Put Things Together

Pham Quang Nhat Minh
(FPT Technology Research Institute)

What we have learnt so far…

- We have learnt many things so far
  - Web crawling with scrapy
  - Text processing with nltk
  - Topic modeling and using lda-c for mining topics from a set of documents
  - Key-phrase extraction by using the TextRank algorithm

What are the next steps?

- Recall to the objectives of the project:
    - Mining topics, key-phrases from a set of documents about some technologies
        - Big-data
        - Internet-of-Things (IoT)
        - Machine Learning
        - Natural Language Processing (NLP)
        - Data Science
- Analyse the result
- Make a Data Visualisation to communicate the results
- Make a web-based demo for the course project

What are the next steps?

- We need to go step by step and put things together

- In this lecture
  - We review steps that we need to do in the project
  - Provide hints/suggestions to complete the project

First step: Web Crawling

- Choose a data source to crawl data
- As an example, I use the url: https://techcrunch.com/tag/big-data
- In last lecture, we have learnt how to get links from that starting url (See the **`code big_dat_spider.py`** on the github)
- Now we need to extract text from these links
  - A simple way to do that is add urls to the start_urls in the code **`web_content_spider.py`**
    - But you should do in **an automatic way**

First step: Web Crawling

- You may need to crawl web documents from multiple data sources
    - Look for newspapers about technology
        - Techmeme, MIT Technology Review, etc
- Data sources may have different structures, so
    - Please customise your web crawler

Next step: Get the text from a file in JSON Line format

- JSON Line format
  - {"link": "https://techcrunch.com/2013/08/15/panorama-education-wants-to-make-polling-parents-students-and-teachers-easier-for-educators/","content": "New Haven-based startup and current Y Combinator Summer 2013 participant \n is looking to address a major pain point for educators, students and parents around school with its polling app targeted at K-12 students….}
  - Note that \n is the newline character

How to read JSON files in python

- In Python, we can do something like this

```
import json
data = []
with open('file') as f:
    for line in f:
        data.append(json.loads(line))
```

Let's try

- Let try on the interactive **`python shell`**
  - Go to the directory technews
  - Type **`ipython`** or **`python`**
  - We will read each line in a JSON-Line file
  - Load each line into a Python object

Now it is time to write a script to get contents from JSON files

- You need to content each line in the JSON file into a document

  - Take care about newline character \n

  - Save documents in plain-text files

Next step: clean the data

- The file may contain URLs, or meaningless characters
    - You need to remove them

Preprocessing

- Will will use nltk to do:
  - Text tokenisation
  - POS Tagging (may be)

Topic mining: convert to LDA format

- Input: A set of documents

- Output:

  - **`vocab.txt`**: vocabulary of unique words in the document collection

  - **`technews.dat`**: data file in lda format

Format required by LDA-C

- `186 0:1 6144:1 3586:2 3:1 4:1 1541:1 8:1 10:1 3927:1 12:7 4621:1 527:1 9232:1 1112:2 20:1 2587:1 6172:1 10269:2 37:1 42:1 3117:1 1582:1 1585:3 435:1 926 8:3 571:2 60:1 61:1 63:2 64:2 5185:1 11:1 4683:1 590:2 1103:2 592:1 5718:1 1623:2 1624:4 89:2 …`
- The data is a file where each line is of the form:

  `[M] [term_1]:[count] [term_2]:[count] ...  [term_N]:[count]`

- where
  - [M] is the number of unique terms in the document, and
  - the [count] associated with each term is how many times that term appeared in the document. Note that [term_1] is an integer which indexes the term; it is not a string.

Let’s do converting with our example data

- First step: Extracting vocabulary from a set of documents
- We use Python dictionary to store uniques in the documents

```
For each document in the set
  For each line in the document
    For each word w in the line
      vocab[w] = 1
Print words (keys of the dict) to the output file
```

LDA Format conversion

- Now using the generated the vocabulary to get .dat file
- We also need to iterate each file in the document collection

Next step: Running LDA tool

- We will use lda-c implementation of LDA
  - http://www.cs.columbia.edu/~blei/lda-c/index.html
- We may want to try different values of parameters

Visualise top words of each topic

- Show top words of each topic