CIS 635 - Knowledge Discovery & Data Mining

Sequence data and modeling introduction

Sequence data

- NLP
  - Machine Translation (MT)
  - Question Answering
  - **Document Classification**
  - Sentiment Classification
  - Document summarization

- DNA Sequencing
  - DNA sequencing
  - **DNA classification**

Sequence data

- **Data/Feature encoding**
  - One-Hot Encoding
  - Label Encoding

- **NLP/DNA sequencing**
  - Tf-idf
  - **CountVectorizer**

CountVectorizer – general idea

A	black	cat	
1	1	1	d_1

“A black cat”

CountVectorizer – general idea

A	black	cat	white	
1	1	1	0	d_1
1	0	1	1	d_2

“A black cat”

“A white cat”

CountVectorizer – general idea

A	black	cat	white	is	as	beautiful	
1	1	1	0	0	0	0	d_1
1	0	1	1	0	0	0	d_2
2	1	1	1	2	2	1	d_3

“A black cat”

“A white cat”

“A black cat is as beautiful as a white cat”

CountVectorizer – general idea

A	black	cat	white	is	as	beautiful	
1	1	1	0	0	0	0	d_1
1	0	1	1	0	0	0	d_2
2	1	1	1	2	2	1	d_3

Corpus
“A black cat” *“A white cat”* *“A black cat is as beautiful as a white cat”*

CountVectorizer – general idea

A	black	cat	white	is	as	beautiful	
1	1	1	0	0	0	0	d_1
1	0	1	1	0	0	0	d_2
2	1	1	1	2	2	1	d_3

Corpus

“A black cat”

“A white cat”

“A black cat is as beautiful as a white cat”

dictionary: *{*
“a”, “is”, “as”,
“cat”, “black”,
“white”, “beautiful”
}

CountVectorizer – general idea

A	black	cat	white	is	as	beautiful	
1	1	1	0	0	0	0	d_1
1	0	1	1	0	0	0	d_2
2	1	1	1	2	2	1	d_3

Corpus

“A black cat”

“A white cat”

“A black cat is as beautiful as a white cat”

Unigram: *[“a”, “is”, “as”, “cat”, “black”, “white”, “beautiful”]*

CountVectorizer – general idea

A	black	cat	white	is	as	beautiful	
1	1	1	0	0	0	0	d_1
1	0	1	1	0	0	0	d_2
2	1	1	1	2	2	1	d_3

Corpus

“A black cat”

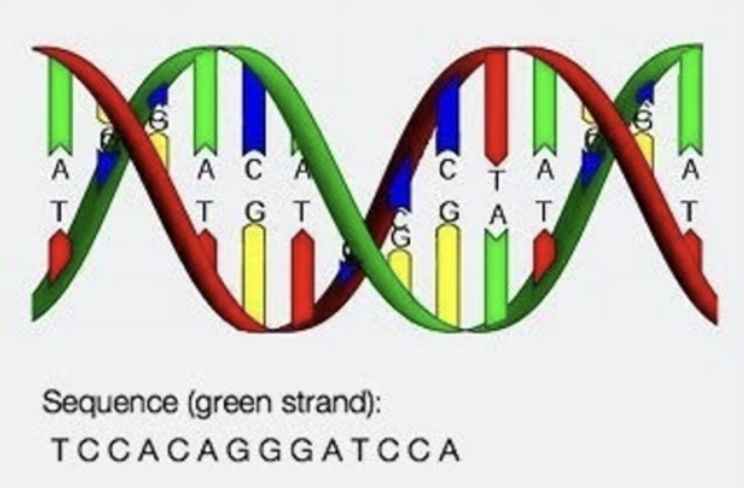
“A white cat”

“A black cat is as beautiful as a white cat”

Unigram: *[“a”, “is”, “as”, “cat”, “black”, “white”, “beautiful”]*

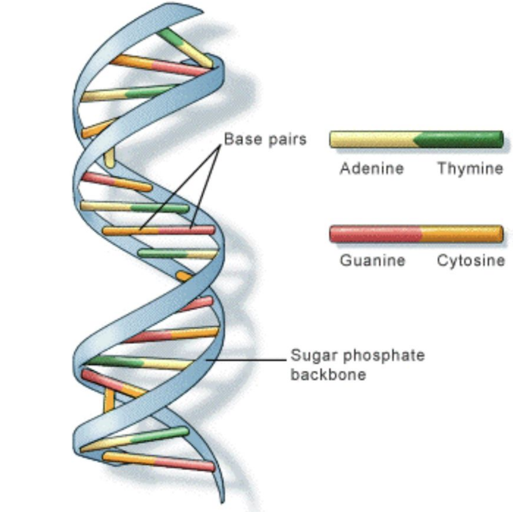
bigram: *[(“a”, “cat”), (“cat”, “a”), (“black”, “cat”), (“cat”, “black”), (“beautiful”, “cat”), (“cat”, “beautiful”]*

DNA Sequence - as a string



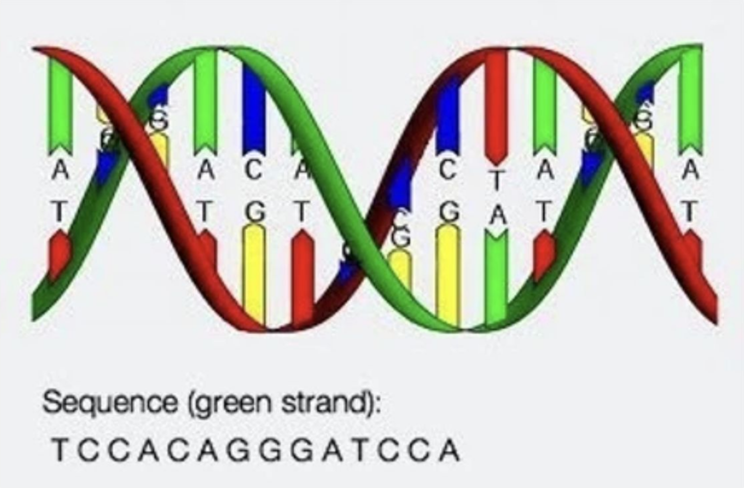
Sequence (green strand):

TCCACAGGGATCCA



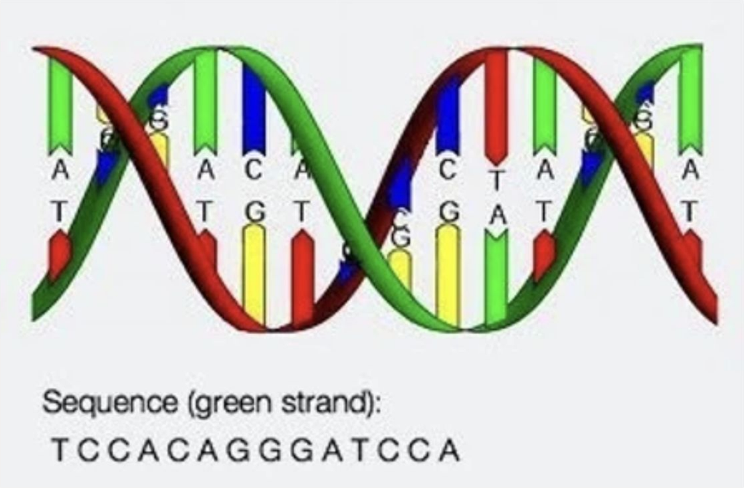
U.S. National Library of Medicine

DNA Sequence - as a string



```
ENST00000435737.5
ATGTTTCGCATCACCAACATTGAGTTTCTTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTCAGCAACAACAAAGGCGGCCTCCTTGTCCACTTTTGGATTGTTTTTGTCATGCCACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTTGAAGGACTCCATCCAGACAAGCATCATAAACCGGACCTCT
GTGGGGAGCTTGCAGGGACTGGCTGTGGACATGGACTCTGTGGTACTAAATGAAGTCCTGGGGCTGACTCTCATT
GTCTGGATTGACTGA
390
ENST00000419127.5
ATGTTTCGCATCACCAACATTGAGTTTCTTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTCAGCAACAACAAAGGCGGCCTCCTTGTCCACTTTTGGATTGTTTTTGTCATGCCACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTTGAAGGACTCCATCCAGACAAGCATCATAAACCGGACCTCT
GTGGGGAGCTTGCAGGGACTGGCTGTGGACATGGACTCTGTGGTACTAAATGACAAAGGCTGCTCTCAGTACTTC
TATGCAGAGCATCTGTCTCTCCACTACCCGCTGGAGATTTCTGCAGCCTCAGGGAGGCTGATGTGTCACTTCAAG
CTGGTGGCCATAGTGGGCTACCTGATTCGTCTCTCAATCAAGTCCATCCAAATCGAAGCCGACAACTGTGTCACT
GACTCCCTGACCATTTACGACTCCCTTTTGCCCATCCGGAGCAGCATCTTGTACAGAATTTGTGAACCCACAAGA
ACATTAATGTCATTTGTTTCTACAAATAATCTCATGTTGGTGACATTTAAGTCTCCTCATATACGGAGGCTCTCA
GGAATCCGGGCATATTTTGAGGTCATTCCAGAACAAAAGTGTGAAAACACAGTGTTGGTCAAAGACATCACTGGC
TTTGAAGGGAAAATTTCAAGCCCATATTACCCGAGCTACTATCCTCCAAAATGCAAGTGTACCTGGAAATTTCAG
ACTTCTCTATCAACTCTTGGCATAGCACTGAAATTCTATAACTATTCAATAACCAAGAAGAGTATGAAAGGCTGT
GAGCATGGATGGTGGGAAATTAATGAGCACATGTACTGTGGCTCCTACATGGATCATCAGACAATTTTTCGAGTG
```

DNA Sequence - as a string



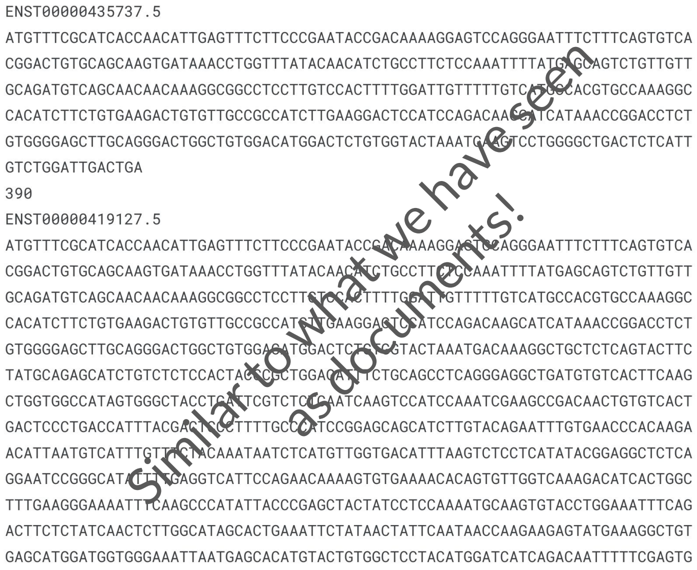
```
ENST00000435737.5
ATGTTTCGCATCACCAACATTGAGTTTCTTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTCAGCAACAACAAAGGCGGCCTCCTTGTCCACTTTTGGATTGTTTTTGTCATGCCACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTTGAAGGACTCCATCCAGACAAGCATCATAAACCGGACCTCT
GTGGGGAGCTTGCAGGGACTGGCTGTGGACATGGACTCTGTGGTACTAAATGAAGTCCTGGGGCTGACTCTCATT
GTCTGGATTGACTGA
390
ENST00000419127.5
ATGTTTCGCATCACCAACATTGAGTTTCTTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTCAGCAACAACAAAGGCGGCCTCCTTGTCCACTTTTGGATTGTTTTTGTCATGCCACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTTGAAGGACTCCATCCAGACAAGCATCATAAACCGGACCTCT
GTGGGGAGCTTGCAGGGACTGGCTGTGGACATGGACTCTGTGGTACTAAATGACAAAGGCTGCTCTCAGTACTTC
TATGCAGAGCATCTGTCTCTCCACTACCCGCTGGACATTTCTGCAGCCTCAGGGAGGCTGATGTGTCACTTCAAG
CTGGTGGCCATAGTGGGCTACCTGATTCGTCTCTCAATCAAGTCCATCCAAATCGAAGCCGACAACTGTGTCACT
GACTCCCTGACCATTTACGACTCCCTTTTGCCCATCCGGAGCAGCATCTTGTACAGAATTTGTGAACCCACAAGA
ACATTAATGTCATTTGTTTCTACAAATAATCTCATGTTGGTGACATTTAAGTCTCCTCATATACGGAGGCTCTCA
GGAATCCGGGCATATTTTGAGGTCATTCCAGAACAAAAGTGTGAAAACACAGTGTTGGTCAAAGACATCACTGGC
TTTGAAGGGAAAATTTCAAGCCCATATTACCCGAGCTACTATCCTCCAAAATGCAAGTGTACCTGGAAATTTCAG
ACTTCTCTATCAACTCTTGGCATAGCACTGAAATTCTATAACTATTCAATAACCAAGAAGAGTATGAAAGGCTGT
GAGCATGGATGGTGGGAAATTAATGAGCACATGTACTGTGGCTCCTACATGGATCATCAGACAATTTTTCGAGTG
```

Similar to what we have seen as documents!

DNA Sequence - as a string

- **Data/Feature encoding**
  - One-Hot Encoding
  - Label Encoding

- **NLP/DNA sequencing**
  - Tf-idf
  - **CountVectorizer**

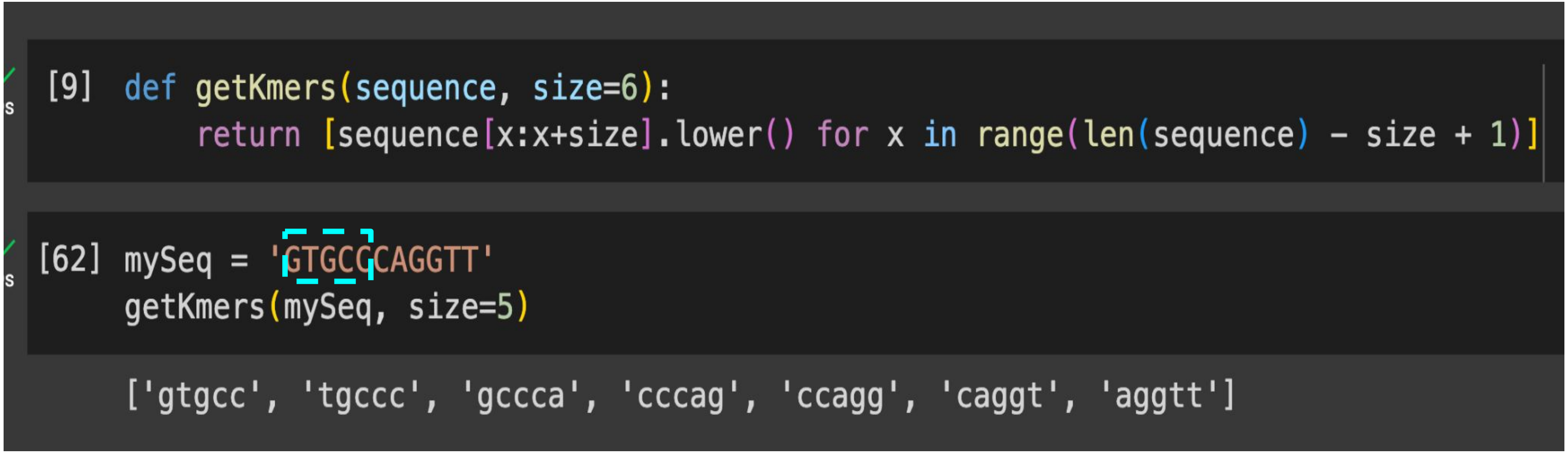


```
ENST00000435737.5
ATGTTTCGCATCACCAACATTGAGTTTCTTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTCAGCAACAACAAAGGCGGCCTCCTTGTCCACTTTTGGATTGTTTTTGTCATGCCACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTTGAAGGACTCCATCCAGACAAGCATCATAAACCGGACCTCT
GTGGGGAGCTTGCAGGGACTGGCTGTGGACATGGACTCTGTGGTACTAAATGAAGTCCTGGGGCTGACTCTCATT
GTCTGGATTGACTGA
390
ENST00000419127.5
ATGTTTCGCATCACCAACATTGAGTTTCTTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTCAGCAACAACAAAGGCGGCCTCCTTGTCCACTTTTGGATTGTTTTTGTCATGCCACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTTGAAGGACTCCATCCAGACAAGCATCATAAACCGGACCTCT
GTGGGGAGCTTGCAGGGACTGGCTGTGGACATGGACTCTGTGGTACTAAATGACAAAGGCTGCTCTCAGTACTTC
TATGCAGAGCATCTGTCTCTCCACTACCCGCTGGACATTTCTGCAGCCTCAGGGAGGCTGATGTGTCACTTCAAG
CTGGTGGCCATAGTGGGCTACCTGATTCGTCTCTCAATCAAGTCCATCCAAATCGAAGCCGACAACTGTGTCACT
GACTCCCTGACCATTTACGACTCCCTTTTGCCCATCCGGAGCAGCATCTTGTACAGAATTTGTGAACCCACAAGA
ACATTAATGTCATTTGTTTCTACAAATAATCTCATGTTGGTGACATTTAAGTCTCCTCATATACGGAGGCTCTCA
GGAATCCGGGCATATTTTGAGGTCATTCCAGAACAAAAGTGTGAAAACACAGTGTTGGTCAAAGACATCACTGGC
TTTGAAGGGAAAATTTCAAGCCCATATTACCCGAGCTACTATCCTCCAAAATGCAAGTGTACCTGGAAATTTCAG
ACTTCTCTATCAACTCTTGGCATAGCACTGAAATTCTATAACTATTCAATAACCAAGAAGAGTATGAAAGGCTGT
GAGCATGGATGGTGGGAAATTAATGAGCACATGTACTGTGGCTCCTACATGGATCATCAGACAATTTTTCGAGTG
```

Similar to what we have seen as documents!

k-mer counting!

DNA sequence as a “language”, known as k-mer counting



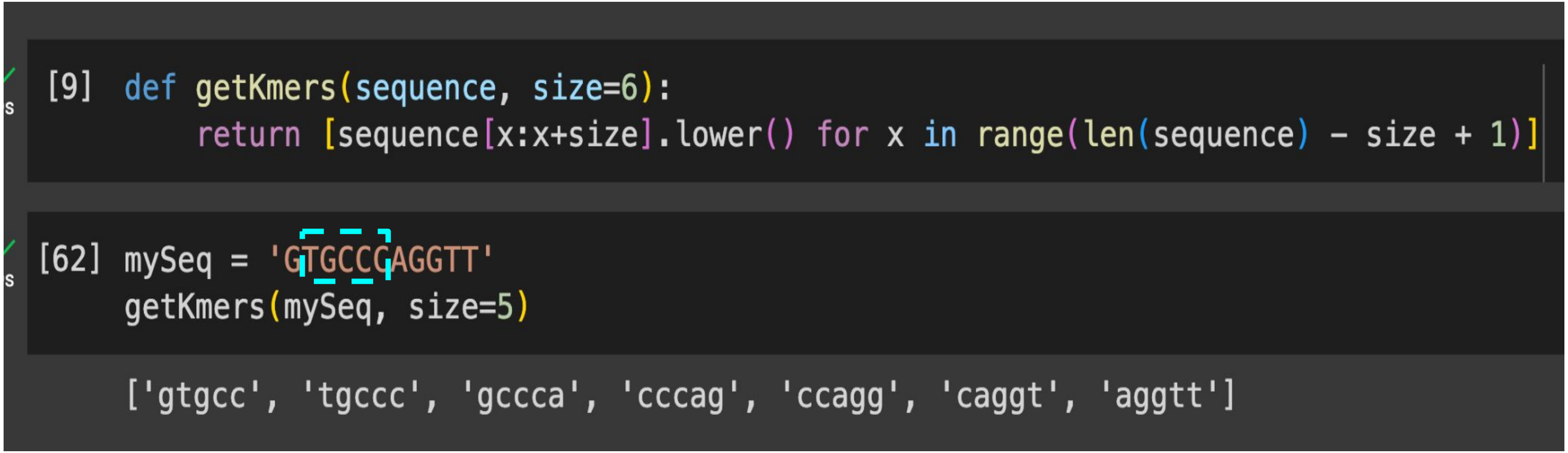
```
[9]  def getKmers(sequence, size=6):
         return [sequence[x:x+size].lower() for x in range(len(sequence) - size + 1)]
```

```
[62] mySeq = 'GTGCCCAGGTT'
     getKmers(mySeq, size=5)
```

```
['gtgcc', 'tgccc', 'gccca', 'cccag', 'ccagg', 'caggt', 'aggtt']
```

k-mer counting!

DNA sequence as a “language”, known as k-mer counting



```
[9]  def getKmers(sequence, size=6):
         return [sequence[x:x+size].lower() for x in range(len(sequence) - size + 1)]
```

```
[62] mySeq = 'GTGCCCAGGTT'
     getKmers(mySeq, size=5)
```

```
     ['gtgcc', 'tgccc', 'gccca', 'cccag', 'ccagg', 'caggt', 'aggtt']
```

Notebook presentation!