MSAI-337: Natural Language Processing

Homework 1 (NLP)

January 24th, 2022

Professor
Dr. David Demeter

Students
Aleksandr Simonyan
Dimitrios Mavrofridis
Donald Baracskay
Wentao Yao

Question 2:

We used regular expressions to replace the various numbers. We created a function that uses a descending order of specificity. For example, the function will catch and replace years before integers, because most 4 digit integers are years, and catching these first ensures that they will be tagged properly as years. The specific order and implementation is: years (catches any isolated 4 digit int), decimal (catches any number of digits followed by a period with any number of digits trailing), day (catches any isolated 2 digit int), other (catches any combination of digits and punctuation [most of these are ISBNs which are of the form 1234-1242-1233-12]), and finally integer (catches any remaining group of digits). This function takes in the text and loops over all tokens with regular expression substitution and then returns the resultant text. In addition, we keep track of all converted numbers and see how many times they have appeared in the text. At the end, we are returning corresponding dictionaries as well to later include in our Summary Statistics.

Question 3:

We first split the corpus into the various articles, via the <end-of-passage> tokens. We figured that it made more sense to split the corpus into articles than into individual tokens. Next, we generated a randomly shuffled list that was the same size as the corpus. We then iterate through the aforementioned list. For every 10 entries, 8 will go to training, 1 to test, and 1 to validation.

Question 4:

Summary statistics

validation data token count	9341				
test token count	9256				
train token count	229534				
number of unk	40910				
out of words	5596				
number of types	string	date	decimal	day	integer
	39529	1308	73	0	1
number of stopwords	209168				
top_words	unk	date	university	new	economic
	42381	11471	1516	894	768
top_n_grams	<unk> <unk> <unk>	<unk> <unk> <date>	<date> <unk> <unk> '	<date> isbn <unk>	world war ii '
	3500	561	483	370	77

Two custom metrics that we have used are “top_words” that shows the number of occurrences of most widespread words and “top_n_grams” that shows how many times most widespread n_grams occur. We have used the first function to understand how our corpus is constructed and whether we have useless additional words. For example, if we see that some particles / propositions appear in top words we can remove them and thus improve our corpus. With “top_n_grams,” we understand what are the words that most likely to co-occur, how much they make sense and how useful their occurrences are in our language modeling tasks. By looking at most frequent n_grams, we can further preprocess our data and improve our corpus.

Question 6:

We apply Hugging Face word-piece tokenizer to construct the word-piece tokenization of the source text. We apply WordPiece() as our Tokenizer and see the unknown token to ‘<unk>’ for later statistics. For the normalizer, we apply NFD(), StripAccents(), NFKC(), and Lowercase(). The parameters for the trainer are vocab_size=5000 and special_tokens=["<unk>", "<end_of_passage>", "<start_of_passage>"]. Vocabulary size is required by the instruction and special_tokens, "<end_of_passage>", "<start_of_passage>" is how we separate the source file as mentioned above in Q3.
The following are the summary statistics: (Please run the Q6.py to see the full summary statistics.)

Trained vocab size:	5000		
Training token count	589749	Validation token count	102849
Test token count 62967	62967		
Number of unk in validation	13	Number of unk in test	9
Number of out of vocabulary words in validation	20	Number of out of vocabulary words in test	13
Number of types in train	'string': 564164, 'number': 25585	Number of types in validation	'string': 99323, 'number': 3526
Number of types in test	'string': 60258, 'number': 2701		
Number of stopwords in train	148182	Number of stopwords in validation	25990
Number of stopwords in test	14744		
top_words in train(part)		the': 22688, ',': 21177, '.': 20681, 'of': 15678, 'in': 11791, 'and': 10604, 'a': 7692, 'to': 7563, '##s': 7424, 'was': 5555, 'he': 5504, '(': 5015, '-': 4498, 'his': 3807, '==': 3642, 'for': 3335, '"':	

	3286, "'": 2969, 'as': 2948, 's': 2889, 'on': 2878, 'at': 2775…
ngram_words in train(part)	'the university of ': 567, '. he was ': 557, '== references == ': 324, 'at the university ': 308, ', he was ': 263, 'member of the ': 252, 'the united states ': 226, '== external links ': 217, 'external links == ': 217, ') was a ': 204, '##s of the ': 202…