CS439 Intro to Data Science – Assignment Vector Representation Report

Dingbang Chen

dc1019

Task1

For task 1 of this assignment, my firstname, which is dingbang, cannot be found in those files. In this case, I choose kobe as my test case to show the nearest 5 words using the respective cosine formula. (I choose glove.6B.50d.txt for example)

Result:

```
chendinangdeAir:Word_Representation DingbangChen$ python Word_Representation.py
-f glove.6B.50d.txt
Enter your firstname:kobe
found it
Here are your 5 nearest neighbours of your first name
('vissel', 0.7551206391465656)
('northridge', 0.6840131228173619)
('shaq', 0.6805795847665419)
('shaquille', 0.6552255590121028)
('hiroshima', 0.6464105957120923)
```

Task 2

For task 2 My sentence S_0 is “computer science is the best”. It will show the nearest 5 words with this sentence by getting cosine scores with words.

Result:

```
Enter your sentence: computer science is the best
Here are your 5 nearest neighbours of your sentence
('as', 0.8785276763275432)
('this', 0.8760077038533459)
('same', 0.8726450107376488)
('.', 0.8663728489372725)
('one', 0.8651825893282021)
```

Task 3

I chose a similar sentence with previous sentence, “computer science is very good”. It got high cosine similarity scores with S_0. Then, I chose a dissimilar sentence with S_0, “Donald Trump rules the world”. It is obvious that the second sentence has nothing to do with S_0. So it got lower scores than S_1. So it shows that cosine similarity scores are reasonable in finding similar words and sentences.

Result: