

4. Assignment IV: Chinese Language Processing

4.1. Question 1

The csv file `dcard-top100.csv` includes top 100 posts from Dcard, which a on-line discussion forum for school life in Taiwan. The texts are in the `content` column.

Please preprocess the data by:

- removing symbols, punctuations, emoticons or other non-linguistic symbols
- removing stopwords (Please use the stopwords list provided in `demo_data/stopwords/tomlinNTUB-chinese-stopwords.txt`)
- performing word segmentation on the corpus using `ckip-transformer`
- creating a word frequency list of this tiny corpus
- including only word tokens which have at least two characters in the frequency list

⚠ Warning

Please note that the preprocessing steps are important. Removal of characters from texts may have a lot to do with the word segmentation performance.

	LEMMA	FREQ
0	真的	115
1	沒有	92
2	覺得	90
3	知道	70
4	看到	67
5	現在	63
6	喜歡	56
7	朋友	54
8	其實	52
9	一直	52
10	不會	51
11	發現	43
12	男友	42
13	一下	41
14	已經	41
15	很多	40
16	時間	40
17	工作	40
18	分享	39
19	感覺	39
20	一起	38

4.2. Question 2

Use `ckip-transformer` to extract all named entities and create a frequency list of the named entities.

In particular, please identify named entities of organizations (`ORG`) and geographical names (`GPE`) and provide their frequencies in the Dcard Corpus.

☰ Contents

[4.1. Question 1](#)

[4.2. Question 2](#)

[4.3. Question 3](#)

[4.4. Question 4](#)

Print to PDF

	LEMMA	FREQ
0	台灣	23
1	日本	18
2	台南	7
3	台	6
4	英國	5
5	台中	4
6	沖繩	4
7	韓國	4
8	聖圭	4
9	台北	3
10	德國	3
11	台大	2
12	杜克大學	2
13	東京	2
14	SHINee	2
15	美	2
16	板橋	2
17	武林	2
18	巴黎	2
19	鹿港	2
20	Celine	2


4.3. Question 3

In this exercise, please work with `spacy` for Chinese processing. (Use the model `zh_core_web_trf`)


Please process the same Dcard Corpus (from the csv file) by:

- performing the word tokenization
- identifying all nouns and verbs (i.e., words whose tags start with N or V)
- identifying all words with at least two characters
- removing all words that contain alphabets or digits
- removing all words that are included in the `stopword_list` (cf. Question 1)

Based on the above text-preprocessing criteria, your goal is to create a word frequency list and visualize the result in a Word Cloud.

 Note

`spacy` uses the `jieba` for Chinese word segmentation. There may be more tagging errors. In the expected results presented below, I did not use any self-defined dictionary. For this exercise, please ignore any tagging errors out of the module for the moment.

 Tip

Please check the module `wordcloud` for the visualization.

	N-V	FREQ
0	知道	70
1	看到	67
2	朋友	55
3	喜歡	55
4	分享	43
5	男友	42
6	沒有	41
7	工作	40
8	很多	39
9	第一	38
10	感覺	38
11	時間	37
12	發現	36
13	==	36
14	希望	35
15	感情	33
16	今天	33
17	蛋糕	32
18	部分	30
19	出去	30
20	想要	30



4.4. Question 4

Following Question 3, after you process each article with `spacy`, please extract all the `subject` + `predicate` word pairs from the corpus.

To simplify the task, please extract word token pairs whose dependency relation is `nsubj`, with the predicate being the head and subject being the dependent.

- Remove words that include alphabets and digits

	SUBJ-PRED	FREQ
0	我_喜歡	21
1	他_說	20
2	我_想	19
3	我_覺	19
4	我_知道	16
5	我_看	14
6	我_看到	11
7	我_用	10
8	我_說	8
9	大家_好	8
10	我_愛	8
11	我_有	7
12	你_有	7
13	我_去	6
14	他_覺	6
15	我_要	6
16	我_決定	6
17	她_說	6
18	我_發現	5
19	我_問	5
20	我_在	5

By Alvin Chen
© Copyright 2020 Alvin Chen.