IDC409 Assignment 2: Keyword identification

Athira Sreejith (MS18033)
Priyansha (MS19094)
Sourav S (MS18084)

October 31, 2021

1 Problem

The top 10 keywords of 4 articles are identified using TF-IDF and TextRank methods available in the Python libraries like Gensim and NLTK. The keywords are then compared with the author-provided keywords.

1.1 Articles used

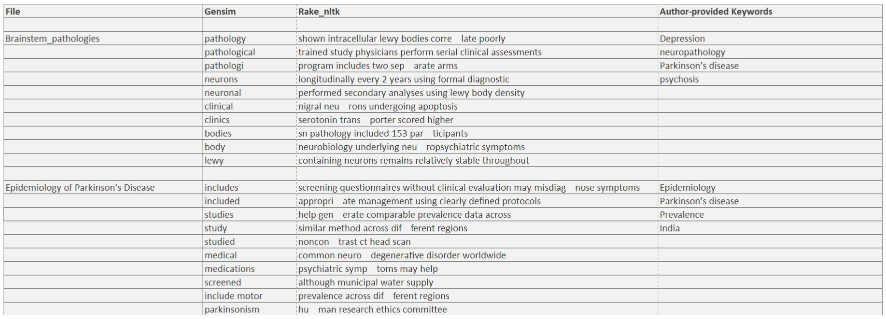
- Brainstem Pathologies Correlate With Depression and Psychosis in Parkinson's Disease
- Epidemiology of Parkinson's Disease in Rural Gujarat, India
- Hydroxytyrosol as anti-parkinsonian molecule: Assessment using in-silico and MPTP-induced Parkinson's disease model
- One Year Trajectory of Caregiver Burden in Parkinson's Disease and Analysis of Gender-Specific Aspects

1.2 Code

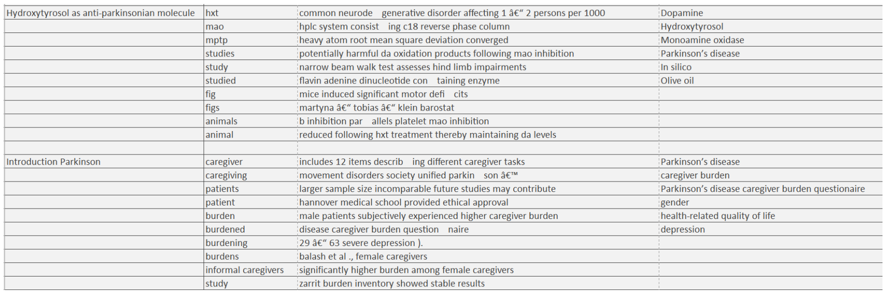
```python
#!/usr/bin/python3

#importing relevant packages
from gensim.summarization import keywords
from rake_nltk import Rake
import csv


#names of the article files
files = ['Brainstem_pathologies','Epidemiology of Parkinsons Disease','Hydroxytyrosol as anti-parkinsonian molecule','Introduction Parkinson']

#top ten keywords of each method gets stored here
gensim_out = []
rake_nltk_out = []


for i in files:

  with open('{}.txt'.format(i),'r',encoding='utf-8') as file:

    text = file.read()

    #using Gensim for keywords
    gensim_out.append(keywords(text).split('\n')[:10])

    #using Rake_nltk for keywords
    nltk_var = Rake()

    nltk_var.extract_keywords_from_text(text)

    rake_nltk_out.append(nltk_var.get_ranked_phrases()[:10])


#writing the output as csv
with open('out.csv','w') as output:

  writer = csv.writer(output)

  writer.writerows([['file'],['gensim'],['rake_nltk'],['']])

  for i in range(len(files)):

    writer.writerows([[files[i]],gensim_out[i],rake_nltk_out[i],[]])
```

1.3 Results and observations

The top 10 Keywords obtained by our code and the keywords provided by the authors are given below.



File	Gensim	Rake_nltk	Author-provided Keywords
Brainstem_pathologies	pathology	shown intracellular lewy bodies corre late poorly	Depression
	pathological	trained study physicians perform serial clinical assessments	neuropathology
	pathologi	program includes two sep arate arms	Parkinson’s disease
	neurons	longitudinally every 2 years using formal diagnostic	psychosis
	neuronal	performed secondary analyses using lewy body density	
	clinical	nigral neu rons undergoing apoptosis	
	clinics	serotonin trans porter scored higher	
	bodies	sn pathology included 153 par ticipants	
	body	neurobiology underlying neu ropsychiatric symptoms	
	lewy	containing neurons remains relatively stable throughout	
Epidemiology of Parkinson's Disease	includes	screening questionnaires without clinical evaluation may misdiag nose symptoms	Epidemiology
	included	appropri ate management using clearly defined protocols	Parkinson’s disease
	studies	help gen erate comparable prevalence data across	Prevalence
	study	similar method across dif ferent regions	India
	studied	noncon trast ct head scan	
	medical	common neuro degenerative disorder worldwide	
	medications	psychiatric symp toms may help	
	screened	although municipal water supply	
	include motor	prevalence across dif ferent regions	
	parkinsonism	hu man research ethics committee	



Hydroxytyrosol as anti-parkinsonian molecule	hxt	common neurode generative disorder affecting 1 â€“ 2 persons per 1000	Dopamine
	mao	hplc system consist ing c18 reverse phase column	Hydroxytyrosol
	mptp	heavy atom root mean square deviation converged	Monoamine oxidase
	studies	potentially harmful da oxidation products following mao inhibition	Parkinson’s disease
	study	narrow beam walk test assesses hind limb impairments	In silico
	studied	flavin adenine dinucleotide con taining enzyme	Olive oil
	fig	mice induced significant motor defi cits	
	figs	martyna â€“ tobias â€“ klein barostat	
	animals	b inhibition par allels platelet mao inhibition	
	animal	reduced following hxt treatment thereby maintaining da levels	
Introduction Parkinson	caregiver	includes 12 items describ ing different caregiver tasks	Parkinson’s disease
	caregiving	movement disorders society unified parkin son â€™	caregiver burden
	patients	larger sample size incomparable future studies may contribute	Parkinson’s disease caregiver burden questionaire
	patient	hannover medical school provided ethical approval	gender
	burden	male patients subjectively experienced higher caregiver burden	health-related quality of life
	burdened	disease caregiver burden question naire	depression
	burdening	29 â€“ 63 severe depression).	
	burdens	balash et al ., female caregivers	
	informal caregivers	significantly higher burden among female caregivers	
	study	zarrit burden inventory showed stable results	

We did not observe any relevant keywords missed by the authors.

2 Contribution

All the group members contributed equally to this task.