Data engineering coursework 2 – Syed Wasti – w1854761

Image collection:

Source: https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset

The images are clothes and I collected 25 of them.

I chose clothes as they each are unique yet similar in their own way.

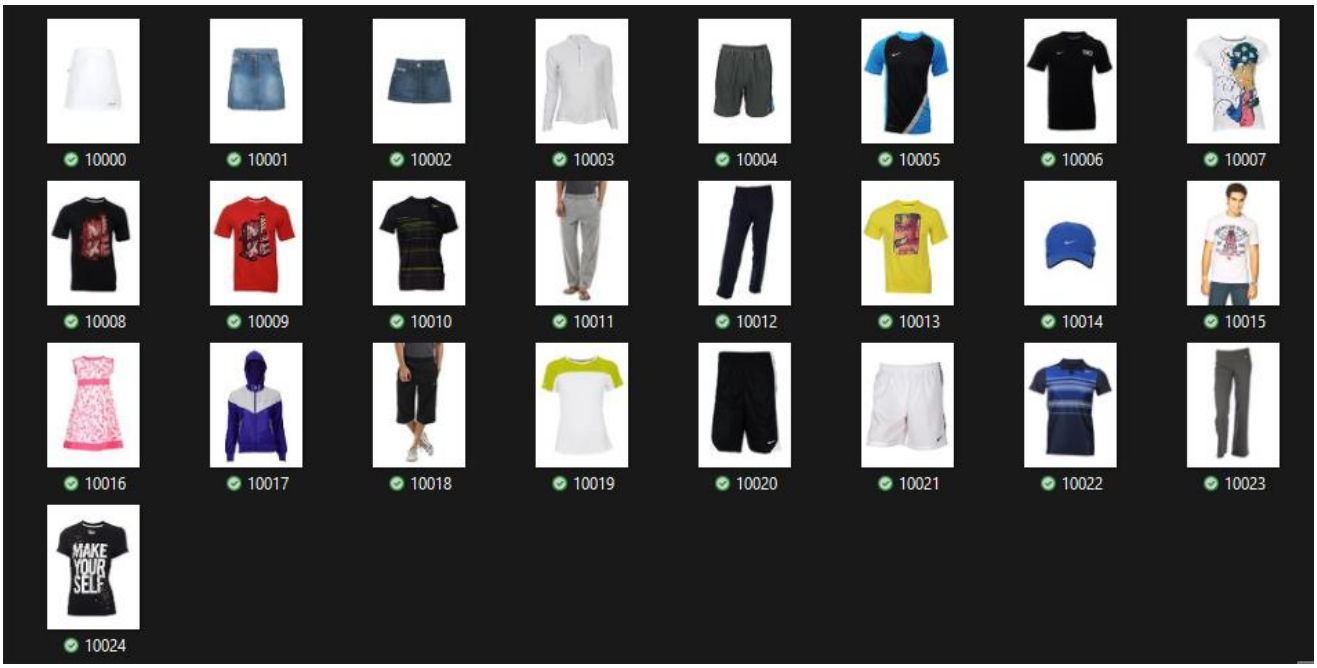


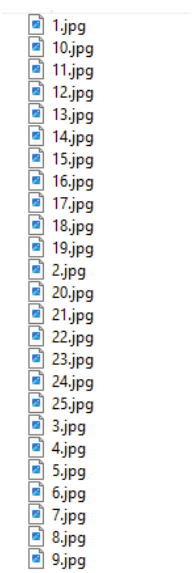
Image processing:

- I started by making a new folder called clean images where I would copy and store them in.

- I now used a loop to see all the images width and height before resizing

```
Dimensions of images before resizing:
Image 1 - Height: 2400, Width: 1800
Image 2 - Height: 2400, Width: 1800
Image 3 - Height: 2400, Width: 1800
Image 4 - Height: 2400, Width: 1800
Image 5 - Height: 1440, Width: 1080
Image 6 - Height: 2400, Width: 1800
Image 7 - Height: 2400, Width: 1800
Image 8 - Height: 2400, Width: 1800
Image 9 - Height: 2400, Width: 1800
Image 10 - Height: 2400, Width: 1800
Image 11 - Height: 2400, Width: 1800
Image 12 - Height: 1440, Width: 1080
Image 13 - Height: 2400, Width: 1800
Image 14 - Height: 2400, Width: 1800
Image 15 - Height: 2400, Width: 1800
Image 16 - Height: 2400, Width: 1800
Image 17 - Height: 2400, Width: 1800
Image 18 - Height: 2400, Width: 1800
Image 19 - Height: 1440, Width: 1080
Image 20 - Height: 2400, Width: 1800
Image 21 - Height: 2400, Width: 1800
Image 22 - Height: 2400, Width: 1800
Image 23 - Height: 2400, Width: 1800
Image 24 - Height: 2400, Width: 1800
Image 25 - Height: 2400, Width: 1800
```

- Next, I used a loop to rename all the images in numerical order 1, 2 and 3 etc.



- Each time they were put in the new directory clean images

Data processing:

- Now I set 2 variables max_hight and max_width each set to 500
- I also made an array for the mean values of each image
- The loop went through each image checking if any had pixels more than 500 x 500 if so, it would then resize it to 500 x 500 this was done so that rerunning the code would not cause overlaps or errors later I also checked each images mean value putting it into an array and after displaying it.

```
Image 1 - Height after resizing: 500, Width after resizing: 500
Image 2 - Height after resizing: 500, Width after resizing: 500
Image 3 - Height after resizing: 500, Width after resizing: 500
Image 4 - Height after resizing: 500, Width after resizing: 500
Image 5 - Height after resizing: 500, Width after resizing: 500
Image 6 - Height after resizing: 500, Width after resizing: 500
Image 7 - Height after resizing: 500, Width after resizing: 500
Image 8 - Height after resizing: 500, Width after resizing: 500
Image 9 - Height after resizing: 500, Width after resizing: 500
Image 10 - Height after resizing: 500, Width after resizing: 500
Image 11 - Height after resizing: 500, Width after resizing: 500
Image 12 - Height after resizing: 500, Width after resizing: 500
Image 13 - Height after resizing: 500, Width after resizing: 500
Image 14 - Height after resizing: 500, Width after resizing: 500
Image 15 - Height after resizing: 500, Width after resizing: 500
Image 16 - Height after resizing: 500, Width after resizing: 500
Image 17 - Height after resizing: 500, Width after resizing: 500
Image 18 - Height after resizing: 500, Width after resizing: 500
Image 19 - Height after resizing: 500, Width after resizing: 500
Image 20 - Height after resizing: 500, Width after resizing: 500
Image 21 - Height after resizing: 500, Width after resizing: 500
Image 22 - Height after resizing: 500, Width after resizing: 500
Image 23 - Height after resizing: 500, Width after resizing: 500
Image 24 - Height after resizing: 500, Width after resizing: 500
Image 25 - Height after resizing: 500, Width after resizing: 500
```

Also shape features.

```
Mean values for each image:
  251.8016
  218.8171
  217.9299
  233.6263
  191.9460
  160.1829
  143.5120
  224.4637
  161.7851
  179.9867
  160.8488
  210.5133
  180.3521
  207.4922
  228.6995
  222.1497
  238.3529
  191.1321
  190.7979
  234.0736
  132.7800
  235.8026
  172.7351
  206.3784
  183.8045
```

Image annotation:

- I started by looping the images again one by one by reading the image displaying it then finally prompting the user to write keywords and descriptions for each image after manually writing them down they were then stored in the variable's keywords and description.

Small sample:

```
Enter keywords: white skirt emboidary
Enter description: white skirt with emboidary
Enter keywords: jean skirt blue
Enter description: blue skirt jeans
Enter keywords: denim jeans shorts
Enter description: jeans shorts
Enter keywords: fullsleave white sports gym
Enter description: full sleave white shirt for gym and sports
Enter keywords: black shorts sports
Enter description: black shorts for sports
```

- I saved them into a Json file named metadata_images

[{"ImageID":"1","Keywords":["white skirt emboidary"],"Description":"white skirt with emboidary"

Feature extraction:

I already did shape features before (height and width)

Next, I started with colour features. I was quite confused if mean(blue), (mean)green and (mean)red count as 3 colour features or mean, std and skewness as 3 colour features for each RGB.

I decided to go with the second option and did mean and std and stored it into a file named feature_extraction_colour using a loop.

[{"file":"1.jpg","red":{"mean":251.17441643518518,"std":8.9758269303908484},

Next, I moved on to texture features but none of the functions worked so I copied the images, changed them to grayscale, moved them to the file named clean_gray_images then tried again but it still did not work.

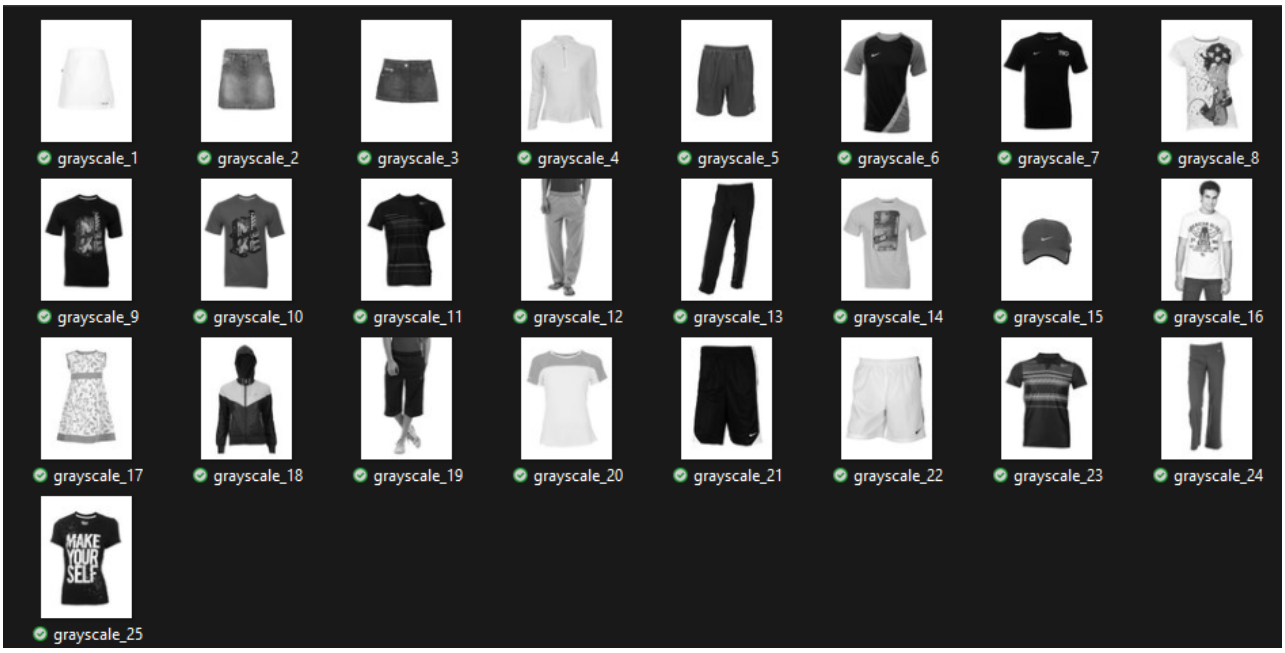


Image Database: N/A

Part 2 Sentiment analysis model:

Data collection:

https://www.kaggle.com/datasets/muhammedabdulazeem/employer-review-about-their-organization

- I picked this review-based Json file because opinions about work can be different despite people being in the same conditions.
- I started by picking 50 random rows in the Json to analyse:

50x1 struct with 5 fields

Fields	ReviewTitle	CompleteReview	URL	Rating	ReviewDetails
1	'Quality assur...	'No compromise in ...	'https://in.i...	'5.0'	'(Former Employ...
2	'Great place to...	'I have had a wonde...	'https://in.i...	'5.0'	'(Current Emplo...
3	'good for fres...	'management impr...	'https://in.i...	'3.0'	'(Former Employ...
4	'Excellent envi...	'Good work culture ...	'https://in.i...	'5.0'	'(Current Emplo...
5	'Not a healthy...	'Very hard to work i...	'https://in.i...	'3.0'	'(Current Emplo...
6	'Decent corpo...	'They have good pr...	'https://in.i...	'4.0'	'(Former Employ...
7	'job culture is ...	'while working with ...	'https://in.i...	'4.0'	'(Former Employ...
8	'Working in v...	'JOB RESPONSIBLITI...	'https://in.i...	'5.0'	'(Former Employ...
9	'Good environ...	'Our day started wit...	'https://in.i...	'4.0'	'(Former Employ...
10	'Product Deve...	'Was working Larse...	'https://in.i...	'4.0'	'(Former Employ...
11	'Productive an...	'Good Place to work...	'https://in.i...	'5.0'	'(Current Emplo...
12	'Productive'	'My project was go...	'https://in.i...	'4.0'	'(Former Employ...
13	'Professional c...	'Good company wit...	'https://in.i...	'5.0'	'(Former Employ...
14	'Awesome ex...	'Nice place to work ...	'https://in.i...	'4.0'	'(Current Emplo...
15	'Good work lif...	'Struggle at the start...	'https://in.i...	'4.0'	'(Former Employ...
16	'Good Salary'	'It is a good place to...	'https://in.i...	'4.0'	'(Former Employ...
17	'Productive'	'Its a nice job for ac...	'https://in.i...	'5.0'	'(Current Emplo...
18	'Good'	'Good place to work...	'https://in.i...	'4.0'	'(Current Emplo...
19	'Awesome Wo...	'Awesome WorkExp...	'https://in.i...	'5.0'	'(Current Emplo...
20	'Best work life ...	'Bosch is a compan...	'https://in.i...	'4.0'	'(Current Emplo...
21	'Good Compa...	'Well received Hr''s ...	'https://in.i...	'4.0'	'(Current Emplo...
22	'Productive an...	'Opportunities are a...	'https://in.i...	'4.0'	'(Current Emplo...

Data preprocessing:

- I first started by separating the review column

50x1 cell

	1	2
1	No compromise in...	
2	I have had a wond...	
3	management impr...	
4	Good work culture...	
5	Very hard to work i...	
6	They have good pr...	
7	while working with...	
8	JOB RESPONSIBLIT...	
9	Our day started wit...	
10	Was working Larse...	
11	Good Place to wor...	
12	My project was go...	
13	Good company wi...	
14	Nice place to work...	
15	Struggle at the star...	
16	It is a good place t...	
17	Its a nice job for ac...	
18	Good place to wor...	
19	Awesome WorkEx...	
20	Bosch is a compan...	
21	Well received Hr's ...	
22	Opportunities are ...	

- Then I lowercased all the letters

50x1 cell

	1
1	no compromise ...
2	i have had a wo...
3	management i...
4	good work cultu...
5	very hard to wor...
6	they have good ...
7	while working w...
8	job responsibliti...
9	our day started ...
10	was working lars...

- Next, I removed all the punctuation from the reviews using erasePunctuatiuon but the function did not work where then I had to install text analysis toolbox after it worked.

50x1 cell

	1
1	no compro...
2	i have had a...
3	manageme...
4	good work ...
5	very hard to...
6	they have g...
7	while worki...
8	job respons...
9	our day star...
10	was workin...

- Next, I tokenized the text and finally used stopWords which I also removed from the reviews

50x1 tokenizedDocument

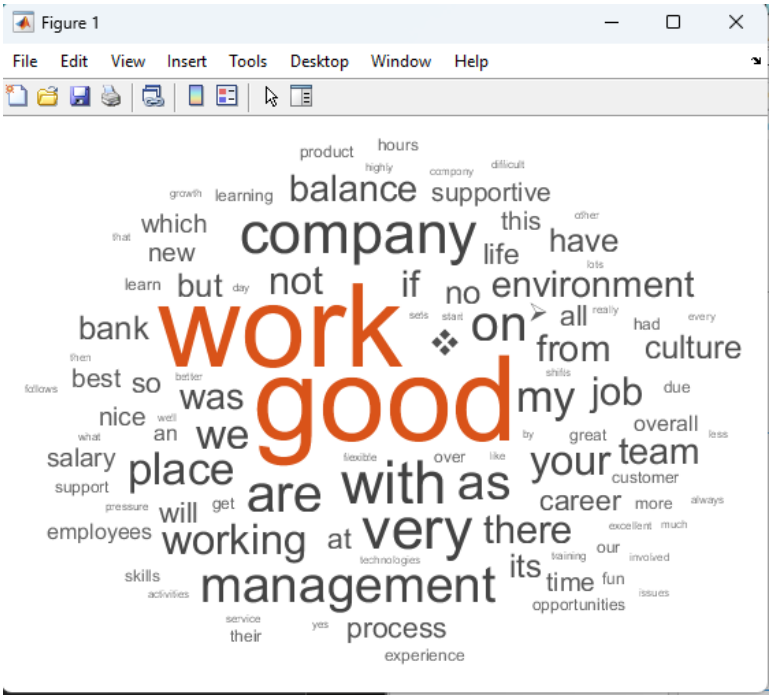
	1
1	1x1 tokenizedDocument
2	1x1 tokenizedDocument
3	1x1 tokenizedDocument
4	1x1 tokenizedDocument
5	1x1 tokenizedDocument
6	1x1 tokenizedDocument
7	1x1 tokenizedDocument
8	1x1 tokenizedDocument
9	1x1 tokenizedDocument
10	1x1 tokenizedDocument
11	1x1 tokenizedDocument
12	1x1 tokenizedDocument
13	1x1 tokenizedDocument
14	1x1 tokenizedDocument
15	1x1 tokenizedDocument
16	1x1 tokenizedDocument
17	1x1 tokenizedDocument
18	1x1 tokenizedDocument

filtered_Tokens(1, 1).Vocabulary

	1	2	3	4
1	no	compromise	quality	at

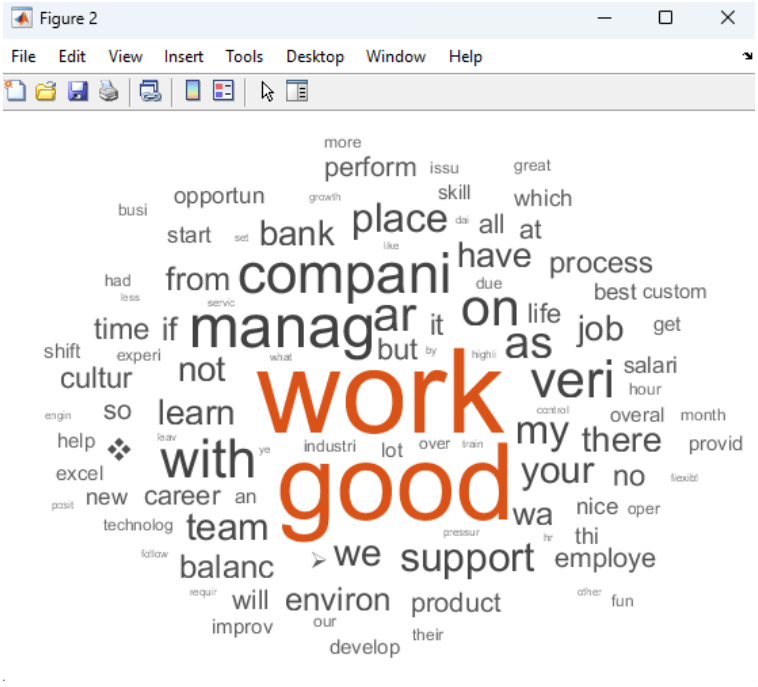
Text vectorization:

- I first started with bag of words by just check the cleaning text

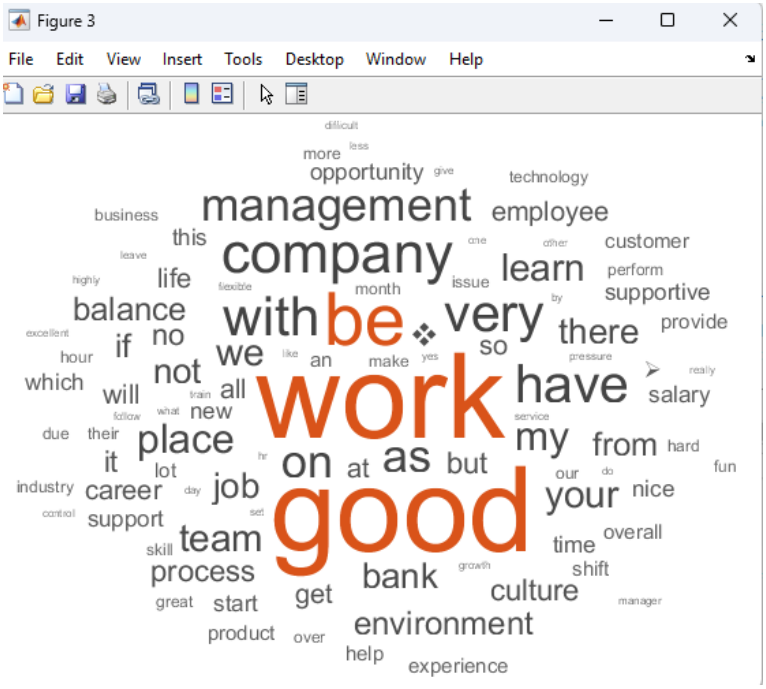


- Then I lemmatized and used stemming on the clean text and made those into bag of words as well

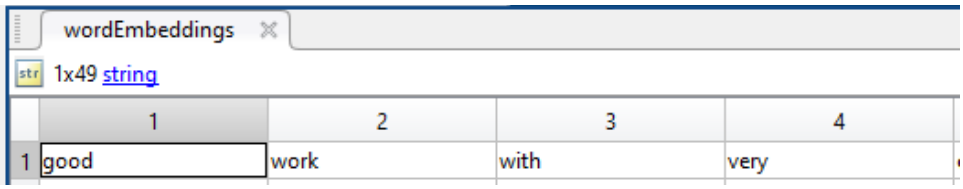
Stemming:



Lemmatisation:



- Finally, I used an already trained word embedder



Metadata and labeling:

- First, I made a Json file with the clean text and another column called sentiment where I then manually read and put a sentiment for all 50 reviews:

[{"Review":"No compromise in quality at Mahindra Ltd. Completely focused on safety first.. Managers and colleuge are very supportive. Got the opportunity to learn what is ISO & CERTIFICATIONS.","Sentiment":[Positive]},{"Review":"I have had a wonderful 3 year plus journey with JLL. The management is very open to innovation and new ideas. The work culture is super flexible and accommodating.","Sentiment":[Positive]},{"Review":"management improvement needed. lack of appreciation from top level management.no time management. negligence towards new staff training. bonus not given on time.","Sentiment":[Negative]},{"Review":"Good work culture\r \r Good environment \r \r No work pressure \r \r Work environment is very nice, start career in best industry in Hp \r Good salary for the region\r Employee fiiendly region\r Flexible work life balance\r Benfits are good","Sentiment":[Positive]},{"Review":"Very hard to work in this bank, very lengthy process and very slow system. Management not supportive, backed process is very slow and not supportive.Very bad workplace culture. Not any enjoyable part of the job.","Sentiment":[Negative]},{"Review":"They have good problem solving approaches by sitting in a board room and having a group discussion. The team is highly skilled and quick on their feet.","Sentiment":[Positive]},{"Review":"while working with yes bank was good ,\r Work culture is nice . \r We get good support from the top management in regards to job profile , your score card in regards to service and sales parameters.\r HR support is very good in yes bank i.e post hiring doc, hr seasons for grooming of job profile.\r \r \r \r \r \r Reason for leaving is relocation as yes bank is having limited options in my home town .","Sentiment":[Neutral]},{"Review":"JOB RESPONSIBLITIES :-\r \r 24hr men power controlling.\r \r Operating overall diesel generator.\r \r Maintenance and service diesel engines sets. \r \r Overall automatic control panel repair DG sets. \r \r DG sets, alternators repair and maintenance.\r \r Monthly report smites in department.","Sentiment":[Neutral]},{"Review":"Our day started with stand up meeting with our scrum team. Nokia follows Agile way of working. We update what we did yesterday and what work we are going to do today to the scrum team.\r \r We pick up the task from Kanban board and start working.\r \r Lot of importance to team work is given in Nokia.","Sentiment":[Neutral]},{"Review":"Was working Larsen & Toubro As Product Development Engineer.\r \r Was working on new products as per customer requirements for customers like steel industries, thermal power plants, cement industries","Sentiment":[Neutral]},

- Then I made another metadata file where I wrote the process for text vectorization

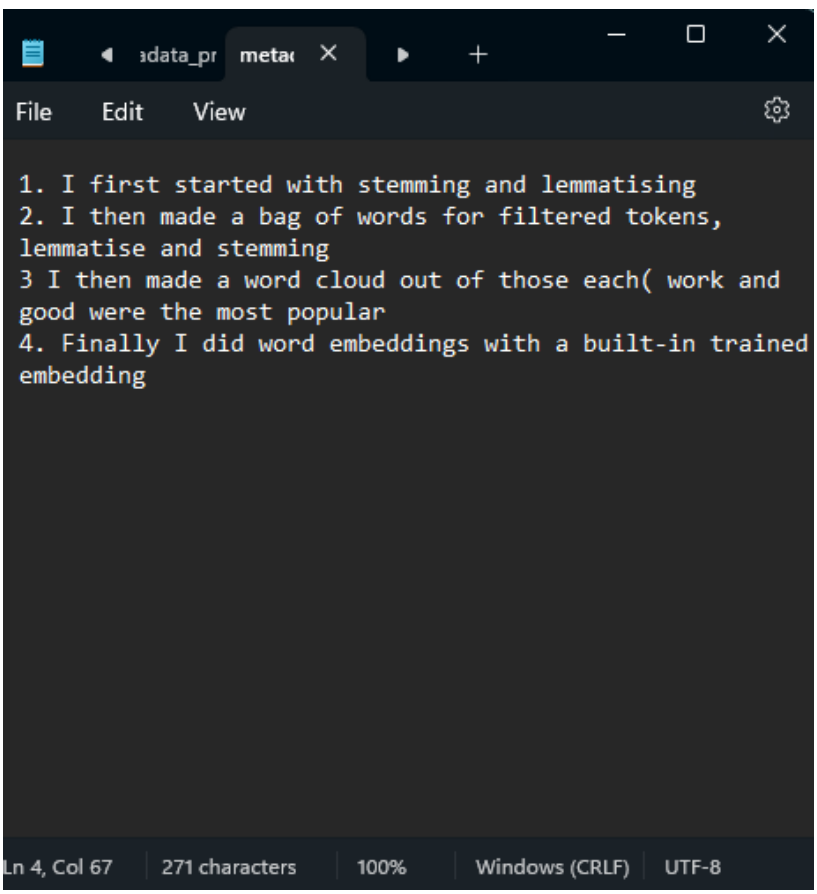


Image database: N/A