE']
['AM', 'SURE', 'OF']
['SURE', 'OF', 'THAT']
['OF', 'THAT', 'HEAUEN']
['THAT', 'HEAUEN', 'KNOWES']
['HEAUEN', 'KNOWES', 'WHAT']
['KNOWES', 'WHAT', 'SHE']
['WHAT', 'SHE', 'HAS']
['SHE', 'HAS', 'KNOWNE']
['HAS', 'KNOWNE', 'LA']
['KNOWNE', 'LA', 'HEERES']
['LA', 'HEERES', 'THE']
['HEERES', 'THE', 'SMELL']
['THE', 'SMELL', 'OF']
['SMELL', 'OF', 'THE']
['OF', 'THE', 'BLOOD']
['THE', 'BLOOD', 'STILL']
['BLOOD', 'STILL', 'ALL']
['STILL', 'ALL', 'THE']
```

I AM SURE OF THAT HEAUEN KNOWES WHAT SHE HAS KNOWNE LA HEERES THE SMELL OF THE BLOOD STILL ALL THE PERFUMES OF ARABIA WILL N
OT SWEETEN THIS LITTLE HAND OH OH OH DOCT WHAT A SIGH IS THERE THE HART IS SORELY CHARGD GENT I WOULD NOT HAUE YOUR ENEMY SA
Y SO NOR SHALL YOU DOE MINE EARE THAT VIOLENCE TO MAKE IT TRUSTER OF YOUR OWNE REPORT AGAINST YOUR SELFE I KNOW YOU ARE NO T
RUANT BUT WHAT IS YOUR AFFAIRE IN ELSENOUR WEEL TEACH YOU TO DRINKE DEEPE ERE YOU DEPART HOR MY LORD THE KING YOUR FATHER HA
M THE

Order 5

```
['I', 'AM', 'SURE', 'OF', 'THAT']
['AM', 'SURE', 'OF', 'THAT', 'HEAUEN']
['SURE', 'OF', 'THAT', 'HEAUEN', 'KNOWES']
['OF', 'THAT', 'HEAUEN', 'KNOWES', 'WHAT']
['THAT', 'HEAUEN', 'KNOWES', 'WHAT', 'SHE']
['HEAUEN', 'KNOWES', 'WHAT', 'SHE', 'HAS']
['KNOWES', 'WHAT', 'SHE', 'HAS', 'KNOWNE']
['WHAT', 'SHE', 'HAS', 'KNOWNE', 'LA']
['SHE', 'HAS', 'KNOWNE', 'LA', 'HEERES']
['HAS', 'KNOWNE', 'LA', 'HEERES', 'THE']
['KNOWNE', 'LA', 'HEERES', 'THE', 'SMELL']
['LA', 'HEERES', 'THE', 'SMELL', 'OF']
['HEERES', 'THE', 'SMELL', 'OF', 'THE']
['THE', 'SMELL', 'OF', 'THE', 'BLOOD']
['SMELL', 'OF', 'THE', 'BLOOD', 'STILL']
['OF', 'THE', 'BLOOD', 'STILL', 'ALL']
['THE', 'BLOOD', 'STILL', 'ALL', 'THE']
['BLOOD', 'STILL', 'ALL', 'THE', 'PERFUMES']
```

I AM SURE OF THAT HEAUEN KNOWES WHAT SHE HAS KNOWNE LA HEERES THE SMELL OF THE BLOOD STILL ALL THE PERFUMES OF ARABIA WILL N
OT SWEETEN THIS LITTLE HAND OH OH OH DOCT WHAT A SIGH IS THERE THE HART IS SORELY CHARGD GENT I WOULD NOT HAUE SUCH A HEART
IN MY BOSOME FOR THE DIGNITY OF THE WHOLE BODY DOCT WELL WELL WELL GENT PRAY GOD IT BE SIR DOCT THIS DISEASE IS BEYOND MY PR
ACTISE YET I HAUE KNOWNE THOSE WHICH HAUE WALKT IN THEIR SLEEP WHO HAUE DYED HOLILY IN THEIR BEDS LAD WASH YOUR HANDS PUT ON
YOUR NIGHTGOWNE

Conclusions

- Best order for the Markov chain was observed to be 3- Because less than 3, the predicted words are random and do not always make sense and greater than 3 makes the predictions too similar to the input corpus.
- Predictions are very dependent on the input corpus.

Future developments

- Extend the predictor to work as the inputs are typed by the user.
- Make the model user specific, i.e. it should be able to predict text depending on the history of the users typed sentences.
- Extend the suggestions to emojis in addition to words since they are so frequently used in messages today.