Course: COSC 4337 Data Science II

Professor: Ricardo Vilalta

TA: Shaila Zaman

Team : 4

Group members: Hieu Trinh , Thanh Le

Machine Translation from English to French

I. Data description

i. Data source and overview

This dataset is created based on the website : http://www.manythings.org/anki/ , which collects a lot of datasets of different language. The author only used sentences that were owned by identified native speakers working on the Tatoeba Project and English sentences that he personally checked and did not reject.

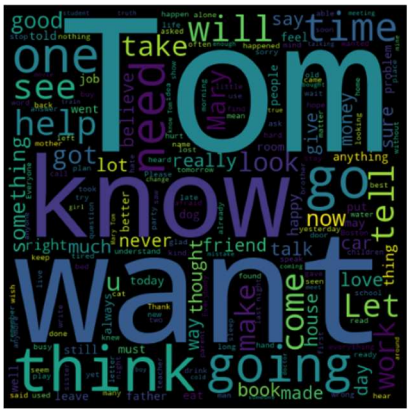
ii. Column and data type

The dataset has two columns : one has English words/sentences and the other has French words/sentences. The data type is string and each column include letters, digits, and punctuations and regular expressions. There are 175,621 rows for both columns as the English words/sentences are matched with the French words/sentences for each row. In addition to that, this is a Machine Translation task, so there is no label in the dataset. The special thing about this dataset is that one English word could be represented in different ways in French, so some words could be repeated in the dataset. For example, “Run!” could be represented as “Cours !” or “Courez !”. There are no null values for both columns, and as we discover along throughout the report, especially when looking at the Exploratory Data Analysis Notebook, we will have a more insightful look of the dataset.

II. Data visualization

We want to see the frequency of each word, so we use the library wordcloud, which is a data visualization technique used for representing text data in which the size of each word indicatew its importance. From the graph, we can clearly see which tokens mostly occur in the dataset.

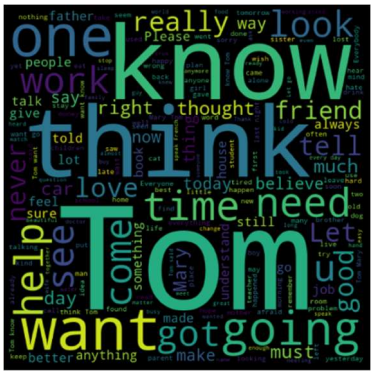
i. Before removing stopwords

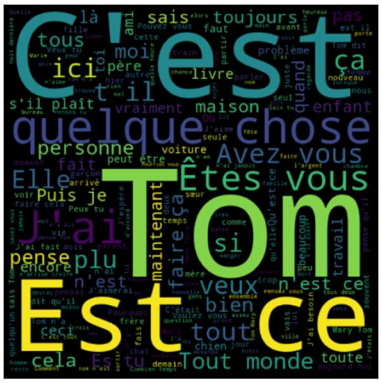


As we can see, a lot of familiar words like “know”, “want”, “think”, “going”, “will”, etc… occur a lot in the dataset. By using a lot of popular words, it will make it easier for the translation because we do not want to interfere with too many complicated words.

In French, some words like Je, que, suis are very popular because they are translated as “I”, “that”, “am”. It is easy to understand because those tokens also occurs a lot in English.

ii. After removing stopwords





We do not see a lot of differences comparing to the original one, so we can assume that our English column is pretty clean even before preprocessing.

In opposite to English, we see a lot of changes in French after removing stop words. We see new tokens like" C'est" ," Est" ," ce" and they are translated as "It is", "is", "this". These are also popular words in English.

III. Exploratory Data Analysis & Feature Engineering:

i. Import Dataset, libraries and clean the dataset

To begin with the preprocessing step, we first import Pandas and built-in functions needed to read the dataset. The dataset has 2 columns with 175,621 rows. We then simply rename two columns as "English" and "French" to make it easier to call. This is a machine translation task, so we do not want to remove any words, even stop words such as "the", "a" , "an", "in", etc… or regular expressions. However, we want to have a better look at our dataset because we try to understand the meaning of words , so we should remove all the regular expressions and punctuations before performing data analysis.

The left table shows the original dataset and the right table shows the dataset after removing the regular expressions and punctuations.

	English	French
0	Hi.	Salut!
1	Run!	Cours !
2	Run!	Courez !
3	Who?	Qui ?
4	Wow!	Ça alors !
...	...	...
175616	Top-down economics never works, said Obama. "T...	« L'économie en partant du haut vers le bas, ç...
175617	A carbon footprint is the amount of carbon dio...	Une empreinte carbone est la somme de pollutio...
175618	Death is something that we're often discourage...	La mort est une chose qu'on nous décourage sou...
175619	Since there are usually multiple websites on a...	Puisqu'il y a de multiples sites web sur chaqu...
175620	If someone who doesn't know your background sa...	Si quelqu'un qui ne connaît pas vos antécédent...

	English	French
0	Hi	Salut
1	Run	Cours
2	Run	Courez
3	Who	Qui
4	Wow	Ça alors
...	...	...
175616	Topdown economics never works said Obama The c...	Léconomie en partant du haut vers le bas ça n...
175617	A carbon footprint is the amount of carbon dio...	Une empreinte carbone est la somme de pollutio...
175618	Death is something that were often discouraged...	La mort est une chose quon nous décourage souv...
175619	Since there are usually multiple websites on a...	Puisquil y a de multiples sites web sur chaque...
175620	If someone who doesnt know your background say...	Si quelquun qui ne connaît pas vos antécédents...

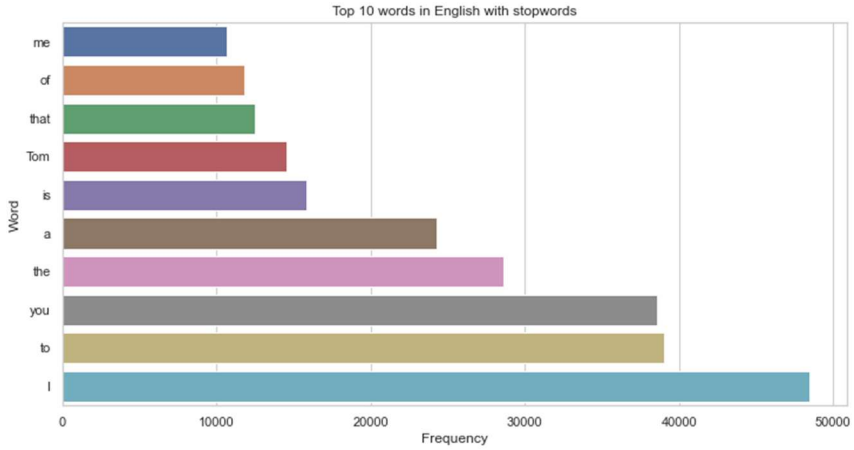
ii. Creating dictionaries for English and French

After performing wordcloud visualization and cleaning the dataset, now we go ahead and create dictionaries for English and French to keep track of the frequency of

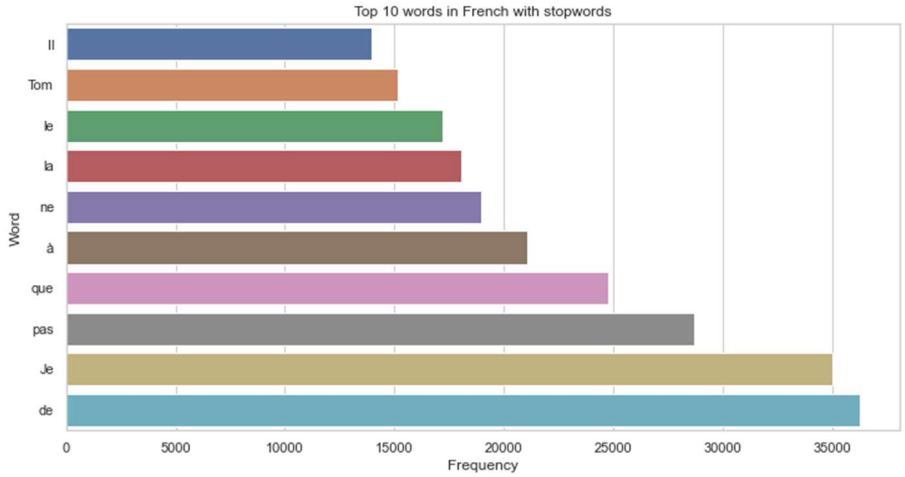
every single token. Then we check if the sentences contain number in English and French column, but we see that sentences from both languages use the numbers but not the similar words like using “one” for “1” , so we decide that we will keep numbers in sentences because it will not affect the translation process. Then, we create English and French dictionaries of our dataset to keep track of the frequency of each word.

Also, we want to understand more about how words are distributed throughout sentences, so we decide to plot the top 10 most frequents words in the bar graph for analysis by using seaborn library.

Before removing stopwords:

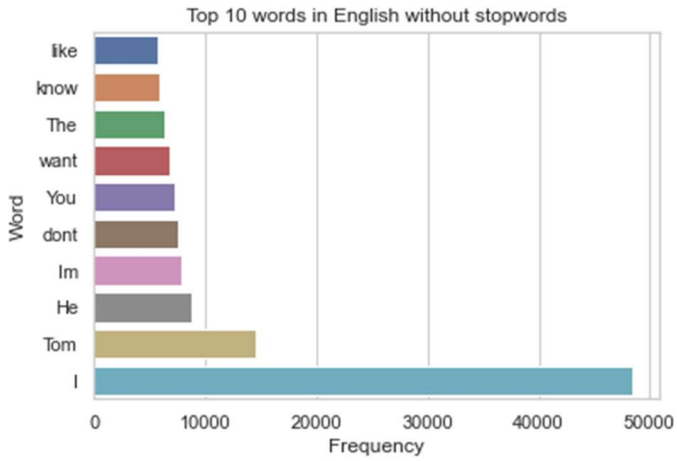


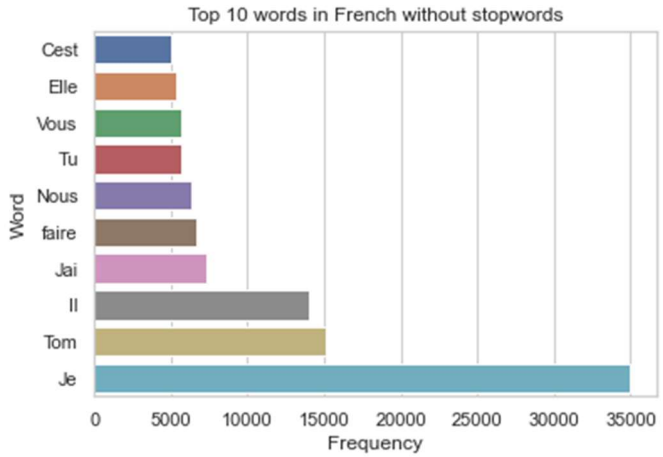
Word	Frequency
me	10676
of	11806
that	12515
Tom	14559
is	15861
a	24319
the	28621
you	38580
to	39059
I	48481



Word	Frequency
Il	13980
Tom	15144
le	17168
la	18035
ne	18979
à	21071
que	24770
pas	28669
Je	34981
de	36253

After removing stopwords:

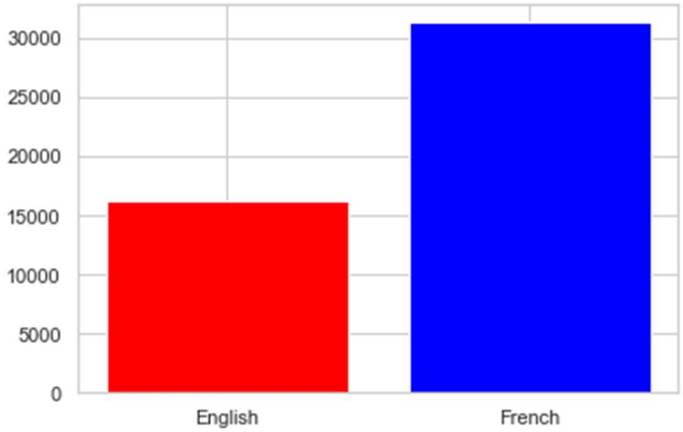




The results before and after removing stop words are so different. Before removing, we can clearly see a lot of words like “I”, “to”, “you”, “the” occur around 30000 to 48000 times in sentences. However, after the removal, the lead automatically belongs to letter “I” and it shows us that stop words have a great impact on every language, and it occurs almost in every sentence.

iii. Uniqueness of English and French

We want have a more insightful look at our dataset to see if English or French vocabulary is more varied and unique because as we describe above, an English word could be translated as two to three French words, so to understand how words are translated is extremely important.



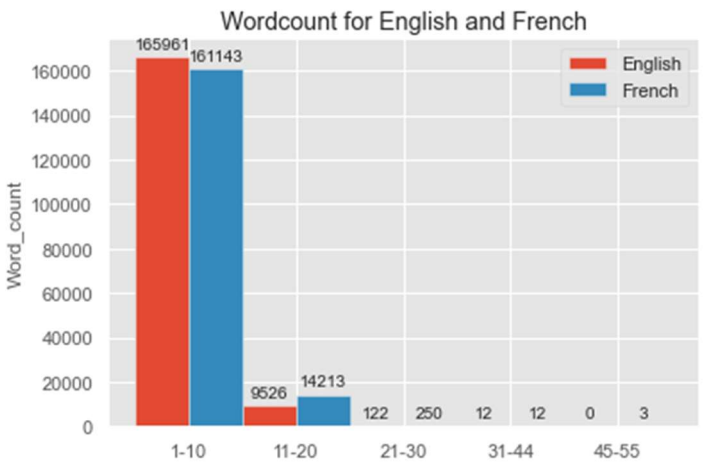
We can easily see that for this particular dataset, French tokens almost double English tokens. This again shows that one English word can be expressed differently in French.

iv. Comparison of words for each row

We want to explore the dataset further, so we count and compare the number of words for each row and try to create different categories depending on the length of sentences. Then, we create bins and plot them into a histogram for visualization The histogram helps us to have an overview the differences in number of words between an English sentence and its translation to French.

The length of sentences is in ascending order, so we display a row for every 10,000 rows to clearly see the difference. We have a new table with three columns: English_wordcount, French_wordcount, Difference and the grouped bar chart between the first two columns in the table.

	English_wordcount	French_wordcount	Difference
0	1	1	0
10000	3	3	0
20000	4	5	1
30000	3	3	0
40000	4	2	-2
50000	5	3	-2
60000	6	5	-1
70000	5	7	2
80000	6	5	-1
90000	6	6	0
100000	6	5	-1
110000	5	6	1
120000	8	6	-2
130000	8	8	0
140000	8	9	1
150000	10	10	0
160000	9	10	1
170000	10	14	4



The observation shows that most paired sentences have less than ten words in comparisons. This is extremely helpful because we focus more on the accuracy of translating each word rather than the accuracy of the long and complicated sentences. We

are aiming at having a machine translation like Google translation that can have a very high accuracy in translating words and short sentences.

v. Understanding Part of Speech Tagging (POS Tag)

POS tag is a special technique to identify the label assigned to each token that we can understand what type of tokens mostly occur in the dataset.

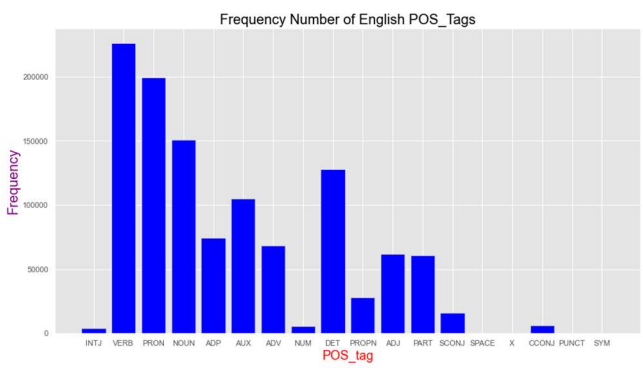
For example: I like you => “like” is a verb with a positive sentiment.

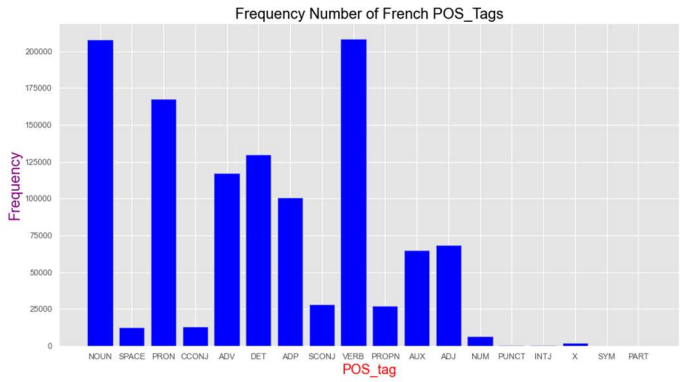
I am like you => “like” is a preposition with a neutral sentiment.

We want to know what type of words the dataset has the most because we want to know the reliability of our machine translation. For example, if our dataset has a lot of nouns and verbs, we should be confident about having every word translating correctly because definitely, nouns and verbs are two of the most popular types of words in grammar for all the countries.

We use the pre-trained models from spaCy to create dictionaries of POS tag for English and French. The results are not surprising when we see a lot of verbs, nouns and pronouns occur a lot in both languages. Once again, it shows the grammar and the structure of sentences between two languages have a lot of similarities.

Here is the bar chart displaying the POS tag of English and French column:

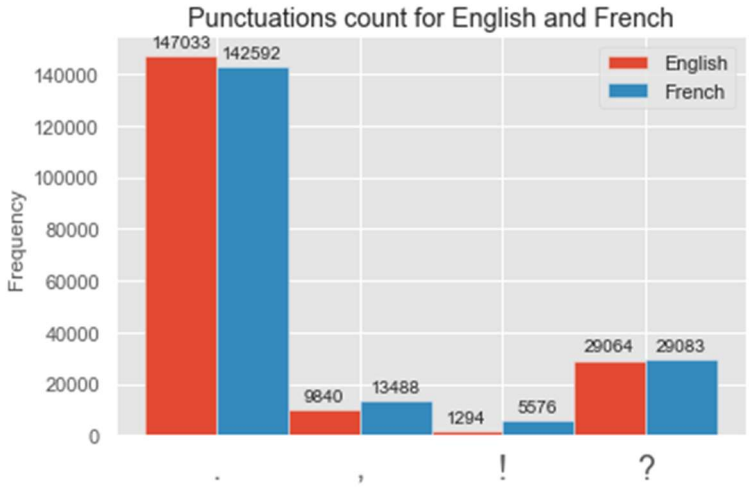




vi. Punctuations count

We will train the model based on the original dataset, so understanding of the number of punctuations of the dataset is essential besides the understanding of different types of words. We first create a dictionary with four popular punctuations: period,

comma, question mark and exclamation. Then we continue counting the number of punctuations and plot it in the bar chart for better visualization. As expected, English and French use a lot of period, but French tends to use more comma and exclamation in sentences. This is very interesting because in French, people probably express feelings more regularly than people using English.



vii. Summary of the dataset

1082094 English words.

16414 unique English words.

10 Most common words in the English dataset:

I	to	you	the	a	is	Tom	that	of	me

1142744 French words.

31487 unique French words.

10 Most common words in the French dataset:

de	Je	pas	que	à	ne	la	le	Tom	Il

IV. Preprocessing Pipeline

1. Tokenization

As deep learning models cannot understand text, we need to convert text into

numerical representation. Therefore, we need to perform a technique called Tokenization, which split sentences into words and encodes these into integers.

For example, there are three sentences : “The quick brown fox jumps over the lazy dog. ”, “I really love my dog.”, “My girlfriend is so beautiful” . We will perform the Tokenization technique to this sequence of sentences

{'the': 1, 'dog': 2, 'my': 3, 'quick': 4, 'brown': 5, 'fox': 6, 'jumps': 7, 'over': 8, 'lazy' : 9, 'i': 10, 'really': 11, 'love': 12, 'girlfriend': 13, 'is': 14, 'so': 15, 'beautiful': 16}

Then, we represent these as sequences from the tokenizer object we created

Sequence 1 in x

Input: The quick brown fox jumps over the lazy dog .

Output: [1, 4, 5, 6, 7, 8, 1, 9, 2]

Sequence 2 in x

Input: I really love my dog .

Output: [10, 11, 12, 3, 2]

Sequence 3 in x

Input: My girlfriend is so beautiful .

Output: [3, 13, 14, 15, 16]

2. **Padding**

In the dataset, the length of sentences is different. However, neural networks models require to have inputs with the same size. For this purpose, padding technique has to be performed.

Sequence 1 in x after padding

Input: [1 4 5 6 7 8 1 9 2]

Output: [1 4 5 6 7 8 1 9 2]

Sequence 2 in x after padding

Input: [10 11 12 3 2]

Output: [10 11 12 3 2 0 0 0 0]

Sequence 3 in x after padding

Input: [3 13 14 15 16]

Output: [3 13 14 15 16 0 0 0 0]

Now, all the sentences have the same length. We will apply this technique for our dataset, and the padding length will be based on the maximum length of the selected column.

V. Project goal and idea

Before training the model, we can use word embedding, a technique that captures the semantic, contextual and syntactic meaning of each word in the sentences, to increase model performance. Our first approach could be using Recurrent Neural Network, and then we want to try out the most recent technique called Sequence to Sequence Learning using Attention. The ultimate goal is we could be able to use Transformer model from HuggingFace library, a very famous library with more than 1,000 pre-trained transformer models, to train the dataset. We are probably going to use PyTorch and Keras for this project.