IRTM PA1

B09705017 資管三 王紹安

Env Setup

Use python 3.10.6 and poetry as dependency management

1. You may install poetry from this link
2. Run the following command under submitted directory.

```
poetry env use python
poetry install
poetry shell
```

3. You may run

```
python PA1.py
```

and see result.

- **Note: the `requirements.txt` file is auto generated by poetry through `poetry export -f requirements.txt -o requirements.txt --without-hashes`, you may try `pip install -r requirements.txt` but the environment can’t be promised to be the same as mine.**

Source Code Logic

1. Get document

- Use python’s `requests` package to retrieve source document.

2. Tokenization

- Use python’s `split` method to split document into tokens, delimiters are `[\r\n, \n, ., ', "]`

3. Lowercasing

- Use python’s built-in method `lower()` to lowercase the tokens.

4. Stemming Using Porter’s Algorithm

- Use `nltk`'s `PorterStemmer` to stem the tokens.

5. Stop words removal

- Stop words are downloaded from `nltk.corpus.stopwords.words('english')`
- Remove stop words by

```
def stopword_removal(words):
    return [word for word in words if word not in stopwords]
```

6. Export Result to result.txt

- Use python’s built-in method `open` and `write` to export results to `result.txt`