Lab 3 VSM Exercise

COMP4901K and MATH 4824B

Fall 2018

Prerequisites

1. You need to have some background knowledge about Vector Space Model (VSM). If not, you can check out:

- https://en.wikipedia.org/wiki/Vector_space_model
- https://en.wikipedia.org/wiki/Tf–idf
- https://en.wikipedia.org/wiki/Cosine_similarity

2. You need to install the Numpy packages:

```
pip3 install --upgrade numpy
```

1 Assignment

You need to download the following file(s) from canvas:

- `lab3_skeleton.py`: the program skeleton.

Q1 Write code to perform the following tasks:

1. Build a vocabulary based on given corpus, and turn raw sentences into Bag-of-words representation.
2. Implement cosine similarity function. And output pairwise similarity of BOW.
3. Implement TFIDF function to turn Bag-of-words representation into TFIDF representation.And output pairwise similarity of TFIDF(hint: the tfidf equation is as follows).

$$\begin{aligned} \text{tf}(t,d) &= f_{t,d} \Big/ \sum_{t' \in d} f(t',d) \\ \text{idf}(t) &= N/n_t \\ \text{tf-idf}(t,d) &= \text{tf}(t,d) \times \text{idf}(t) \end{aligned}$$

2 Submission

You need to submit two files, program output and your python script. After you finished the assignments, make sure you include the header information in the beginning of your code

```
# author: Your_name
# student_id: Your_student_ID
```

Copy all the program output in to text file named `StudentID_lab3_output.txt`, and submit with your python script solution named `StudentID_lab3.py` to canvas.