



CIS 678 Machine Learning

Document & DNA classification



Sequence data

- NLP
 - Machine Translation (MT)
 - Question Answering
 - Document Classification
 - Sentiment Classification
 - Document summarization
- DNA Sequencing
 - DNA sequencing
 - DNA classification



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

What are the challenges?



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
1	0	1	1

d 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
1	0	1	1	0	0	0
2	1	2	1	2	2	1

d_1
d_2
d_3

“A black cat”

“A white cat”

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

CountVectorizer – general idea

A	black	cat	white	is	as	beautiful
1	1	1	0	0	0	0
1	0	1	1	0	0	0
2	1	2	1	2	2	1

d_1
d_2
d_3

Corpus
<i>"A black cat"</i>
<i>"A white cat"</i>
<i>"A black cat is as beautiful as a white cat is"</i>

CountVectorizer – general idea

A	black	cat	white	is	as	beautiful
1	1	1	0	0	0	0
1	0	1	1	0	0	0
2	1	2	1	2	2	1

d ₁
d ₂
d ₃

Corpus
<i>"A black cat"</i>
<i>"A white cat"</i>
<i>"A black cat is as beautiful as a white cat is"</i>
<pre>dictionary: { "a", "is", "as", "cat", "black", "white", "beautiful" }</pre>

CountVectorizer – general idea

A	black	cat	white	is	as	beautiful
1	1	1	0	0	0	0
1	0	1	1	0	0	0
2	1	2	1	2	2	1

d ₁
d ₂
d ₃

Corpus
<i>"A black cat"</i>
<i>"A white cat"</i>
<i>"A black cat is as beautiful as a white cat is"</i>
Unigram: ["a", "is", "as", "cat", "black", "white", "beautiful"]

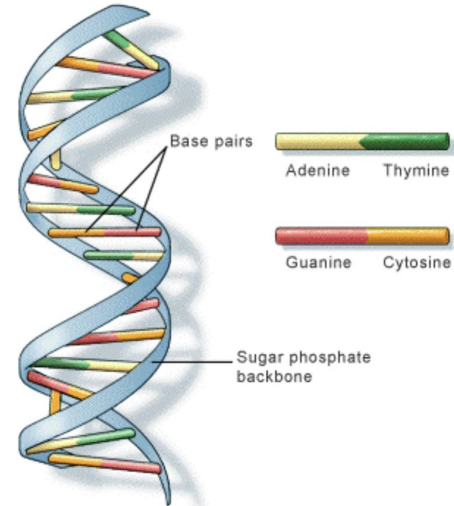
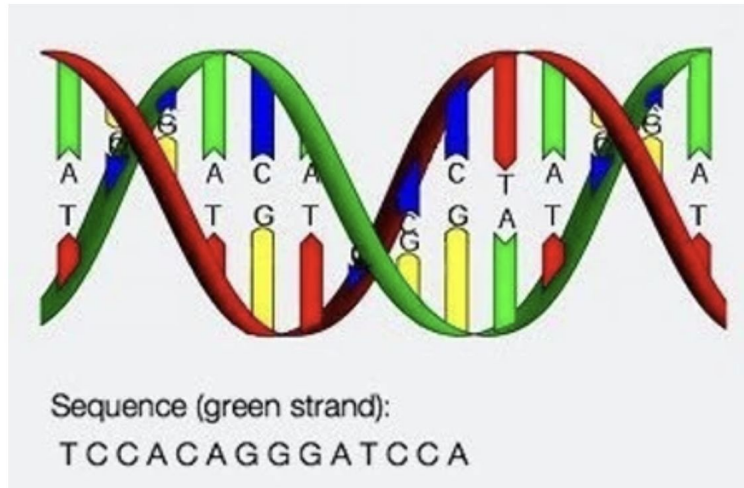
CountVectorizer – general idea

A	black	cat	white	is	as	beautiful
1	1	1	0	0	0	0
1	0	1	1	0	0	0
2	1	2	1	2	2	1

d ₁
d ₂
d ₃

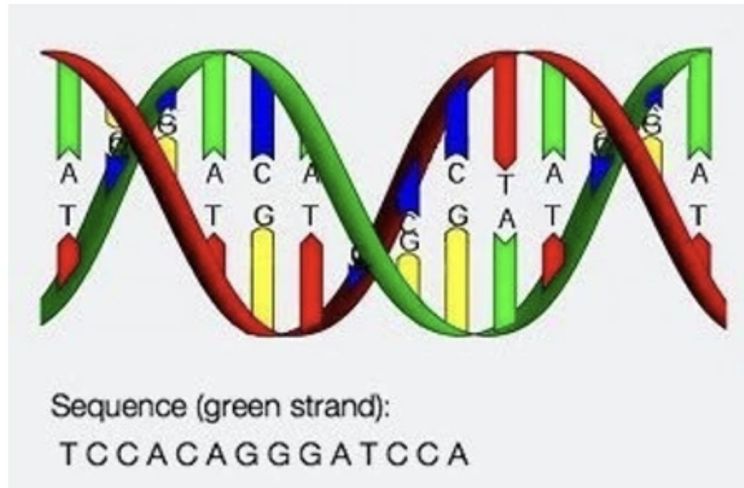
Corpus
<i>"A black cat"</i>
<i>"A white cat"</i>
<i>"A black cat is as beautiful as a white cat is"</i>
Unigram: ["a", "is", "as", "cat", "black", "white", "beautiful"] bigram: [("a", "cat"), ("cat", "a"), ("black", "cat"), ("cat", "black"), ("beautiful", "cat"), ("cat", "beautiful")]

DNA Sequence - as a string



U.S. National Library of Medicine

DNA Sequence - as a string



ENST00000435737.5

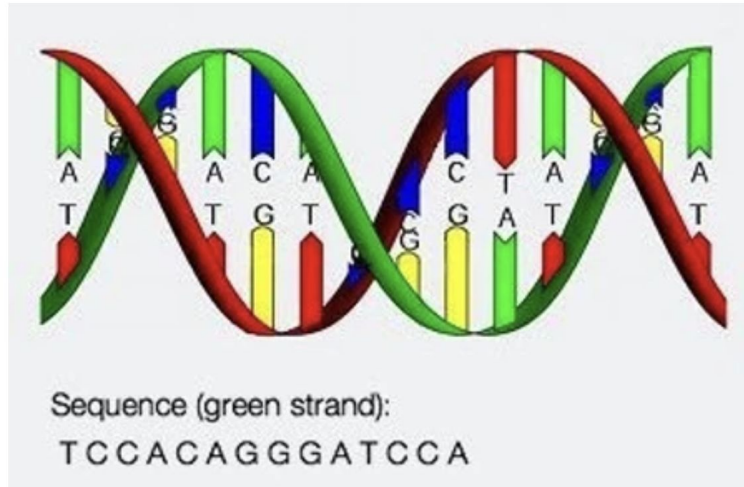
```
ATGTTTCGCATCACCAACATTGAGTTTCTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTCAGCAACAACAAAGGCGGCCTCCTTGCCACTTTTGGATTGTTTTGTCATGCCACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTTGAAGGACTCCATCCAGACAAGCATATAAACCGGACCTCT
GTGGGGAGCTTGACAGGACTGGCTGTGGACATGGACTCTGTTACTAAATGAAGTCCGTGGGGCTGACTCTCATT
GTCTGGATTGACTGA
```

390

ENST00000419127.5

```
ATGTTTCGCATCACCAACATTGAGTTTCTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTCAGCAACAACAAAGGCGGCCTCCTTGCCACTTTTGGATTGTTTTGTCATGCCACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTTGAAGGACTCCATCCAGACAAGCATATAAACCGGACCTCT
GTGGGGAGCTTGACAGGACTGGCTGTGGACATGGACTCTGTTACTAAATGACAAAGGCTGCTCTCAGTACTTC
TATGCAGAGCATCTGTCTCTCCACTACCGCTGGAGATTTCTGCAGCCTCAGGGAGGCTGATGTGTCACTTCAAG
CTGGTGGCCATAGTGGGCTACCTGATTCTCTCAATCAAGTCCATCCAAATCGAAGCCGACAACTGTGTCACT
GACTCCCTGACCATTTCAGACTCCCTTTTGGCCATCCGGAGCAGCATCTTGACAGAAATTTGTGAACCCACAAGA
ACATAATATGTCATTTGTTTCTACAAATAATCTCATGTTGGTGACATTTAAGTCTCCTCATATACGGAGGCTCTCA
GGAATCCGGGCATATTTTGAGGTCATTTCCAGAACAAAAGTGTGAAACACAGTGTGGTCAAAGACATCACTGGC
TTTGAAGGGAAAATTTCAAGCCCATATTACCGAGCTACTATCCTCCAAATGCAAGTGTACCTGGAAATTTTCAG
ACTTCTCTATCAACTCTTGGCATAGCACTGAAATTTCTAATACTATTCAATAACCAAGAGAGTATGAAAGGCTGT
GAGCATGGATGGTGGGAAATTAATGAGCACATGTACTGTGGCTCTACATGGATCATCAGACAATTTTCGAGTG
```

DNA Sequence - as a string



ENST00000435737.5

ATGTTTCGCATCACCAACATTGAGTTTCTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTCAGCAACAACAAAGGCGGCCTCCTTGTCCACTTTTGGATTGTTTTGTCTGACACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTGAAGGACTCCATCCAGACAAAGTCATATAACCGGACCTCT
GTGGGGAGCTTGACAGGACTGGCTGTGGACATGGACTCTGTGGTACTAAATGAGTCTTGGGGCTGACTCTCATT
GTCTGGATTGACTGA

398

ENST00000419127.5

ATGTTTCGCATCACCAACATTGAGTTTCTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTCAGCAACAACAAAGGCGGCCTCCTTGTCCACTTTTGGATTGTTTTGTCTGACACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTGAAGGACTCCATCCAGACAAAGTCATATAACCGGACCTCT
GTGGGGAGCTTGACAGGACTGGCTGTGGACATGGACTCTGTGGTACTAAATGACAAAGGCTGCTCTCAGTACTTC
TATGCAGAGCATCTGTCTCTCCACTAGCTGGTGGACATCTGTCAGCCTCAGGGAGGCTGATGTGCTCACTTCAAG
CTGGTGGCCATAGTGGGCTACCTCTGCTCTGCAATCAAGTCCATCCAAATCGAAGCCGACAACTGTGTCACT
GACTCCCTGACCATTTACGACTCTCTTTTGGCTCTCCGGAGCAGCATCTTGTACAGAAATTTGTGAACCCACAAGA
ACATAATGTCATTTGTTTACACAAATAATCTCATGTTGGTGACATTTAAGTCTCCTCATATACGGAGGCTCTCA
GGAATCCGGGCATATTTGAGGTCATTCCAGAACAAAAGTGTGAAAACACAGTGTGGTCAAAGACATCACTGGC
TTTGAAGGGAAAATTTCAAGCCCATATTACCGAGCTACTATCCTCCAAAATGCAAGTGTACCTGGAAAATTCAG
ACTTCTCTATCAACTCTTGGCATAGCACTGAAATTCATAACTATTCAATAACCAAGAGTATGAAAGGCTGT
GAGCATGGATGGTGGAAAATTAATGAGCACATGACTGTGGCTCTACATGGATCATCAGACAATTTTCGAGTG

DNA Sequence - as a string

- **Data/Feature encoding**
 - One-Hot Encoding
 - Label Encoding
- **NLP/DNA sequencing**
 - Tf-idf
 - **CountVectorizer**

```
ENST00000435737.5
ATGTTTCGCATCACCAACATTGAGTTTCTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTGAGCAACAACAAAGGCGGCCTCCTTGCCACTTTTGGATTGTTTTTGTCTGCGACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTTGAAGGACTCCATCCAGACAAAGTCATAAACCGGACCTCT
GTGGGGAGCTTGCAAGGACTGGCTGTGGACATGGACTCTGTGGTACTAAATGAGTCTCGGGGCTGACTCTCATT
GTCTGGATTGACTGA
390
ENST00000419127.5
ATGTTTCGCATCACCAACATTGAGTTTCTCCCGAATACCGACAAAAGGAGTCCAGGGAATTTCTTTCAGTGTCA
CGGACTGTGCAGCAAGTGATAAACCTGGTTTATACAACATCTGCCTTCTCCAAATTTTATGAGCAGTCTGTTGTT
GCAGATGTGAGCAACAACAAAGGCGGCCTCCTTGCCACTTTTGGATTGTTTTTGTCTGCGACGTGCCAAAGGC
CACATCTTCTGTGAAGACTGTGTTGCCGCCATCTTGAAGGACTCCATCCAGACAAAGTCATAAACCGGACCTCT
GTGGGGAGCTTGCAAGGACTGGCTGTGGACATGGACTCTGTGGTACTAAATGACAAAGGCTGCTCTCAGTACTTC
TATGCAGAGCATCTGTCTCCACTAGCGCTGGACATCTGCGAGCCTCAGGGAGGCTGATGTGCTCACTTCAAG
CTGGTGGCCATAGTGGGCTACCTGCTCTCGAATCAAGTCCATCCAAATCGAAGCCGACAACTGTGTCACT
GACTCCCTGACCATTACGACTCTCTTTTGGCTCTCCGGAGCAGCATCTTGACAGAAATTTGTGAACCCACAAGA
ACATAATGTCATTTGTTTACACAAATAATCTCATGTTGGTGACATTTAAGTCTCCTCATATACGGAGGCTCTCA
GGAATCCGGGCATATCTGAGGTCATTCCAGAACAAAAGTGTGAAAACACAGTGTGGTCAAGACATCACTGGC
TTTGAAGGGAAAATTTCAAGCCCATATTACCGAGCTACTATCCTCCAAATGCAAGTGTACCTGGAAAATTCAG
ACTTCTCTATCAACTCTTGGCATAGCACTGAAATTCATAACTATTCAATAACCAAGAGAGTATGAAAGGCTGT
GAGCATGGATGGTGGGAAATTAATGAGCACATGTACTGTGGCTCTACATGGATCATCAGACAATTTTCGAGTG
```




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 = 'GTGCCAGGTT'  
      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 = 'GTGCCGAGGTT'  
      getKmers(mySeq, size=5)
```

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



Notebook presentation!