**Text mining by using Python: Application to patent documents**

（2025 Project description）

Instructor: Kazuyuki Motohashi ([motohashi@tmi.t.u-tokyo.ac.jp](mailto:motohashi@tmi.t.u-tokyo.ac.jp)) and his lab members

1. Introduction

This project is for understanding practices of applying NLP techniques for technology management analysis using patent text information. The project starts with a course work for one week with some preparatory activities, followed by individual project works (mainly topic modeling of some particular type of technology) for some weeks.

1. Setting
   * PC with Python: You need to install Python 3.X “anaconda” environment. We use “Jupyter notebook” inside “anaconda” for this project. It can be downloaded freely.
   * I will let you access to the shared drive of Google Suite of my laboratory (motolabo.net). Python\_Text\_Mining holder contains
     + Some files used in this project (including this file)
     + Coursera\_video directory: Michigan U, “Applied Text Mining by Python”, referred as “VIDEO” here after
     + Data for topic modeling project

* Please check whether you could access to the following holder. If it is not the case, report to prof. Motohashi with your name and e-mail address.

<https://drive.google.com/drive/folders/1eIbsqmaCnY8tjryYRnriZ2Wu_stQl7g9?usp=shari>ng

The folder of “Internahip\_files” is only for those who did the phase 2 of this project, “innovation discovery from patent and web contents information”

1. Preparatory activities (Will be introduced at the orientation day)
   * Preparatory activities to start course work is provided in the appendix 1.
2. Whole structure

* Part 1 : June 9- June 13 : course works with practical exercise
* Part 2 : June 16 – August 1 : internship project with real data

**[Phase 1]: Learning text mining technique by small sample data : June 9 – June 13 (Orientation will take place on June 2, as follows)**

Venue : Zoom Conference Room : https://u-tokyo-ac-jp.zoom.us/j/99083411670?pwd=U0dicHlONXNiUjBmUFU3Y1h6NWFZdz09

**Orientation Day: June 2, 10:00-12:00 Thai& Vietnam Time (12:00-14:00, JST)**

* Explanation of home works etc (Appendix 1)
* Web crawling introduction by Suchit

**Day1: June 9, 10:00-12:00 Thai& Vietnam Time (12:00-14:00, JST)**

* Review of **preparatory activities**
  + How to convert technology contents info (in patent document) to numeric (vector) information? : Keyword extraction, TF-IDF, Word/Document embedding (such as Word2Vec)
  + **HOMEWORK1**: Extract three **keywords by using TF-IDF score from 25 AI patents abstracts**, supplied as a home work file. TF-IDF score can be obtained easily by Python module like “scikit-learn” and “genism”. But I would advice you to do it without such modules for your understanding of the concept. There are many hints over internet such as

<https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3>

**Day 2: June 11, 10:00-12:00 Thai& Vietnam Time (12:00-14:00, JST)**

* Review of homework of Day 1
  + What are ML applications of patent text information?
  + **Preprocessing of text**: tokenization regularization (lowering character), stemming/lemmatization and stop words exclusion
  + **Text processing of tf-idf vectors:** 
    1. Create dictionary: mapping every word to a number
    2. Corpus (list of bags of words) : a list of number of words occurring in each documents
  + **HOMEWORK2**: Calculation of **similarity measures** across each documents, and find out the closed documents of each of 25 AI patents : Now please use genism module, again there are many hints over internet such as “How do I compare document similarity using Python?” by Jonathan Mugan

<https://www.oreilly.com/content/how-do-i-compare-document-similarity-using-python/>

**Day 3: June 13 10:00-12:00 Thai& Vietnam Time (12:00-14:00, JST)**

* + Review of homework of Day2
  + Understanding **technology trend by patent information (clustering vs topic modeling)**
  + Watch VIDEO 16&17
    1. Understanding the concept of topic modeling
    2. LDA model by using genism module
  + ~~Introduction for your~~ **~~project work~~** ~~: Pick up one of subject areas shown in the appendix 2 to construct LDA topic model, then understand the contents of technological development of that field. More instructions will be provided later with individual consultation of works to be done.~~
  + Some references
    1. There is a very easy to read document of LDA model available in the shared drive. (Beginners\_Guide\_Topic\_Modeling.doc)
    2. Good reference about topic modeling by genism <https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/>

**[Phase 2]: Project work with final presentation : June 16 – August 1**

Orientation Day: June 16, 10:00-12:00 Thai& Vietnam Time (12:00-14:00, JST)

* Orientation of phase 2
* Dual Attention Model

**(Outline of works, details to be announced)**

* This process is using NLP techniques above to real world exercises, such as technology opportunity discovery tasks. Please refer to the following Youtube video for this research project.

<https://www.youtube.com/watch?v=Y6TIHxfKsmM&t=581s>

* Major tasks are scraping company’s web pages and using such text information together with patent information to predict market opportunities based on particular type of technology.
* Web scraping task will start after June 2 orientation (with a short course of web scraping)
* Some research assistant honorarium will be offered to this works.

**Appendix1 Preparatory Homework**

1. Python environment and understanding on basic command

* You need to bring your own PC with Python 3.X anaconda environment, particularly Jupyter notebook (download for free)
* Familiarize yourself with basic Python function dealing with text, such as making word list from sentence, word count and distribution
* Playing with NLTK library of Python

1. Tasks
   * Watch **Video 01,05 and 06** in “Coursera\_Video” directory) for the following contents
     + What is National Language Processing?
     + Playing with NLTK (Natural Language Processing Tool Kit)
   * Use the patent abstract data (25 AI related patent abstracts, provided as “homework.csv”), to calculate
     + How many sentences, words, unique words
     + List of words occurring no less than 10 times **and** no less than 5 character lengths
     + After regularization (lowering characters), then comparing the case of Porter Stemmer and Lenmatizer. Compare the results and which one would you think it is better and why?
     + Any other observations?
2. Notes

* You may go ahead to watching other videos, but I will give you some general directions.
  + - **VIDEO 02 and 03** are for regular expression, working with Pandas module of python (concept of “data framework”) : No need to worry about them for now, but it will be useful for **rule based text manipulations**.
    - **VIDEO 04** is for character code information for various languages (Python’s default character set is UFT-8, no need to worry about this for now either.
    - Watching VIDEO 07 is optional (could be helpful for your subsequent tasks)