ECE 60146 (Homework 11)

Kinjol Barua

Email: kbarua@purdue.edu

Introduction:

In this homework, we will compare the BERT tokenizer with the tokenizer that comes bundled with babyGPT.

Programming Tasks:

4.1 Word-level tokenization:

The word level tokenization was performed both on the given 4 sentences and 5 custom sentences.

Word-level tokenization performed on given 4 sentences:

Input sentence 1:

The GeoSolutions technology will leverage Benefon 's GPS solutions by providing Location Based Search Technology, a Communities Platform, location relevant multimedia content and a new and powerful commercial model .

Word-level tokenization of Input sentence 1:

['The', 'GeoSolutions', 'technology', 'will', 'leverage', 'Benefon', "'s", 'GPS', 'solutions', 'by', 'providing', 'Location', 'Based', 'Search', 'Technology', ',', 'a', 'Communities', 'Platform', ',', 'location', 'relevant', 'multimedia', 'content', 'and', 'a', 'new', 'and', 'powerful', 'commercial', 'model', '.']

Input sentence 2:

$ESI on lows, down $1.50 to $2.50 BK a real possibility

Word-level tokenization of Input sentence 2:

['$ESI', 'on', 'lows,', 'down', '$1.50', 'to', '$2.50', 'BK', 'a', 'real', 'possibility']

Input sentence 3:

For the last quarter of 2010, Componenta 's net sales doubled to EUR131m from EUR76m for the same period a year earlier, while it moved to a zero pre-tax profit from a pre-tax loss of EUR7m .

Word-level tokenization of Input sentence 3:

['For', 'the', 'last', 'quarter', 'of', '2010', ',', 'Componenta', "'s", 'net', 'sales', 'doubled', 'to', 'EUR131m', 'from', 'EUR76m', 'for', 'the', 'same', 'period', 'a', 'year', 'earlier', ',', 'while', 'it', 'moved', 'to', 'a', 'zero', 'pre-tax', 'profit', 'from', 'a', 'pre-tax', 'loss', 'of', 'EUR7m', '.']

Input sentence 4:

According to the Finnish-Russian Chamber of Commerce, all the major construction companies of Finland are operating in Russia .

Word-level tokenization of Input sentence 4:

['According', 'to', 'the', 'Finnish-Russian', 'Chamber', 'of', 'Commerce', ',', 'all', 'the', 'major', 'construction', 'companies', 'of', 'Finland', 'are', 'operating', 'in', 'Russia', '.']

Word-level tokenization performed on custom 5 sentences:

Custom input sentence 1:

Fibroblasts, once perceived as a uniform cell type, are now recognized as a mosaic of distinct populations with specialized roles in tissue homeostasis and pathology. Here we provide a global overview of the expanding compendium of fibroblast cell types and states, their diverse lineage origins and multifaceted functions across various human organs.

Word-level tokenization of custom input sentence 1:

['Fibroblasts,', 'once', 'perceived', 'as', 'a', 'uniform', 'cell', 'type,', 'are', 'now', 'recognized', 'as', 'a', 'mosaic', 'of', 'distinct', 'populations', 'with', 'specialized', 'roles', 'in', 'tissue', 'homeostasis', 'and', 'pathology.', 'Here', 'we', 'provide', 'a', 'global', 'overview', 'of', 'the', 'expanding', 'compendium', 'of', 'fibroblast', 'cell', 'types', 'and', 'states,', 'their', 'diverse', 'lineage', 'origins', 'and', 'multifaceted', 'functions', 'across', 'various', 'human', 'organs.']

Custom input sentence 2:

Modern imaging technologies are widely based on classical principles of light or electromagnetic wave propagation. They can be remarkably sophisticated, with recent successes ranging from single-molecule microscopy to imaging far-distant galaxies.

Word-level tokenization of custom input sentence 2:

['Modern', 'imaging', 'technologies', 'are', 'widely', 'based', 'on', 'classical', 'principles', 'of', 'light', 'or', 'electromagnetic', 'wave', 'propagation.', 'They', 'can', 'be', 'remarkably', 'sophisticated,', 'with', 'recent', 'successes', 'ranging', 'from', 'single-molecule', 'microscopy', 'to', 'imaging', 'far-distant', 'galaxies.']

Custom input sentence 3:

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically *improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics.*

Word-level tokenization of custom input sentence 3:

['Deep', 'learning', 'allows', 'computational', 'models', 'that', 'are', 'composed', 'of', 'multiple', 'processing', 'layers', 'to', 'learn', 'representations', 'of', 'data', 'with', 'multiple', 'levels', 'of', 'abstraction.', 'These', 'methods', 'have', 'dramatically', 'improved', 'the', 'state-of-the-art', 'in', 'speech', 'recognition,', 'visual', 'object', 'recognition,', 'object', 'detection', 'and', 'many', 'other', 'domains', 'such', 'as', 'drug', 'discovery', 'and', 'genomics.']

Custom input sentence 4:

The rotation period of Uranus was estimated to be 17.24?h in 1986 from radio auroral measurements during the brief Voyager 2 flyby. This value is the basis for the Uranian SIII longitude system still in use. However, the poor period uncertainty limited its validity to a few years, after which the orientation of the magnetic axis was lost.

Word-level tokenization of custom input sentence 4:

['The', 'rotation', 'period', 'of', 'Uranus', 'was', 'estimated', 'to', 'be', '17.24?h', 'in', '1986', 'from', 'radio', 'auroral', 'measurements', 'during', 'the', 'brief', 'Voyager', '2', 'flyby.', 'This', 'value', 'is', 'the', 'basis', 'for', 'the', 'Uranian', 'SIII', 'longitude', 'system', 'still', 'in', 'use.', 'However,', 'the', 'poor', 'period', 'uncertainty', 'limited', 'its', 'validity', 'to', 'a', 'few', 'years,', 'after', 'which', 'the', 'orientation', 'of', 'the', 'magnetic', 'axis', 'was', 'lost.']

Custom input sentence 5:

Fuelled by increasing computer power and algorithmic advances, machine learning techniques have become powerful tools for finding patterns in data. Quantum systems produce atypical patterns that classical systems are thought not to produce efficiently, so it is reasonable to postulate that quantum computers may outperform classical computers on machine learning tasks.

Word-level tokenization of custom input sentence 5:

['Fuelled', 'by', 'increasing', 'computer', 'power', 'and', 'algorithmic', 'advances,', 'machine', 'learning', 'techniques', 'have', 'become', 'powerful', 'tools', 'for', 'finding', 'patterns', 'in', 'data.', 'Quantum', 'systems', 'produce', 'atypical', 'patterns', 'that', 'classical', 'systems', 'are', 'thought', 'not', 'to', 'produce', 'efficiently,', 'so', 'it', 'is', 'reasonable', 'to', 'postulate', 'that', 'quantum', 'computers', 'may', 'outperform', 'classical', 'computers', 'on', 'machine', 'learning', 'tasks.']

4.2 Sub-word-level tokenization:

The sub-word level tokenization was performed both on the given 4 sentences and 5 custom sentences.

Sub-Word-level tokenization performed on given 4 sentences:

Input sentence 1:

The GeoSolutions technology will leverage Benefon 's GPS solutions by providing Location Based Search Technology, a Communities Platform, location relevant multimedia content and a new and powerful commercial model .

Sub-word-level tokenization of Input sentence 1:

['[CLS]', 'the', 'geo', '##sol', '##ution', '##s', 'technology', 'will', 'leverage', 'ben', '##ef', '##on', "'", 's', 'gps', 'solutions', 'by', 'providing', 'location', 'based', 'search', 'technology', ',', 'a', 'communities', 'platform', ',', 'location', 'relevant', 'multimedia', 'content', 'and', 'a', 'new', 'and', 'powerful', 'commercial', 'model', '[SEP]']

Input sentence 2:

$ESI on lows, down $1.50 to $2.50 BK a real possibility

Sub-word-level tokenization of Input sentence 2:

['[CLS]', '$', 'es', '##i', 'on', 'low', '##s', ',', 'down', '$', '1', '.', '50', 'to', '$', '2', '.', '50', 'bk', 'a', 'real', 'possibility', '[SEP]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]']

Input sentence 3:

For the last quarter of 2010, Componenta 's net sales doubled to EUR131m from EUR76m for the same period a year earlier, while it moved to a zero pre-tax profit from a pre-tax loss of EUR7m .

Sub-word-level tokenization of Input sentence 3:

['[CLS]', 'for', 'the', 'last', 'quarter', 'of', '2010', ',', 'component', '##a', "'", 's', 'net', 'sales', 'doubled', 'to', 'eu', '##r', '##13', '##1', '##m', 'from', 'eu', '##r', '##7', '##6', '##m', 'for', 'the', 'same', 'period', 'a', 'year', 'earlier', ',', 'while', 'it', 'moved', '[SEP]']

Input sentence 4:

According to the Finnish-Russian Chamber of Commerce, all the major construction companies of Finland are operating in Russia .

Sub-word-level tokenization of Input sentence 4:

```
['[CLS]', 'according', 'to', 'the', 'finnish', '-', 'russian', 'chamber', 'of', 'commerce', ',', 'all', 'the',
'major', 'construction', 'companies', 'of', 'finland', 'are', 'operating', 'in', 'russia', '.', '[SEP]',
'[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]',
'[PAD]', '[PAD]', '[PAD]', '[PAD]']
```

Sub-Word-level tokenization performed on custom 5 sentences:

Custom input sentence 1:

Fibroblasts, once perceived as a uniform cell type, are now recognized as a mosaic of distinct populations with specialized roles in tissue homeostasis and pathology. Here we provide a global overview of the expanding compendium of fibroblast cell types and states, their diverse lineage origins and multifaceted functions across various human organs.

Sub-word-level tokenization of custom input sentence 1:

```
['[CLS]', 'fi', '##bro', '##bla', '##sts', ',', 'once', 'perceived', 'as', 'a', 'uniform', 'cell', 'type', ',', 'are',
'now', 'recognized', 'as', 'a', 'mosaic', 'of', 'distinct', 'populations', 'with', 'specialized', 'roles', 'in',
'tissue', 'home', '##osta', '##sis', 'and', 'pathology', '.', 'here', 'we', 'provide', 'a', 'global', 'overview',
'of', 'the', 'expanding', 'com', '##pen', '##dium', 'of', 'fi', '##bro', '##bla', '##st', 'cell', 'types', 'and',
'states', ',', 'their', '[SEP]']
```

Custom input sentence 2:

Modern imaging technologies are widely based on classical principles of light or electromagnetic wave propagation. They can be remarkably sophisticated, with recent successes ranging from single-molecule microscopy to imaging far-distant galaxies.

Sub-word-level tokenization of custom input sentence 2:

```
['[CLS]', 'modern', 'imaging', 'technologies', 'are', 'widely', 'based', 'on', 'classical', 'principles',
'of', 'light', 'or', 'electromagnetic', 'wave', 'propagation', '.', 'they', 'can', 'be', 'remarkably',
'sophisticated', ',', 'with', 'recent', 'successes', 'ranging', 'from', 'single', '-', 'molecule', 'microscopy',
'to', 'imaging', 'far', '-', 'distant', 'galaxies', '.', '[SEP]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]',
'[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]', '[PAD]',
'[PAD]', '[PAD]']
```

Custom input sentence 3:

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics.

Sub-word-level tokenization of custom input sentence 3:

['[CLS]', 'deep', 'learning', 'allows', 'computational', 'models', 'that', 'are', 'composed', 'of', 'multiple', 'processing', 'layers', 'to', 'learn', 'representations', 'of', 'data', 'with', 'multiple', 'levels', 'of', 'abstraction', '.', 'these', 'methods', 'have', 'dramatically', 'improved', 'the', 'state', '-', 'of', '-', 'the', '-', 'art', 'in', 'speech', 'recognition', ',', 'visual', 'object', 'recognition', ',', 'object', 'detection', 'and', 'many', 'other', 'domains', 'such', 'as', 'drug', 'discovery', 'and', 'gen', '[SEP]']

Custom input sentence 4:

The rotation period of Uranus was estimated to be 17.24?h in 1986 from radio auroral measurements during the brief Voyager 2 flyby. This value is the basis for the Uranian SIII longitude system still in use. However, the poor period uncertainty limited its validity to a few years, after which the orientation of the magnetic axis was lost.

Sub-word-level tokenization of custom input sentence 4:

['[CLS]', 'the', 'rotation', 'period', 'of', 'ur', '##anus', 'was', 'estimated', 'to', 'be', '17', '.', '24', '?', 'h', 'in', '1986', 'from', 'radio', 'aurora', '##l', 'measurements', 'during', 'the', 'brief', 'voyager', '2', 'fly', '##by', '.', 'this', 'value', 'is', 'the', 'basis', 'for', 'the', 'ur', '##anian', 'si', '##ii', 'longitude', 'system', 'still', 'in', 'use', '.', 'however', ',', 'the', 'poor', 'period', 'uncertainty', 'limited', 'its', 'validity', '[SEP]']

Custom input sentence 5:

Fuelled by increasing computer power and algorithmic advances, machine learning techniques have become powerful tools for finding patterns in data. Quantum systems produce atypical patterns that classical systems are thought not to produce efficiently, so it is reasonable to postulate that quantum computers may outperform classical computers on machine learning tasks.

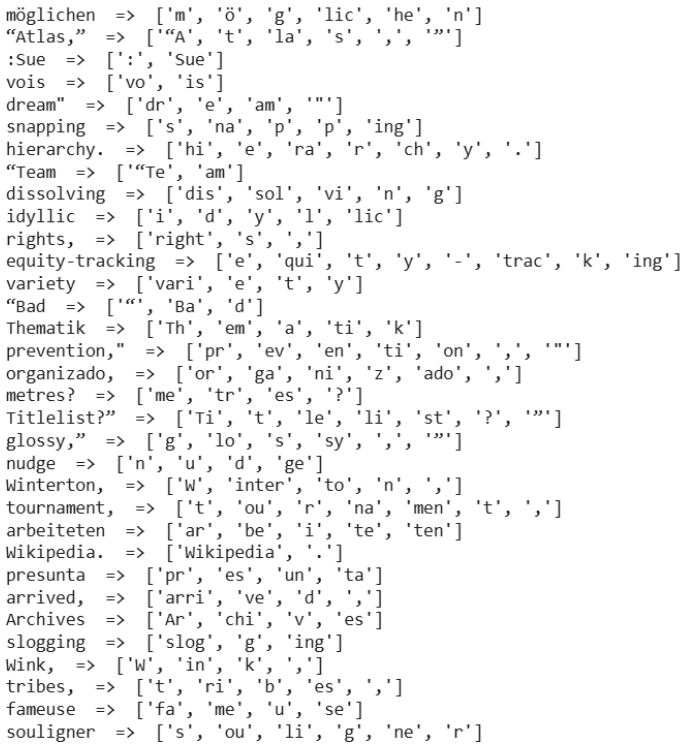
Sub-word-level tokenization of custom input sentence 5:

['[CLS]', 'fuel', '##led', 'by', 'increasing', 'computer', 'power', 'and', 'algorithm', '##ic', 'advances', ',', 'machine', 'learning', 'techniques', 'have', 'become', 'powerful', 'tools', 'for', 'finding', 'patterns', 'in', 'data', '.', 'quantum', 'systems', 'produce', 'at', '##yp', '##ical', 'patterns', 'that', 'classical', 'systems', 'are', 'thought', 'not', 'to', 'produce', 'efficiently', ',', 'so', 'it', 'is', 'reasonable', 'to', 'post', '##ulate', 'that', 'quantum', 'computers', 'may', 'out', '##per', '##form', 'classical', '[SEP]']

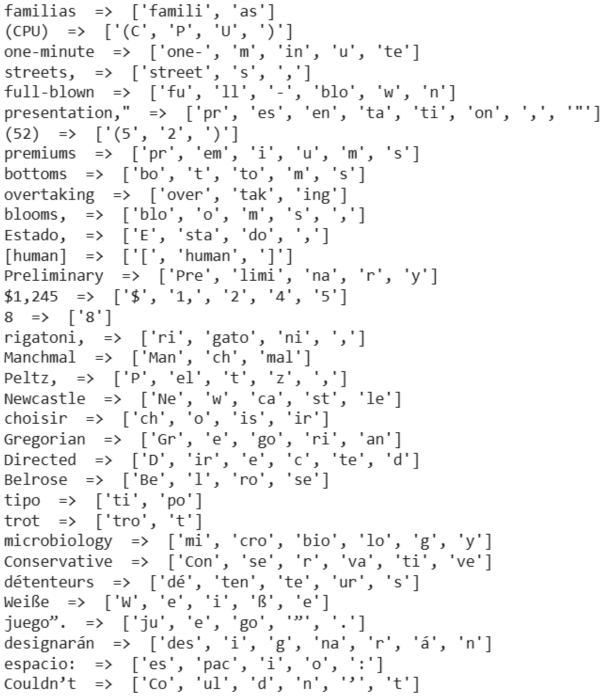
4.3 baby-GPT tokenizer

To train babyGPT tokenizer, I have used the text files in the directory “saved_articles_dir_12M”. The size of the vocabulary I used is 10,000. Here are some screenshots during the training process

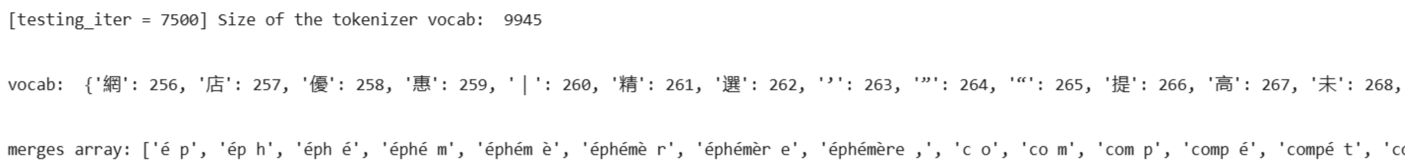
Training Process:



```
möglichen  =>  ['m', 'ö', 'g', 'lic', 'he', 'n']
“Atlas,”  =>  ['“A', 't', 'la', 's', ',', '”']
:Sue  =>  [':', 'Sue']
vois  =>  ['vo', 'is']
dream"  =>  ['dr', 'e', 'am', '"']
snapping  =>  ['s', 'na', 'p', 'p', 'ing']
hierarchy.  =>  ['hi', 'e', 'ra', 'r', 'ch', 'y', '.']
“Team  =>  ['“Te', 'am']
dissolving  =>  ['dis', 'sol', 'vi', 'n', 'g']
idyllic  =>  ['i', 'd', 'y', 'l', 'lic']
rights,  =>  ['right', 's', ',']
equity-tracking  =>  ['e', 'qui', 't', 'y', '-', 'trac', 'k', 'ing']
variety  =>  ['vari', 'e', 't', 'y']
“Bad  =>  ['“', 'Ba', 'd']
Thematik  =>  ['Th', 'em', 'a', 'ti', 'k']
prevention,"  =>  ['pr', 'ev', 'en', 'ti', 'on', ',', '"']
organizado,  =>  ['or', 'ga', 'ni', 'z', 'ado', ',']
metres?  =>  ['me', 'tr', 'es', '?']
Titlelist?”  =>  ['Ti', 't', 'le', 'li', 'st', '?', '”']
glossy,”  =>  ['g', 'lo', 's', 'sy', ',', '”']
nudge  =>  ['n', 'u', 'd', 'ge']
Winterton,  =>  ['W', 'inter', 'to', 'n', ',']
tournament,  =>  ['t', 'ou', 'r', 'na', 'men', 't', ',']
arbeiteten  =>  ['ar', 'be', 'i', 'te', 'ten']
Wikipedia.  =>  ['Wikipedia', '.']
presunta  =>  ['pr', 'es', 'un', 'ta']
arrived,  =>  ['arri', 've', 'd', ',']
Archives  =>  ['Ar', 'chi', 'v', 'es']
slogging  =>  ['slog', 'g', 'ing']
Wink,  =>  ['W', 'in', 'k', ',']
tribes,  =>  ['t', 'ri', 'b', 'es', ',']
fameuse  =>  ['fa', 'me', 'u', 'se']
souligner  =>  ['s', 'ou', 'li', 'g', 'ne', 'r']
```



```
familias  =>  ['famili', 'as']
(CPU)  =>  ['(C', 'P', 'U', ')']
one-minute  =>  ['one-', 'm', 'in', 'u', 'te']
streets,  =>  ['street', 's', ',']
full-blown  =>  ['fu', 'll', '-', 'blo', 'w', 'n']
presentation,"  =>  ['pr', 'es', 'en', 'ta', 'ti', 'on', ',', '"']
(52)  =>  ['(5', '2', ')']
premiums  =>  ['pr', 'em', 'i', 'u', 'm', 's']
bottoms  =>  ['bo', 't', 'to', 'm', 's']
overtaking  =>  ['over', 'tak', 'ing']
blooms,  =>  ['blo', 'o', 'm', 's', ',']
Estado,  =>  ['E', 'sta', 'do', ',']
[human]  =>  ['[', 'human', ']']
Preliminary  =>  ['Pre', 'limi', 'na', 'r', 'y']
$1,245  =>  ['$', '1,', '2', '4', '5']
8  =>  ['8']
rigatoni,  =>  ['ri', 'gato', 'ni', ',']
Manchmal  =>  ['Man', 'ch', 'mal']
Peltz,  =>  ['P', 'el', 't', 'z', ',']
Newcastle  =>  ['Ne', 'w', 'ca', 'st', 'le']
choisir  =>  ['ch', 'o', 'is', 'ir']
Gregorian  =>  ['Gr', 'e', 'go', 'ri', 'an']
Directed  =>  ['D', 'ir', 'e', 'c', 'te', 'd']
Belrose  =>  ['Be', 'l', 'ro', 'se']
tipo  =>  ['ti', 'po']
trot  =>  ['tro', 't']
microbiology  =>  ['mi', 'cro', 'bio', 'lo', 'g', 'y']
Conservative  =>  ['Con', 'se', 'r', 'va', 'ti', 've']
détenteurs  =>  ['dé', 'ten', 'te', 'ur', 's']
Weiße  =>  ['W', 'e', 'i', 'ß', 'e']
juego”.  =>  ['ju', 'e', 'go', '”', '.']
designarán  =>  ['des', 'i', 'g', 'na', 'r', 'á', 'n']
espacio:  =>  ['es', 'pac', 'i', 'o', ':']
Couldn’t  =>  ['Co', 'ul', 'd', 'n', '’', 't']
```



```
[testing_iter = 7500] Size of the tokenizer vocab:  9945

vocab:  {'網': 256, '店': 257, '優': 258, '惠': 259, '｜': 260, '精': 261, '選': 262, '’': 263, '”': 264, '“': 265, '提': 266, '高': 267, '末': 268,

merges array: ['é p', 'ép h', 'éph é', 'éphé m', 'éphém è', 'éphémè r', 'éphémèr e', 'éphémère ,', 'c o', 'co m', 'com p', 'comp é', 'compé t', 'c
```

babyGPT tokenization performed on given 4 sentences:

Input sentence 1:

The GeoSolutions technology will leverage Benefon 's GPS solutions by providing Location Based Search Technology, a Communities Platform, location relevant multimedia content and a new and powerful commercial model .

babyGPT tokenization of Input sentence 1:

The Ge o S ol uti on s tech no lo gy will l ever age Ben ef on ' s GPS solution s by provi ding Lo ca ti on Base d Se ar ch Te ch no lo gy , a Com m unit i es Pla t for m , lo ca ti on relev ant multi medi a content an d a new an d power ful commercial model .

Input sentence 2:

$ESI on lows, down $1.50 to $2.50 BK a real possibility

babyGPT tokenization of Input sentence 2:

$ E SI on low s , down $ 1 . 50 to $ 2 . 50 B K a real possi bi lit y

Input sentence 3:

For the last quarter of 2010, Componenta 's net sales doubled to EUR131m from EUR76m for the same period a year earlier, while it moved to a zero pre-tax profit from a pre-tax loss of EUR7m .

babyGPT tokenization of Input sentence 3:

For the last quarter of 2010 , Com pon en ta ' s net sale s dou ble d to EU R 13 1m from EU R 76 m for the sa me peri od a year earlier , whi le it mov ed to a zero pre - tax pro fit from a pre - tax loss of EU R 7 m .

Input sentence 4:

According to the Finnish-Russian Chamber of Commerce, all the major construction companies of Finland are operating in Russia .

babyGPT tokenization of Input sentence 4:

Accor ding to the Fin ni sh - Russian Cham be r of Com merce , all the major construction compani es of Fin land are operat ing in Russia .

babyGPT tokenization performed on custom 5 sentences:

Custom input sentence 1:

Fibroblasts, once perceived as a uniform cell type, are now recognized as a mosaic of distinct populations with specialized roles in tissue homeostasis and pathology. Here we provide a global overview of the expanding compendium of fibroblast cell types and states, their diverse lineage origins and multifaceted functions across various human organs.

babyGPT tokenization of custom input sentence 1:

Fi bro bla st s , once per ce iv ed as a uniform cel l ty pe , are now re co g ni ze d as a mo sai c of dist inc t population s with specialize d role s in t issue home o sta si s an d pa tho lo gy . He re we provide a glo bal over vie w of the exp an ding com pen di um of fi bro bla st cel l ty pe s an d state s , the ir di ver se line age origin s an d multifaceted functions across vari ou s human organ s .

Custom input sentence 2:

Modern imaging technologies are widely based on classical principles of light or electromagnetic wave propagation. They can be remarkably sophisticated, with recent successes ranging from single-molecule microscopy to imaging far-distant galaxies.

babyGPT tokenization of custom input sentence 2:

Mo der n imaging tech no lo gi es are widely base d on classi cal principle s of light or electro mag net ic wa ve propagation . They can be remarkabl y so p hist i cate d , with recent successes ran g ing from single - mole cul e micr o sco py to imaging far - di stan t gal ax i es .

Custom input sentence 3:

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics.

babyGPT tokenization of custom input sentence 3:

Deep lear nin g allo w s com pu ta ti o nal models that are com po se d of multi ple proce s sing layer s to learn repres en ta ti on s of data with multi ple level s of ab st rac ti on . The se methods have dra mat i cal ly improve d the state - of - the - art in speech re co g nit i on , visual object re co g nit i on , object detecti on an d m any other do main s such as drug discove r y an d ge no mic s .

Custom input sentence 4:

The rotation period of Uranus was estimated to be 17.24?h in 1986 from radio auroral measurements during the brief Voyager 2 flyby. This value is the basis for the Uranian SIII *longitude system still in use. However, the poor period uncertainty limited its validity to a few years, after which the orientation of the magnetic axis was lost.*

babyGPT tokenization of custom input sentence 4:

Custom input sentence 5:

Fuelled by increasing computer power and algorithmic advances, machine learning techniques have become powerful tools for finding patterns in data. Quantum systems produce atypical patterns that classical systems are thought not to produce efficiently, so it is reasonable to postulate that quantum computers may outperform classical computers on machine learning tasks.

babyGPT tokenization of custom input sentence 5:

The ro ta ti on peri od of Ur an us was e sti mat ed to be 17 . 24 ? h in 198 6 from radio au r or al mea su rem ent s du ring the bri ef Vo ya ge r 2 fl y by . T his value is the basi s for the Ur an i an SI II lon gi tu de system still in use . How ever , the poor peri od un certa int y limited it s va li di ty to a fe w years , after whi ch the ori en ta ti on of the mag net ic ax is was lo st .

Comparison:

By empirical observation, we can write some qualitative comparison between word level, sub-word level and babyGPT level tokenization.

Word Tokenization	**Sub-word Tokenization**	**babyGPT Tokenization**
Vocabulary size is very large	Vocabulary size is moderate (BPE/wordpiece)	Smaller vocabulary size, smaller than sub-word tokenization
Punctuation stays with the word	Punctuation is separate token	Punctuation is separate token
Granularity is coarse, picks each individual word with punctuation.	Granularity is medium, typically common word pieces	Breaks in phonemes and syllables
The number of unknown words is small	Rare words are also broken into sub words	Rare words are decomposed further into syllables and phonemes
Tokens can be interpreted easily	Tokens can be interpreted moderately	Tokens are difficult to interpret
Used in simple NLP models	Used in LLM models like BERT and GPT	May be used in models where syllables and phonemes are more focused
Used white space to separate words	Does not typically use white space to separate words	Use syllables and phonemes to separate words

Curation of 5 custom sentences and their sources:

“Fibroblasts, once perceived as a uniform cell type, are now recognized as a mosaic of distinct populations with specialized roles in tissue homeostasis and pathology. Here we provide a global overview of the expanding compendium of fibroblast cell types and states, their diverse lineage origins and multifaceted functions across various human organs.” Torregrossa et al., “**Effects of embryonic origin, tissue cues and pathological signals on fibroblast diversity in humans**”, *Nature Cell Biology*, (2025)

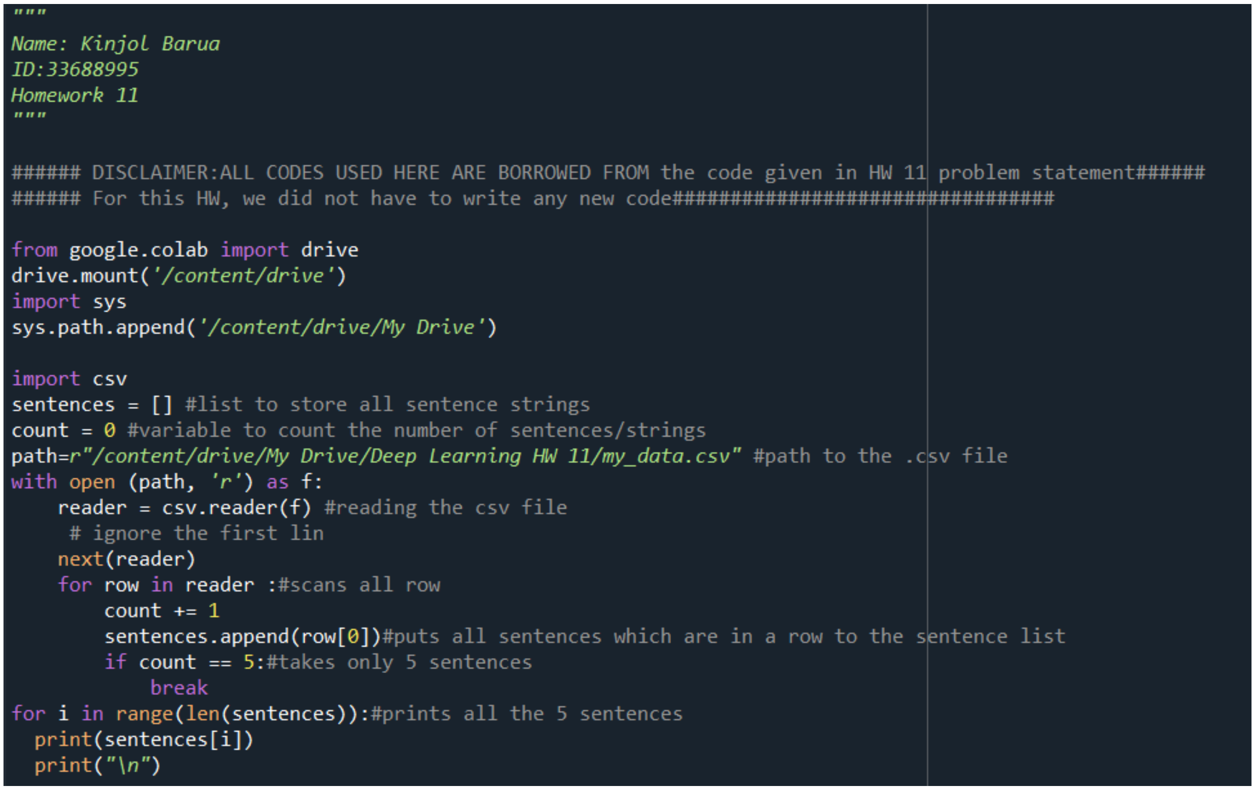
“Modern imaging technologies are widely based on classical principles of light or electromagnetic wave propagation. They can be remarkably sophisticated, with recent successes ranging from single-molecule microscopy to imaging far-distant galaxies.” Defienne et al., “**Advances in quantum imaging**”, *Nature Photonics*, **18**, 1024-1036 (2024)

“Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics.” LeCun et al., “**Deep Learning**”, *Nature*, **521**, 436-444 (2015)

“The rotation period of Uranus was estimated to be 17.24 ± 0.01 h in 1986 from radio auroral measurements during the brief Voyager 2 flyby. This value is the basis for the Uranian SIII longitude system still in use. However, the poor period uncertainty limited its validity to a few years, after which the orientation of the magnetic axis was lost.” Lamy et al., “**A new rotation period and longitude system for Uranus**”, *Nature Astronomy*, (2025)

“Fuelled by increasing computer power and algorithmic advances, machine learning techniques have become powerful tools for finding patterns in data. Quantum systems produce atypical patterns that classical systems are thought not to produce efficiently, so it is reasonable to postulate that quantum computers may outperform classical computers on machine learning tasks.” Biamonte et al., “**Quantum Machine Learning**”, *Nature*, **549**, 195-202 (2017)

Source Code:

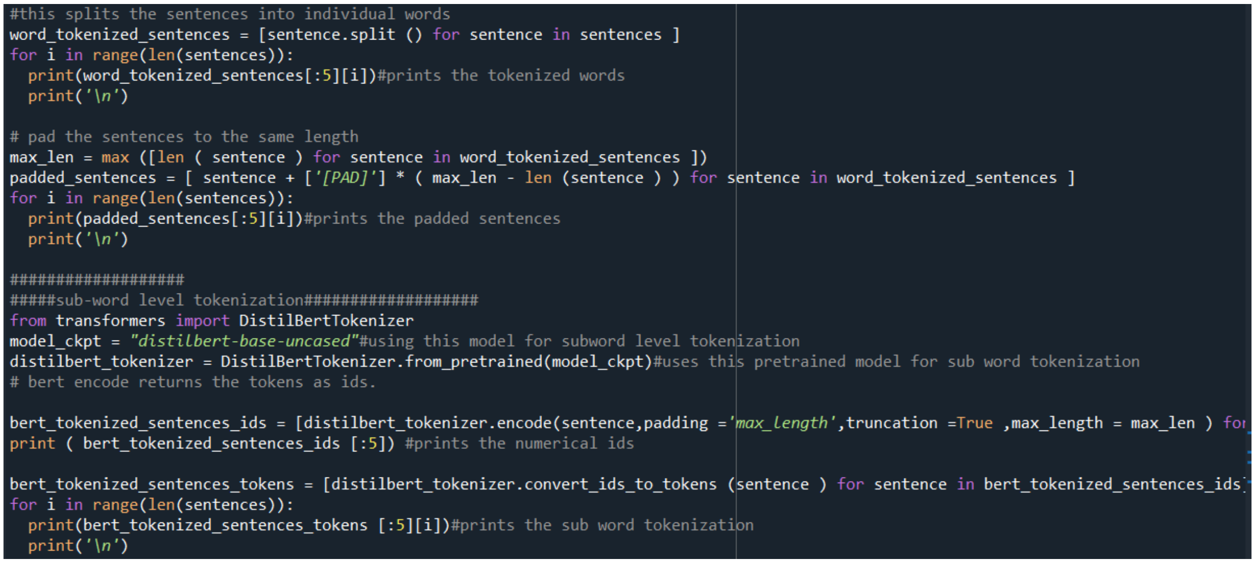


```
"""
Name: Kinjol Barua
ID:33688995
Homework 11
"""

###### DISCLAIMER:ALL CODES USED HERE ARE BORROWED FROM the code given in HW 11 problem statement######
###### For this HW, we did not have to write any new code##################################

from google.colab import drive
drive.mount('/content/drive')
import sys
sys.path.append('/content/drive/My Drive')

import csv
sentences = [] #list to store all sentence strings
count = 0 #variable to count the number of sentences/strings
path=r"/content/drive/My Drive/Deep Learning HW 11/my_data.csv" #path to the .csv file
with open (path, 'r') as f:
    reader = csv.reader(f) #reading the csv file
     # ignore the first lin
    next(reader)
    for row in reader :#scans all row
        count += 1
        sentences.append(row[0])#puts all sentences which are in a row to the sentence list
        if count == 5:#takes only 5 sentences
            break
for i in range(len(sentences)):#prints all the 5 sentences
  print(sentences[i])
  print("\n")
```



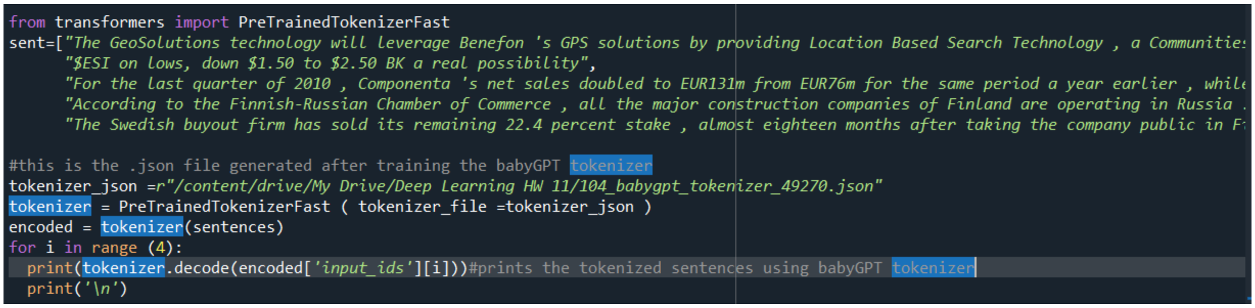
```
#this splits the sentences into individual words
word_tokenized_sentences = [sentence.split () for sentence in sentences ]
for i in range(len(sentences)):
  print(word_tokenized_sentences[:5][i])#prints the tokenized words
  print('\n')

# pad the sentences to the same length
max_len = max ([len ( sentence ) for sentence in word_tokenized_sentences ])
padded_sentences = [ sentence + ['[PAD]'] * ( max_len - len (sentence ) ) for sentence in word_tokenized_sentences ]
for i in range(len(sentences)):
  print(padded_sentences[:5][i])#prints the padded sentences
  print('\n')

###################
#####sub-word level tokenization##################
from transformers import DistilBertTokenizer
model_ckpt = "distilbert-base-uncased"#using this model for subword level tokenization
distilbert_tokenizer = DistilBertTokenizer.from_pretrained(model_ckpt)#uses this pretrained model for sub word tokenization
# bert encode returns the tokens as ids.

bert_tokenized_sentences_ids = [distilbert_tokenizer.encode(sentence,padding ='max_length',truncation =True ,max_length = max_len ) for
print ( bert_tokenized_sentences_ids [:5]) #prints the numerical ids

bert_tokenized_sentences_tokens = [distilbert_tokenizer.convert_ids_to_tokens (sentence ) for sentence in bert_tokenized_sentences_ids
for i in range(len(sentences)):
  print(bert_tokenized_sentences_tokens [:5][i])#prints the sub word tokenization
  print('\n')
```



```python
from transformers import PreTrainedTokenizerFast
sent=["The GeoSolutions technology will leverage Benefon 's GPS solutions by providing Location Based Search Technology , a Communitie
      "$ESI on lows, down $1.50 to $2.50 BK a real possibility",
      "For the last quarter of 2010 , Componenta 's net sales doubled to EUR131m from EUR76m for the same period a year earlier , whil
      "According to the Finnish-Russian Chamber of Commerce , all the major construction companies of Finland are operating in Russia .
      "The Swedish buyout firm has sold its remaining 22.4 percent stake , almost eighteen months after taking the company public in F

#this is the .json file generated after training the babyGPT tokenizer
tokenizer_json =r"/content/drive/My Drive/Deep Learning HW 11/104_babygpt_tokenizer_49270.json"
tokenizer = PreTrainedTokenizerFast ( tokenizer_file =tokenizer_json )
encoded = tokenizer(sentences)
for i in range (4):
  print(tokenizer.decode(encoded['input_ids'][i]))#prints the tokenized sentences using babyGPT tokenizer
  print('\n')
```