COMP30810 - INTRO TO TEXT ANALYTICS

> LECTURER: BINH THANH LE

> LAB05: N-Grams

TASKS:

1) Read the Harry Potter Series using the CSV of raw data from previous lab (LAB04)

2) Extract the new feature "Tokens" by using the token analysis as in previous lab.

3) Get the sample for Tokens and text (using the HP1)

Sample\_tokens = df\_handle['Tokens'].iloc[0]

Sample\_text = df\_handle['Content'].iloc[0]

4) First Bi-Gram analysis:

4.1. Generate bigram for HP1 text.

4.2. Should N-gram be generated by RAW TEXT or TOKENS? What do you think? Explain your opinions?

4.3 Get frequency of bigram of HP1 text. What do you think about the result?

4.4. Extract the noun-bigram using the grammar (NOUN, NOUN) and (ADJ, NOUN). What can you observe from the result? How can we get better

4.5. Why should we split the gram from text by using the space (" ")?

5) Generate Trigram for HP1 text. What do you think about the trigram?

6) Combine the all steps above to be one single function to extract n-gram from text.

Function: ngramExtraction

Input: Text: the raw text

N: number of grams

...

Output: result: vector of n-grams

7) Please apply the ngramExtraction for the rest books of Harry Potter to extract bigram only.