And so sayeth the Lord...

Utkarsh Kumar	130050022
Sivaprasad S	130050085
Shantanu Thakoor	13D100003

Problem Statement

We aim to capture the grammar and writing style of the author of several sample text. Based on this we intend to create a random text which may not already exist in the corpus, but will make sense at least locally. Coherence within a sentence is difficult, within a paragraph is almost impossible (with this method).

Deliverables

Input

- A large enough text for us build an accurate model, preferably by the same author on the same topic
- Currently we are handling only reported text, as handling quotations etc. requires global knowledge
- Used Bible as source text

Output

- Sample text that resembles the writing styles of the author

Unigram Approach

- We have taken an HMM based approach
- For now each represents a word
- A transition indicates a word followed by another
- No coherence

- The transition from w_{i-1} and w_i is taken with probability

  $P(W_i|W_{i-1}) = c(W_i)/\sum c(W_j)$

Bigram Approach

- For now each node represents a word and a transition indicates a word followed by another
- These transitions will have probabilities which we have learned from our corpus

- $P(W_i|W_{i-1}) = c(W_{i-1},W_i)/c(W_{i-1})$

Trigram Approach

- Every node represents a bigram rather than a word, improves the locality of coherence
- However doing this increases the length, as the probability of a period decreases

- $P(W_i|W_{i-1}W_{i-2}) = c(W_{i-2},W_{i-1},W_i)/c(W_{i-2},W_{i-1})$

Problems with these approaches

- JOHN READ MOBY DICK

  MARY READ A DIFFERENT BOOK

  SHE READ A BOOK BY CHER
- $p(\text{JOHN READ A BOOK}) = p(\text{JOHN}|\bullet)\ p(\text{READ}|\text{JOHN})\ p(\text{A}|\text{READ})\ p(\text{BOOK}|\text{A})\ p(\bullet|\text{BOOK}) \approx 0.06$
- $p(\text{CHER READ A BOOK}) = p(\text{CHER}|\bullet)\ p(\text{READ}|\text{CHER})\ p(\text{A}|\text{READ})\ p(\text{BOOK}|\text{A})\ p(\bullet|\text{BOOK}) = 0$

Approach used

- Combination of all the three
- Better than earlier three

- $P(W_i|W_{i-1}W_{i-2}) = \lambda_1 * c(W_{i-2},W_{i-1},W_i)/c(W_{i-2},W_{i-1}) + \lambda_2 * c(W_{i-1,} W_i)/c(W_{i-1}) + \lambda_3 * c(W_i)/\sum c(W_j)$, where $\lambda_1+\lambda_2+\lambda_3=1$.

Code Overview

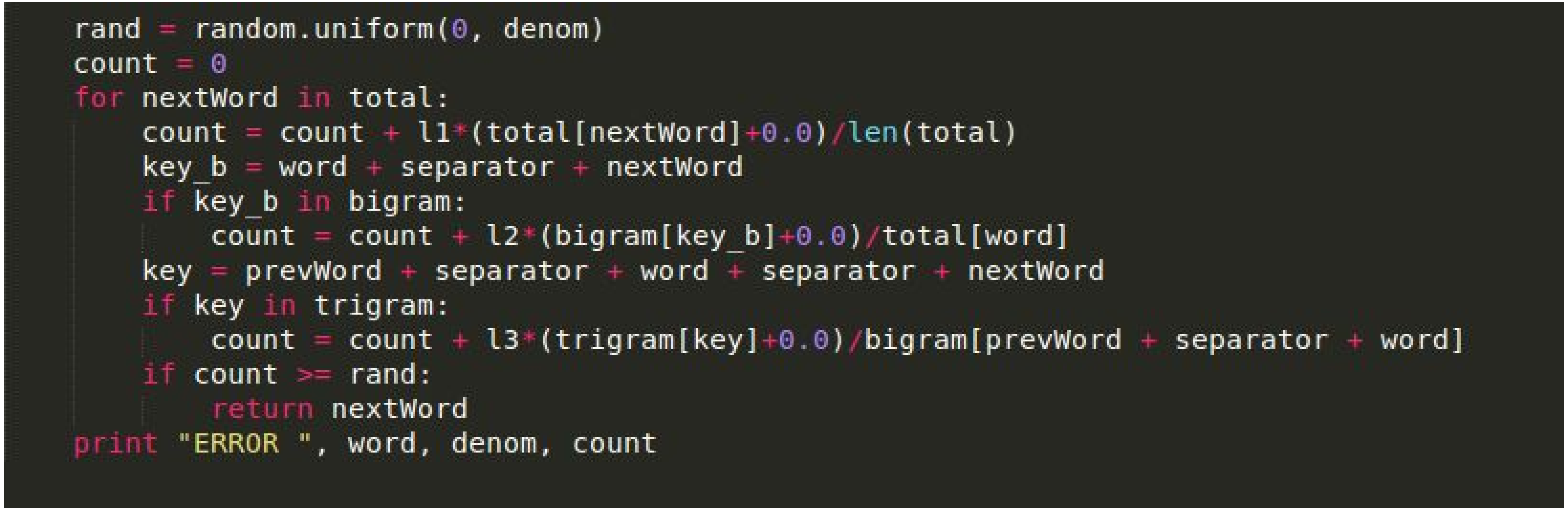
generateModel.py

- Read from the corpus and populate a hash table with the number of occurrences of each bigram and word; handle punctuations smartly
- Write the generated model in json format on disk to be read by executeModel.py

executeModel.py

- Load the Markov model from disk and unpack it
- Choose the next word based on the probability calculated as earlier
- Generate the document sentence by sentence until a given word limit is reached

Code Overview



```
    rand = random.uniform(0, denom)
    count = 0
    for nextWord in total:
        count = count + l1*(total[nextWord]+0.0)/len(total)
        key_b = word + separator + nextWord
        if key_b in bigram:
            count = count + l2*(bigram[key_b]+0.0)/total[word]
        key = prevWord + separator + word + separator + nextWord
        if key in trigram:
            count = count + l3*(trigram[key]+0.0)/bigram[prevWord + separator + word]
        if count >= rand:
            return nextWord
    print "ERROR ", word, denom, count
```

Try guessing…

From one came to declare unto him the officers in thy days and the land of Moab, and the temple of Simon, I bring thee the seas, and the strength in the office.

Of the silver, I spare them that were the chariots, as much abroad into the king.

*Casing has been edited

They hatch cockatrice' eggs, and weave the spider's web: he that eateth of their eggs dieth, and that which is crushed breaketh out into a viper.

For my people is foolish, they have not known me; they are sottish children, and they have none understanding: they are wise to do evil, but to do good they have no knowledge

*Casing has been edited

Thus they changed their glory into the similitude of an ox that eateth grass. They forgat God their saviour, which had done great things in Egypt;

And as we have in heaven. I desired mercy, that the Eunuch saw him eat at thy right hand of the lowest of the wicked, and for all Israel, that we which live are always delivered unto the people, which he alloweth.

*Casing has been edited

She had a leprous men of Gershon, therefore hearken unto her people out against thy dwelling without the soul, had wholly followed hard by a time.

And to the angel of the church in Philadelphia write; These things saith he that is holy, he that is true, he that hath the key of David, he that openeth, and no man shutteth, and shutteth, and no man openeth.

*Casing has been edited

And thou wentest to the king with ointment, and didst increase thy perfumes, and didst send thy messengers far off, and didst debase thyself even unto hell.

We have mourned to you and your mighty sins: they must walk to ten cubits. Arise thou therefore, and ye dwell. I will cause the blessing from the avenger; that we may know that my soul is escaped, which is by Ibleam.

*Casing has been edited

And the woman said unto the serpent, We may eat of the fruit of the trees of the garden:

Yea, though thou wash my feet from falling, that thou do unto this day. Therefore by the space of one language; Judah shall abide on his breast and his own house.

*Casing has been edited

From one came to declare unto him the officers in thy days and the land of Moab, and the temple of Simon, I bring thee the seas, and the strength in the office.

Of the silver, I spare them that were the chariots, as much abroad into the king.

*Casing has been edited

They hatch cockatrice' eggs, and weave the spider's web: he that eateth of their eggs dieth, and that which is crushed breaketh out into a viper.

For my people is foolish, they have not known me; they are sottish children, and they have none understanding: they are wise to do evil, but to do good they have no knowledge

*Casing has been edited

Thus they changed their glory into the similitude of an ox that eateth grass. They forgat God their saviour, which had done great things in Egypt;

And as we have in heaven. I desired mercy, that the Eunuch saw him eat at thy right hand of the lowest of the wicked, and for all Israel, that we which live are always delivered unto the people, which he alloweth.

*Casing has been edited

She had a leprous men of Gershon, therefore hearken unto her people out against thy dwelling without the soul, had wholly followed hard by a time.

And to the angel of the church in Philadelphia write; These things saith he that is holy, he that is true, he that hath the key of David, he that openeth, and no man shutteth, and shutteth, and no man openeth.

*Casing has been edited

And thou wentest to the king with ointment, and didst increase thy perfumes, and didst send thy messengers far off, and didst debase thyself even unto hell.

We have mourned to you and your mighty sins: they must walk to ten cubits. Arise thou therefore, and ye dwell. I will cause the blessing from the avenger; that we may know that my soul is escaped, which is by Ibleam.

*Casing has been edited

And the woman said unto the serpent, We may eat of the fruit of the trees of the garden:

Yea, though thou wash my feet from falling, that thou do unto this day. Therefore by the space of one language; Judah shall abide on his breast and his own house.

*Casing has been edited

Statistics and Testing

- We asked 12 people the above 12 passages and asked them to identify whether they were from the Bible or generated by us
- Out of a total 144 responses, 88 were incorrect
- This gives us an estimate of 61.1% as our accuracy
- Here, accuracy was defined simply as #incorrect answers / #total answers
- A more useful estimation might be: a histogram depicting how much one person is likely to go wrong (removes outliers who are very familiar/unfamiliar with the Bible verses)

Individual Contributions

- Equal :)

Learning from the project

- Training an HMM
- Identifying features out of writing style
- Cleaning data

Future Work

- Handling punctuations and spoken speech
- Learning λ_is
- Changing parameters based on number of words already made in sentence

References

- http://nlp.stanford.edu/~wcmac/papers/20050421-smoothing-tutorial.pdf
- https://pdos.csail.mit.edu/archive/scigen/
- http://www.decontextualize.com/teaching/rwet/n-grams-and-markov-chains/

Demo