Name: Gerosh George

19BCE1403

22/9/2021

Inverted Document Indexing

DOC1:

On February 14, 2019, a terror attack was carried out in Pulwama in Jammu and Kashmir by a suicide bomber resulting in the death of 40 CRPF personnel.

DOC2:

The Indian Air Force early Tuesday morning destroyed a major terror training camp of the Jaish-e-Mohammed in Balakot, Pakistan. Several terror trainers and suicide bombers were killed in the attack that lasted just eight minutes.

DOC3:

IAF releases proofs of shooting down Pakistan's F-16 fighter jet.

PROGRAM CODE:

```
import re

documents =['doc1','doc2','doc3']
index={}

for id,doc in enumerate(documents):
    filename = doc+".txt"

    with open(filename,'r') as fp:
        data = "".join(fp.readlines())
        data = data.lower()
        ext_words = re.findall(r"([a-z0-9-]+)",data)
        for pos,word in enumerate(ext_words):

            if word[-1]=='s':
                if word[:-1] in index:
                    word = word[:-1]
                elif word[:-2] in index:
                    word = word[:-2]

            if word not in index:
                index[word]={   "freq":1,
                                "listing": [(id+1,pos)]
                            }
            else:
                index[word]['freq']+=1
                index[word]['listing'].append((id+1,pos))


from collections import OrderedDict
index = OrderedDict(sorted(index.items()))
with open("inverted.txt",'w') as fp:
    for key in index:
        print(f"{key} : {index[key]}")
        fp.write(f"{key} : {index[key]}\n")
```

OUTPUT:

```
(venv) PS C:\Users\Gerosh\Desktop\VIT\Third Year\Web Minning\Programs> python inverted.py
14 : {'freq': 1, 'listing': [(1, 2)]}
2019 : {'freq': 1, 'listing': [(1, 3)]}
40 : {'freq': 1, 'listing': [(1, 25)]}
a : {'freq': 3, 'listing': [(1, 4), (1, 17), (2, 8)]}
air : {'freq': 1, 'listing': [(2, 2)]}
and : {'freq': 2, 'listing': [(1, 14), (2, 22)]}
attack : {'freq': 2, 'listing': [(1, 6), (2, 29)]}
balakot : {'freq': 1, 'listing': [(2, 17)]}
bomber : {'freq': 2, 'listing': [(1, 19), (2, 24)]}
by : {'freq': 1, 'listing': [(1, 16)]}
camp : {'freq': 1, 'listing': [(2, 12)]}
carried : {'freq': 1, 'listing': [(1, 8)]}
crpf : {'freq': 1, 'listing': [(1, 26)]}
death : {'freq': 1, 'listing': [(1, 23)]}
destroyed : {'freq': 1, 'listing': [(2, 7)]}
down : {'freq': 1, 'listing': [(3, 5)]}
early : {'freq': 1, 'listing': [(2, 4)]}
eight : {'freq': 1, 'listing': [(2, 33)]}
f-16 : {'freq': 1, 'listing': [(3, 8)]}
february : {'freq': 1, 'listing': [(1, 1)]}
fighter : {'freq': 1, 'listing': [(3, 9)]}
force : {'freq': 1, 'listing': [(2, 3)]}
iaf : {'freq': 1, 'listing': [(3, 0)]}
in : {'freq': 5, 'listing': [(1, 10), (1, 12), (1, 21), (2, 16), (2, 27)]}
indian : {'freq': 1, 'listing': [(2, 1)]}
jaish-e-mohammed : {'freq': 1, 'listing': [(2, 15)]}
jammu : {'freq': 1, 'listing': [(1, 13)]}
jet : {'freq': 1, 'listing': [(3, 10)]}
just : {'freq': 1, 'listing': [(2, 32)]}
kashmir : {'freq': 1, 'listing': [(1, 15)]}
killed : {'freq': 1, 'listing': [(2, 26)]}
lasted : {'freq': 1, 'listing': [(2, 31)]}
major : {'freq': 1, 'listing': [(2, 9)]}
minutes : {'freq': 1, 'listing': [(2, 34)]}
morning : {'freq': 1, 'listing': [(2, 6)]}
```

```
of : {'freq': 3, 'listing': [(1, 24), (2, 13), (3, 3)]}
on : {'freq': 1, 'listing': [(1, 0)]}
out : {'freq': 1, 'listing': [(1, 9)]}
pakistan : {'freq': 2, 'listing': [(2, 18), (3, 6)]}
personnel : {'freq': 1, 'listing': [(1, 27)]}
proofs : {'freq': 1, 'listing': [(3, 2)]}
pulwama : {'freq': 1, 'listing': [(1, 11)]}
releases : {'freq': 1, 'listing': [(3, 1)]}
resulting : {'freq': 1, 'listing': [(1, 20)]}
s : {'freq': 1, 'listing': [(3, 7)]}
several : {'freq': 1, 'listing': [(2, 19)]}
shooting : {'freq': 1, 'listing': [(3, 4)]}
suicide : {'freq': 2, 'listing': [(1, 18), (2, 23)]}
terror : {'freq': 3, 'listing': [(1, 5), (2, 10), (2, 20)]}
that : {'freq': 1, 'listing': [(2, 30)]}
the : {'freq': 4, 'listing': [(1, 22), (2, 0), (2, 14), (2, 28)]}
trainers : {'freq': 1, 'listing': [(2, 21)]}
training : {'freq': 1, 'listing': [(2, 11)]}
tuesday : {'freq': 1, 'listing': [(2, 5)]}
was : {'freq': 1, 'listing': [(1, 7)]}
were : {'freq': 1, 'listing': [(2, 25)]}
(venv) PS C:\Users\Gerosh\Desktop\VIT\Third Year\Web Minning\Programs>
```