***Mikhail Terentev***

**CSTU Natural Language Processing Homework**

**A. Answer any 10 Short answer questions: Answer in less than 10 lines**

**1. What are stop words?**

Stop words are the words in a vocabulary which are of little value when considering word frequencies in text. Stop words are those words that do not contribute to the deeper meaning of the phrase.

**2. List any two real-life applications of Natural Language Processing.**

“Application Tracking Systems” and “Grammarly” are based on NLP technology.

1. **Language translation apps** (Google Translate, for example).
2. **Email filters** - spam filters, uncovering certain words or phrases that signal a spam message. But filtering has upgraded, just like early adaptations of NLP.

**3. What is TF-IDF?**

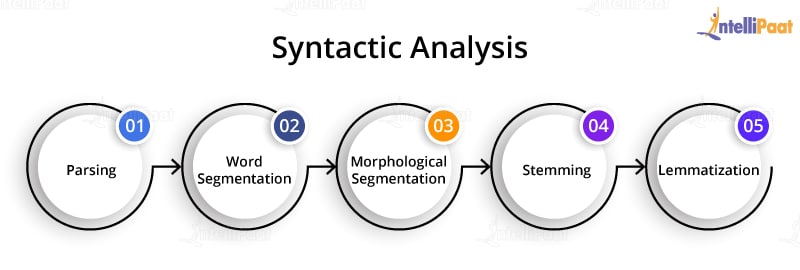
TF-IDF stands for “Term Frequency – Inverse Document Frequency” which are the components of the resulting scores assigned to each word.

* Term Frequency (TF) - summarizes how often a given word appears within a document. TF helps calculate the ratio of the frequency of a term in a document and the total number of terms.
* Inverse Document Frequency - downscales words that appear a lot across documents. IDF denotes the importance of the term in a document.

**4. What is Syntactic Analysis?**

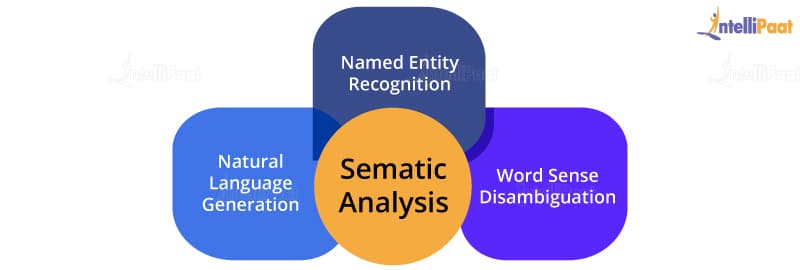
Syntactic analysis or parsing or syntax analysis is the third phase of NLP. Syntax analysis checks the text for meaningfulness comparing to the rules of formal grammar. For example, the sentence like “hot ice-cream” would be rejected by semantic analyzer.

In context of Syntactic analysis explain Parsing, Word Segmentation, Morphological segmentation, stemming, lemmatization.

 The techniques used for syntactic analysis are as follows:

1. Parsing - it helps in deciding the structure of a sentence or text in a document. It helps analyze the words in the text based on the grammar of the language.
2. Word segmentation - the segmentation of words segregates the text into small significant units.
3. Morphological segmentation - the purpose of morphological segmentation is to break words into their base form.
4. Stemming - it is the process of removing the suffix from a word to obtain its root word.
5. Lemmatization - the task of removing inflectional endings only and to return the base dictionary form of a word - meaningful base form (the ‘lemma’).

**5. What is Semantic Analysis? In Context of Semantic Analysis Explain Named Entity Recognition, Natural Language Generation and Word Sense Disambiguation.**



Semantic analysis describes the process of understanding natural language – the way that humans communicate–based on meaning and context. The semantic analysis of natural language content starts by reading all the words in content to capture the real meaning of any text. It identifies the text elements and assigns them to their logical and grammatical role, analyzes context in the surrounding text and the text structure to accurately disambiguate the proper meaning of words that have more than one definition.

1. Named entity recognition NER - given a stream of text, determine which items in the text map to proper names, such as people or places, and what the type of each such name is (e.g. person, location, etc.)
2. Natural language generation NLG - it is a process used by the software to convert the structured data into human spoken languages. By using NLG, organizations can automate content for custom reports.
3. Word sense disambiguation WSD - it helps identify the sense of a word used in different sentences and which have more than one meaning.
   1. **What is NLTK? Name few NLTK Libraries**

NLTK - Natural Language Toolkit. It is a collection of libraries for natural language processing. Some examples of the NLTK libraries are corpus, parse, sentiment, and tokenize. Some NLTK libraries: nltk.parse; nltk.tag; nltk.stem, nltk.test

**7. Explain how we can do parsing. What is Unigram, bigram, trigram?**

Parsing in NLP is the method to identify and understand the syntactic structure of a text. It is done by analyzing the individual elements of the text. Specific example of the n-gram includes the unigram (1-gram), bigram (2-gram) and trigram (3- gram).

**8. Explain Stemming with the help of an example.**

In NLP stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.  
For example, we can reduce 'CONNECTING' to ‘connect’ and ‘playing’ to ‘play’  
# -\*- coding: utf-8 -\*-

"""

Created on Sun Nov 6 21:07:30 2020

@ Mikhail Terentev

8. Explain Stemming with the help of an example.

"""

from nltk.stem import PorterStemmer

porter = PorterStemmer()

print(porter.stem('CONNECTING'))

print(porter.stem('playing'))

Output:

connect

play

**9. Explain Lemmatization with the help of an example.**

Lemmatization is the process of converting a word to its base form. The difference between stemming and lemmatization is, lemmatization considers the context and converts the word to its meaningful base form, whereas stemming just removes the last few characters, often leading to incorrect meanings and spelling errors.

# -\*- coding: utf-8 -\*

"""

Created on Sun Nov 6 21:07:30 2020

@ Mikhail Terentev

9. Explain Lemmatization

"""

from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

print(lemmatizer.lemmatize("studies")

print(lemmatizer.lemmatize("feet"))

Output:

study

foot

**10. What is Parts-of-speech Tagging?**

It is a process of converting a sentence to forms – list of words, list of tuples (where each tuple is having a form (word, tag)). The tag in case of is a part-of-speech tag, and signifies whether the word is a noun, adjective, verb, and so on.

# -\*- coding: utf-8 -\*-

"""

Created on Sun Nov 6 21:07:30 2020

@ Mikhail Terentev

10. Explain Parts-of-speech Tagging

"""

import nltk

from nltk.tokenize import word\_tokenize

text = word\_tokenize("How to check word similarity using the spacy package? Review and run following code and explain.")

# Split text into words

tokens = nltk.pos\_tag(text)

print (tokens)

Output:

[('How', 'WRB'), ('to', 'TO'), ('check', 'VB'), ('word', 'NN'), ('similarity', 'NN'), ('using', 'VBG'), ('the', 'DT'), ('spacy', 'NN'), ('package', 'NN'), ('?', '.'), ('Review', 'NNP'), ('and', 'CC'), ('run', 'VB'), ('following', 'VBG'), ('code', 'NN'), ('and', 'CC'), ('explain', 'NN'), ('.', '.')]

**11. Explain Named Entity Recognition by implementing it.**

Named entity recognition (NER) ‒ is an technique that automatically identifies named entities in a text and classifies them into predefined categories. Entities can be names of people, organizations, locations, times, quantities, monetary values, percentages, and more. Extracting these entities helps easily analyze huge amounts of unstructured data, like emails, open-ended survey responses, social media conversations, etc.

# -\*- coding: utf-8 -\*-

"""

Created on Sun Nov 6 21:07:30 2020

@ Mikhail Terentev

11. Explain NER

"""

import spacy

nlp = spacy.load("en\_core\_web\_sm")

text = "Netflix leads on downloads, but YouTube Kids grabs more hours in Washington"

doc = nlp(text)

for ent in doc.ents:

print(ent.text, ent.start\_char, ent.end\_char, ent.label\_)

Output:

YouTube Kids 32 44 PERSON

more hours 51 61 TIME

Washington 65 75 GPE

**12. How to check word similarity using the spacy package? Review and run following code and explain.**

# -\*- coding: utf-8 -\*-

"""

Created on Sun Nov 6 21:07:30 2020

@ Mikhail Terentev

12. Review and run following code and explain

"""

import spacy

nlp = spacy.load('en\_core\_web\_md')

print("Enter the words")

input\_words = input()

tokens = nlp(input\_words)

for i in tokens:

print(i.text, i.has\_vector, i.vector\_norm, i.is\_oov)

token\_1, token\_2 = tokens[0], tokens[1]

print("Similarity between words:", token\_1.similarity(token\_2))

Output:

hot cold

hot True 6.7365236 False

Similarity between words: 0.59720457

cold True 6.5119066 False

Similarity between words: 0.59720457

This means that the similarity between the words ‘hot’ and ‘cold’ is only 59 %.

**13. What are components of Natural Language Processing. Explain Entity Extraction, syntactic analysis, pragmatic analysis, morphological and lexical analysis.**

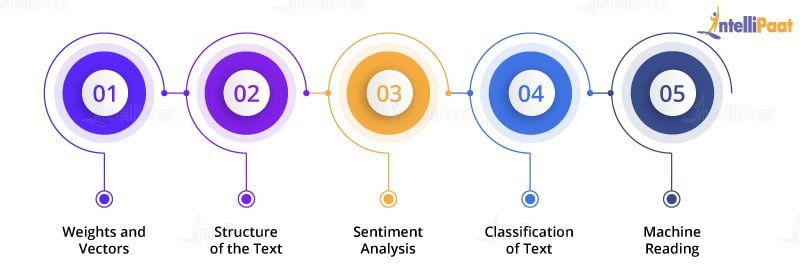
The major components of Natural Language processing are:

* **Entity extraction - e**ntity extraction refers to the retrieval of information such as place, person, organization, etc. by the segmentation of a sentence. It helps in the recognition of an entity in a text.
* **Morphological and Lexical Analysis** is a vocabulary that includes its words and expressions. It depicts analyzing, identifying and description of the structure of words. It includes dividing a text into paragraphs, words and the sentences. Individual words are analyzed into their components, and nonword tokens such as punctuations are separated from the words.
* **Semantic Analysis** is a structure created by the syntactic analyzer which assigns meanings. This component transfers linear sequences of words into structures. It shows how the words are associated with each other. Semantics focuses only on the literal meaning of words, phrases, and sentences. This only abstracts the dictionary meaning or the real meaning from the given context. The structures assigned by the syntactic analyzer always have assigned meaning
* **Pragmatic Analysis** deals with the overall communicative and social content and its effect on interpretation. It means abstracting or deriving the meaningful use of language in situations. In this analysis, the focus always on what was said in reinterpreted on what is meant.
* **Syntax analysis** - the words are commonly accepted as being the smallest units of syntax. The syntax refers to the principles and rules that govern the sentence structure of any individual languages. Syntax focus about the proper ordering of words which can affect its meaning. This involves analysis of the words in a sentence by following the grammatical structure of the sentence. The words are transformed into the structure to show how the word are related to each other.

**14. Define the terminology in NLP and interpretation of NLP in context of weights and vectors, structure of text, sentiment analysis, classification of text, machine reading.**

This is one of the most often asked NLP interview questions.

The interpretation of Natural Language Processing depends on various factors, and they are:



* **Weights and Vectors** are simply vectors of numbers, that represent the meaning of a word.**a word vector is a row of real valued numbers***where***each point captures a dimension of the word’s meaning***and***where semantically similar words have similar vectors**.This means that words such as wheel and engine should have similar word vectors to the word car

# **Structure of the Text - Language Syntax and Structure.** For any language, syntax and structure usually go hand in hand, where a set of specific rules, conventions, and principles govern the way words are combined into phrases; phrases get combines into clauses; and clauses get combined into sentences.

* **Sentiment Analysis -** The key aspect of sentiment analysis is to analyze a body of text for understanding the opinion expressed by it. Typically, we quantify this sentiment with a positive or negative value, called **polarity**. The **overall sentiment** is often inferred as positive, neutral or negative from the sign of the polarity score.
* **Classification of Text - Text classification** is the process of assigning tags or categories to text according to its content. It’s one of the fundamental tasks in natural language processing with broad applications such as sentiment analysis, topic labeling, spam detection, and intent detection.

# **Machine Reading Comprehension - MRC** scans documents and extracts meaning from the text, just like a human reader. You can ask MRC questions about a document and it will use different parts of the content until an answer is formed.

**15. What are the steps involved in solving an NLP problem?**

Step 1: Gather your data. ...

Step 2: Clean your data. ...

Step 3: Find a good data representation. ...

Step 4: Classification. ...

Step 5: Inspection. ...

Step 6: Accounting for vocabulary structure. ...

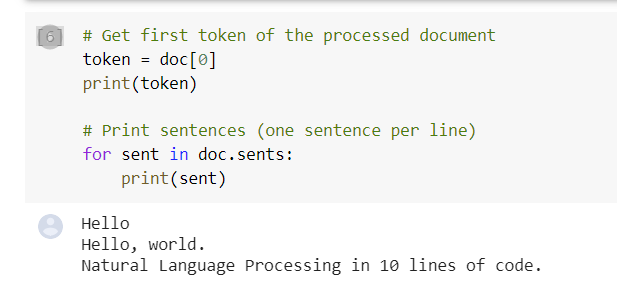
Step 7: Leveraging semantics. ...

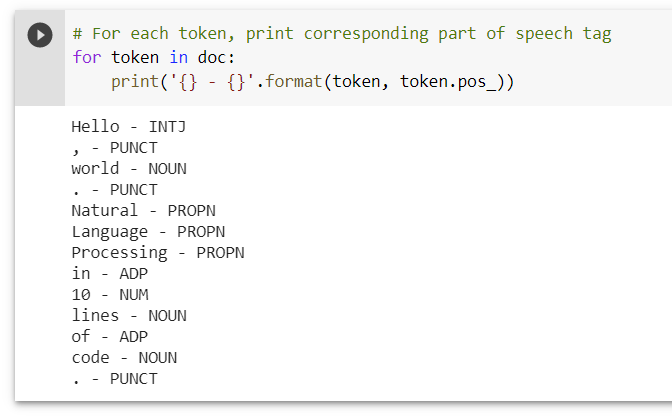
Step 8: Leveraging syntax using end-to-end approaches.

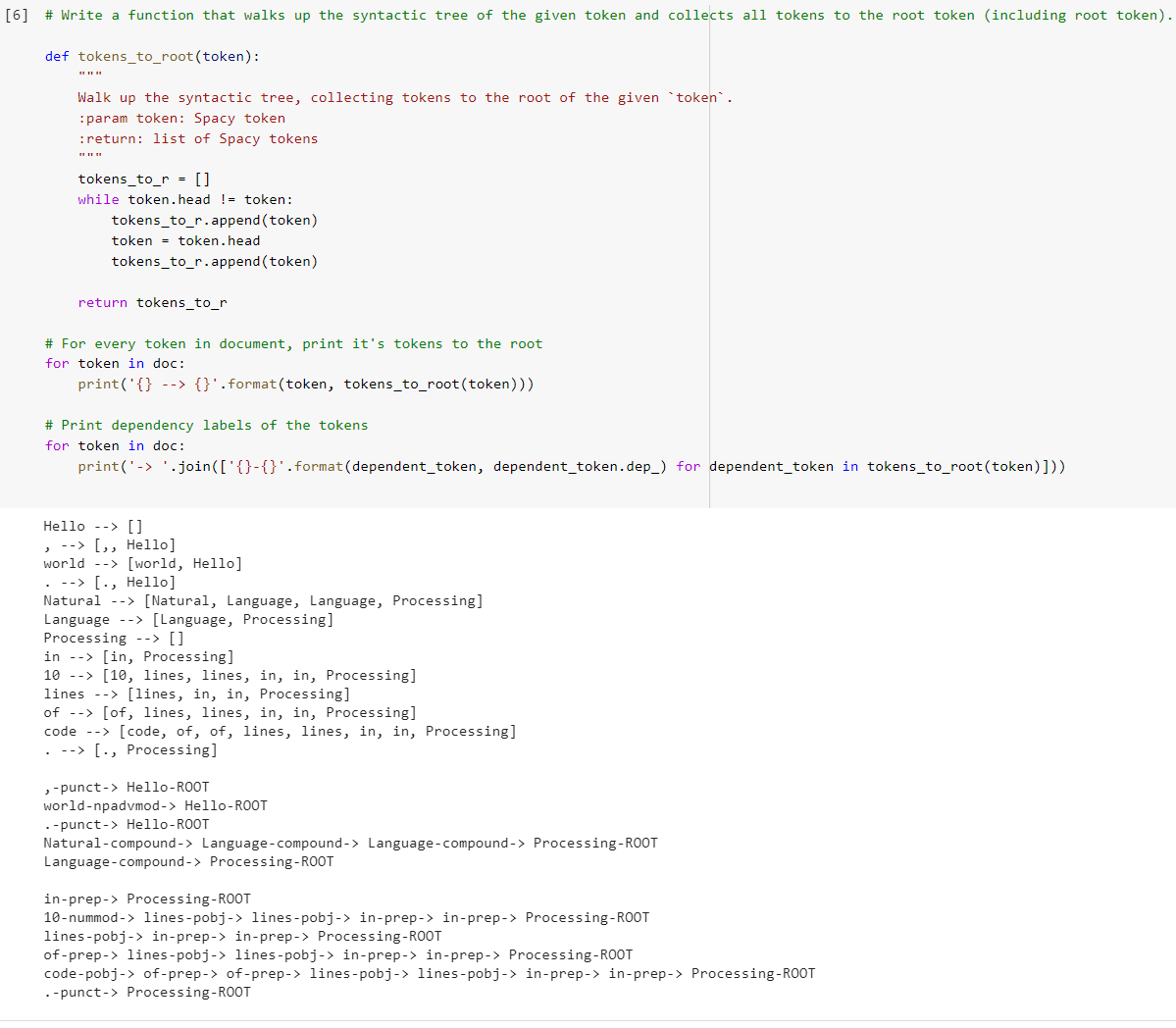
**B. Run any 3 following colab notebooks and submit your output.**

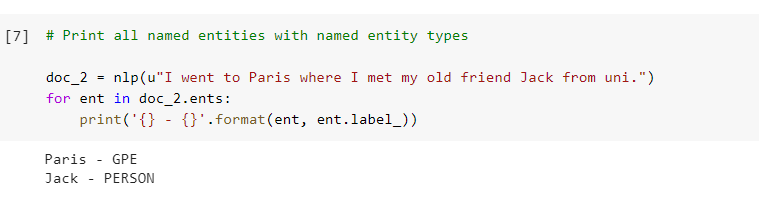
**0. Intro to Spacy**

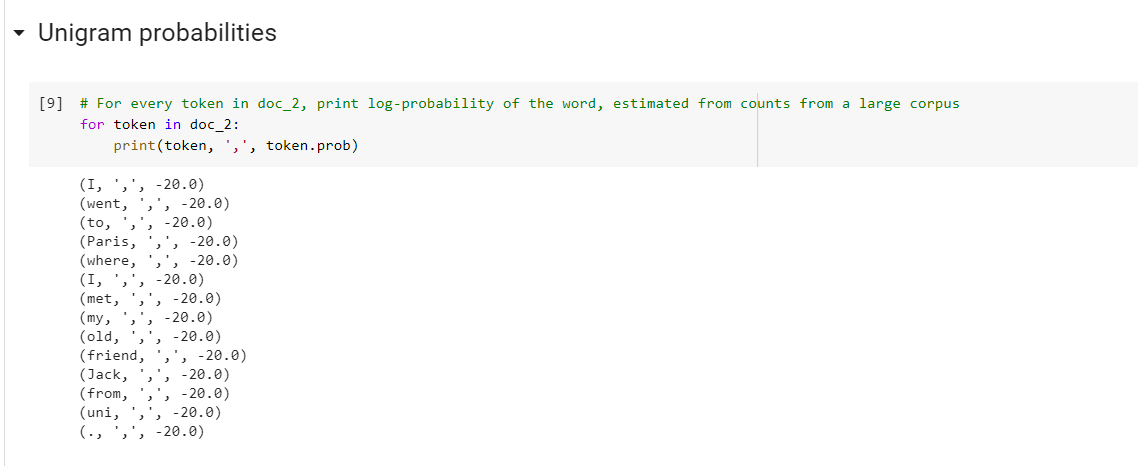
[**https://colab.research.google.com/github/explosion/spacy-notebooks/blob/master/notebooks/conference\_notebooks/pycon\_nlp/00\_spacy\_intro.ipynb**](https://colab.research.google.com/github/explosion/spacy-notebooks/blob/master/notebooks/conference_notebooks/pycon_nlp/00_spacy_intro.ipynb)

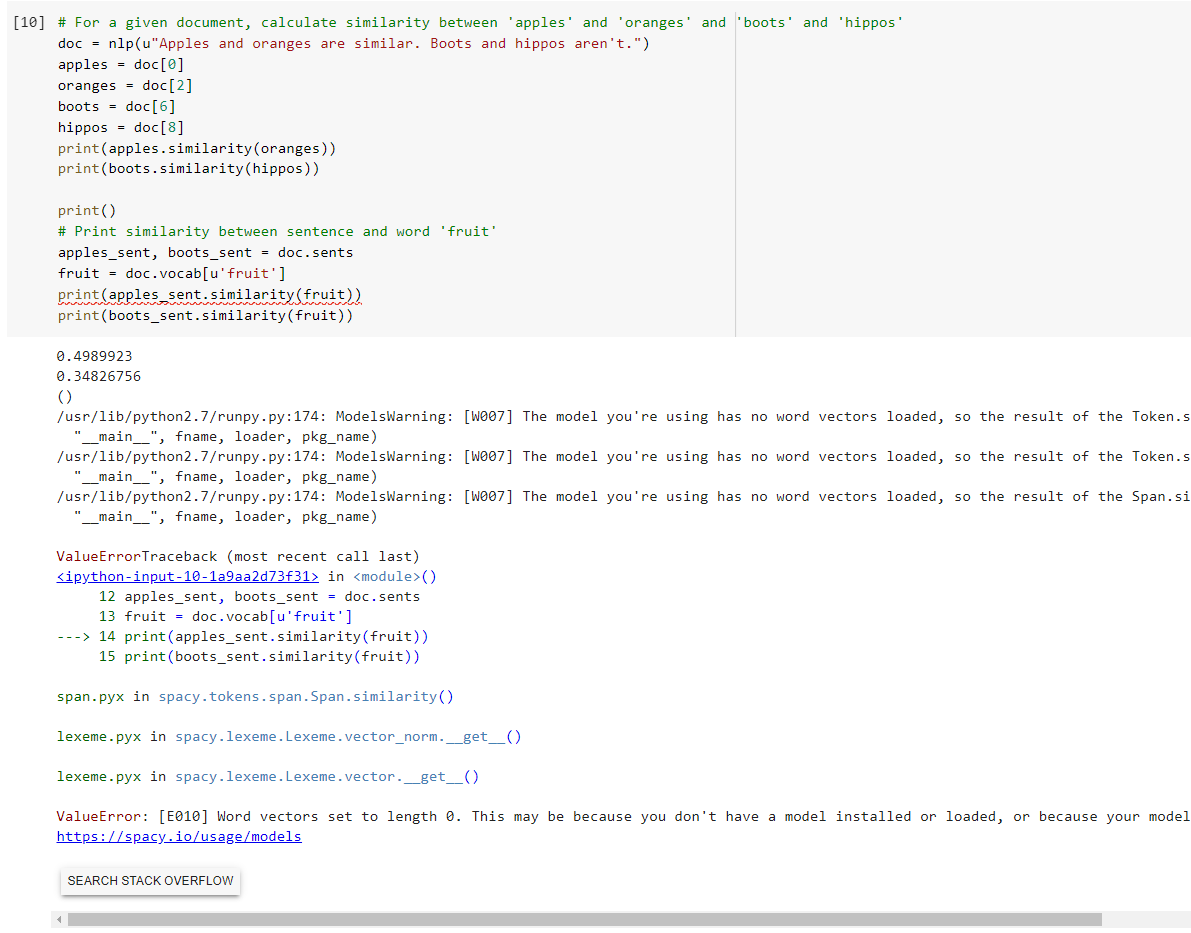




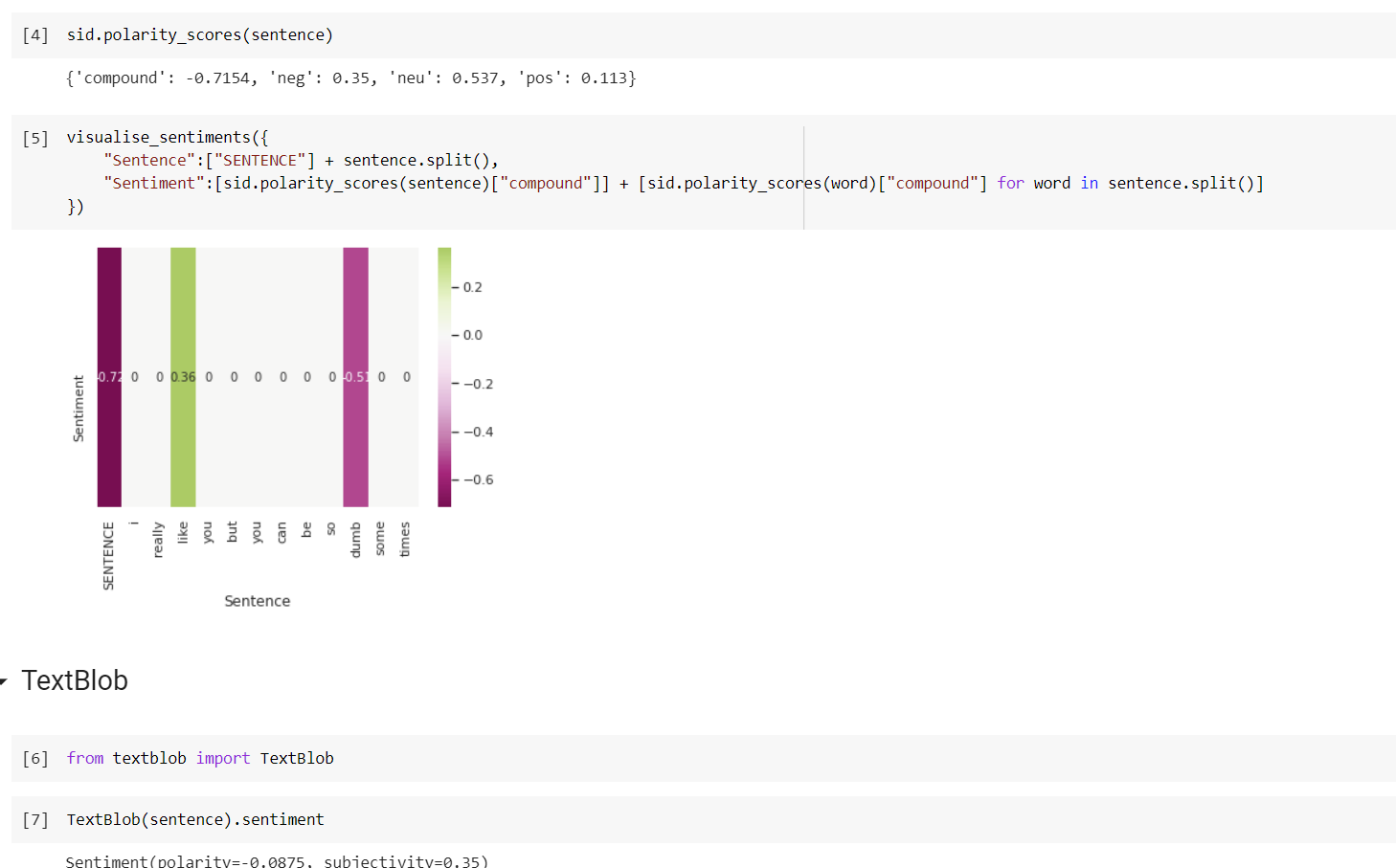


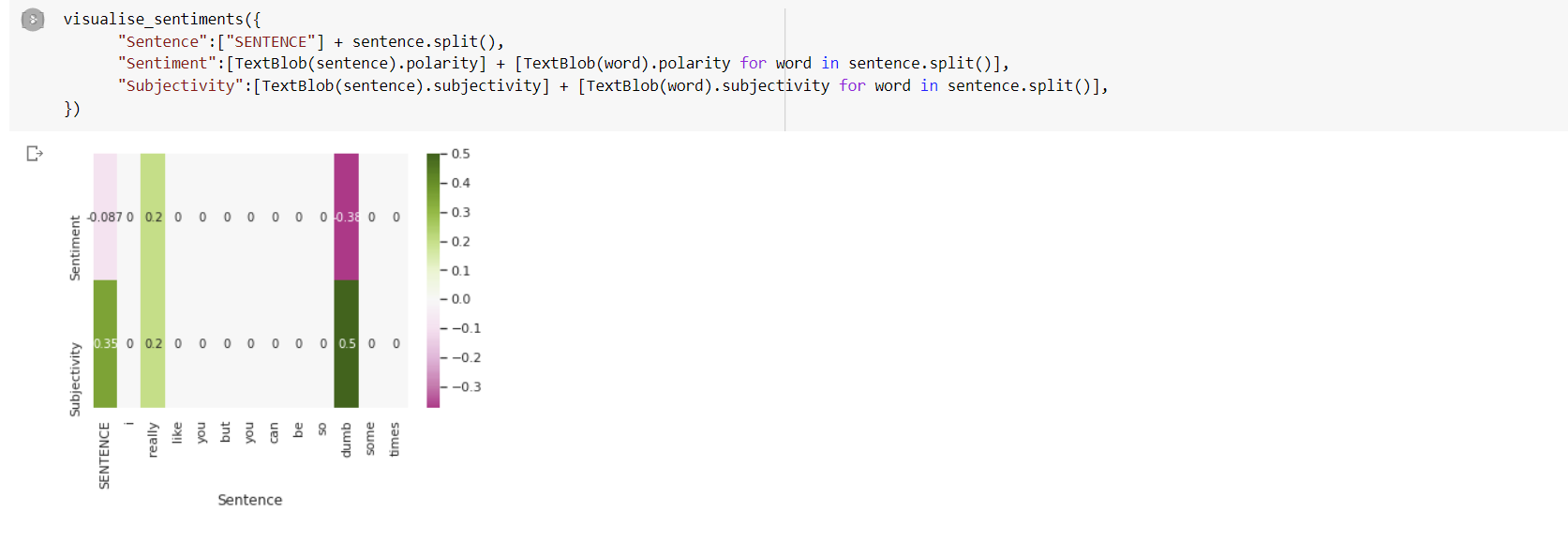


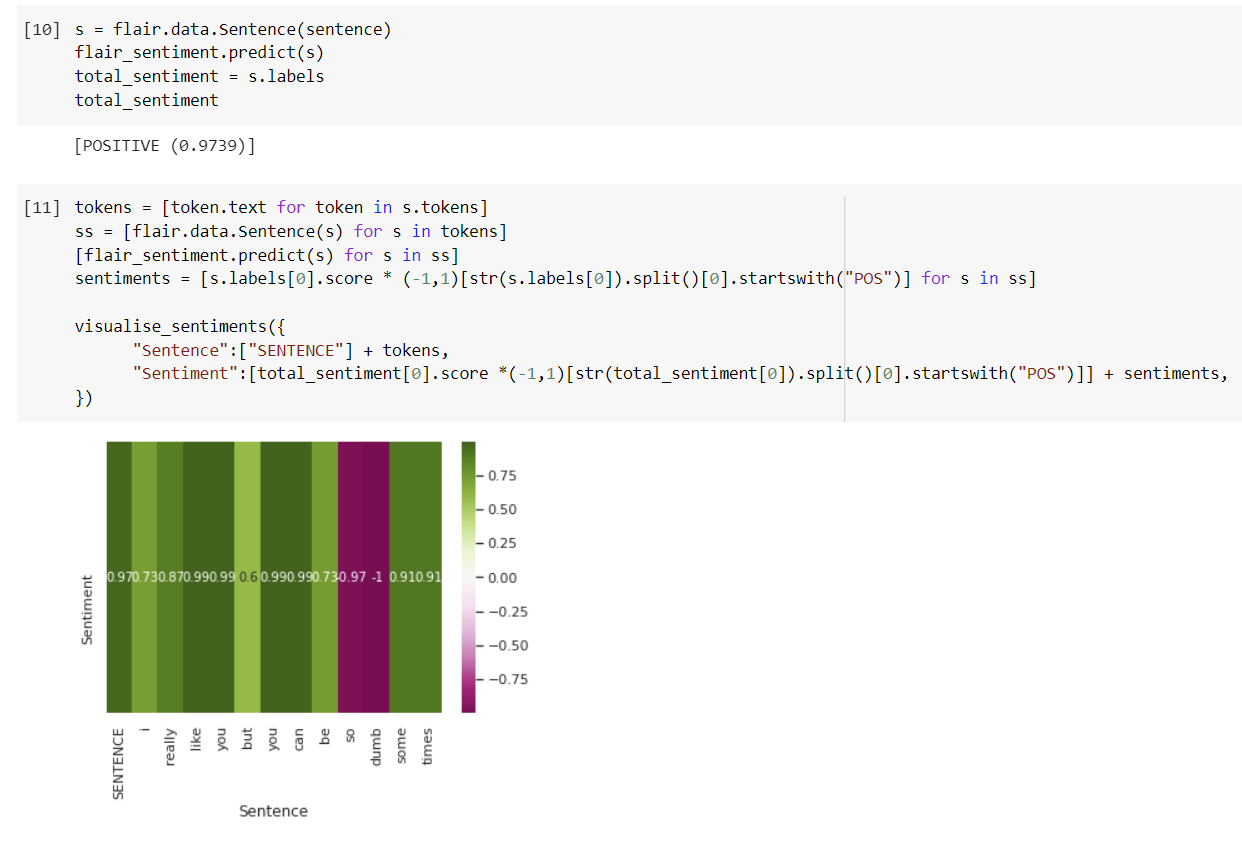




1. **NLP sentiment analysis:** [**https://colab.research.google.com/github/mohammedterry/NLP\_for\_ML/blob/master/Sentiment\_Analysis.ipynb**](https://colab.research.google.com/github/mohammedterry/NLP_for_ML/blob/master/Sentiment_Analysis.ipynb)

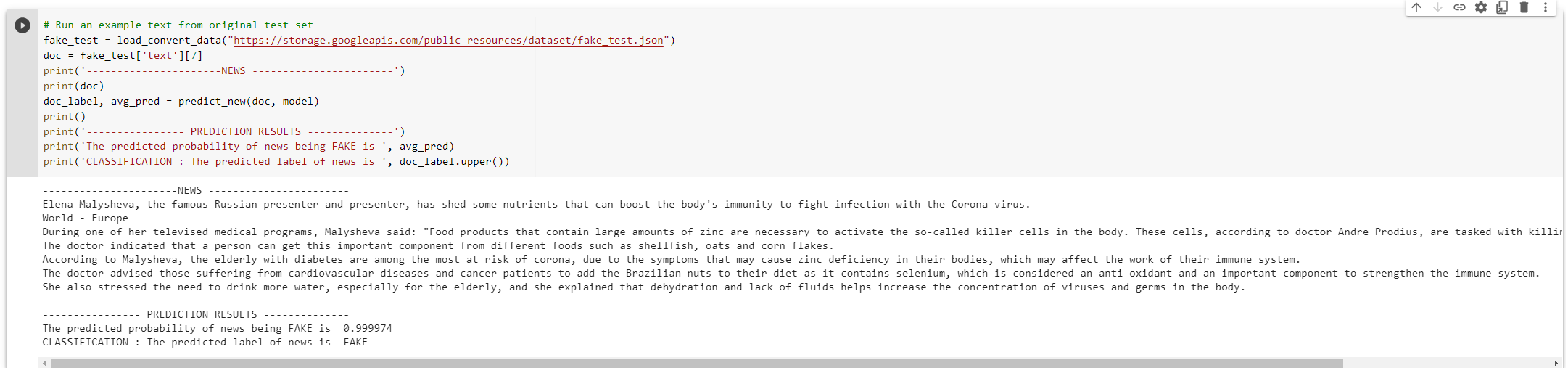






1. **A. Fake News detection with BERT**

[**https://colab.research.google.com/github/singularity014/BERT\_FakeNews\_Detection\_Challenge/blob/master/Detect\_fake\_news.ipynb#scrollTo=XzJ7vfkM0A1d**](https://colab.research.google.com/github/singularity014/BERT_FakeNews_Detection_Challenge/blob/master/Detect_fake_news.ipynb#scrollTo=XzJ7vfkM0A1d)



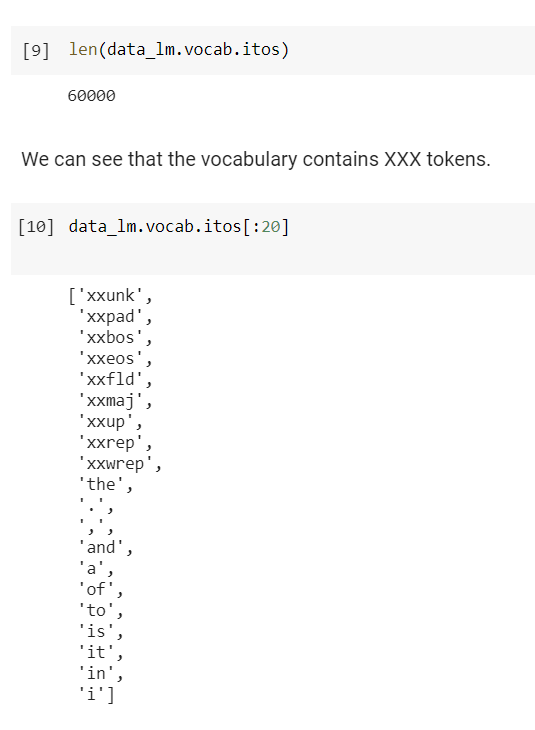
# **B. T5 for Sentiment Span Extraction**

[**https://colab.research.google.com/github/enzoampil/t5-intro/blob/master/t5\_qa\_training\_pytorch\_span\_extraction.ipynb#scrollTo=y1vih3VDqP\_p**](https://colab.research.google.com/github/enzoampil/t5-intro/blob/master/t5_qa_training_pytorch_span_extraction.ipynb#scrollTo=y1vih3VDqP_p)

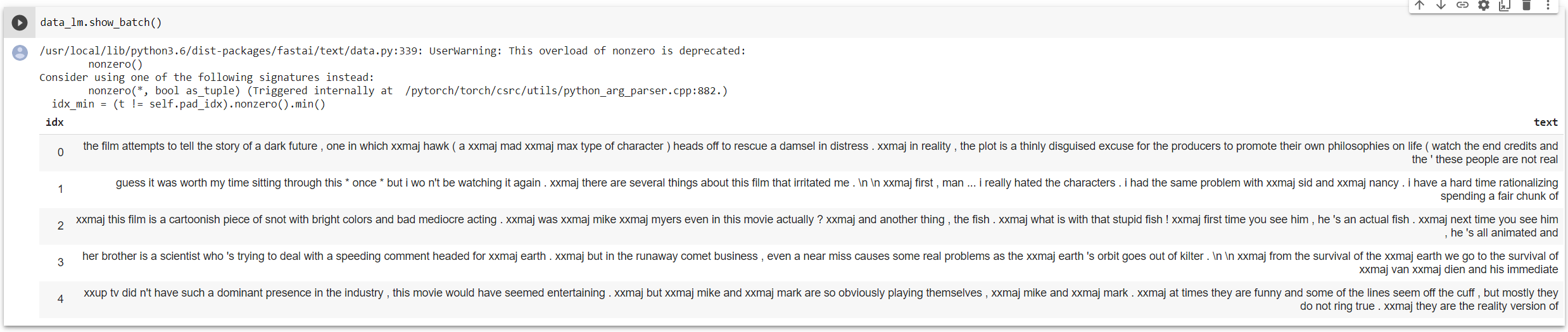


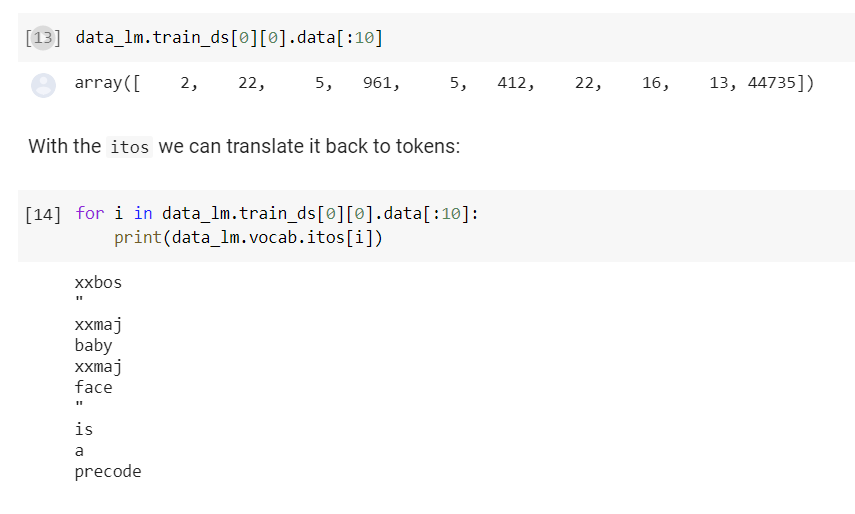
**3. NLP and Deep Learning**

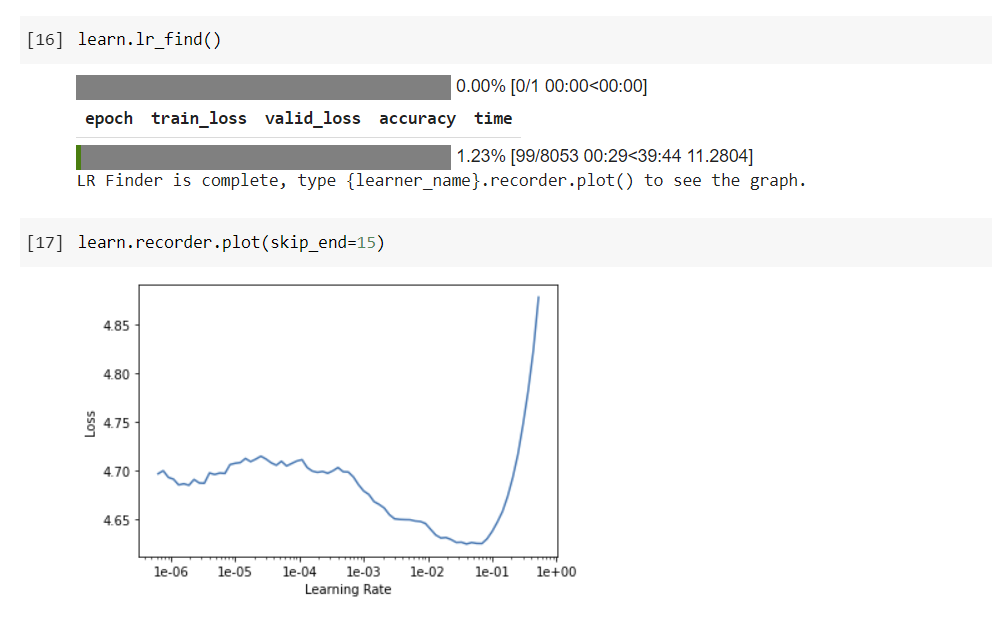
[**https://colab.research.google.com/github/lewtun/dslectures/blob/master/notebooks/lesson13\_nlp-deep.ipynb#scrollTo=cP7lpESaaBOX**](https://colab.research.google.com/github/lewtun/dslectures/blob/master/notebooks/lesson13_nlp-deep.ipynb#scrollTo=cP7lpESaaBOX)











**Failure because of usage limit of GPU / TPU**