**Homework 2: Exploring Common Topics in Health Research News  
By Shivani Sanjay Mahaddalkar**

**Introduction**

The Health Research News Headlines (HW2\_headlines\_10000.csv) is a data set that sampled 10,000 headlines from health research press releases posted on the EurekAlert! Website. This analysis aims to understand broader topics the headlines seem to fall under. This explores the common theme of health research news.

**Method 1: Cosine similarity**

On comparing headlines’ cosine similarities, we can find out which headlines are most like the others. We want to find the ten headlines that have the highest cosine similarity with the third headline. Using the cosine\_similarity package, we find the cosine similarity between each headline with the third headline in the data. For health related terms, it makes more sense to use ngram range from 1 to 2, as a lot of common scientific terms are two word terms.

On using a simple vectorizer that counts the term frequency we get the following result.

The third headline in the data is: ‘Anti-inflammatory drugs do not improve cognitive function in older adults’

The ten headlines having the highest cosine similarity are:

1. The link between cognitive function and sexuality in older adults
2. Anti-inflammatory drugs ineffective for prevention of Alzheimer's disease
3. Plant compounds may boost brain function in older adults, study says
4. Colorectal cancer prevention: A proven benefit of nonsteroidal anti-inflammatory drug
5. Non-steroidal anti-inflammatory drugs linked to increased risk of erectile dysfunction
6. Self-administered light therapy may improve cognitive function after traumatic brain injury
7. Hazelnuts improve older adults' micronutrient level
8. Can training improve memory, thinking abilities in older adults with cognitive impairment?
9. Aspirin and similar drugs may be associated with brain microbleeds in older adult
10. Are preventive drugs overused at the end of life in older adults with cancer?

On using a vectorizer that takes term frequency- inverse document frequency in account, we get the following result.

The ten headlines having the highest cosine similarity are:

1. Anti-inflammatory drugs ineffective for prevention of Alzheimer's disease
2. The link between cognitive function and sexuality in older adults
3. Non-steroidal anti-inflammatory drugs linked to increased risk of erectile dysfunction
4. Colorectal cancer prevention: A proven benefit of nonsteroidal anti-inflammatory drugs
5. Self-administered light therapy may improve cognitive function after traumatic brain injury
6. Plant compounds may boost brain function in older adults, study says
7. Anti-inflammatory therapy cuts risk of lung cancer
8. Anti-inflammatory medicine can have a beneficial effect on depression
9. Anti-inflammatory plant-based diet helps reduce gingivitis
10. Cancer and chemobrain: Cancer diagnosis affects cognitive function

The three main aspects in the third headline are anti-inflammatory drugs, cognitive function and older adults. The top ten headlines have some combination of these. In the TFIDF vectorizer it looks like anti-inflammatory drugs have a larger weight and comes up with other headlines that put emphasis on anti-inflammatory drugs.

**Method 2: K Means Clustering**

For the term frequency vectorizer and the tfidf vectorizer, using the elbow method, we get 3 and 5 number of clusters respectively.

From the clustering the term frequency vectorizer gives a better clustering. The elbow graph is given below:

Chart, line chart

Description automatically generated

Cluster 1: This cluster has headlines related to drugs and treatment. Some of the headlines in the cluster are:  
1. New form of ECT is as effective as older types but without cognitive side effects  
2. Anti-inflammatory drugs do not improve cognitive function in older adults  
3. Many men with low testosterone levels do not receive treatment  
4. Cannabis reduces OCD symptoms by half in the short-term  
5. Childhood lead exposure associated with criminal behavior in adulthood

Cluster 2: This cluster has headlines related to cancer. Some of the headlines in the cluster are:  
1. Prevalence of pre-cancerous masses in the colon same in patients in their 40s and 50s  
2. Researchers uncover higher prevalence of periodontal disease in rheumatoid arthritis patients  
3. Specialist nurses can play a key role in supporting patients having radiotherapy  
4. Genomic study of 6000 NCI-MATCH cancer patients leads to new clinical trial benchmarks  
5. Catheter ablation linked to reduced risk of dementia in patients with atrial fibrillation

Cluster 3: This cluster has headlines relating to studies conducted. Some of the headlines in the cluster are:  
1. Cosmetic surgery procedures to exceed 55 million in 2015, ASPS study predicts  
2. Case Western Reserve University study finds caregivers of spouses with dementia enjoy life less  
3. Study examines lung cancer among lifelong nonsmokers  
4. Study finds older adults using cannabis to treat common health conditions  
5. COVID-19 heightens urgency of advanced care planning, according to WVU study

**Method 3: LDA Topic Modeling**

On running different numbers of topics and observing the key words, it looks like eight different topics makes sense. The following are the different topics and their keywords.

Topic 0: Environmental conditions  
pancreas populations holidays cardiac death artificial pancreas sufficient ibs exercise help raising weather

Topic 0: Cancer treatment in women  
risk cancer study patients new linked disease heart women treatment

Topic 1: Covid 19  
covid 19 covid 19 brain injury patients study kidney injury cardiac acute

Topic 2: not sure  
osteoarthritis knee osteoarthritis metabolic storm knee metabolic syndrome environmental disease linked temperature alcoholic

Topic 3: Smoking and its impact  
fatty liver smoke fatty liver disease secondhand secondhand smoke insulin resistance postmenopausal liver uc

Topic 4: Diseases  
bipolar bipolar disorder eye disease eye patients diabetes game happiness kidney transplant disorder mark

Topic 5: Health in children  
study health mental new children long air disease term long term

Topic 6: Reporting and dashboarding  
neck head head neck review efficiency improves cancer dashboard improves digital dashboard dashboard case review cancer case

Topic 7: Cancer and the treatment  
cervical cervical cancer determines test determines cancers respond respond treatment new experimental determines pancreatic experimental blood live longer

**Conclusion:**

From K Means and LDA, the results and not very consistent. However, both the methods require human interpretation and topics regarding health can be classified in very different ways. It is possible that the K Means clustering has only picked the higher level of classification, whereas in LDA it was easier to comprehend deeper layers of the classification.