1. Describe all columns in the data:

‘dataset\_id’: Which dataset to use. ‘dvqa\_train’ means the dvqa training dataset has been used

‘image\_id’: Image id associated to the question

‘passage’: Facts provided along with the image

‘question’: Question for answer

‘answer’: Answer to the question

1. Task assigned:

Using “Add” method to generate new data samples in dvqa dataset

The tasks involved:

1. Familiar with the dataset
2. Design the following temple for the data generation (utilizing the nltk and gensim to find similar words)

Templates: (*red* lines indicate the generated text)

<a> Metadata\_table:

len(metadata[‘table’]) == 2:

len(metadata[‘table’]) > 2:

<b> Facts:

if a1:

*<metadata> is <rand number> times <larger> than <similar word>*

(1f) numerical expression: similar word(value) =1/n \* metadata(value)

*<metadata> is <rand number> times <smaller> than <similar word>*

(2f) numerical expression: similar word(value) = n \* metadata(value)

*<metadata> is <rand[metadata(label,value),value\_ticks(lo,hi)]> <greater> than <similar word>*

(3f) numerical expression: similar word(value) = metadata(value) - n

*<metadata> is <rand[metadata(label,value),value\_ticks(lo,hi)]> <smaller> than <similar word>*

(4f) numerical expression: similar word(value) = metadata(value) + n

(If len(metadata) >=2 :) *<similar word> is equals to the value of <metadata\_1> and <metadata\_2> combined*

(5f) numerical expression: similar word(value) = metadata\_1(value) + metadata\_2(value)

Else:

*<metadata> is <rand number> times <larger> than <similar word> in <table[1][0]>*

(1f) numerical expression: similar word(value) =1/n \* metadata(value)

*<metadata> is <rand number> times <smaller> than <similar word> in <table[1][0]>*

(2f) numerical expression: similar word(value) = n \* metadata(value)

*<metadata> is <rand[metadata(label,value),value\_ticks(lo,hi)]> <greater> than <similar word> in <table[1][0]>*

(3f) numerical expression: similar word(value) = metadata(value) - n

*<metadata> is <rand[metadata(label,value),value\_ticks(lo,hi)]> <smaller> than <similar word> in <table[1][0]>*

(4f) numerical expression: similar word(value) = metadata(value) + n

(If len(metadata) >=2 :) *<similar word> is equal to the value of <metadata\_1> and <metadata\_2> combined in <table[1][0]>*

(5f) numerical expression: similar word(value) = metadata\_1(value) + metadata\_2(value)

<c> Question:

{\*} only applies if len(metadata[‘table’]) > 2

*What’s the value of <similar word> {in <table[1][0]>}?*

(If len(metadata) >=2 and Facts are type(1) or type(2):) *Is the value of <similar word> <greater> than <metadata\_2> {in <table[1][0]>}?*

(If len(metadata) >=2 and Facts are type(1) or type(2):) *Is the value of <similar word> <smaller> than <metadata\_2> {in <table[1][0]>}?*

*How many bars have value <greater> than rand(value\_ticks(lo,hi)) {in <table[1][0]>}?*

How many bars have value <smaller> than rand(value\_ticks(lo,hi)) {in <table[1][0]>}?

Which bar has the greatest value {in <table[1][0]>}?

Which bar has the smallest value {in <table[1][0]>}?

What is the sum of all values {in <table[1][0]>}?

What is the difference between the largest and the smallest value of {<table[1][0]} in the chart?

<d> Answer:

similar\_word(value), metadata(value), value\_ticks, rand(value\_ticks(lo,hi))

c(1):

(1a) similar\_word(value)

c(2):

(2a) if similar\_word(value) > metadata\_2(value): Yes

else: No

c(3):

(3a) if similar\_word(value) < metadata\_2(value): Yes

else: No

c(4):

(4a) count([value\_ticks,similar\_word(value)] > rand(value\_ticks(lo,hi))

c(5):

(5a) count([value\_ticks,similar\_word(value)] < rand(value\_ticks(lo,hi))

c(6):

(6a) max(value\_ticks, similar\_word(value))

c(7):

(7a) min(value\_ticks, similar\_word(value))

c(8):xm

(8a) sum(value\_ticks,similar\_word(value))

c(9):

(9a) max(value\_ticks, similar\_word(value)) - min(value\_ticks, similar\_word(value))

1. Implement the algorithm to generate 500 new data points
2. How to run the code

Two packages are needed for the program:

**nltk** and **gensim**

**!python -m pip install --user -U nltk**

**!python -m pip install -U gensim**

After installing the package, the program can run as:

python dvqa\_add\_script.py