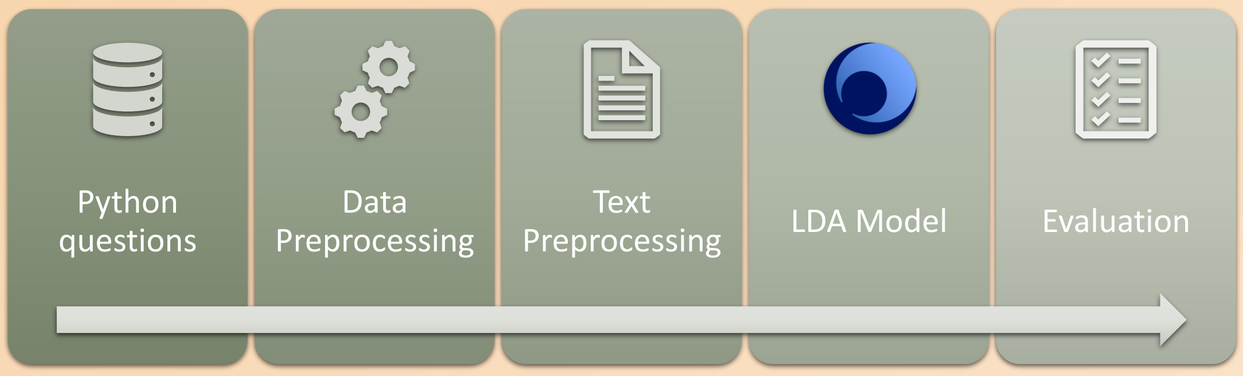
A Text Analytics Approach to Study Python Questions Posted on stackoverflow

Problem

1. How does the topics of Python questions posted on Stack Overflow **change** over the years?
2. What are the topics of Python questions with **high score** on Stack Overflow?

Data source: Kaggle

Proposed Solution

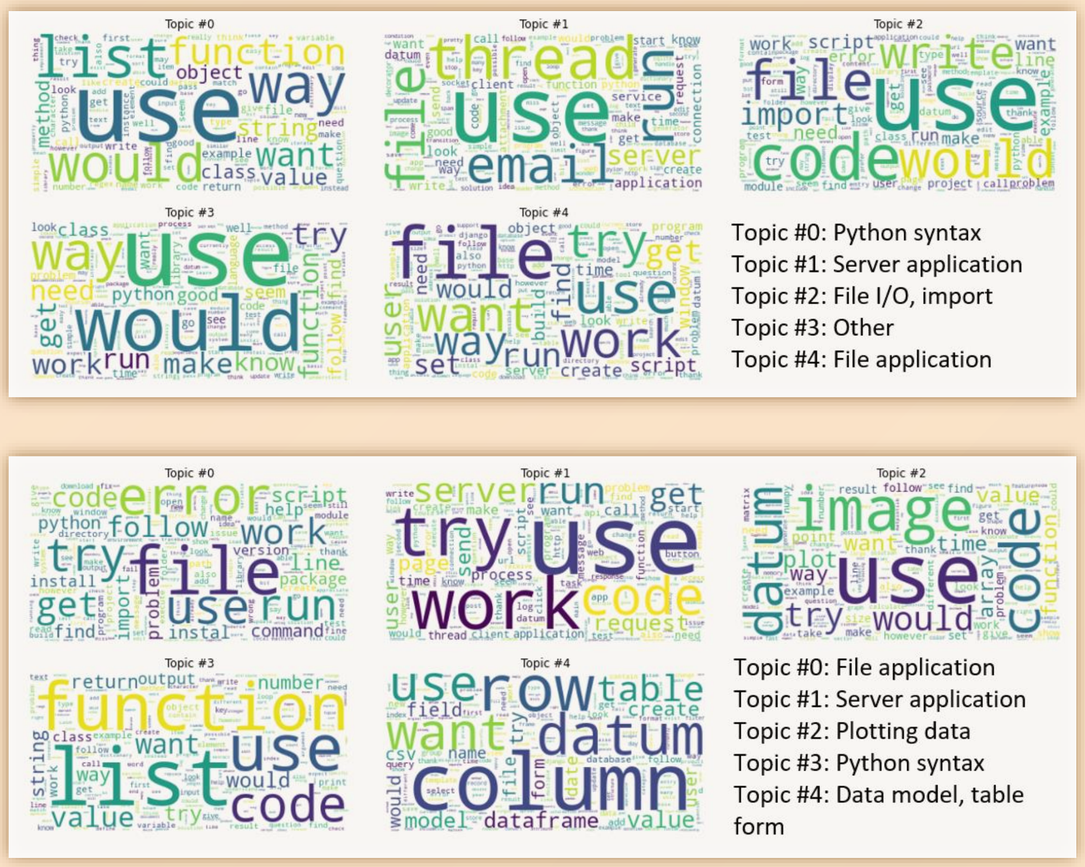


Text Preprocessing:
Remove punctuations, lowercase, HTML to normal text, tokenization, remove stop words, lemmatization.

Code in Python

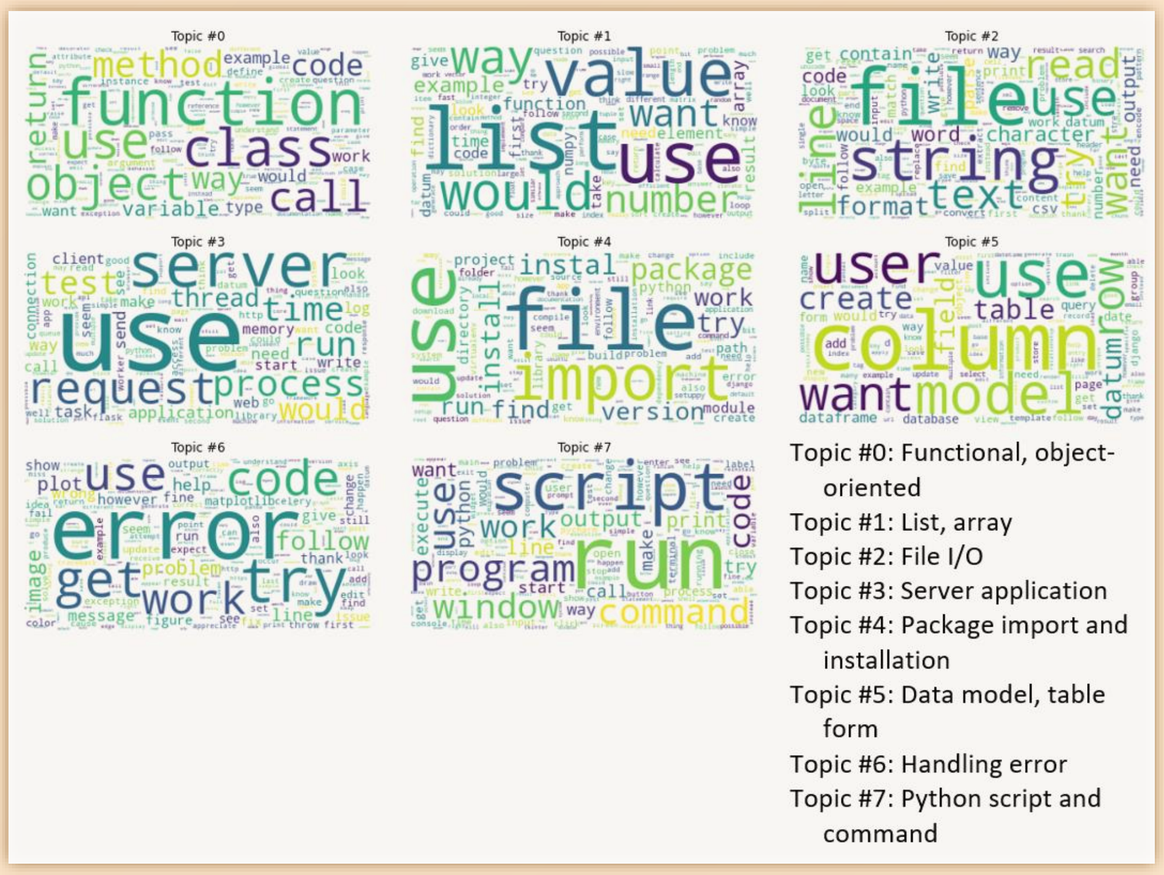
Experimental Results

Topic Model by Year (Top: 2008; Bottom: 2016)



Coherence score: **0.3752** (2008); **0.4432** (2016)

Topic Model for High Score Questions (2008-2016)



Coherence score: **0.4482**

Analysis and Findings

1. Topics shift towards data-related questions from 2008 – 2016.
2. LDA model with 8 topics can capture the topics among high score questions from 2008 – 2016.
3. The word “use” appears in all topics.

Conclusion

Topic modelling is an **unsupervised approach** to extract information from textual data.

Future works: remove domain-specific stop words, tune more hyperparameters to search for better topic model, study questions of different programming languages such as R.

Lee Yong Meng (P-COM0012/20) | Soo Yin Yi (P-COM0123/19)
School of Computer Sciences, Universiti Sains Malaysia, Penang.