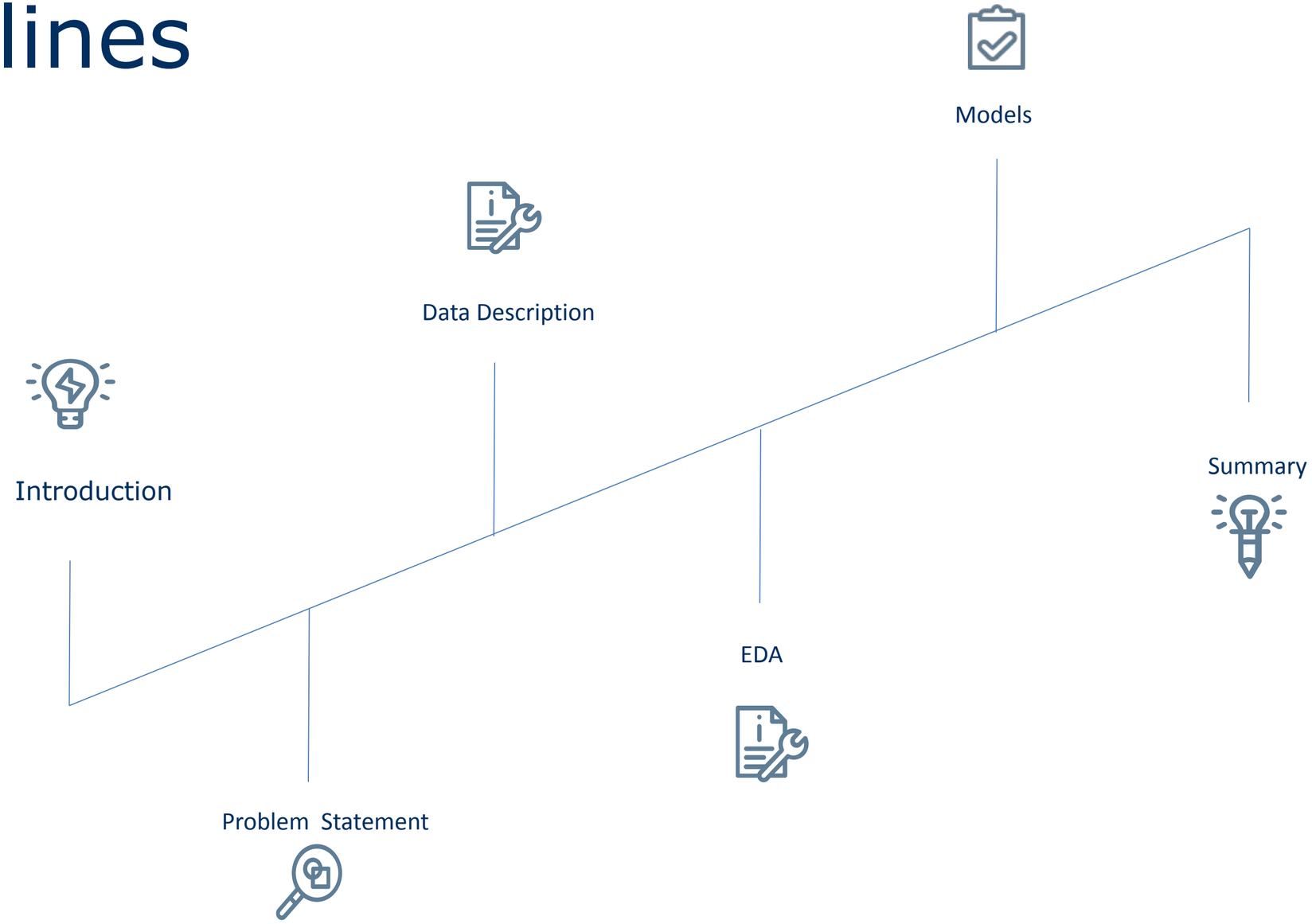
News
Articles
&
Essays

Topic Modeling and Clustering

By:Abdulrahman ,Naser and Abdullah

Outlines


Introduction
Problem Statement
Data Description
EDA
Models
Summary

Introduction

- **News is information about current event**
- **This may be provided through many different media: word of mouth, printing, postal systems, broadcasting, electronic communication…**
- **Common topic for news reports include war government, politics, education, health…etc**

Problem Statement

- News come in different formats, different types and different categories.
- Here we attempt to use Topic modeling and Clustering to get answers on what each content contains
- There are many topics so we've decided to let our models to define the news by only 5 general topics.

Data Description

The data is acquired from : https://components.one/datasets/all-the-news-articles-dataset

The Raw data contains 12 features : id, title, author, date, content, year, month, publication, category, digital, section, url .

The features we are using are only the 'title' and 'content

EDA

Messy Data

Explore

We explored and dropped unwanted columns

Preprocess Data

1. Stopword removal
2. Lemmatization
3. Stemming.

Mapping docs

Mapping documents to index: (index, [xx,yy,kk,...])

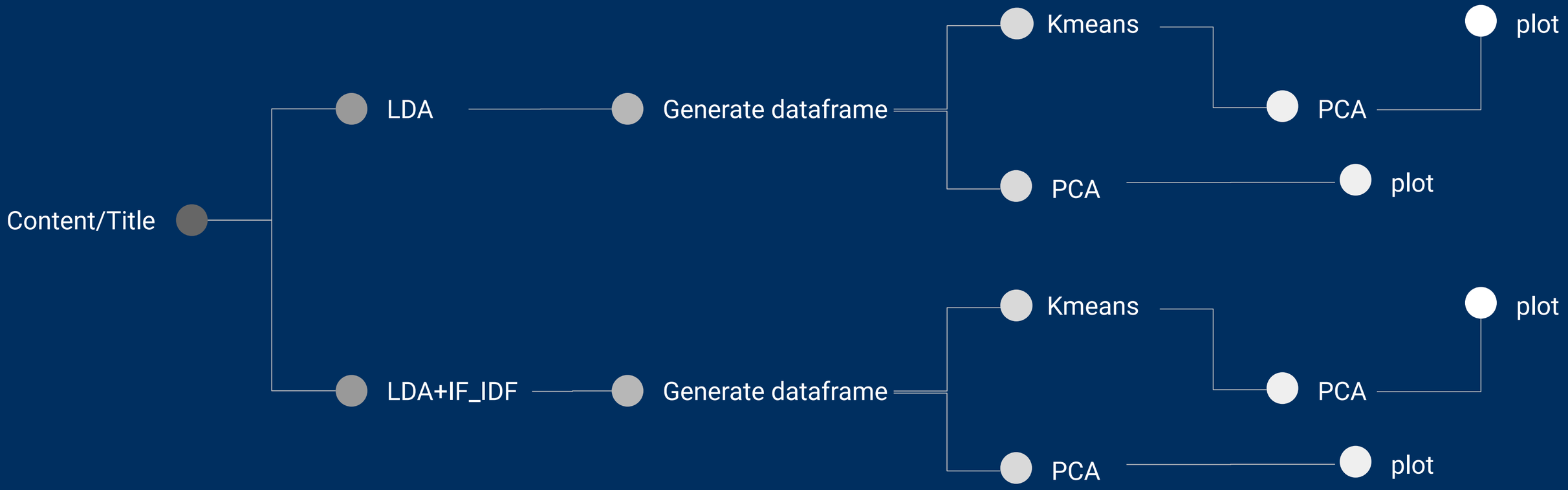
Bag Of Words

1. Bag of words for all documents
2. Tokenize
3. Remove extremes
4. Bag of words for each document

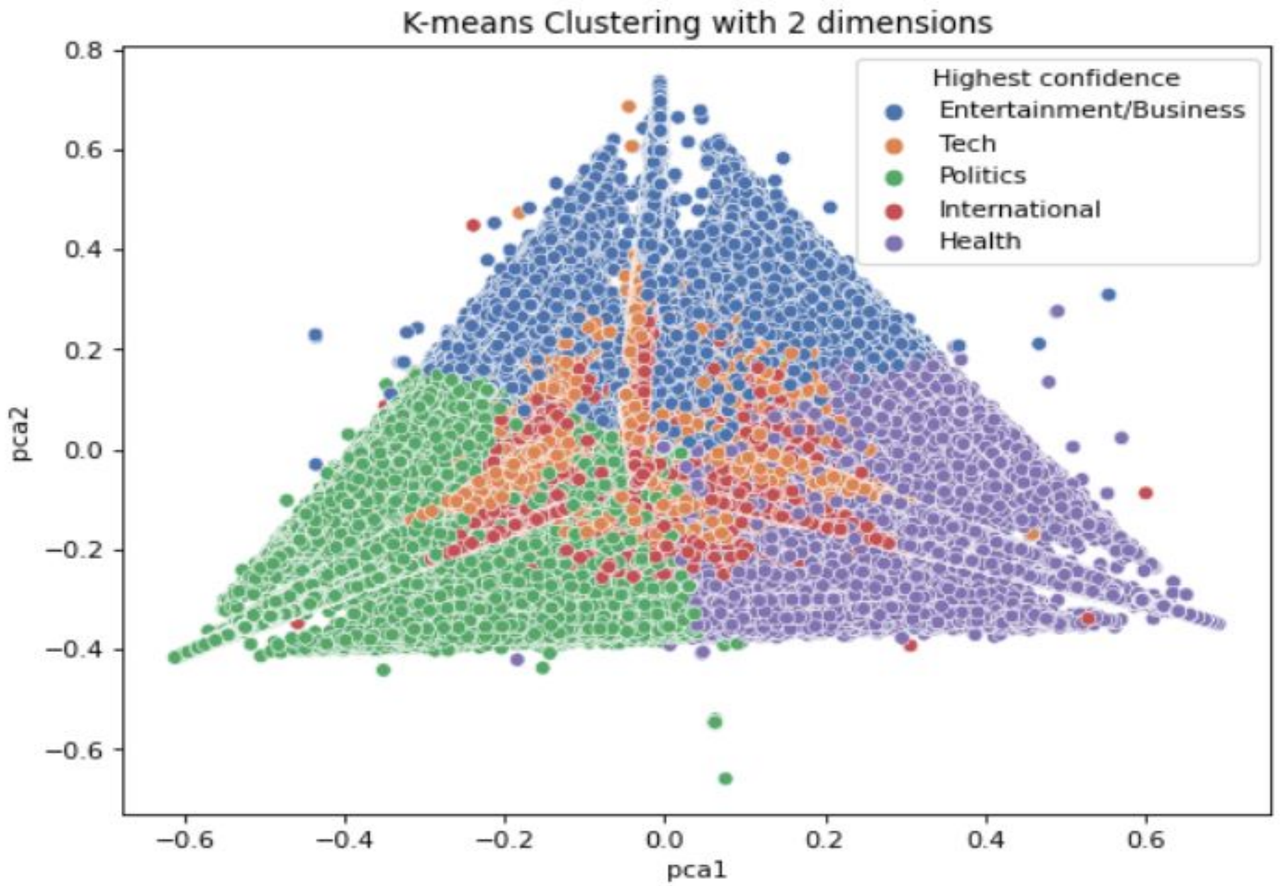
Vectorize

- Vectorize
- Vectorize using TF-IDF

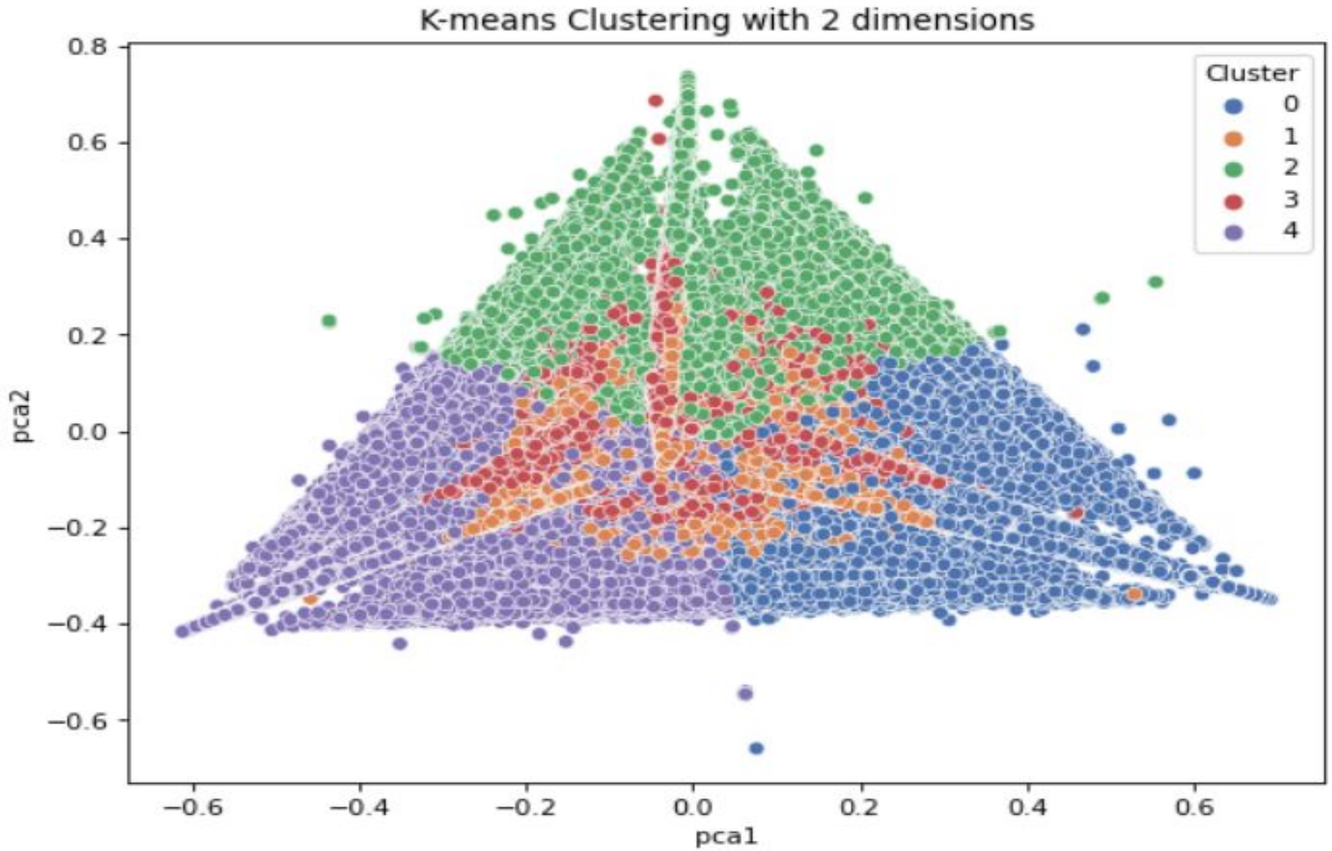
Ready Data


Content/Title
LDA
Generate dataframe
Kmeans
PCA
plot
PCA
plot
LDA+IF_IDF
Generate dataframe
Kmeans
PCA
plot
PCA
plot

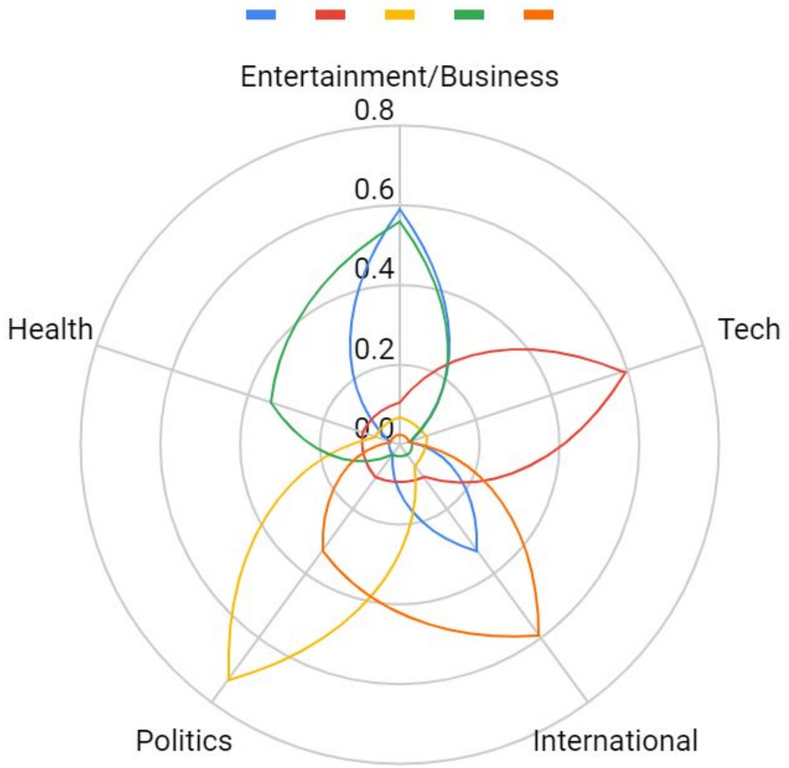
Topic Modeling



Clustering


K-means Clustering with 2 dimensions
Cluster
0
1
2
3
4
pca2
pca1
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

Spider Chart



Summary

- **The model based on the title performed worse than the model based on content in Topic modeling.**
- **The model based on the title performed better than the model based on content in clustering.**

Thank You