 # Problem Statement:

Design a recommendation system that can

1) suggest personalized news articles or blog posts to users based on their reading history and interests.
2) The system should be able to learn and adapt to the user's preferences over time and provide relevant and engaging content.

 # Dataset

https://www.kaggle.com/datasets/rmisra/news-category-dataset

> This dataset contains around 210k news headlines from 2012 to 2022 from HuffPost.
> Columns:
category: category in which the article was published.
headline: the headline of the news article.
authors: list of authors who contributed to the article.
link: link to the original news article.
short_description: Abstract of the news article.
date: publication date of the article.

 # Preprocessing

> Joined the ‘Headline’ column and ‘Short Description’ column and made one ‘Text’ column out of it.
> Replaced missing values in the 'authors' column with 'Unknown'.
> Drop the rows with null values in the 'text' column.

```
df
```

	link	category	authors	date	text
0	https://www.huffpost.com/entry/covid-boosters-...	U.S. NEWS	Carla K. Johnson, AP	2022-09-23	Over 4 Million Americans Roll Up Sleeves For O...
1	https://www.huffpost.com/entry/american-airlin...	U.S. NEWS	Mary Papenfuss	2022-09-23	American Airlines Flyer Charged, Banned For Li...
2	https://www.huffpost.com/entry/funniest-tweets...	COMEDY	Elyse Wanshel	2022-09-23	23 Of The Funniest Tweets About Cats And Dogs ...
3	https://www.huffpost.com/entry/funniest-parent...	PARENTING	Caroline Bologna	2022-09-23	The Funniest Tweets From Parents This Week (Se...
4	https://www.huffpost.com/entry/amy-cooper-lose...	U.S. NEWS	Nina Golgowski	2022-09-22	Woman Who Called Cops On Black Bird-Watcher Lo...
...	...	...	...	...	...
209522	https://www.huffingtonpost.com/entry/rim-ceo-t...	TECH	Reuters, Reuters	2012-01-28	RIM CEO Thorsten Heins' 'Significant' Plans Fo...
209523	https://www.huffingtonpost.com/entry/maria-sha...	SPORTS	Unknown	2012-01-28	Maria Sharapova Stunned By Victoria Azarenka I...
209524	https://www.huffingtonpost.com/entry/super-bow...	SPORTS	Unknown	2012-01-28	Giants Over Patriots, Jets Over Colts Among M...
209525	https://www.huffingtonpost.com/entry/aldon-smi...	SPORTS	Unknown	2012-01-28	Aldon Smith Arrested: 49ers Linebacker Busted ...
209526	https://www.huffingtonpost.com/entry/dwight-ho...	SPORTS	Unknown	2012-01-28	Dwight Howard Rips Teammates After Magic Loss ...

189814 rows × 5 columns

> Removed Stopwords and did Lemmatization using nltk package.

> Vectorization of ‘Text’ column using Google's pre-trained Word2Vec model from gensim package.

 # Model Building

- In order to build a recommendation system that recommends news articles based on their similarity to a given news article.
- We first use Word2Vec, a natural language processing technique, to create vector representations of words in article headlines.
- Then, we calculate the similarity between different headlines based on the average of the Word2Vec vectors of their words.
- For recommendations based on single feature ie using only ‘Text’ Column - we write a function that returns the "num_similar_items" most similar articles to a given article, based on the Euclidean distance between their average Word2Vec vectors. The recommended articles are displayed in a DataFrame along with their publishing dates and similarities to the given article.
- For recommendations based on multiple features ie ‘Text’ column and ‘Category’ column. We first do one hot encoding on ‘Category’ column and then based on formula: **weighted_couple_dist = (w1 * w2v_dist + w2 * category_dist)/float(w1 + w2),** we get we recommended articles that are from the same category as the queried article category due to high w2 weight.
- For recommendations based on multiple features ie ‘Text’ column, ‘Category’ column and ‘Authors’ column. We do one hot encoding on ‘Authors’ column and then based on formula: **weighted_couple_dist = (w1 * w2v_dist + w2 * category_dist + w3 * authors_dist)/float(w1 + w2 + w3),** we get recommended articles that are from the same author as the queried article author due to high w3 weight.

 # Summary till now

Since we only have data on news articles and no user data, we have built a recommendation system based on Content-Based Filtering Method. This model when given a query article name, it will give out a certain number of similar articles based on similar headlines/categories/authors.

 # Future Work

> Building a simple UI/ Website for our content-based recommendation system.
> Using this built website of ours, we will record the user interaction with our recommendation system and built up **User data** on the news articles like ratings, time spent on reading the article, search history etc.
> Using this **User data**, we will then implement Collaborative Filtering on our recommendation system and get much Robust Recommendation System that will recommend articles not only based on new article features but also based on similar users.

Team Member’s
Saksham and Merin - Model Building.
Gautam - Research on content based filtering.
Satyam - Research on collaborative filtering and Github update.
Lomesh Soni - Research on web scraping.