mathjax2mobi：將 MathJax HTML 轉換為電子書

項目簡介

先大致講講項目情況。

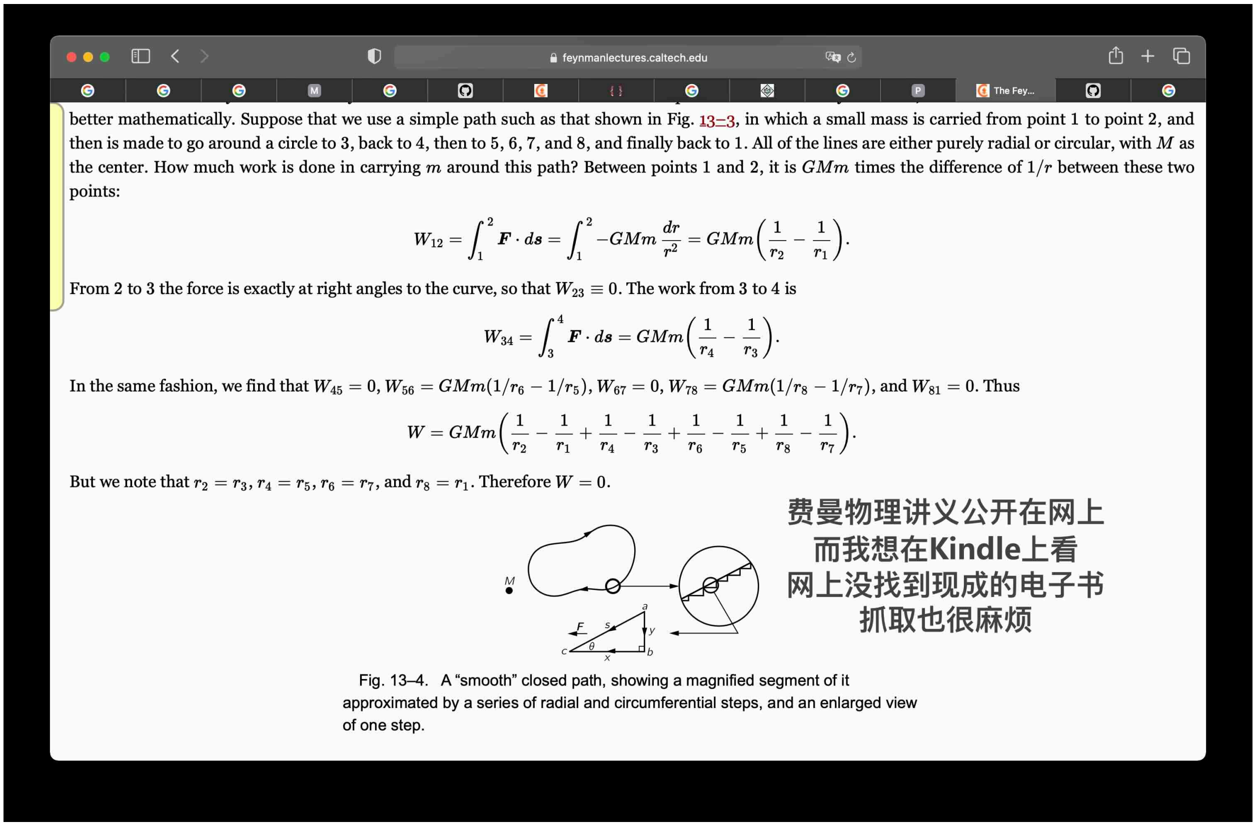


Figure 1: feynman_online

做完項目後，有點開心。寫下了這樣一段話。

寫了一天代碼，終於得到了漂亮的費曼物理講義電子書！費曼物理講義公開在網上，是用 `latex` 渲染的。人們常用 `latex` 來寫論文，它對數學公式的渲染很棒。而公開在網上，用到了 `mathjax` 這個庫。它把 `latex` 源碼變成了 `html` 代碼，生成了很多的 `div` 和 `span` 標籤。電子書卻不支持這種方式。這時，想法是抓取網頁，逆向 `mathjax` 渲染，接著替換成 `svg` 圖片。出現了不少問題，一個是源碼有很多的 `latex` 自定義宏，需要加上；第二個是內嵌很多 `svg` 會有問題。如果是單個 `svg` 倒沒問題，很多的時候會出現問題。大概是瀏覽器和 `svg` 的詭異 Bug。這時只要把 `svg` 保存為文件，用 `img` 標籤引入進來即可。公式也分為兩種，一種是文本中間的公式，一種是單行的公式。所以，最後就得到了漂亮的電子書！

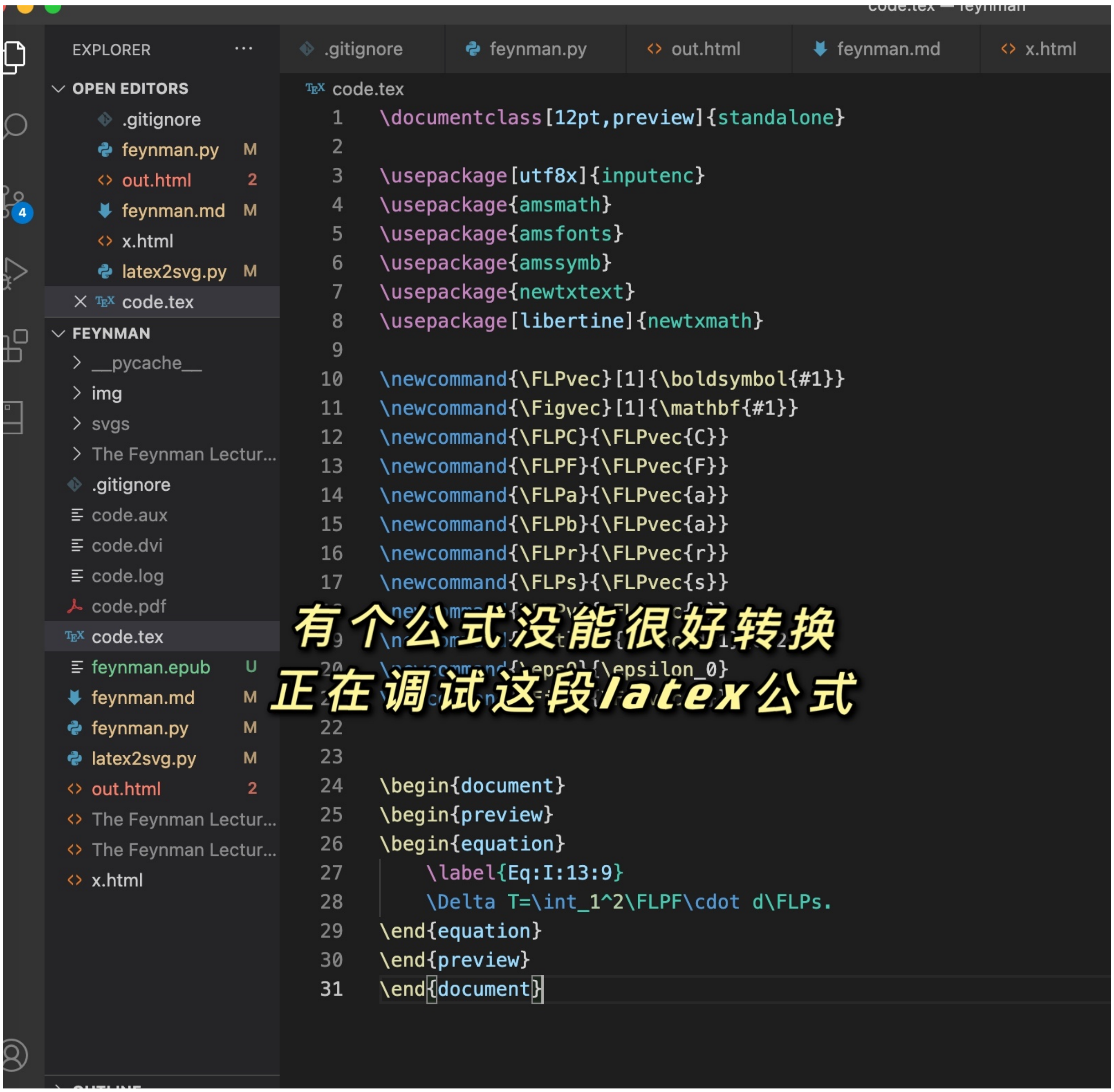
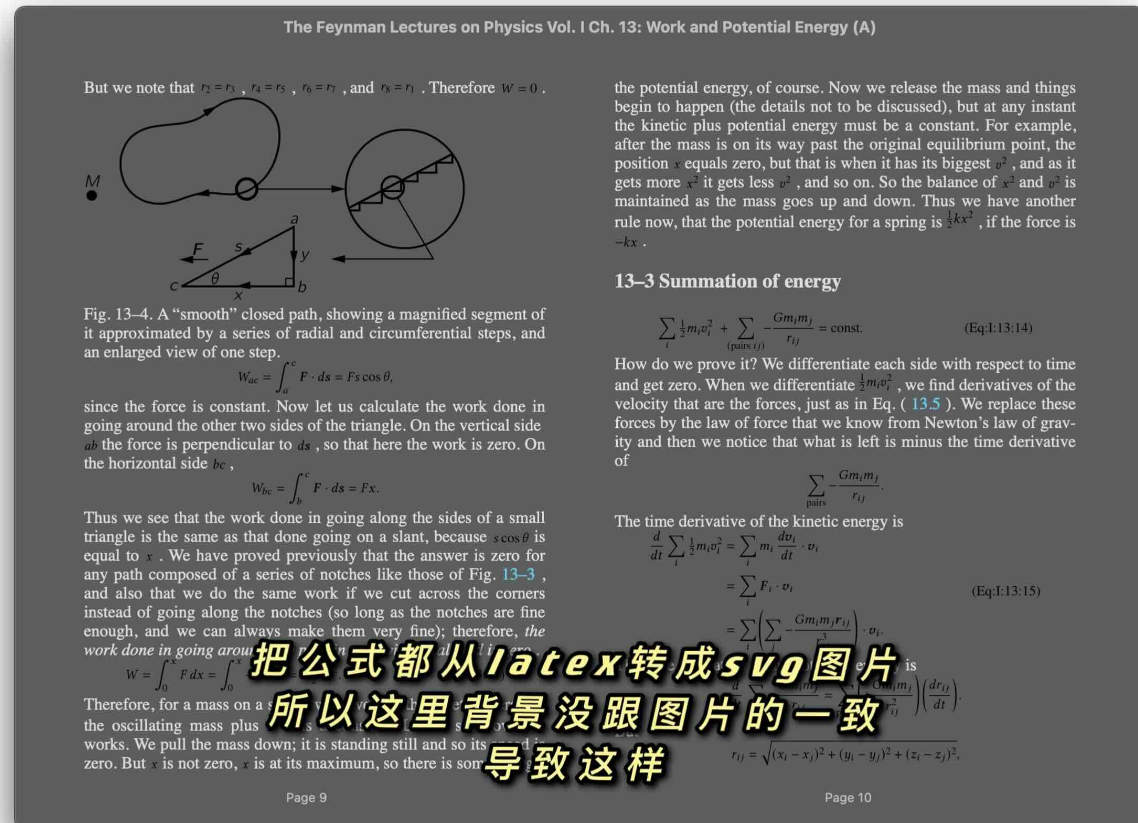


Figure 2: latex



The Feynman Lectures on Physics Vol. I Ch. 13: Work and Potential Energy (A)

But we note that $r_2 = r_3$, $r_4 = r_5$, $r_6 = r_7$, and $r_8 = r_1$. Therefore $W = 0$.

Fig. 13–4. A “smooth” closed path, showing a magnified segment of it approximated by a series of radial and circumferential steps, and an enlarged view of one step.

$$W_{ac} = \int_a^c \boldsymbol{F} \cdot d\boldsymbol{s} = Fs\cos\theta,$$

since the force is constant. Now let us calculate the work done in going around the other two sides of the triangle. On the vertical side ab the force is perpendicular to $d\boldsymbol{s}$, so that here the work is zero. On the horizontal side bc ,

$$W_{bc} = \int_b^c \boldsymbol{F} \cdot d\boldsymbol{s} = Fx.$$

Thus we see that the work done in going along the sides of a small triangle is the same as that done going on a slant, because $s\cos\theta$ is equal to x . We have proved previously that the answer is zero for any path composed of a series of notches like those of Fig. 13–3 , and also that we do the same work if we cut across the corners instead of going along the notches (so long as the notches are fine enough, and we can always make them very fine); therefore, *the work done in going arou*

$$W = \int_0^x F\,dx = \int_0^x$$

Therefore, for a mass on a s

the oscillating mass plus

works. We pull the mass down; it is standing still and so its

zero. But x is not zero, x is at its maximum, so there is som

Page 9

the potential energy, of course. Now we release the mass and things begin to happen (the details not to be discussed), but at any instant the kinetic plus potential energy must be a constant. For example, after the mass is on its way past the original equilibrium point, the position x equals zero, but that is when it has its biggest v^2 , and as it gets more x^2 it gets less v^2 , and so on. So the balance of x^2 and v^2 is maintained as the mass goes up and down. Thus we have another rule now, that the potential energy for a spring is $\frac{1}{2}kx^2$, if the force is $-kx$.

13–3 Summation of energy

$$\sum_i \tfrac{1}{2} m_i v_i^2 + \sum_{(\text{pairs } ij)} -\frac{Gm_im_j}{r_{ij}} = \text{const.} \qquad \text{(Eq:I:13:14)}$$

How do we prove it? We differentiate each side with respect to time and get zero. When we differentiate $\frac{1}{2}m_iv_i^2$, we find derivatives of the velocity that are the forces, just as in Eq. (13.5). We replace these forces by the law of force that we know from Newton’s law of gravity and then we notice that what is left is minus the time derivative of

$$\sum_{\text{pairs}} -\frac{Gm_im_j}{r_{ij}}.$$

The time derivative of the kinetic energy is

$$\begin{aligned} \frac{d}{dt}\sum_i \tfrac{1}{2}m_iv_i^2 &= \sum_i m_i\frac{d\boldsymbol{v}_i}{dt}\cdot\boldsymbol{v}_i \\ &= \sum_i \boldsymbol{F}_i\cdot\boldsymbol{v}_i \qquad \text{(Eq:I:13:15)} \\ &= \sum_i\left(\sum_j -\frac{Gm_im_j\boldsymbol{r}_{ij}}{r_{ij}^3}\right)\cdot\boldsymbol{v}_i. \end{aligned}$$

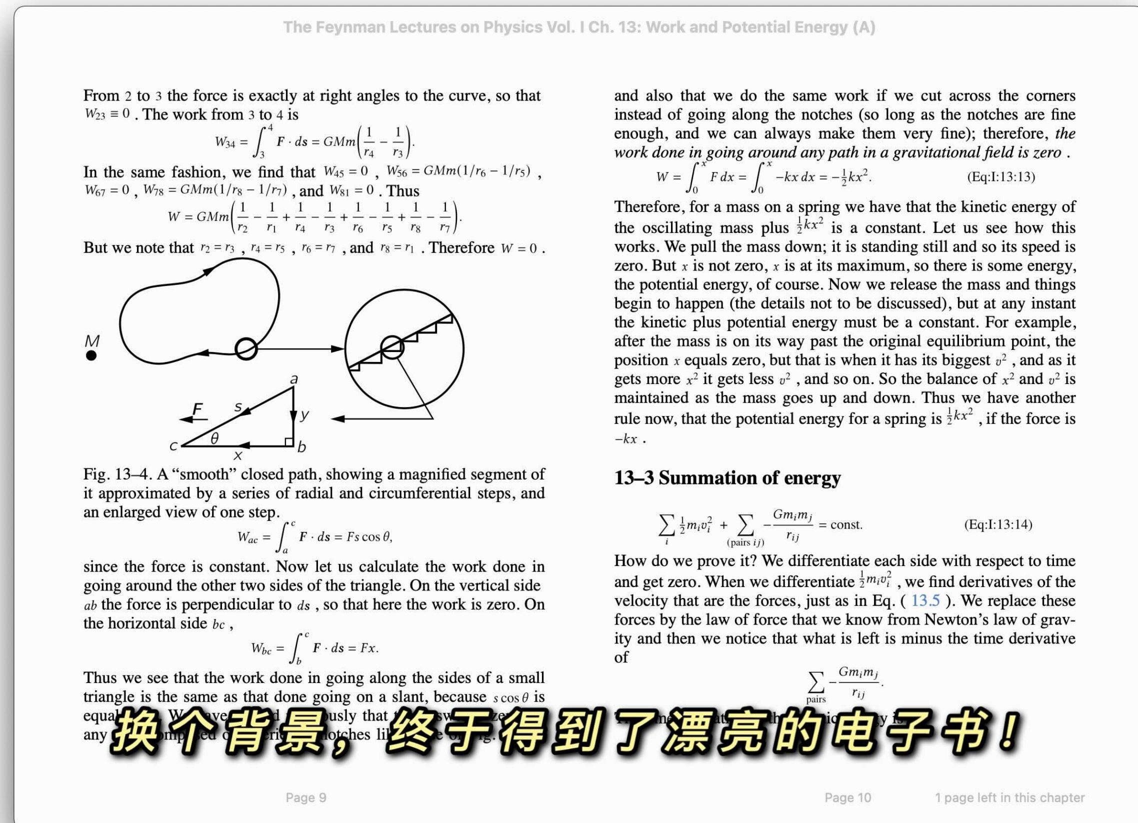
is

$$\left(\frac{Gm_im_j}{r_{ij}^2}\right)\left(\frac{dr_{ij}}{dt}\right).$$

$$r_{ij} = \sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2},$$

Page 10

Figure 3: epub_black



From 2 to 3 the force is exactly at right angles to the curve, so that $W_{23} \equiv 0$. The work from 3 to 4 is

$$W_{34} = \int_3^4 \boldsymbol{F} \cdot d\boldsymbol{s} = GMm\left(\frac{1}{r_4} - \frac{1}{r_3}\right).$$

In the same fashion, we find that $W_{45} = 0$, $W_{56} = GMm(1/r_6 - 1/r_5)$, $W_{67} = 0$, $W_{78} = GMm(1/r_8 - 1/r_7)$, and $W_{81} = 0$. Thus

$$W = GMm\left(\frac{1}{r_2} - \frac{1}{r_1} + \frac{1}{r_4} - \frac{1}{r_3} + \frac{1}{r_6} - \frac{1}{r_5} + \frac{1}{r_8} - \frac{1}{r_7}\right).$$

But we note that $r_2 = r_3$, $r_4 = r_5$, $r_6 = r_7$, and $r_8 = r_1$. Therefore $W = 0$.

Fig. 13–4. A "smooth" closed path, showing a magnified segment of it approximated by a series of radial and circumferential steps, and an enlarged view of one step.

$$W_{ac} = \int_a^c \boldsymbol{F} \cdot d\boldsymbol{s} = Fs\cos\theta,$$

since the force is constant. Now let us calculate the work done in going around the other two sides of the triangle. On the vertical side ab the force is perpendicular to ds , so that here the work is zero. On the horizontal side bc ,

$$W_{bc} = \int_b^c \boldsymbol{F} \cdot d\boldsymbol{s} = Fx.$$

Thus we see that the work done in going along the sides of a small triangle is the same as that done going on a slant, because $s\cos\theta$ is

and also that we do the same work if we cut across the corners instead of going along the notches (so long as the notches are fine enough, and we can always make them very fine); therefore, *the work done in going around any path in a gravitational field is zero* .

$$W = \int_0^x F\,dx = \int_0^x -kx\,dx = -\tfrac{1}{2}kx^2. \quad \text{(Eq:I:13:13)}$$

Therefore, for a mass on a spring we have that the kinetic energy of the oscillating mass plus $\frac{1}{2}kx^2$ is a constant. Let us see how this works. We pull the mass down; it is standing still and so its speed is zero. But x is not zero, x is at its maximum, so there is some energy, the potential energy, of course. Now we release the mass and things begin to happen (the details not to be discussed), but at any instant the kinetic plus potential energy must be a constant. For example, after the mass is on its way past the original equilibrium point, the position x equals zero, but that is when it has its biggest v^2 , and as it gets more x^2 it gets less v^2 , and so on. So the balance of x^2 and v^2 is maintained as the mass goes up and down. Thus we have another rule now, that the potential energy for a spring is $\frac{1}{2}kx^2$, if the force is $-kx$.

13–3 Summation of energy

$$\sum_i \tfrac{1}{2}m_i v_i^2 + \sum_{(\text{pairs } ij)} -\frac{Gm_i m_j}{r_{ij}} = \text{const.} \quad \text{(Eq:I:13:14)}$$

How do we prove it? We differentiate each side with respect to time and get zero. When we differentiate $\frac{1}{2}m_i v_i^2$, we find derivatives of the velocity that are the forces, just as in Eq. (13.5). We replace these forces by the law of force that we know from Newton's law of gravity and then we notice that what is left is minus the time derivative of

$$\sum_{\text{pairs}} -\frac{Gm_i m_j}{r_{ij}}.$$

Figure 4: epub_beautiful

查詢的資料

這裡記錄了解決項目過程中訪問的資料。因為這是一個教程，所以向學生展示一下大概做一個項目是怎麼樣的體驗。

開始項目

費曼物理講義已經在公開在網上可以閱讀。我想在 `Kindle` 上看它。然而因為它有挺多的數學公式。它最初的稿子應該是用 `latex` 做的。它用 `mathjax` 這個庫來把 `latex` 格式的內容顯示在網頁上。

舉個例子。

```
<span class="MathJax_Preview" style="color: inherit; display: none;">
</span>
<div class="MathJax_Display">
    <span class="MathJax MathJax_FullWidth" id="MathJax-Element-10-Frame" tabindex="0" style="">
              <span class="mi" id="MathJax-Span-159" style="font-family: MathJax_Math-italic;">d<span s
                </span>
    </span>
</div>
<script type="math/tex; mode=display" id="MathJax-Element-10">\begin{equation}
\label{Eq:I:13:3}
dT/dt = Fv.
\end{equation}
</script>
```

上面是截取的一段 `html` 代碼。這一塊 `html` 代碼中。`script` 標籤下是 `latex` 的原樣文本。`mathjax` 把它變成很多的 `span`。來顯示它。

我們現在有個思路。就是把 `mathjax` 的顯示方法改成 `svg` 圖片。

從 GitHub 上找到一個項目 `tuxu/latex2svg`。

```
from latex2svg import latex2svg
out = latex2svg(r'\( e^{i \pi} + 1 = 0 \)')
print(out['depth'])
print(out['svg'])
```

試著運行，但出錯了。

```
    raise RuntimeError('latex not found')
RuntimeError: latex not found
```

看看代碼。

```python
    # Run LaTeX and create DVI file
    try:
        ret = subprocess.run(shlex.split(params['latex_cmd']+' code.tex'),
                             stdout=subprocess.PIPE, stderr=subprocess.PIPE,
                             cwd=working_directory)
        ret.check_returncode()
    except FileNotFoundError:
        raise RuntimeError('latex not found')
```

原來這也依賴於 `latex` 命令。

安裝一下。

```
brew install --cask mactex
==> Caveats
You must restart your terminal window for the installation of MacTex CLI tools to take effect.
Alternatively, Bash and Zsh users can run the command:
  eval "$(/usr/libexec/path_helper)"
==> Downloading http://mirror.ctan.org/systems/mac/mactex/mactex-20200407.pkg
==> Downloading from https://mirrors.aliyun.com/CTAN/systems/mac/mactex/mactex-20200407.pkg
######################################################################## 100.0%
All formula dependencies satisfied.
==> Installing Cask mactex
==> Running installer for mactex; your password may be necessary.
installer: Package name is MacTeX
installer: choices changes file '/private/tmp/choices20210315-4643-5884ro.xml' applied
installer: Installing at base path /
installer: The install was successful.
   mactex was successfully installed!
```

安裝成功。

```
% latex
```

```
This is pdfTeX, Version 3.14159265-2.6-1.40.21 (TeX Live 2020) (preloaded format=latex)
 restricted \write18 enabled.
**
```

```python
out = latex2svg(r'\( e^{i \pi} + 1 = 0 \)')
print(out['depth'])
print(out['svg'])

svg = open('1.svg', 'w')
svg.write(out['svg'])
svg.close()
```

可以生成 svg 了。

所以試試把 mathjax 中得到的 latex 文本都生成一下。

```python
from bs4 import BeautifulSoup
from latex2svg import latex2svg

file = open('The Feynman Lectures on Physics Vol. I Ch. 13_ Work and Potential Energy (A).html')
content = file.read()

soup = BeautifulSoup(content)

mathjaxs = soup.findAll('script', {'type': 'math/tex'})
for mathjax in mathjaxs:
    print(mathjax.string)
    out = latex2svg(mathjax.string)
    print(out['svg'])
```

可惜出錯了。

```
    raise CalledProcessError(self.returncode, self.args, self.stdout,
subprocess.CalledProcessError: Command '['latex', '-interaction', 'nonstopmode', '-halt-on-error', 'code
```

具體哪個公式錯了呢。

```latex
\tfrac{1}{2}mv^2
```

latex

來學習一下 latex。

```latex
\documentclass[12pt]{article}
\usepackage{lingmacros}
\usepackage{tree-dvips}
\begin{document}

\section*{Notes for My Paper}

Don't forget to include examples of topicalization.
They look like this:

{\small
\enumsentence{Topicalization from sentential subject:\\
\shortex{7}{a John$_i$ [a & kltukl & [el &
  {\bf l-}oltoir & er & ngii$_i$ & a Mary]]}
{ & {\bf R-}clear & {\sc comp} &
  {\bf IR}.{\sc 3s}-love   & P & him & }
{John, (it's) clear that Mary loves (him).}}
}

\subsection*{How to handle topicalization}

I'll just assume a tree structure like (\ex{1}).

{\small
\enumsentence{Structure of A$'$ Projections:\\ [2ex]
\begin{tabular}[t]{cccc}
    & \node{i}{CP}\\ [2ex]
    \node{ii}{Spec} &   &\node{iii}{C$'$}\\ [2ex]
        &\node{iv}{C} & & \node{v}{SAgrP}
\end{tabular}
\nodeconnect{i}{ii}
\nodeconnect{i}{iii}
\nodeconnect{iii}{iv}
```

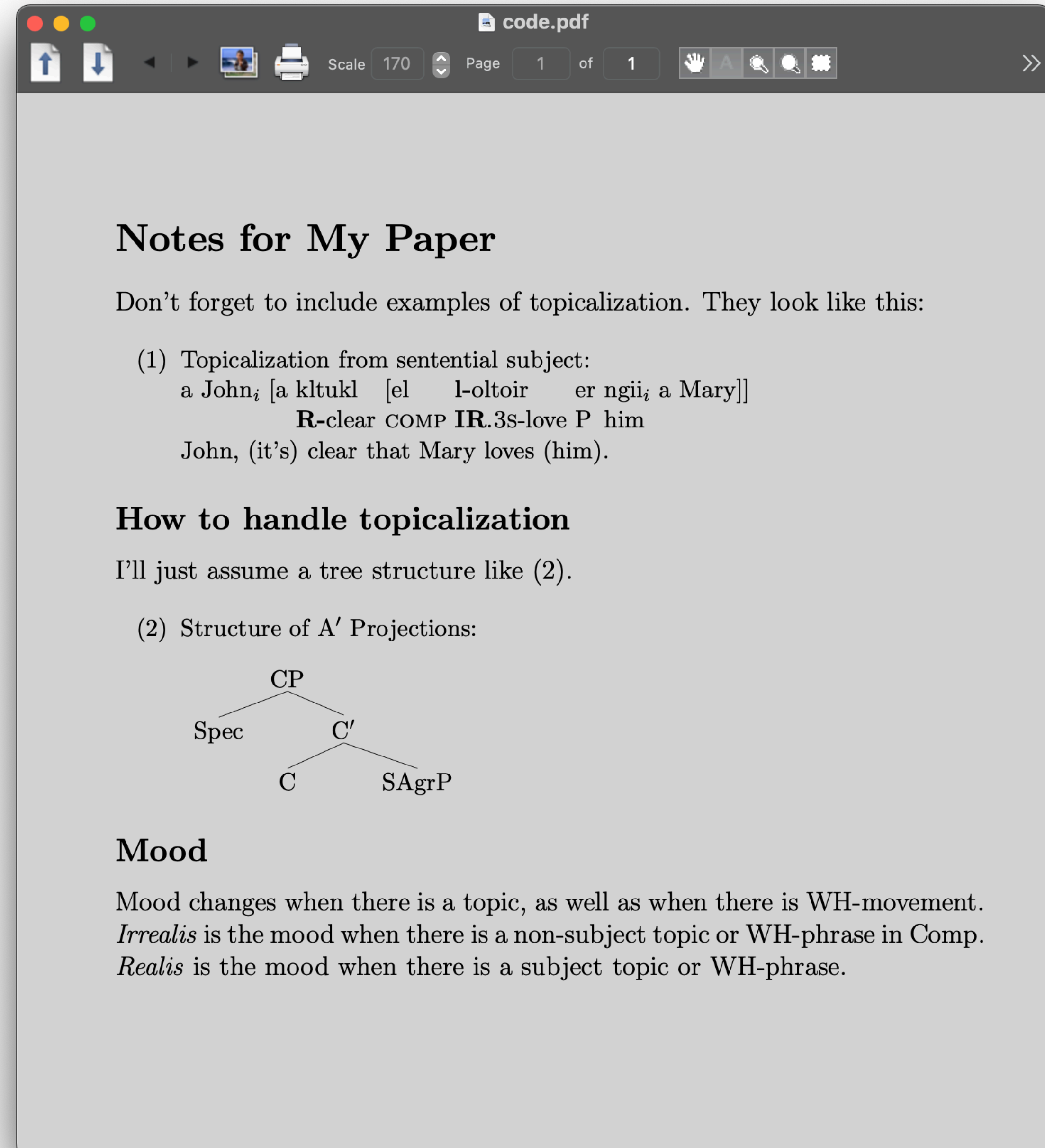
```
\nodeconnect{iii}{v}
}
}

\subsection*{Mood}

Mood changes when there is a topic, as well as when
there is WH-movement.  \emph{Irrealis} is the mood when
there is a non-subject topic or WH-phrase in Comp.
\emph{Realis} is the mood when there is a subject topic
or WH-phrase.

\end{document}
```

網上找到一段樣例的 `latex` 源碼。

```
% latex code.tex
This is pdfTeX, Version 3.14159265-2.6-1.40.21 (TeX Live 2020) (preloaded format=latex)
 restricted \write18 enabled.
entering extended mode
(./code.tex
LaTeX2e <2020-02-02> patch level 5
L3 programming layer <2020-03-06>
(/usr/local/texlive/2020/texmf-dist/tex/latex/base/article.cls
Document Class: article 2019/12/20 v1.4l Standard LaTeX document class
(/usr/local/texlive/2020/texmf-dist/tex/latex/base/size12.clo))
(/usr/local/texlive/2020/texmf-dist/tex/latex/tree-dvips/lingmacros.sty)
(/usr/local/texlive/2020/texmf-dist/tex/latex/tree-dvips/tree-dvips.sty
tree-dvips version .91 of May 16, 1995
) (/usr/local/texlive/2020/texmf-dist/tex/latex/l3backend/l3backend-dvips.def)
(./code.aux) [1] (./code.aux) )
Output written on code.dvi (1 page, 3416 bytes).
Transcript written on code.log.
```

來對著源碼和渲染後的效果，看看能學到什麼。

```
\begin{document}
\end{document}
```



Notes for My Paper

Don't forget to include examples of topicalization. They look like this:

(1) Topicalization from sentential subject:
a John$_i$ [a kltukl [el **l**-oltoir er ngii$_i$ a Mary]]
R-clear COMP **IR**.3S-love P him
John, (it's) clear that Mary loves (him).

How to handle topicalization

I'll just assume a tree structure like (2).

(2) Structure of A$'$ Projections:

CP
Spec C$'$
C SAgrP

Mood

Mood changes when there is a topic, as well as when there is WH-movement. *Irrealis* is the mood when there is a non-subject topic or WH-phrase in Comp. *Realis* is the mood when there is a subject topic or WH-phrase.

Figure 5: latex

這樣來把文檔裹起來。

```
\section*{Notes for My Paper}
```

這表示 `section` 標題開頭。

```
\subsection*{How to handle topicalization}
```

這表示子標題。

```
\shortex{7}{a John$_i$ [a & kltukl & [el &
  {\bf l-}oltoir & er & ngii$_i$ & a Mary]]}
```

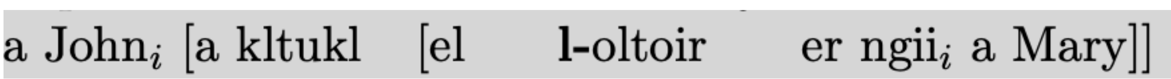


Figure 6: shortex

可見 `$_i$` 來表示下標。`{\bf l-}` 來表示加粗。

```
\enumsentence{Structure of A$'$ Projections:\\ [2ex]
\begin{tabular}[t]{cccc}
    & \node{i}{CP}\\ [2ex]
    \node{ii}{Spec} &   &\node{iii}{C$'$}\\ [2ex]
        &\node{iv}{C} & & \node{v}{SAgrP}
\end{tabular}
\nodeconnect{i}{ii}
\nodeconnect{i}{iii}
\nodeconnect{iii}{iv}
\nodeconnect{iii}{v}
}
```

注意到 `nodeconnect` 來表示連線。

latex 轉換成 svg

繼續項目。

```
\documentclass[16pt]{article}
\usepackage{amsmath}
\begin{document}

\[\tfrac{1}{2}mv^2\]

\end{document}
```

這樣可以正確地被渲染。在代碼裡無法被渲染，可能是因為沒有加上`\usepackage{amsmath}`。

```
\documentclass[12pt,preview]{standalone}

\usepackage[utf8x]{inputenc}
\usepackage{amsmath}
\usepackage{amsfonts}
\usepackage{amssymb}
\usepackage{newtxtext}
\usepackage[libertine]{newtxmath}

\begin{document}
\begin{preview}
\tfrac{1}{2}mv^2
\end{preview}
\end{document}
```

```
! Missing $ inserted.
<inserted text>
                $
l.12 \tfrac{1}{2}
                 mv^2
```

這樣出錯了。而改成一下這樣就可以。

```
\[\tfrac{1}{2}mv^2\]
```

進行各種試探。

```python
from bs4 import BeautifulSoup
from latex2svg import latex2svg

file = open('The Feynman Lectures on Physics Vol. I Ch. 13_ Work and Potential Energy (A).html')
content = file.read()

soup = BeautifulSoup(content, features="lxml")

mathjaxs = soup.findAll('script', {'type': 'math/tex'})
for mathjax in mathjaxs:
    print(mathjax.string)
    wrap = '$' + mathjax.string + '$'
    # if 'frac' in mathjax.string:
    #     wrap = '$' + mathjax.string + '$'
    if 'FLP' in mathjax.string:
        continue
    elif 'Fig' in mathjax.string:
        continue
    elif 'eps' in mathjax.string:
        continue
    out = latex2svg(wrap)
    # print(out)
    node = BeautifulSoup(out['svg'], features="lxml")
    svg = node.find('svg')
    mathjax.insert_after(svg)
    # print(out['svg'])
    # break
    # mathjax.replaceWith(out['svg'])

    # print(dir(mathjax))
    # break

    # out = latex2svg(wrap)
    # print(out['svg'])

# print(len(soup.contents))
```

```
output_file = open('out.html', 'w')
output_file.write(soup.prettify())
output_file.close()
# print(soup.contents)

# out = latex2svg(r'\( e^{i \pi} + 1 = 0 \)')
# print(out['depth'])
# print(out['svg'])

# svg = open('1.svg', 'w')
# svg.write(out['svg'])
# svg.close()
```

這些我都在試探什麼呢。

```
    if 'FLP' in mathjax.string:
        continue
    elif 'Fig' in mathjax.string:
        continue
    elif 'eps' in mathjax.string:
        continue
```

這裡當解析到有 FLP、Fig、eps 在 latex 源碼的時候，轉換的過程出錯了。

例如，在 HTML 中，有這樣的腳本：

```
<script type="math/tex" id="MathJax-Element-11">\FLPF\cdot\FLPv</script>
```

解析拿到：

```
\FLPF\cdot\FLPv
```

當在代碼裡轉換的時候出錯了。也即，latex2svg.py 出錯了。這裡就是用 latex 程序來轉換。

code.tex:

```
\documentclass[12pt,preview]{standalone}
```

```
\usepackage[utf8x]{inputenc}
\usepackage{amsmath}
\usepackage{amsfonts}
\usepackage{amssymb}
\usepackage{newtxtext}
\usepackage[libertine]{newtxmath}

\begin{document}
\begin{preview}
\begin{equation}
    \FLPF\cdot\FLPv
\end{equation}
\end{preview}
\end{document}
```

```
$latex code.tex
! Undefined control sequence.
l.13     \FLPF
              \cdot\FLPv
?
```

這到底是什麼問題。我後來才注意到在 html 中的這段代碼。

```
<script type="text/x-mathjax-config;executed=true">
      MathJax.Hub.Config({
        TeX: {
          Macros: {
            FLPvec: ["\\boldsymbol{#1}", 1], Figvec: ["\\mathbf{#1}", 1], FLPC: ["\\FLPvec{C}", 0], FLPI
          }
        }
      });
</script>
```

這表示網頁在渲染的時候，給 MathJax 設置上了宏。所以我們的 latex 轉換源碼裡也應該加上。來加上它們。

```
\documentclass[12pt,preview]{standalone}
```

```
\usepackage[utf8x]{inputenc}
\usepackage{amsmath}
\usepackage{amsfonts}
\usepackage{amssymb}
\usepackage{newtxtext}
\usepackage[libertine]{newtxmath}

\newcommand{\FLPvec}[1]{\boldsymbol{#1}}
\newcommand{\Figvec}[1]{\mathbf{#1}}
\newcommand{\FLPC}{\FLPvec{C}}
\newcommand{\FLPF}{\FLPvec{F}}
\newcommand{\FLPa}{\FLPvec{a}}
\newcommand{\FLPb}{\FLPvec{a}}
\newcommand{\FLPr}{\FLPvec{r}}
\newcommand{\FLPs}{\FLPvec{s}}
\newcommand{\FLPv}{\FLPvec{v}}
\newcommand{\ddt}[2]{\frac{d#1}{d#2}}
\newcommand{\eps0}{\epsilon_0}
\newcommand{\FigC}{\Figvec{C}}
\begin{document}
\begin{preview}
\begin{equation}
    \FLPF\cdot\FLPv
\end{equation}
\end{preview}
\end{document}
```

這樣就對了。

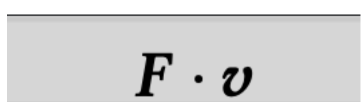


Figure 7: fv1

分析代碼

來看看最後的代碼。

“ ‘python import subprocess from bs4 import BeautifulSoup from latex2svg import latex2svg

def clean_mathjax(soup, name, cls): previews