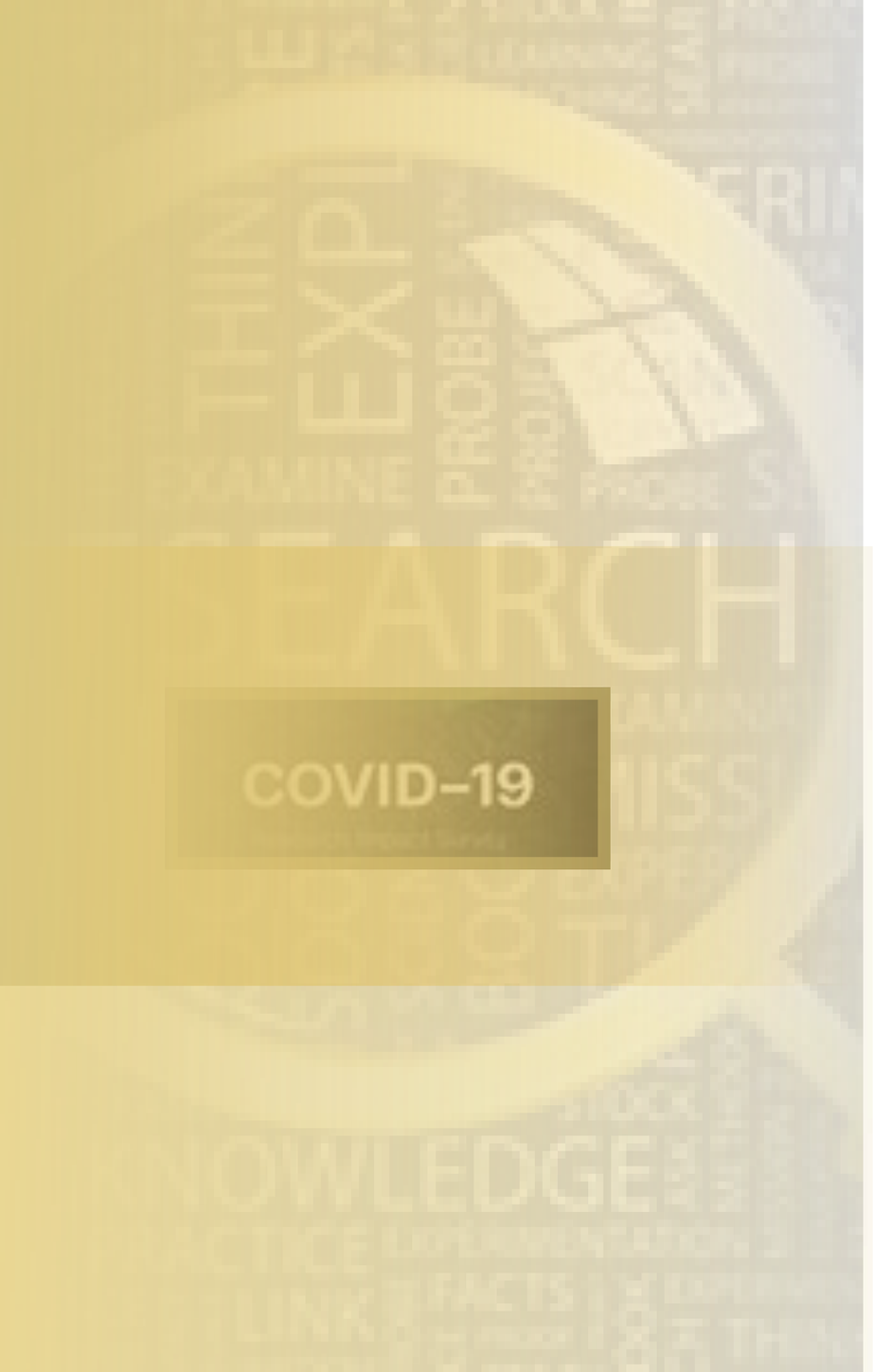


COVID-19 Open Research Dataset Challenge

NATURAL LANGUAGE PROCESSING AND TOPIC MODELING: FINDING RELATED ARTICLES

Team members:

Tahani Alqahtani – Amani Albalawi.

Outline:

- Objectives
- Workflow
- Dataset
- Tools
- Data preprocessing
- Topic Modelling
- Clustering Data

Objective:

In this project we will find related articles by using Topic modelling. Here I am using, LSA , NMF , CorEx, and LDA. All this topic models are used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model.

WORKFLOW

FROM START TO FINISH

Read Data

from Kaggle

Data Preprocessing

preprocessing
the raw data

Topic Modelling

NMF, LSA, LDA,
and CoreX

clustering

K-mens

Dataset

We have made this dataset available on Kaggle and we use abstract column to make my model .

Tools & libraries:

Libraries

Technologies

Data Preprocessing

1

Remove Numbers, Capital litters and Punctuation

2

Remove stop word

3

Tokenizing and Lemmatizing

Topic Modelling

NMF

Topics(10)
CV & TFIDF

LSA

Topics(10)
CV & TFIDF

CorEx

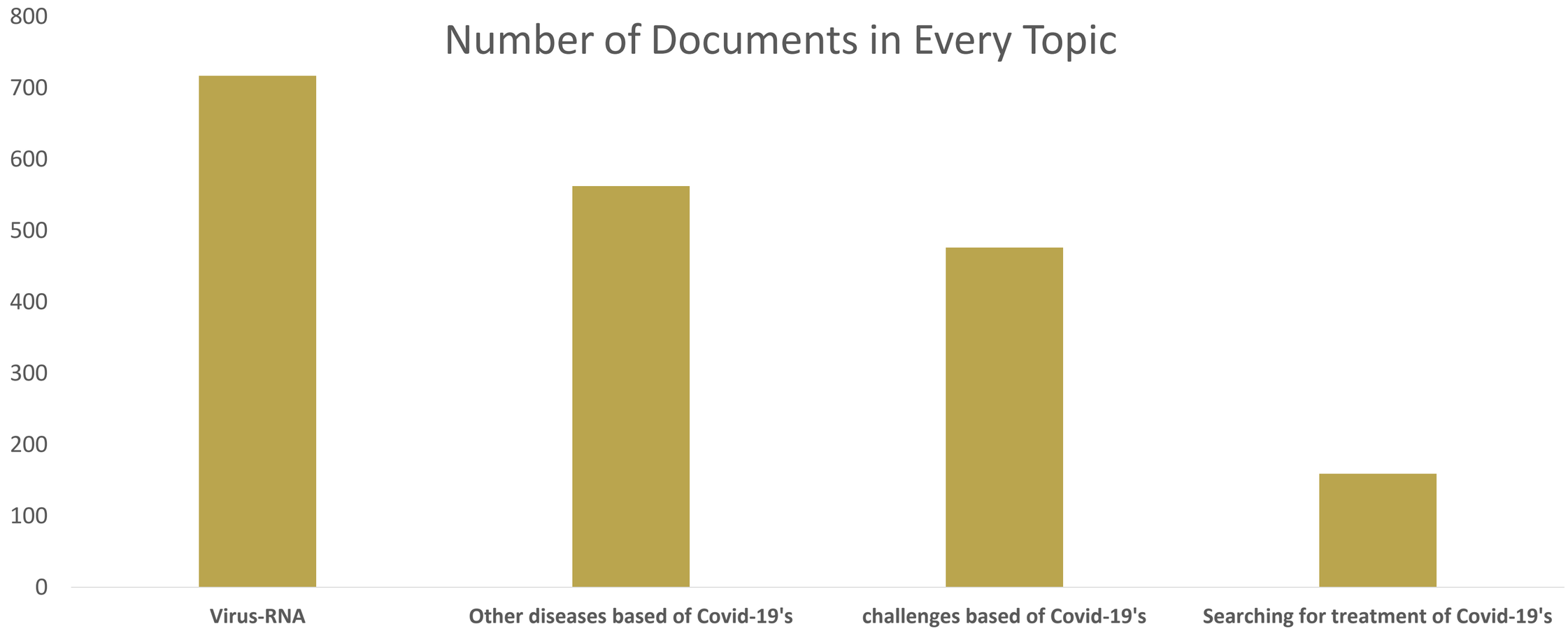
Topics(5)
CV & TFIDF

LDA

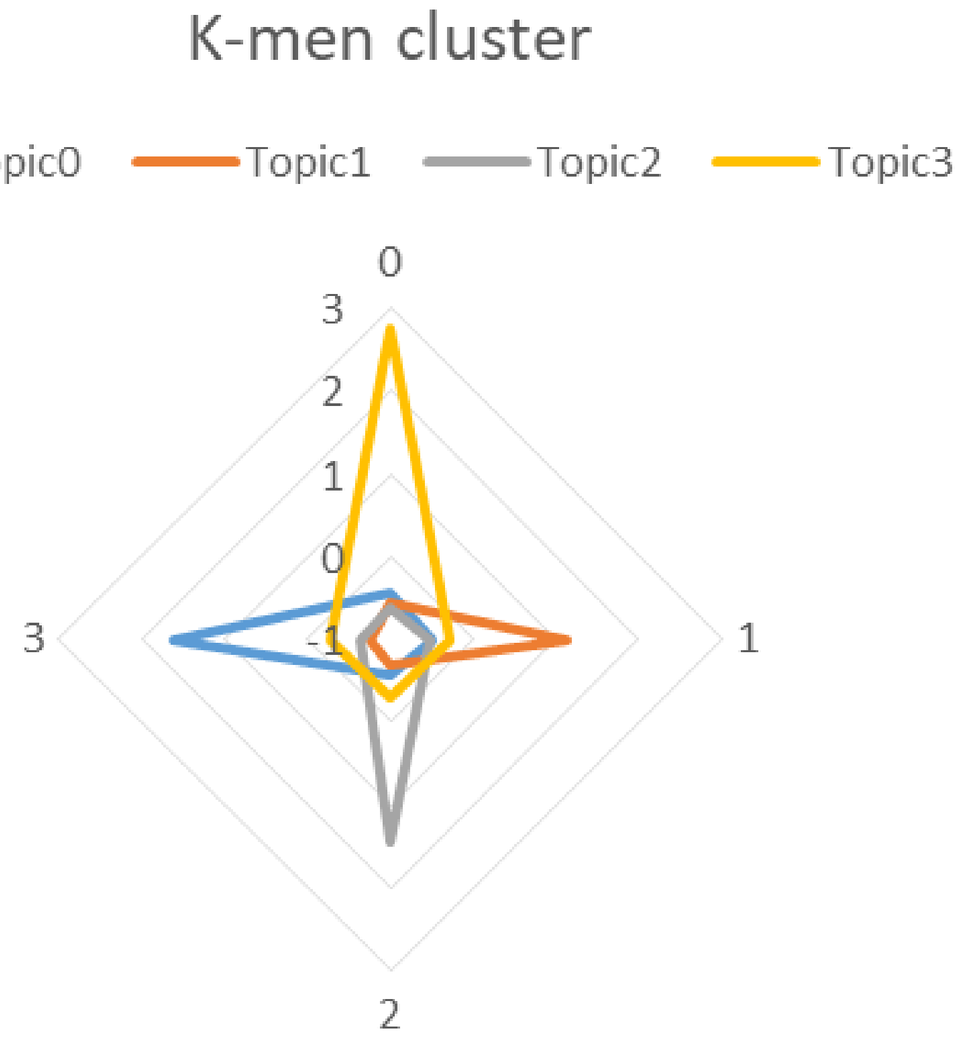
Topics(2-10)
CV & TFIDF

Best model was Count Vectorizer LDA with 4 topics

LDA


Number of Documents in Every Topic
800
700
600
500
400
300
200
100
0
Virus-RNA
Other diseases based of Covid-19's
challenges based of Covid-19's
Searching for treatment of Covid-19's

Clustering Model



Thanks