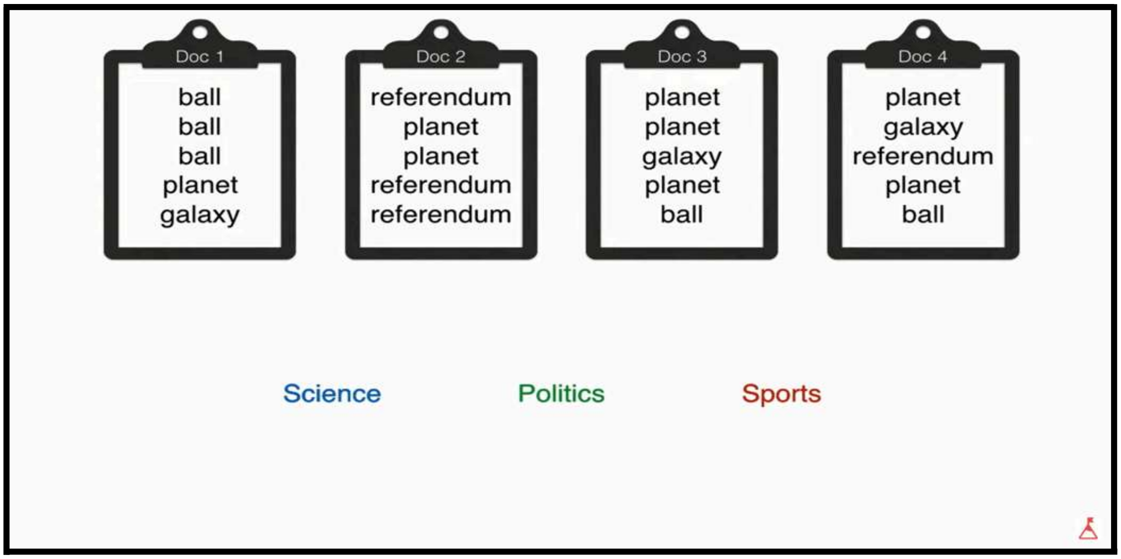
ASSIGNMENT 8 - MINI PROJECT

- **Problem Statement:** Building a personalized news recommender system.

- **Introduction:** News consumption is very subjective in nature. The articles that one prefers to read depends on their personal interests, hobbies and profession. What we have attempted to build through this project is a primitive news recommendation system that sorts random news articles into various fields of interest, and based on the user's choice, recommends other news articles from a similar topic.

Sorting articles into topics:



Doc 1
ball
ball
ball
planet
galaxy

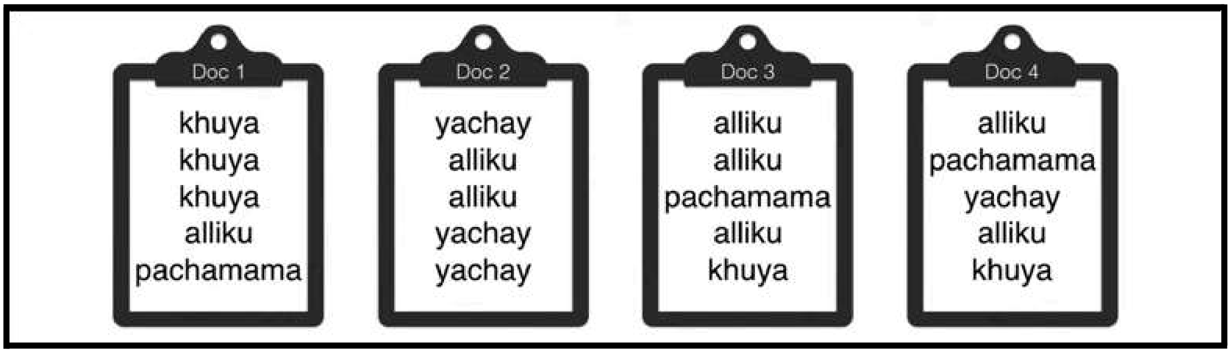
Doc 2
referendum
planet
planet
referendum
referendum

Doc 3
planet
planet
galaxy
planet
ball

Doc 4
planet
galaxy
referendum
planet
ball

Science Politics Sports

Consider 4 documents, containing a mixture of 4 words: ball, planet, galaxy and referendum. Using historical knowledge and the theory of association, we humans can classify the documents into the three given categories, based on the frequency of the words used in the document. However, the computer cannot understand English. The documents will look like a compilation of gibberish, the way the following documents seem to us:



All that the computer knows is that the given documents have to be sorted into 3 different topics, each document containing a combination of words that can help indicate the topic of the document.

This is where LDA comes into picture. Latent Dirichlet Allocation is like a machine that is used to sort documents into n topics, based on two parameters - alpha and beta- which are indicative of the importance of the topic to the document and the importance of the word to the topic, respectively.

We have used LDA for topic modeling, and as an alternative, have also presented how KNN classification can be used to sort the news articles into relevant topics.

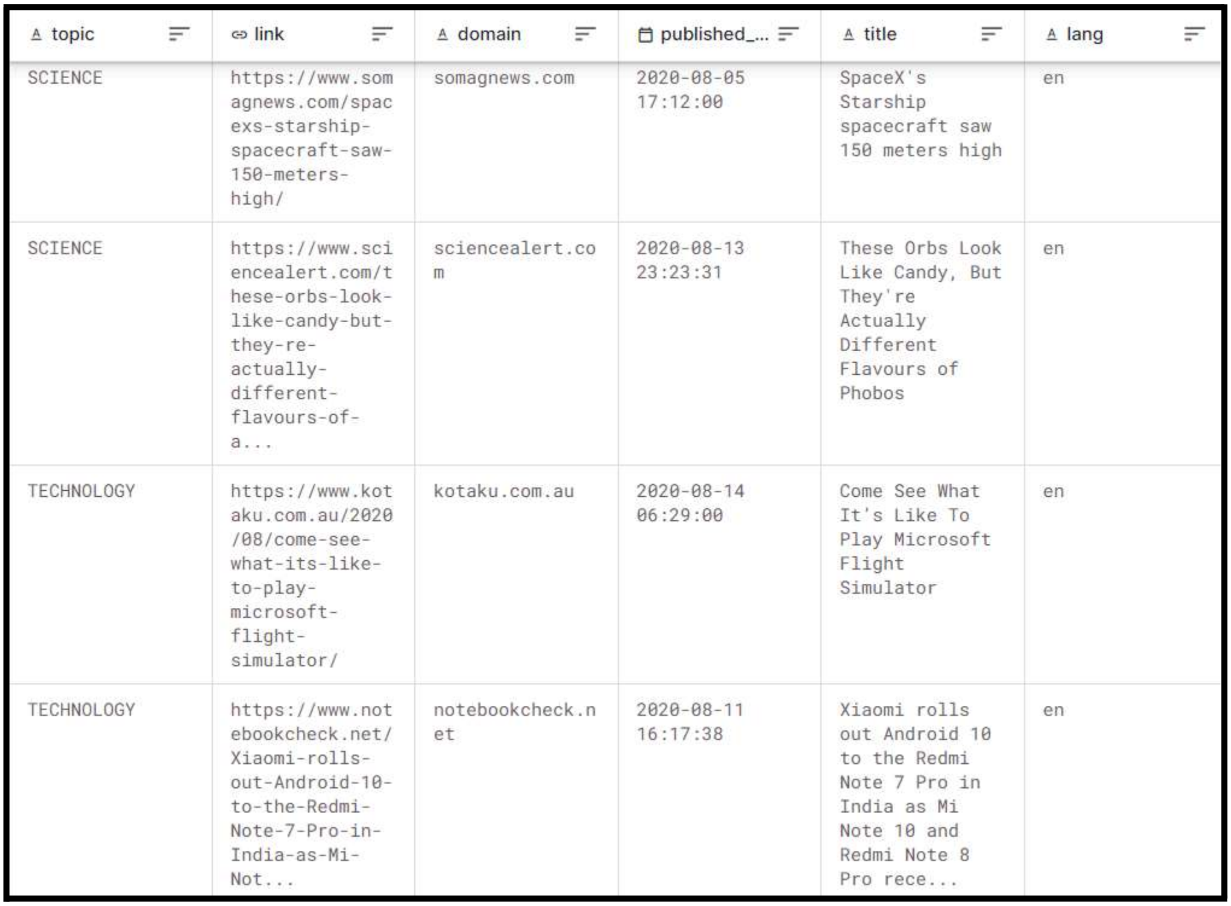
Subsequently, using cosine similarity, we have recommended relevant news to the user, based on their previous article selection.

- **Dataset Information:**

Link:
https://www.kaggle.com/datasets/kotartemiy/topic-labeled-news-dataset

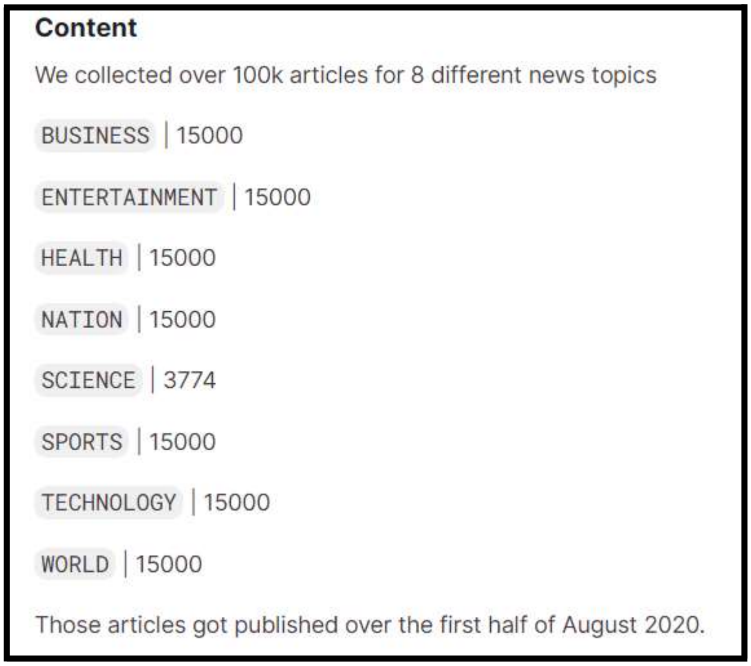
We have used a subset of 145 entries from this dataset

Snippet of the dataset:



topic	link	domain	published_...	title	lang
SCIENCE	https://www.somagnews.com/spacexs-starship-spacecraft-saw-150-meters-high/	somagnews.com	2020-08-05 17:12:00	SpaceX's Starship spacecraft saw 150 meters high	en
SCIENCE	https://www.sciencealert.com/these-orbs-look-like-candy-but-they-re-actually-different-flavours-of-a...	sciencealert.com	2020-08-13 23:23:31	These Orbs Look Like Candy, But They're Actually Different Flavours of Phobos	en
TECHNOLOGY	https://www.kotaku.com.au/2020/08/come-see-what-its-like-to-play-microsoft-flight-simulator/	kotaku.com.au	2020-08-14 06:29:00	Come See What It's Like To Play Microsoft Flight Simulator	en
TECHNOLOGY	https://www.notebookcheck.net/Xiaomi-rolls-out-Android-10-to-the-Redmi-Note-7-Pro-in-India-as-Mi-Not...	notebookcheck.net	2020-08-11 16:17:38	Xiaomi rolls out Android 10 to the Redmi Note 7 Pro in India as Mi Note 10 and Redmi Note 8 Pro rece...	en

Note by the publishers:



Content

We collected over 100k articles for 8 different news topics

BUSINESS	15000
ENTERTAINMENT	15000
HEALTH	15000
NATION	15000
SCIENCE	3774
SPORTS	15000
TECHNOLOGY	15000
WORLD	15000

Those articles got published over the first half of August 2020.

Conclusion:

We have demonstrated two approaches of classifying news articles based on their topics using their content and titles:

1.Using the concept of tfidf+knn classification(supervised)

Advantages-

a.Simplicity: KNN is straightforward to understand and implement. It doesn't make assumptions about the underlying data distribution, making it suitable for various types of data, including text.

b.No Training Phase: KNN is a lazy learning algorithm, meaning it doesn't require a training phase. Instead, it stores all the training data and makes predictions based on similarity measures during the testing phase. This can be advantageous when dealing with large datasets or when data is continuously updated.

Disadvantages-

a.Computation and Memory Intensity: KNN requires storing all training instances in memory, making it memory-intensive, especially for large datasets. It also requires computing distances between the query instance and all training instances during prediction, making it computationally expensive, especially as the dataset grows.

b.Sensitivity to Irrelevant Features: KNN considers all features when computing distances between instances, which can lead to degraded performance if there are irrelevant or noisy features in the dataset.

2.Latent Dirichlet Allocation(unsupervised):

Advantages-

1.LDA learns probabilistic topic distributions from the corpus and uses these distributions to classify or recommend documents, which is typically less memory and computationally intensive compared to KNN.

2.LDA automatically identifies the latent topics in the corpus and represents each document as a distribution over these topics, effectively reducing the dimensionality and focusing on the most relevant aspects of the text.

Implementation of LDA was again demonstrated by two approaches:

i.Implementation using LDA model from sklearn

ii.Built-in implementation of LDA using ktrain: A lightweight wrapper for the deep learning library Keras to help build, train, and deploy neural networks

This was followed by selection of a random news article and recommending 3 articles most similar to the selected article from the same topic using the method of cosine similarity in both the user-defined and built-in(k-train) approaches.

References:

https://analyticsindiamag.com/a-complete-guide-to-ktrain-a-wrapper-for-tensorflow-keras/

https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

https://www.youtube.com/watch?v=T05t-SqKArY
https://www.youtube.com/watch?v=BaM1uiCpj_E