text_processing

February 21, 2020

1 Text Processing

1.1 Capturing Text Data

1.1.1 Plain Text

In [1]:
```
import os

# Read in a plain text file
with open(os.path.join("data", "hieroglyph.txt"), "r") as f:
    text = f.read()
    print(text)
```

```
Hieroglyphic writing dates from c. 3000 BC, and is composed of hundreds of symbols. A hieroglyph
```

1.1.2 Tabular Data

In [2]:
```
import pandas as pd

# Extract text column from a dataframe
df = pd.read_csv(os.path.join("data", "news.csv"))
df.head()[['publisher', 'title']]

# Convert text column to lowercase
df['title'] = df['title'].str.lower()
df.head()[['publisher', 'title']]
```

Out[2]:

	publisher	title
0	Livemint	fed's charles plosser sees high bar for change...
1	IFA Magazine	us open: stocks fall after fed official hints ...
2	IFA Magazine	fed risks falling 'behind the curve', charles ...
3	Moneynews	fed's plosser: nasty weather has curbed job gr...
4	NASDAQ	plosser: fed may have to accelerate tapering pace

1.1.3 Online Resource

```
In [8]: import requests
        import json

        # Fetch data from a REST API
        r = requests.get(
            "https://quotes.rest/qod.json")
        res = r.json()
        print(json.dumps(res, indent=4))
```


```
{
    "success": {
        "total": 1
    },
    "contents": {
        "quotes": [
            {
                "quote": "Don't count the days; make the days count.",
                "length": "42",
                "author": "Mohamad Ali",
                "tags": {
                    "0": "inspire",
                    "1": "life",
                    "2": "productive",
                    "4": "tod"
                },
                "category": "inspire",
                "language": "en",
                "date": "2020-02-21",
                "permalink": "https://theysaidso.com/quote/mohamad-ali-dont-count-the-days-make-
                "id": "zdTLGFsZSu3_FtmJw7XbxQeF",
                "background": "https://theysaidso.com/img/qod/qod-inspire.jpg",
                "title": "Inspiring Quote of the day"
            }
        ]
    },
    "baseurl": "https://theysaidso.com",
    "copyright": {
        "year": 2022,
        "url": "https://theysaidso.com"
    }
}
```


```
In [7]: # Extract relevant object and field
        q = res["contents"]["quotes"][0]
        print(q["quote"], "\n--", q["author"])
```

```
Don't count the days; make the days count.
-- Mohamad Ali
```

1.2 Cleaning

```
In [9]: import requests

        # Fetch a web page
        r = requests.get("https://news.ycombinator.com")
        print(r.text)
```

```
<html op="news"><head><meta name="referrer" content="origin"><meta name="viewport" content="widt
        <link rel="shortcut icon" href="favicon.ico">
          <link rel="alternate" type="application/rss+xml" title="RSS" href="rss">
        <title>Hacker News</title></head><body><center><table id="hnmain" border="0" cellpadding
        <tr><td bgcolor="#ff6600"><table border="0" cellpadding="0" cellspacing="0" width="100%"
                  <td style="line-height:12pt; height:10px;"><span class="pagetop"><b class="hnn
              <a href="newest">new</a> | <a href="front">past</a> | <a href="newcomments">commen
                             <a href="login?goto=news">login</a>
                          </span></td>
              </tr></table></td></tr>
<tr id="pagespace" title="" style="height:10px"></tr><tr><td><table border="0" cellpadding="0" c
              <tr class='athing' id='22386960'>
      <td align="right" valign="top" class="title"><span class="rank">1.</span></td>      <td va
        <span class="score" id="score_22386960">80 points</span> by <a href="user?id=ProAm" clas
      <tr class="spacer" style="height:5px"></tr>
                <tr class='athing' id='22385491'>
      <td align="right" valign="top" class="title"><span class="rank">2.</span></td>      <td va
        <span class="score" id="score_22385491">331 points</span> by <a href="user?id=hhs" class
      <tr class="spacer" style="height:5px"></tr>
                <tr class='athing' id='22383746'>
      <td align="right" valign="top" class="title"><span class="rank">3.</span></td>      <td va
        <span class="score" id="score_22383746">280 points</span> by <a href="user?id=aty268" cl
      <tr class="spacer" style="height:5px"></tr>
                <tr class='athing' id='22382618'>
      <td align="right" valign="top" class="title"><span class="rank">4.</span></td>      <td va
        <span class="score" id="score_22382618">538 points</span> by <a href="user?id=chris_over
      <tr class="spacer" style="height:5px"></tr>
                <tr class='athing' id='22373906'>
      <td align="right" valign="top" class="title"><span class="rank">5.</span></td>      <td va
        <span class="score" id="score_22373906">59 points</span> by <a href="user?id=DyslexicAth
      <tr class="spacer" style="height:5px"></tr>
                <tr class='athing' id='22386440'>
      <td align="right" valign="top" class="title"><span class="rank">6.</span></td>      <td va
        <span class="score" id="score_22386440">166 points</span> by <a href="user?id=bjourne" c
      <tr class="spacer" style="height:5px"></tr>
                <tr class='athing' id='22374794'>
```

```
<td align="right" valign="top" class="title"><span class="rank">7.</span></td>      <td va
  <span class="score" id="score_22374794">15 points</span> by <a href="user?id=mmoez" clas
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22386096'>
<td align="right" valign="top" class="title"><span class="rank">8.</span></td>      <td va
  <span class="score" id="score_22386096">26 points</span> by <a href="user?id=evo_9" clas
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22386892'>
<td align="right" valign="top" class="title"><span class="rank">9.</span></td>      <td></
  <span class="age"><a href="item?id=22386892">27 minutes ago</a></span> | <a href="hide?i
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22381861'>
<td align="right" valign="top" class="title"><span class="rank">10.</span></td>      <td v
  <span class="score" id="score_22381861">556 points</span> by <a href="user?id=r_singh" c
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22386417'>
<td align="right" valign="top" class="title"><span class="rank">11.</span></td>      <td v
  <span class="score" id="score_22386417">16 points</span> by <a href="user?id=lawrenceyan
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22385408'>
<td align="right" valign="top" class="title"><span class="rank">12.</span></td>      <td v
  <span class="score" id="score_22385408">47 points</span> by <a href="user?id=rudnek" cla
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22386388'>
<td align="right" valign="top" class="title"><span class="rank">13.</span></td>      <td v
  <span class="score" id="score_22386388">17 points</span> by <a href="user?id=mot2ba" cla
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22385370'>
<td align="right" valign="top" class="title"><span class="rank">14.</span></td>      <td v
  <span class="score" id="score_22385370">79 points</span> by <a href="user?id=funkaster"
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22382337'>
<td align="right" valign="top" class="title"><span class="rank">15.</span></td>      <td v
  <span class="score" id="score_22382337">265 points</span> by <a href="user?id=ColinWrigh
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22372847'>
<td align="right" valign="top" class="title"><span class="rank">16.</span></td>      <td v
  <span class="score" id="score_22372847">99 points</span> by <a href="user?id=lelf" class
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22384680'>
<td align="right" valign="top" class="title"><span class="rank">17.</span></td>      <td v
  <span class="score" id="score_22384680">62 points</span> by <a href="user?id=malisper" c
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22383205'>
<td align="right" valign="top" class="title"><span class="rank">18.</span></td>      <td v
  <span class="score" id="score_22383205">48 points</span> by <a href="user?id=dsr_" class
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22384356'>
```

```
<td align="right" valign="top" class="title"><span class="rank">19.</span></td>      <td v
  <span class="score" id="score_22384356">333 points</span> by <a href="user?id=rauhl" cla
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22383841'>
<td align="right" valign="top" class="title"><span class="rank">20.</span></td>      <td v
  <span class="score" id="score_22383841">68 points</span> by <a href="user?id=johtela" cl
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22362123'>
<td align="right" valign="top" class="title"><span class="rank">21.</span></td>      <td v
  <span class="score" id="score_22362123">26 points</span> by <a href="user?id=apollinaire
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22381719'>
<td align="right" valign="top" class="title"><span class="rank">22.</span></td>      <td v
  <span class="score" id="score_22381719">17 points</span> by <a href="user?id=ascertain"
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22386320'>
<td align="right" valign="top" class="title"><span class="rank">23.</span></td>      <td v
  <span class="score" id="score_22386320">5 points</span> by <a href="user?id=blopeur" cla
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22382606'>
<td align="right" valign="top" class="title"><span class="rank">24.</span></td>      <td v
  <span class="score" id="score_22382606">170 points</span> by <a href="user?id=mikro2nd"
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22361780'>
<td align="right" valign="top" class="title"><span class="rank">25.</span></td>      <td v
  <span class="score" id="score_22361780">66 points</span> by <a href="user?id=luu" class=
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22382942'>
<td align="right" valign="top" class="title"><span class="rank">26.</span></td>      <td v
  <span class="score" id="score_22382942">219 points</span> by <a href="user?id=mmoez" cla
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22381919'>
<td align="right" valign="top" class="title"><span class="rank">27.</span></td>      <td v
  <span class="score" id="score_22381919">99 points</span> by <a href="user?id=mxschumache
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22382691'>
<td align="right" valign="top" class="title"><span class="rank">28.</span></td>      <td v
  <span class="score" id="score_22382691">70 points</span> by <a href="user?id=Tomte" clas
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22373516'>
<td align="right" valign="top" class="title"><span class="rank">29.</span></td>      <td v
  <span class="score" id="score_22373516">123 points</span> by <a href="user?id=jordybg" c
<tr class="spacer" style="height:5px"></tr>
          <tr class='athing' id='22366638'>
<td align="right" valign="top" class="title"><span class="rank">30.</span></td>      <td v
  <span class="score" id="score_22366638">139 points</span> by <a href="user?id=rodrigo975
<tr class="spacer" style="height:5px"></tr>
      <tr class="morespace" style="height:10px"></tr><tr><td colspan="2"></td><td class="t
```

```
  </table>
</td></tr>
<tr><td><img src="s.gif" height="10" width="0"><table width="100%" cellspacing="0" cellpadding="
        Applications are open for YC Summer 2020
      </a></center><br><center><span class="yclinks"><a href="newsguidelines.html">Guidelines</a
        | <a href="newsfaq.html">FAQ</a>
        | <a href="mailto:hn@ycombinator.com">Support</a>
        | <a href="https://github.com/HackerNews/API">API</a>
        | <a href="security.html">Security</a>
        | <a href="lists">Lists</a>
        | <a href="bookmarklet.html" rel="nofollow">Bookmarklet</a>
        | <a href="http://www.ycombinator.com/legal/">Legal</a>
        | <a href="http://www.ycombinator.com/apply/">Apply to YC</a>
        | <a href="mailto:hn@ycombinator.com">Contact</a></span><br><br><form method="get" actio
          <input type="text" name="q" value="" size="17" autocorrect="off" spellcheck="false" au
            </center></td></tr>
      </table></center></body><script type='text/javascript' src='hn.js?37WgrVym4z6Hej6XAKqR'></
```

```
In [10]: import re

         # Remove HTML tags using RegEx
         pattern = re.compile(r'<.*?>')  # tags look like <...>
         print(pattern.sub('', r.text))  # replace them with blank
```

```
        Hacker News

            Hacker News
          new | past | comments | ask | show | jobs | submit
                        login


      1.      Amazon Let a Fraudster Keep My Sony A7R IV and Refunded Him $2,900 (petapixel.com)
        80 points by ProAm 21 minutes ago  | hide | 11 comments


      2.      More bosses give four-day workweek a try (npr.org)
        331 points by hhs 3 hours ago  | hide | 182 comments


      3.      Google resists demands from states in digital-ad probe (wsj.com)
        280 points by aty268 6 hours ago  | hide | 130 comments
```

4. Radical hydrogen-boron reactor leapfrogs current nuclear fusion tech (newatlas.com
538 points by chris_overseas 9 hours ago | hide | 211 comments

5. An Open Guide to Equity Compensation (github.com)
59 points by DyslexicAtheist 2 hours ago | hide | 8 comments

6. Single-payer healthcare would save $450B and 68k lives a year: study (thelancet.co
166 points by bjourne 1 hour ago | hide | 132 comments

7. SQL query to generate the Mandelbrot Set as ASCII-art (sqlite.org)
15 points by mmoez 1 hour ago | hide | discuss

8. JP Morgan economists warn of 'catastrophic' climate change (bbc.com)
26 points by evo_9 2 hours ago | hide | 2 comments

9. Smarking (YC W15) is hiring Back end Tech Lead to scale urban mobility tech infra
27 minutes ago | hide

10. How to Write Usefully (paulgraham.com)
556 points by r_singh 12 hours ago | hide | 229 comments

11. Improved protein structure prediction using potentials from deep learning (nature
16 points by lawrenceyan 1 hour ago | hide | 7 comments

12. Show HN: MDAnki convert Markdown to Anki cards (github.com)
47 points by rudnek 3 hours ago | hide | 25 comments

13. Neat URL cleans URLs, removing parameters such as fbclid utm parameters (github.c
17 points by mot2ba 1 hour ago | hide | 6 comments

14. QEMU for iOS (github.com)
79 points by funkaster 3 hours ago | hide | 43 comments

15. Gilbert Strang Teaches Linear Algebra (mit.edu)
265 points by ColinWright 10 hours ago | hide | 59 comments

16. Does register selection matter to performance on x86 CPUs? (fiigii.com)
99 points by lelf 6 hours ago | hide | 30 comments

17. Launch HN: Freshpaint (YC S19) an automated, retroactive Segment alternative
62 points by malisper 4 hours ago | hide | 31 comments

18. Crabs: The bitmap terror (1985) [pdf] (lucacardelli.name)
48 points by dsr_ 5 hours ago | hide | 6 comments

19. Discord is not an acceptable choice for free software projects (sneak.berlin)
333 points by rauhl 5 hours ago | hide | 233 comments

20. Show HN: Build your own Vim emulation for VS Code (johtela.github.io)
68 points by johtela 6 hours ago | hide | 40 comments

21. When Einstein Was Bohemian (scientificamerican.com)
26 points by apollinaire 4 hours ago | hide | 12 comments

22. Wood wide web: Trees' social networks are mapped (2019) (bbc.co.uk)
17 points by ascertain 3 hours ago | hide | discuss

23. Detecting manuscripts and publications from paper mills (wiley.com)
5 points by blopeur 1 hour ago | hide | discuss

24. Chrome deploys deep-linking in latest build despite privacy concerns (theregister
170 points by mikro2nd 9 hours ago | hide | 129 comments

25. Web Programming in SWI Prolog (monolune.com)
66 points by luu 8 hours ago | hide | 11 comments

26. Maryam Mirzakhani (wikipedia.org)
219 points by mmoez 8 hours ago | hide | 29 comments

27. Domain Logic and SQL (2003) (martinfowler.com)
99 points by mxschumacher 12 hours ago | hide | 49 comments

```
28.      Stuff I said at Kansas City Startup Weekend that sounded smart (2011) (apenwarr.c
  70 points by Tomte 8 hours ago  | hide | 36 comments


29.      How communist Bulgaria became a leader in technology, robotics and sci-fi (2018)
  123 points by jordybg 12 hours ago  | hide | 55 comments


30.      Unix Toolbox (2008) (cb.vu)
  139 points by rodrigo975 12 hours ago  | hide | 24 comments

      More



  Applications are open for YC Summer 2020
Guidelines
  | FAQ
  | Support
  | API
  | Security
  | Lists
  | Bookmarklet
  | Legal
  | Apply to YC
  | ContactSearch:
```

```
In [11]: from bs4 import BeautifulSoup

         # Remove HTML tags using Beautiful Soup library
         soup = BeautifulSoup(r.text, "html5lib")
         print(soup.get_text())
```

```
  Hacker News

          Hacker News
      new | past | comments | ask | show | jobs | submit
                    login
```

1. Amazon Let a Fraudster Keep My Sony A7R IV and Refunded Him $2,900 (petapixel.com)
   80 points by ProAm 21 minutes ago | hide | 11ăcomments

2. More bosses give four-day workweek a try (npr.org)
   331 points by hhs 3 hours ago | hide | 182ăcomments

3. Google resists demands from states in digital-ad probe (wsj.com)
   280 points by aty268 6 hours ago | hide | 130ăcomments

4. Radical hydrogen-boron reactor leapfrogs current nuclear fusion tech (newatlas.com
   538 points by chris_overseas 9 hours ago | hide | 211ăcomments

5. An Open Guide to Equity Compensation (github.com)
   59 points by DyslexicAtheist 2 hours ago | hide | 8ăcomments

6. Single-payer healthcare would save $450B and 68k lives a year: study (thelancet.co
   166 points by bjourne 1 hour ago | hide | 132ăcomments

7. SQL query to generate the Mandelbrot Set as ASCII-art (sqlite.org)
   15 points by mmoez 1 hour ago | hide | discuss

8. JP Morgan economists warn of 'catastrophic' climate change (bbc.com)
   26 points by evo_9 2 hours ago | hide | 2ăcomments

9. Smarking (YC W15) is hiring Back end Tech Lead to scale urban mobility tech infra
   27 minutes ago | hide

10. How to Write Usefully (paulgraham.com)
    556 points by r_singh 12 hours ago | hide | 229ăcomments

11. Improved protein structure prediction using potentials from deep learning (nature
    16 points by lawrenceyan 1 hour ago | hide | 7ăcomments

12. Show HN: MDAnki convert Markdown to Anki cards (github.com)
47 points by rudnek 3 hours ago | hide | 25ăcomments

13. Neat URL cleans URLs, removing parameters such as fbclid utm parameters (github.c
17 points by mot2ba 1 hour ago | hide | 6ăcomments

14. QEMU for iOS (github.com)
79 points by funkaster 3 hours ago | hide | 43ăcomments

15. Gilbert Strang Teaches Linear Algebra (mit.edu)
265 points by ColinWright 10 hours ago | hide | 59ăcomments

16. Does register selection matter to performance on x86 CPUs? (fiigii.com)
99 points by lelf 6 hours ago | hide | 30ăcomments

17. Launch HN: Freshpaint (YC S19) an automated, retroactive Segment alternative
62 points by malisper 4 hours ago | hide | 31ăcomments

18. Crabs: The bitmap terror (1985) [pdf] (lucacardelli.name)
48 points by dsr_ 5 hours ago | hide | 6ăcomments

19. Discord is not an acceptable choice for free software projects (sneak.berlin)
333 points by rauhl 5 hours ago | hide | 233ăcomments

20. Show HN: Build your own Vim emulation for VS Code (johtela.github.io)
68 points by johtela 6 hours ago | hide | 40ăcomments

21. When Einstein Was Bohemian (scientificamerican.com)
26 points by apollinaire 4 hours ago | hide | 12ăcomments

22. Wood wide web: Trees' social networks are mapped (2019) (bbc.co.uk)
17 points by ascertain 3 hours ago | hide | discuss

23. Detecting manuscripts and publications from paper mills (wiley.com)
5 points by blopeur 1 hour ago | hide | discuss

24. Chrome deploys deep-linking in latest build despite privacy concerns (theregister
170 points by mikro2nd 9 hours ago | hide | 129ăcomments

25. Web Programming in SWI Prolog (monolune.com)
66 points by luu 8 hours ago | hide | 11ăcomments

26. Maryam Mirzakhani (wikipedia.org)
219 points by mmoez 8 hours ago | hide | 29ăcomments

27. Domain Logic and SQL (2003) (martinfowler.com)
99 points by mxschumacher 12 hours ago | hide | 49ăcomments

28. Stuff I said at Kansas City Startup Weekend that sounded smart (2011) (apenwarr.c
70 points by Tomte 8 hours ago | hide | 36ăcomments

29. How communist Bulgaria became a leader in technology, robotics and sci-fi (2018)
123 points by jordybg 12 hours ago | hide | 55ăcomments

30. Unix Toolbox (2008) (cb.vu)
139 points by rodrigo975 12 hours ago | hide | 24ăcomments

More

```
In [12]: # Find all articles
         summaries = soup.find_all("tr", class_="athing")
         summaries[0]
```

```
Out[12]: <tr class="athing" id="22386960">
               <td align="right" class="title" valign="top"><span class="rank">1.</span></td>
```

```
In [13]: # Extract title
         summaries[0].find("a", class_="storylink").get_text().strip()
```

```
Out[13]: 'Amazon Let a Fraudster Keep My Sony A7R IV and Refunded Him $2,900'
```

```
In [14]: # Find all articles, extract titles
         articles = []
         summaries = soup.find_all("tr", class_="athing")
         for summary in summaries:
             title = summary.find("a", class_="storylink").get_text().strip()
             articles.append((title))


         print(len(articles), "Article summaries found. Sample:")
         print(articles[0])
```

```
30 Article summaries found. Sample:
Amazon Let a Fraudster Keep My Sony A7R IV and Refunded Him $2,900
```

1.3 Normalization

1.3.1 Case Normalization

```
In [15]: # Sample text
         text = "The first time you see The Second Renaissance it may look boring. Look at it at
         print(text)
```

```
The first time you see The Second Renaissance it may look boring. Look at it at least twice and
```

```
In [16]: # Convert to lowercase
         text = text.lower()
         print(text)
```

```
the first time you see the second renaissance it may look boring. look at it at least twice and
```

1.3.2 Punctuation Removal

```
In [17]: import re

         # Remove punctuation characters
         text = re.sub(r"[^a-zA-Z0-9]", " ", text)
         print(text)
```

```
the first time you see the second renaissance it may look boring  look at it at least twice and
```

1.4 Tokenization

```
In [18]: # Split text into tokens (words)
         words = text.split()
         print(words)
```

```
['the', 'first', 'time', 'you', 'see', 'the', 'second', 'renaissance', 'it', 'may', 'look', 'bor
```

1.4.1 NLTK: Natural Language ToolKit

```
In [19]: import os
         import nltk
         nltk.data.path.append(os.path.join(os.getcwd(), "nltk_data"))
```

```
In [20]: # Another sample text
         text = "Dr. Smith graduated from the University of Washington. He later started an anal
         print(text)
```

```
Dr. Smith graduated from the University of Washington. He later started an analytics firm called
```

```
In [21]: from nltk.tokenize import word_tokenize

         # Split text into words using NLTK
         words = word_tokenize(text)
         print(words)
```

```
['Dr.', 'Smith', 'graduated', 'from', 'the', 'University', 'of', 'Washington', '.', 'He', 'later
```

```
In [22]: from nltk.tokenize import sent_tokenize

         # Split text into sentences
         sentences = sent_tokenize(text)
         print(sentences)
```

```
['Dr. Smith graduated from the University of Washington.', 'He later started an analytics firm c
```

```
In [23]: # List stop words
         from nltk.corpus import stopwords
         print(stopwords.words("english"))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll
```

```
In [24]: # Reset text
         text = "The first time you see The Second Renaissance it may look boring. Look at it at

         # Normalize it
         text = re.sub(r"[^a-zA-Z0-9]", " ", text.lower())

         # Tokenize it
         words = text.split()
         print(words)
```

```
['the', 'first', 'time', 'you', 'see', 'the', 'second', 'renaissance', 'it', 'may', 'look', 'bor
```

```
In [25]: # Remove stop words
         words = [w for w in words if w not in stopwords.words("english")]
         print(words)
```

```
['first', 'time', 'see', 'second', 'renaissance', 'may', 'look', 'boring', 'look', 'least', 'twi
```

1.4.2 Sentence Parsing

```
In [26]: import nltk

         # Define a custom grammar
         my_grammar = nltk.CFG.fromstring("""
         S -> NP VP
         PP -> P NP
         NP -> Det N | Det N PP | 'I'
         VP -> V NP | VP PP
         Det -> 'an' | 'my'
         N -> 'elephant' | 'pajamas'
         V -> 'shot'
         P -> 'in'
         """)
         parser = nltk.ChartParser(my_grammar)

         # Parse a sentence
         sentence = word_tokenize("I shot an elephant in my pajamas")
         for tree in parser.parse(sentence):
             print(tree)
```

```
(S
  (NP I)
  (VP
    (VP (V shot) (NP (Det an) (N elephant)))
    (PP (P in) (NP (Det my) (N pajamas)))))
(S
  (NP I)
```

```
  (VP
    (V shot)
    (NP (Det an) (N elephant) (PP (P in) (NP (Det my) (N pajamas))))))
```

1.5 Stemming & Lemmatization

1.5.1 Stemming

```
In [27]: from nltk.stem.porter import PorterStemmer

         # Reduce words to their stems
         stemmed = [PorterStemmer().stem(w) for w in words]
         print(stemmed)
```

```
['first', 'time', 'see', 'second', 'renaiss', 'may', 'look', 'bore', 'look', 'least', 'twice', '
```

1.5.2 Lemmatization

```
In [28]: from nltk.stem.wordnet import WordNetLemmatizer

         # Reduce words to their root form
         lemmed = [WordNetLemmatizer().lemmatize(w) for w in words]
         print(lemmed)
```

```
['first', 'time', 'see', 'second', 'renaissance', 'may', 'look', 'boring', 'look', 'least', 'twi
```

```
In [29]: # Lemmatize verbs by specifying pos
         lemmed = [WordNetLemmatizer().lemmatize(w, pos='v') for w in lemmed]
         print(lemmed)
```

```
['first', 'time', 'see', 'second', 'renaissance', 'may', 'look', 'bore', 'look', 'least', 'twice
```

```
In [ ]:
```