NLP AstraZeneca HAckathon

Soorya Prakash K
3rd Year ECE with Specialization in Biomedical Engineering
VIT Vellore ,Tamil Nadu ,India.

The Need For Simple Solution

- This Task is intended to classify generic documents.
- Word Embedding will be not be a best feature because of technical terms that the model may encounter.
- There won't be any significant difference in the words used in the heading and section so embeddings is not a good choice.

How a Human will think to do this task

Factors human will consider :

- The Difference in the word and the character length between the above and below sentences.
- Capitalized or Not.
- New Lines used between the Sentences.

Generation of Features according to thinking of human.

	index	lines	char_length	no_of_words	index_	caps	space	next_char	next_words	next_next_char	next_next_words	next_caps	prev_char	prev_words	prev_prev_char	prev_prev_words	label	label2
0	0	Symphyotrichum lateriflorum	27	2	0	0.037037	1	0	0	37	5	0.000000	0	0	0	0	1.0	0.0
1	1	NaN	0	0	1	0.000000	0	37	5	0	0	0.054054	27	2	0	0	0.0	0.0
2	2	From Wikipedia, the free encyclopedia	37	5	2	0.054054	2	0	0	32	5	0.000000	0	0	27	2	0.0	0.0
3	3	NaN	0	0	3	0.000000	0	32	5	0	0	0.062500	37	5	0	0	0.0	0.0
4	4	Jump to navigationJump to search	32	5	4	0.062500	2	0	0	27	2	0.000000	0	0	37	5	0.0	0.0

Process

- 6 Wikpedia documents were taken and features were generated for them .
- Dataframe of documents were concatenated and fed to XGBoost Model to classify whether the sentence is heading or section.
- Output File is Generated.

Implementation Procedure

Open astra_pipeline and change the input and the output file names. Tadah!

Task Accomplished : Classifying Headings and Section in a document.