N-Gram Modelling

Introduction to Markov Chains, bigrams and trigrams

TF-IDF Model – Model

Words/ Documents	going	to	today	i	am	it	is	rain
Document 1	0	0.07	0.07	0	0	0.17	0.17	0.17
Document 2	0	0	0.07	0.07	0.07	0	0	0
Document 3	0	0.05	0	0.05	0.05	0	0	0

Markov Chains – States

- A is a state.
- B is a state.

Markov Chains – States

- A is a state.
- B is a state.

- Probability of A→B : 50%
- Probability of A→A : 50%
- Probability of B→A : 50%
- Probability of B→B : 50%

Markov Chains – States

- A is a state.
- B is a state.

A

- Probability of A→B : 50%
- Probability of A→A : 50%
- Probability of B→A : 50%
- Probability of B→B : 50%

Markov Chains – States

- A is a state.
- B is a state.

AA

- Probability of A→B : 50%
- Probability of A→A : 50%
- Probability of B→A : 50%
- Probability of B→B : 50%

Markov Chains – States

- A is a state.
- B is a state.

AAB

- Probability of A→B : 50%
- Probability of A→A : 50%
- Probability of B→A : 50%
- Probability of B→B : 50%

Markov Chains – States

- A is a state.
- B is a state.

AABA

- Probability of A→B : 50%
- Probability of A→A : 50%
- Probability of B→A : 50%
- Probability of B→B : 50%

N-Gram Modelling – What are N-Grams?

“An N-gram is a contiguous sequence of n items from a given sample of text or speech” – Wikipedia

Items refer to states in Markov Chains

Items can be “characters”, ”words”, “sentences” etc.

N-Gram Modelling – What are N-Grams?

N = 2, then bigrams

N = 3, then trigrams

and so on

N-Gram Modelling – Character Grams

Characters are the states of Markov Chains

N-Gram Modelling – Character Grams

“the bird is flying on the blue sky”

N-Gram Modelling – Character Grams

“the bird is flying on the blue sky”

N = 2

Bigrams = ‘th’, ‘he’, ‘e ’, ‘ b’, ‘bi’, ‘ir’, ‘rd’, ‘d ’, ‘ i’ etc.

N = 3

Trigrams = ‘the’, ’he ’, ’e b’, ‘ bi’, ‘bir’, ‘ird’, ‘rd ’, ‘d i’ etc

N-Gram Modelling – Character Grams

“the bird is flying on the blue sky”

Consider only trigrams,

N-Gram Modelling – Character Grams

“the bird is flying on the blue sky”

Consider only trigrams,

Window size = N = 3

N-Gram Modelling – Character Grams

Trigrams

“the bird is flying on the blue sky”

N-Gram Modelling – Character Grams

Trigrams

the bird is flying on the blue sky

© Bijoyan Das

N-Gram Modelling – Character Grams

Trigrams

the

the bird is flying on the blue sky

© Bijoyan Das

N-Gram Modelling – Character Grams

Trigrams

the

he

the bird is flying on the blue sky

© Bijoyan Das

N-Gram Modelling – Character Grams

Trigrams
the
he
e b

the bird is flying on the blue sky

N-Gram Modelling – Character Grams

Trigrams

the

he

e b

bi

the bird is flying on the blue sky

© Bijoyan Das

N-Gram Modelling – Character Grams

the bird is flying on the blue sky

Trigrams

the

he

e b

bi

bir

N-Gram Modelling – Character Grams

the bird is flying on the blue sky

Trigrams

the

he

e b

bi

bir

ird

N-Gram Modelling – Character Grams

the bird is flying on the blue sky

Trigrams	Next
the	
he	b
e b	i
bi	r
bir	d
ird	

N-Gram Modelling – Word Grams

Words are the states of Markov Chains

N-Gram Modelling – Word Grams

“the bird is flying on the blue sky”

N-Gram Modelling – Character Grams

“the bird is flying on the blue sky”

Consider only trigrams,

Window size = N = 3 (Words)

N-Gram Modelling – Word Grams

Trigrams

“the bird is flying on the blue sky”

© Bijoyan Das

N-Gram Modelling – Word Grams

Trigrams
the bird is

the bird is flying on the blue sky

© Bijoyan Das

N-Gram Modelling – Word Grams

Trigrams
the bird is
bird is flying

the bird is flying on the blue sky

N-Gram Modelling – Word Grams

Trigrams
the bird is
bird is flying
is flying on

the bird is flying on the blue sky

© Bijoyan Das

N-Gram Modelling – Word Grams

Trigrams
the bird is
bird is flying
is flying on
flying on the

the bird is flying on the blue sky

© Bijoyan Das

N-Gram Modelling – Word Grams

the bird is flying on the blue sky

Trigrams

the bird is

bird is flying

is flying on

flying on the

on the blue

© Bijoyan Das

N-Gram Modelling – Word Grams

the bird is flying on the blue sky

Trigrams

the bird is

bird is flying

is flying on

flying on the

on the blue

the blue sky

© Bijoyan Das

N-Gram Modelling – Word Grams

the bird is flying on the blue sky

Trigrams	Next
the bird is	flying
bird is flying	on
is flying on	the
flying on the	blue
on the blue	sky
the blue sky	

N-Gram Modelling – Usage

Trigrams	Next
the bird is	flying
bird is flying	on
is flying on	the
flying on the	blue
on the blue	sky
the blue sky	

N-Gram Modelling – Usage

Trigrams	Next
the bird is	[flying, eating, sleeping]
bird is flying	[on, through, on]
is flying on	the
flying on the	[blue, orange]
on the blue	sky
the blue sky	[.]

N-Gram Modelling – Usage

"the bird is"

N-Gram Modelling – Usage

Trigrams	Next
the bird is	[flying, eating, sleeping]
bird is flying	[on, through, on]
is flying on	the
flying on the	[blue, orange]
on the blue	sky
the blue sky	[.]

N-Gram Modelling – Usage

“the bird is flying”

N-Gram Modelling – Usage

“the bird is flying”

© Bijoyan Das

N-Gram Modelling – Usage

Trigrams	Next
the bird is	[flying, eating, sleeping]
bird is flying	[on, through, on]
is flying on	the
flying on the	[blue, orange]
on the blue	sky
the blue sky	[.]

N-Gram Modelling – Usage

“the bird is flying on”

N-Gram Modelling – Usage

“the bird is flying on”

N-Gram Modelling – Usage

Trigrams	Next
the bird is	[flying, eating, sleeping]
bird is flying	[on, through, on]
is flying on	the
flying on the	[blue, orange]
on the blue	sky
the blue sky	[.]

N-Gram Modelling – Usage

“the bird is flying on the”

N-Gram Modelling – Usage

“the bird is flying on the”

N-Gram Modelling – Usage

Trigrams	Next
the bird is	[flying, eating, sleeping]
bird is flying	[on, through, on]
is flying on	the
flying on the	[blue, orange]
on the blue	sky
the blue sky	[.]

N-Gram Modelling – Usage

“the bird is flying on the blue”

N-Gram Modelling – Usage

"the bird is flying on the blue"

© Bijoyan Das

N-Gram Modelling – Usage

Trigrams	Next
the bird is	[flying, eating, sleeping]
bird is flying	[on, through, on]
is flying on	the
flying on the	[blue, orange]
on the blue	sky
the blue sky	[.]

N-Gram Modelling – Usage

“the bird is flying on the blue sky”

N-Gram Modelling – Usage

“the bird is flying on the blue sky”

© Bijoyan Das

N-Gram Modelling – Usage

Trigrams	Next
the bird is	[flying, eating, sleeping]
bird is flying	[on, through, on]
is flying on	the
flying on the	[blue, orange]
on the blue	sky
the blue sky	[.]

N-Gram Modelling – Usage

“the bird is flying on the blue sky.”

N-Gram Modelling – Usage

“the bird is flying on the blue sky.”

© Bijoyan Das

N-Gram Modelling – Usage

Trigrams	Next
the bird is	[flying, eating, sleeping]
bird is flying	[on, through, on]
is flying on	the
flying on the	[blue, orange]
on the blue	sky
the blue sky	[.]

N-Gram Modelling – Usage

“the bird is flying on the blue sky.”

N-Gram Modelling – Usage

“the bird is sleeping”

“the bird is flying on the orange”

N-Gram Modelling – Additional Reading

Text Mining, Analytics & More – What are N-grams?

http://text-analytics101.rxnlp.com/2014/11/what-are-n-grams.html