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

- OA is a state.
- OB is a state.

- OA is a state.
- OB is a state.

- \bigcirc Probability of A \rightarrow B: 50%
- OProbability of A→A:50%
- OProbability of B→A: 50%
- OProbability of B→B: 50%

- OA is a state.
- OB is a state.

- \bigcirc Probability of A \rightarrow B: 50%
- OProbability of A→A:50%
- OProbability of B→A: 50%
- OProbability of B→B: 50%

A

- OA is a state.
- OB is a state.

OProbability of A→B: 50%

OProbability of A→A:50%

OProbability of B→A: 50%

OProbability of B→B: 50%

AA

- OA is a state.
- OB is a state.

AAB

- Probability of A→B: 50%
- \bigcirc Probability of $A \rightarrow A : 50\%$
- OProbability of B→A: 50%
- OProbability of B→B: 50%

- OA is a state.
- OB is a state.

AABA

- Probability of A→B: 50%
- \bigcirc Probability of $A \rightarrow A : 50\%$
- OProbability of B→A: 50%
- OProbability 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

Characters are the states of Markov Chains

"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

"the bird is flying on the blue sky" Consider only trigrams,

"the bird is flying on the blue sky"

Consider only trigrams,

Window size = N = 3

Trigrams

Trigrams

Trigrams the

Trigrams the he

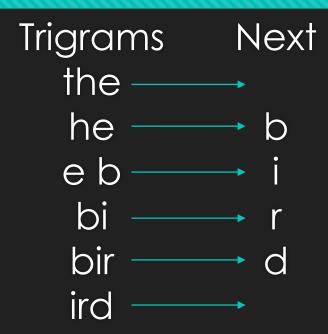
```
Trigrams
the
he
e b
```

```
Trigrams
the
he
e b
bi
```

```
Trigrams
the
he
e b
bi
bir
```

the b<mark>ird</mark> is flying on the blue sky

```
Trigrams
the
he
e b
bi
bir
ird
```



Words are the states of Markov Chains

"the bird is flying on the blue sky"

Consider only trigrams,

Window size = N = 3 (Words)

Trigrams

Trigrams the bird is

Trigrams
the bird is
bird is flying

Trigrams
the bird is
bird is flying
is flying on

the bird is <mark>flying on the</mark> blue sky

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

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 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



N-Gram Modelling – Usage

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

"the bird is"

"the bird is flying"

"the bird is flying"

"the bird is flying on"

"the bird is flying on"

"the bird is flying on the"

"the bird is flying on the"

"the bird is flying on the blue"

"the bird is flying on the blue"

"the bird is flying on the blue sky"

"the bird is flying on the blue sky"

"the bird is flying on the blue sky."

"the bird is flying on the blue sky."

"the bird is flying on the blue sky."

"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