Content based News Recommendation System

Priya Sathish

Online platforms support so many of our daily activities that we have become dependent on them in our personal and professional lives. We rely on them to buy and sell goods and services, to find information online and to keep in touch with each other. These platforms could help consumers by recommending items as per their interest and preference by just analyzing your past interaction or behavior with the system. From Amazon to LinkedIn, Netflix to Spotify, Facebook, recommender systems are most extensively used to suggest "Similar items", "Relevant jobs", "preferred foods", "Movies of interest" etc. to their users. Recommender system with appropriate item suggestions helps in boosting sales, increasing revenue, retaining customers and also adds competitive advantage. Recommender systems use a number of different technologies and can be classified into two broad groups as Content based recommendation and Collaborative filtering.

On a day to day basis, the internet has a lot of sources that generate immense amount of daily news diversified in subject matter. There is continuous demand for new information to be available immediately and with ease by the consumers. So, it is crucial that the news is classified and targets the needs and requirements of the user effectively and efficiently. News services have attempted to identify articles of interest to readers based on the articles that they have read in the past. The similarity might be based on the similarity of important words in the documents or on the articles that are read by people with similar reading tastes. The same principles apply to recommending blogs from among the millions of blogs available or other sites where content is provided regularly.

This project focuses on content-based recommendation using News category dataset. The goal is to recommend news articles which are similar to the already read article by using attributes like article headline, short description, category, author and publishing date.

Client

News readers, blog readers, news agencies, bloggers, retailers and several online platforms.

Data and approach

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

This dataset contains around 200k news headlines from the year 2012 to 2018 obtained from HuffPost. News in this dataset belongs to 41 different categories. Each news record consists of a headline with a short description in our analysis. In addition, we will combine attributes 'headline' and 'short description' into a single attribute 'text' as the input for classification and proceed with developing a deep learning model to build the recommender system.

Citation: "https://rishabhmisra.github.io/publications/"

Data Wrangling

Acquired the news category dataset from kaggle as a json file and converted it to a pandas dataframe with a shape of (200853, 6) for analysis. When grouped by category we can see that the dataset contains 41 categories of news articles.

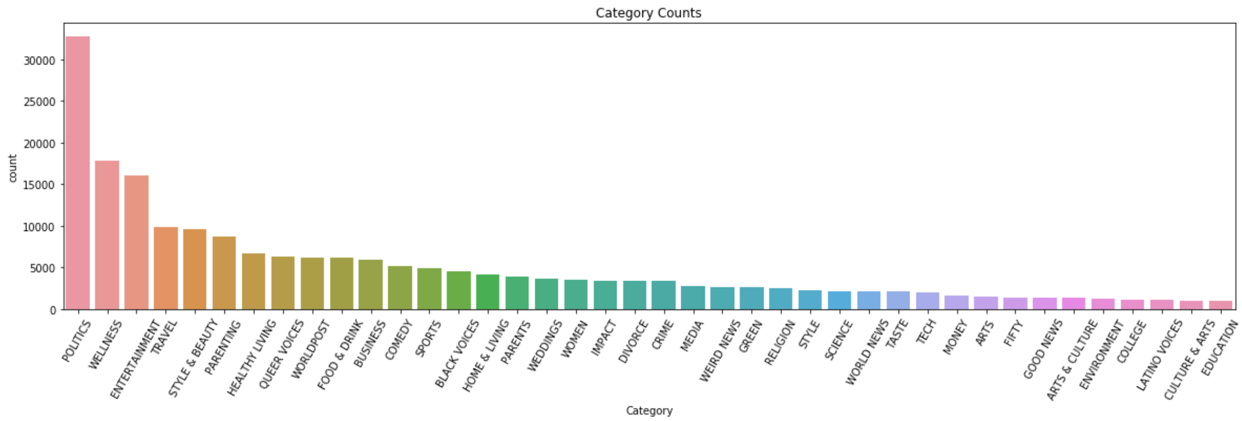
'THE WORLDPOST' and 'WORLDPOST' should be the same category, so we change 'THE WORLDPOST' to 'WORLDPOST' and merge them.

After which we have,
Total number of articles: 200853
Total number of authors: 27993
Total number of unique categories: 40

Top 5 categories include:
category
POLITICS 32739
WELLNESS 17827
ENTERTAINMENT 16058
TRAVEL 9887
STYLE & BEAUTY 9649

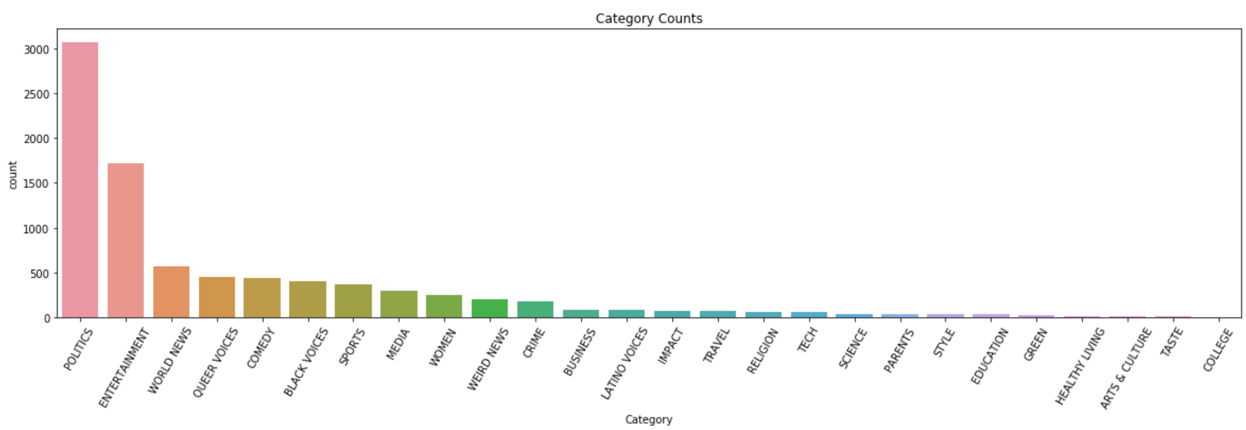
Also, we will check for any missing data,
category 0
headline 0
authors 0
link 0

short_description   0
date   0
text   0
words   0
word_length   0



From the plot above we observe that politics, wellness, entertainment, travel and beauty form the top 5 categories of news article headlines during the period of 2014-2018.

We will consider only the latest articles from the year 2018 as the size of the dataset is quite large and processing may consume too much time. So, we filter the dataset to contain only row from the year 2018 and will proceed with text preprocessing.



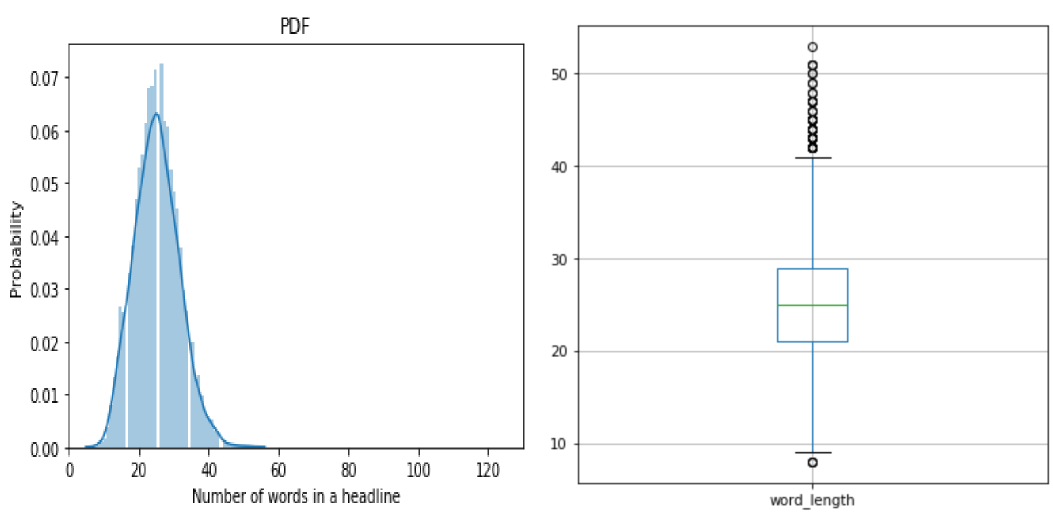
From the plot above we can see that politics, entertainment, world news, queer voices and comedy rates the top 5 categories in the year 2018. Looking at the plots, trend is consumers are interested in social awareness (like Politics, Entertainment, World news etc.) that energizes news media for popularity and maintains the consumers interest.

We will be using headlines and short description as input X. So, we will combine the short description and headline by concatenation and create a new column.

We tokenize the headlines and then we delete some empty and short data with word length less than 5. We take a look at the word length distribution:

```
Count   8583.000000
mean     25.176861
std       6.418315
min       8.000000
25%      21.000000
50%      25.000000
75%      29.000000
max      53.000000
Name: word_length, dtype: float64
```

We create a PDF and box plots to view the word length distribution.



From the plot above we can clearly see that the word length ranges between 8 and 53 with a mean value of 25 and standard deviation of about 6 over 8583 news articles.

Text Preprocessing

We will clean and process the text data so that it is ready for modeling using the Natural Language Toolkit or NLTK Python library.

Tokenize:
We will split the news headline text into tokens based on white space or punctuation. This is considered as a base step for stemming and lemmatization. Once we use word_tokenize() on the headline text we can use the output for stop words removal.

Stop words removal:
Stop words are not much helpful in analysis and also their inclusion consumes much time during processing so let's remove these. We will use the default NLTK corpus to remove the unwanted stop words.

Lemmatize:
Lemmatization is the algorithmic process of finding the lemma of the word depending on the meaning. We will use the WordNetLemmatizer() from the NLTK Python library to remove inflectional endings.

To perform all of the above-mentioned operations on our headline text we will define a function to process:

```
def process_headlines(main_text):
   headlines_without_numbers = re.sub('[^a-zA-Z]', ' ', main_text)
   words = word_tokenize(headlines_without_numbers.lower())
   stop_words_english = set(stopwords.words('english'))
   final_words = [lemmatizer.lemmatize(word) for word in words if word not in stop_words_english]
   return(' '.join(final_words))
```

Bag Of Words – Count Vectorize

To extract features from the text documents and create a vocabulary of all the unique words in all of the news headlines we will use CountVectorizer() from the NLTK Python library. The BOW model only considers if a known word occurs in a document or not. It does not care about meaning, context and order in which they appear. This gives the insight that similar documents will have word counts similar to each other. In other words, the more similar the words in two documents, the more similar the documents can be. However, there are some limitations using this method such as it doesn’t take the semantic meaning or context into account and also if the vector size is huge it might result in lot of computation and time.

Preliminary model evaluation using default parameters

Now that we have preprocessed our text, we will try several kinds of classifiers and compare them with their default parameters. The problem we have here is that the algorithm may not perform well right away but might perform really well with the right set of hyperparameters. To get a preliminary understanding of which types of classifiers will inherently work better we will compare them.

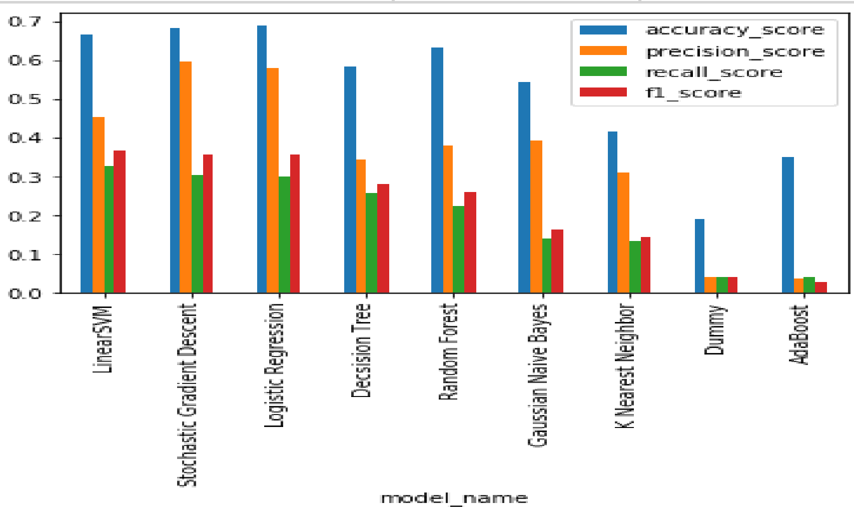
We will take a look at 8 different classifiers along with sklearn's dummy classifier which is just a random baseline. The classifiers we will compare include:
Dummy Classifier, Stochastic Gradient Descent, RandomForestClassifier, DecisionTreeClassifier, AdaBoostClassifier, Gaussian Naive Bayes, LogisticRegression, LinearSVM and K Nearest Neighbor.

The metrics we will use to evaluate the different classifiers are:

- Accuracy – the fraction of samples predicted correctly
- Precision – the ratio of true positives to false positives (ability of the classifier not to label a positive as a negative sample)
- Recall – the ratio of true positives to false negatives (ability of the classifier to find all the positive samples)
- F1 Score – the harmonic average of precision and recall (we will use macro averaging which will compute the average of the F1 scores)

Comparison Results:

model_name	accuracy_score	precision_score	recall_score	f1_score
LinearSVM	0.666078	0.452625	0.328205	0.36622
Stochastic Gradient Descent	0.682316	0.59672	0.305721	0.358697
Logistic Regression	0.689375	0.581914	0.301341	0.356729
Decsision Tree	0.584539	0.345634	0.257183	0.28229
Random Forest	0.633251	0.379824	0.225787	0.259581
Gaussian Naive Bayes	0.545358	0.395019	0.141648	0.164515
K Nearest Neighbor	0.417226	0.311188	0.133848	0.144641
Dummy	0.190964	0.0400646	0.0404579	0.040161



From the above plot and the F1 scores we can see that Linear SVM (0.37), Stochastic Gradient (0.36) Descent and Logistic Regression (0.36) performed better than all other classifiers. The F1 scores for the top three classifiers are also not that great though the real test is how they perform on unseen articles.

The scores were calculated only the news articles from the year 2018. The classifiers might have performed well if we had used the news articles from the years 2014-2018.

Deep Learning Models and comparison

As our next step, we will implement several deep learning models, tune them and compare them with evaluation metrics. We considered only the data from the year 2018 for our preliminary preprocessing and models. For deep learning models we will use the full dataset from years 2014 to 2018 for better performance and evaluation metrics.

A word embedding is an approach to provide a dense vector representation of words that capture something about their meaning. Word embeddings are an improvement over simpler bag-of-word model word encoding schemes like word counts and frequencies that result in large and sparse vectors (mostly 0 values) that describe documents but not the meaning of the words. Word embeddings work by using an algorithm to train a set of fixed-length dense and continuous-valued vectors based on a large corpus of text. Each word is represented by a point in the embedding space and these points are learned and moved around based on the words that surround the target word. It is defining a word by the company that it keeps that allows the word embedding to learn something about the meaning of words. The vector space representation of the words provides a projection where words with similar meanings are locally clustered within the space. The use of word embeddings over other text representations is one of the key methods that has led to breakthrough performance with deep neural networks on problems like machine translation.

GloVe Embedding

Global Vectors for Word Representation:

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. The smallest GloVe pre-trained model is from the GloVe website. It an 822 Megabyte zip file with 4 different models (50, 100, 200 and 300-dimensional vectors) trained on Wikipedia data with 6 billion tokens and a 400,000-word vocabulary.

Convolution Neural Network (CNN) - GloVe embedding

Traditionally CNN is popular is for identifying objects inside images. It can also be extended for text classification with the help of word embeddings. CNN has been found effective for text in search query retrieval, sentence modelling and other traditional NLP (Natural Language Processing) tasks. Once an image is converted to vectorized representation or text is converted to embedding, it looks similar to machine as shown in picture below. In case of image each cell in the represents raw intensity of specific channel whereas in case of text, each row of table represents a word. Just like in traditional CNN, lower level layers help in identifying edges, parts of bigger objects and successive layers identifies objects, in case of text classification, lower layer tried to find association between words whereas higher layer tries to find association between group of words. These groups can be sentences, paragraphs or smaller subgroups.

A typical convolution layer network architecture has multiple layers. CNN is supervised ML algorithm. The training set is first converted to word embeddings using glove embeddings. We first pass it through a series of convolution and pooling layer to extract lower levels features first and then learn higher level features from lower level features.

In our training first we split input dataset into 80% training and 20% validation set. We feed input dataset for training to CNN. we are going to apply 64 such filters on training dataset. Each filter will be applied to 2,3,4 words at a time. After convolution the output is passed through RELU activation layer to remove negative samples and keep only positive samples. Output of RELU is passed through max pooling layer to retain most important information.

We then pass the output through a dropout layer to prevent overfitting. We then pass it through another set of 1D convolution, RELU and max pooling.

Finally, the last layer in CNN is typically feed forward neural network that learns to map the pooling function output to output categories in terms of softmax probabilities.

Now that our network architecture is up, we train the model for 20 epochs and measure its performance over validation set.

Model Summary:

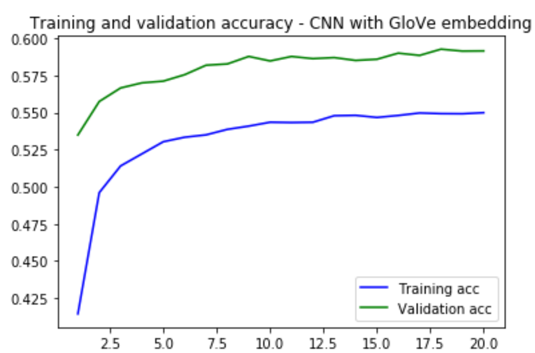
```
Model: "model_1"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_1 (InputLayer)            (None, 50)           0
__________________________________________________________________________________________________
embedding_1 (Embedding)         (None, 50, 100)      11661800    input_1[0][0]
__________________________________________________________________________________________________
conv1d_1 (Conv1D)               (None, 50, 64)       12864       embedding_1[0][0]
__________________________________________________________________________________________________
conv1d_2 (Conv1D)               (None, 50, 64)       19264       embedding_1[0][0]
__________________________________________________________________________________________________
conv1d_3 (Conv1D)               (None, 50, 64)       25664       embedding_1[0][0]
__________________________________________________________________________________________________
max_pooling1d_1 (MaxPooling1D)  (None, 16, 64)       0           conv1d_1[0][0]
```

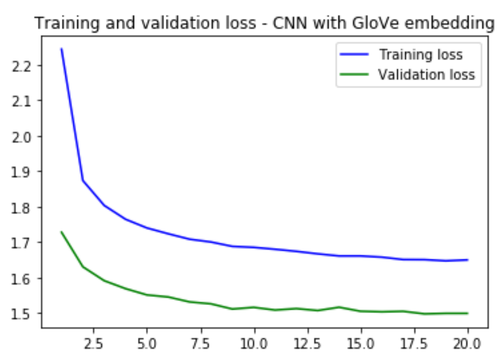
```
____________________________________________________________________________________________________
max_pooling1d_2 (MaxPooling1D)  (None, 16, 64)      0        conv1d_2[0][0]
____________________________________________________________________________________________________
max_pooling1d_3 (MaxPooling1D)  (None, 16, 64)      0        conv1d_3[0][0]
____________________________________________________________________________________________________
dropout_1 (Dropout)          (None, 16, 64)      0        max_pooling1d_1[0][0]
____________________________________________________________________________________________________
dropout_2 (Dropout)          (None, 16, 64)      0        max_pooling1d_2[0][0]
____________________________________________________________________________________________________
dropout_3 (Dropout)          (None, 16, 64)      0        max_pooling1d_3[0][0]
____________________________________________________________________________________________________
concatenate_1 (Concatenate)     (None, 16, 192)     0        dropout_1[0][0]
                                                    dropout_2[0][0]
                                                    dropout_3[0][0]
____________________________________________________________________________________________________
flatten_1 (Flatten)          (None, 3072)       0        concatenate_1[0][0]
____________________________________________________________________________________________________
dropout_4 (Dropout)          (None, 3072)       0        flatten_1[0][0]
____________________________________________________________________________________________________
dense_1 (Dense)            (None, 40)        122920     dropout_4[0][0]
====================================================================================================
Total params: 11,842,512
Trainable params: 180,712
Non-trainable params: 11,661,800
```

Model Score:

Validation loss: 1.4993746934662053
Validation accuracy: 0.5915763974189758

Model Metrics Plots:





Recurrent Neural Network (RNN) - Long Short Term Memory (LSTM) using Keras - without Embedding

Vectorize news headlines, by turning each text into either a sequence of integers or into a vector. Limit the data set to the top 1000 words. SpatialDropout1D performs variational dropout in NLP models. The next layer is the LSTM layer with 512 memory units. The output layer must create 40 output values, one for each class. Activation function is softmax for multi-class classification. Because it is a multi-class classification problem, categorical_crossentropy is used as the loss function.

Model Summary:

Model: "sequential_11"

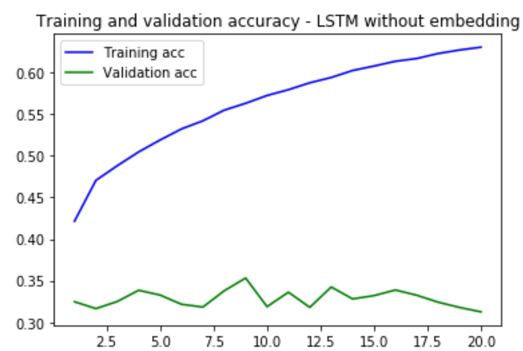
Layer (type)	Output Shape	Param #
dense_14 (Dense)	(None, 512)	512512
activation_7 (Activation)	(None, 512)	0
dropout_10 (Dropout)	(None, 512)	0
dense_15 (Dense)	(None, 40)	20520
activation_8 (Activation)	(None, 40)	0

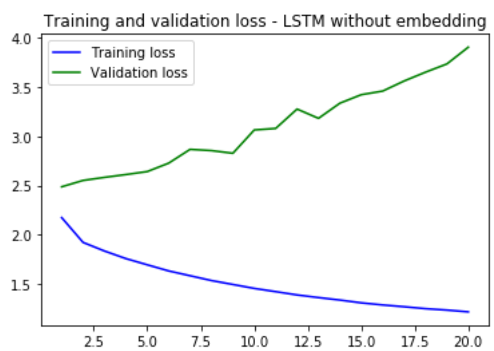
Total params: 533,032
Trainable params: 533,032
Non-trainable params: 0

Model Score:

Validation Loss: 3.9045065682964952
Validation accuracy: 0.31256356835365295

Model Metrics Plots:





Recurrent Neural Network (RNN) - Long Short Term Memory (LSTM) architecture - using Keras Embedding

Vectorize news headlines, by turning each text into either a sequence of integers or into a vector. Limit the data set to the top 1000 words. The first layer is the embedded layer that uses 100 length vectors to represent each word. SpatialDropout1D performs variational dropout in NLP models. The next layer is the LSTM layer with 100 memory units. The output layer must create 40 output values, one for each class. Activation function is softmax for multi-class classification. Because it is a multi-class classification problem, categorical_crossentropy is used as the loss function. Keras inbuilt embedding is used here.

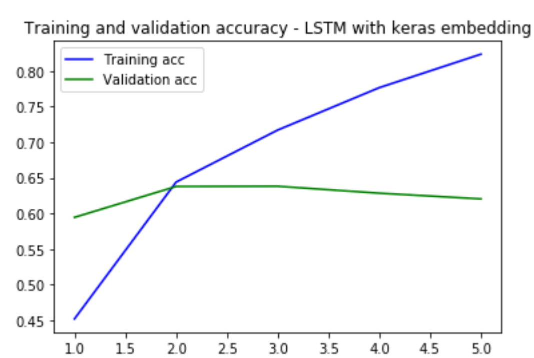
Model Summary:

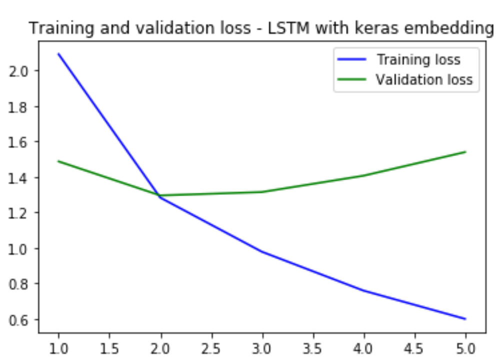
```
Model: "sequential_7"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_5 (Embedding)      (None, 50, 100)           11661800
_________________________________________________________________
spatial_dropout1d_3 (Spatial (None, 50, 100)           0
_________________________________________________________________
lstm_3 (LSTM)                (None, 100)               80400
_________________________________________________________________
dense_10 (Dense)             (None, 40)                4040
=================================================================
Total params: 11,746,240
Trainable params: 11,746,240
Non-trainable params: 0
```

Model Score:

Validation Loss: 1.5391094215074308
Validation Accuracy: 0.620488703250885

Model Metrics Plots:





Recurrent Neural Network (RNN) - Long Short Term Memory

LSTM architecture - with gensim Word2Vec

Word2vec, like doc2vec, belongs to the text preprocessing phase. Specifically, to the part that transforms a text into a row of numbers. Word2vec is a type of mapping that allows words with similar meaning to have similar vector representation. The idea behind Word2vec is rather simple: we want to use the surrounding words to represent the target words with a Neural Network whose hidden layer encodes the word representation. First, we load a word2vec model. It has been pre-trained by Google on a 100 billion-word Google News corpus. Google's pre-trained model(1.5GB!) includes word vectors for a vocabulary of 3 million words and phrases that they trained on roughly 100 billion words from a Google News dataset. The vector length is 300 features. Gensim allocates a big matrix to hold all of the word vectors.

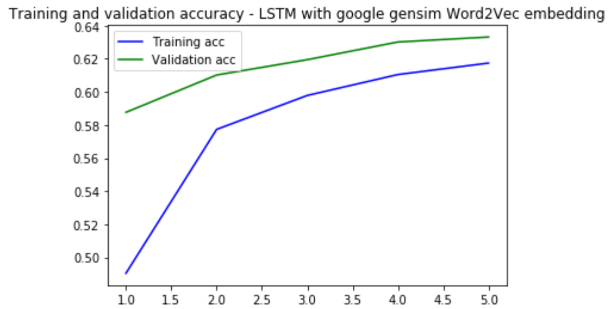
Model Summary:

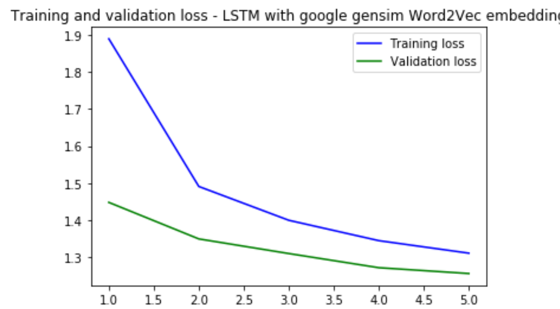
```
Model: "sequential_8"
_________________________________________________________________
Layer (type)              Output Shape           Param #
=================================================================
embedding_6 (Embedding)      (None, 50, 300)        34985400
_________________________________________________________________
spatial_dropout1d_4 (Spatial (None, 50, 300)        0
_________________________________________________________________
lstm_4 (LSTM)            (None, 100)            160400
_________________________________________________________________
dense_11 (Dense)           (None, 40)             4040
=================================================================
Total params: 35,149,840
Trainable params: 164,440
Non-trainable params: 34,985,400
```

Model Score:

Validation loss: 1.2570988957271256
Validation accuracy: 0.6331190466880798

Model Metrics Plots:





Build neural network with LSTM + CNN
- with gensim Word2Vec embedding

The LSTM model worked well. However, it takes forever to train five epochs. One way to speed up the training time is to improve the network adding “Convolutional” layer. Convolutional Neural Networks (CNN) come from image processing. They pass a “filter” over the data and calculate a higher-level representation. They have been shown to work surprisingly well for text, even though they have none of the sequence processing ability of LSTMs.

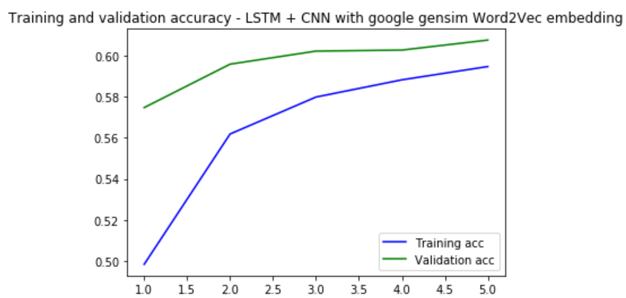
Model Summary:

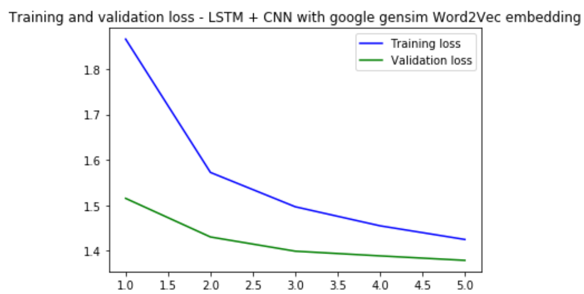
```
Model: "sequential_10"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_6 (Embedding)      (None, 50, 300)           34985400
_________________________________________________________________
dropout_9 (Dropout)          (None, 50, 300)           0
_________________________________________________________________
conv1d_5 (Conv1D)            (None, 46, 64)            96064
_________________________________________________________________
max_pooling1d_5 (MaxPooling1 (None, 11, 64)            0
_________________________________________________________________
lstm_6 (LSTM)                (None, 100)               66000
_________________________________________________________________
dense_13 (Dense)             (None, 40)                4040
=================================================================
Total params: 35,151,504
Trainable params: 166,104
Non-trainable params: 34,985,400
```

Model Score:

Validation loss: 1.3792227489106357
Validation accuracy: 0.6077082753181458

Model Metrics Plots:

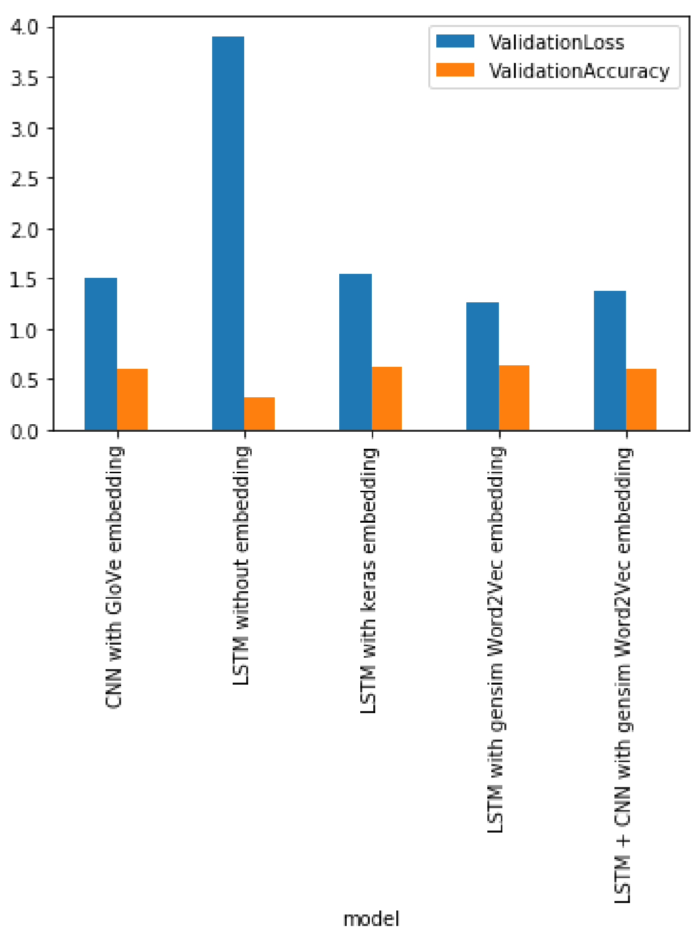




Display models and metrics

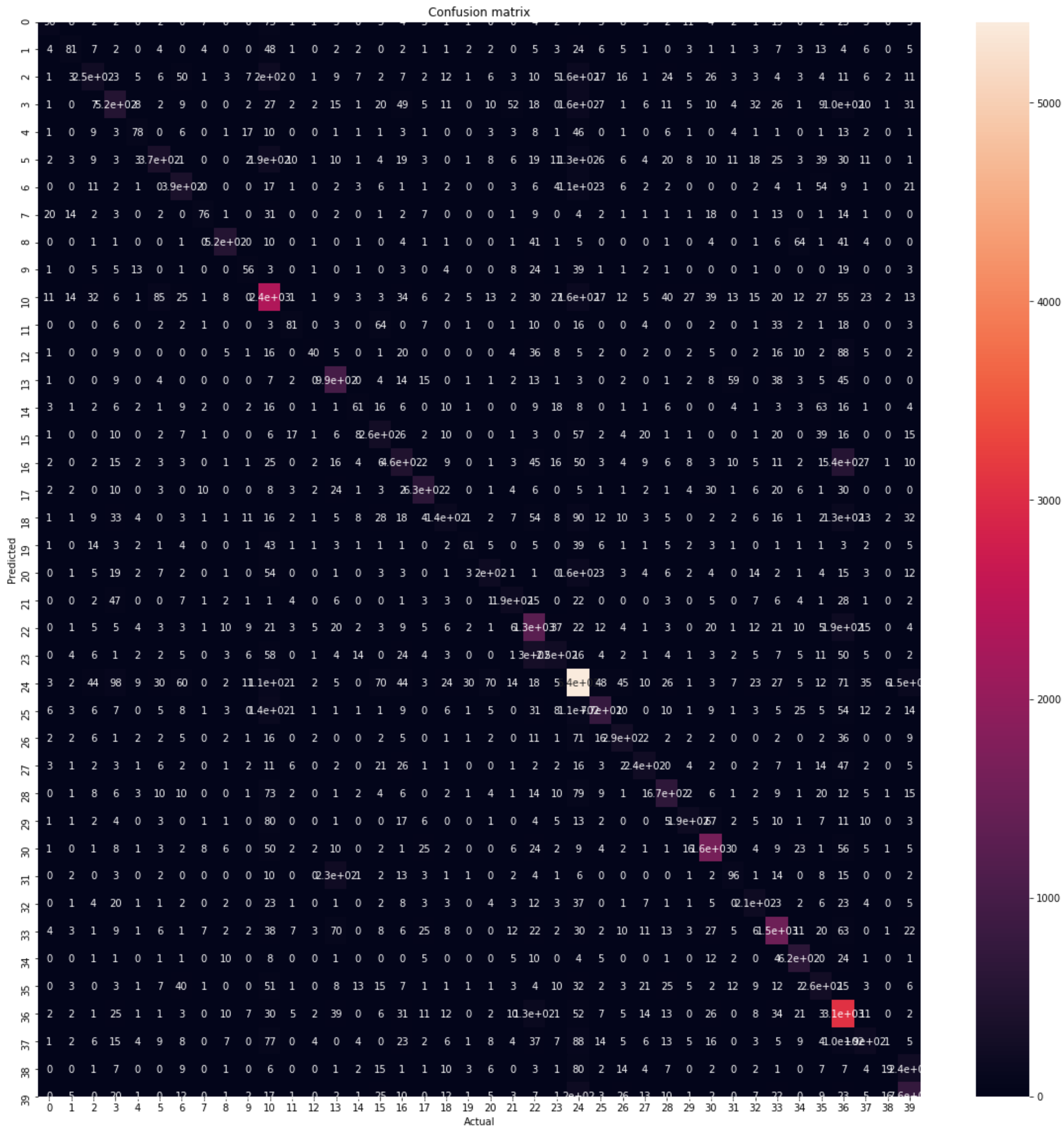
model	ValidationLoss	ValidationAccuracy	
0	CNN with GloVe embedding	1.499375	0.591576
1	LSTM without embedding	3.904507	0.312564
2	LSTM with keras embedding	1.539109	0.620489
3	LSTM with gensim Word2Vec embedding	1.257099	0.633119
4	LSTM + CNN with gensim Word2Vec embedding	1.379223	0.607708

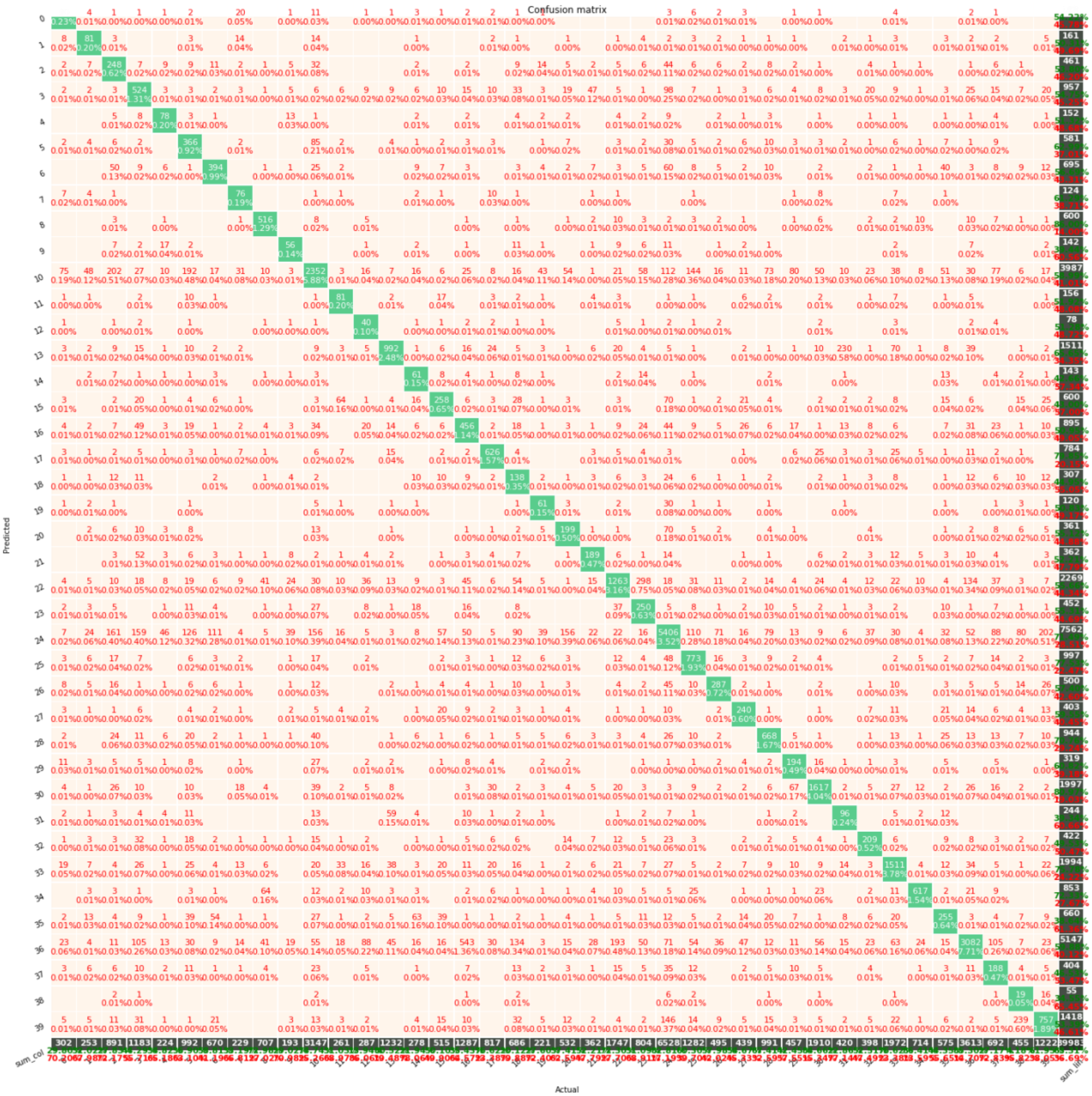
Plot models and metrics



Confusion Matrix - LSTM with gensim Word2Vec embedding

From the above data and plot we clearly see that LSTM model with gensim Word2Vec embedding has given the maximum accuracy (63.3%) and lowest loss (1.26%) compared to other implemented models. Let's take a look at the confusion matrix for the same.

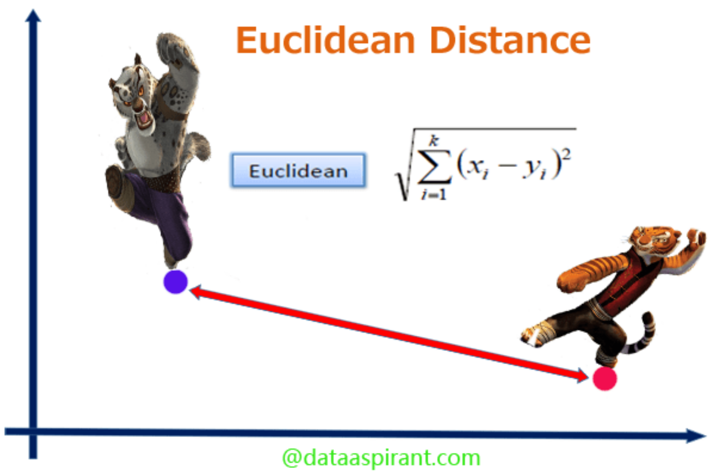



Confusion matrix
Predicted
Actual

Recommendation System

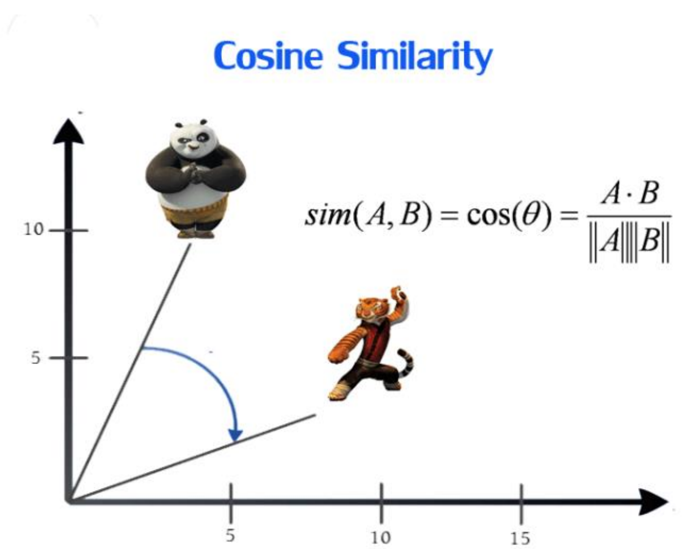
Let us consider only the latest articles from the year 2018 as the size of the dataset is quite large and processing may consume too much time to build our recommendation system.
To find the similarity among sentences or documents which in this project will be the headlines, we will implement Euclidian distance and cosine distance to check for similarity.

Euclidian Distance:



Comparing the shortest distance among two objects. Score means the distance between two objects. If it is 0, it means that both objects are identical.

Cosine Distance:



Determine the angle between two objects is the calculation method to the find similarity. The range of score is 0 to 1. If score is 0, it means that they are same in orientation.

Recommend news articles based on sklearn pairwise_distances with its defaut Euclidean distance metric - BagOfWords

```
News Headline:  Woman Accused Of Poisoning Friend With Cheesecake In Identity Theft Plot

Recommended articles based on the above news headline:
```

	Publish_date	Category	Headline	Euclidean similarity
1	2018-01-03	HEALTHY LIVING	I Was Ghosted By My Best Friend	3.000000
2	2018-01-06	WORLD NEWS	Why I Accused Israel Of Cultural Genocide	3.162278
3	2018-02-21	LATINO VOICES	All They Will Call You Will Be Deportees	3.162278
4	2018-02-01	POLITICS	The Next Financial Crisis -- Not If, But When	3.316625
5	2018-05-03	POLITICS	Is The Left Having A Senior Moment?	3.316625
6	2018-04-16	ENTERTAINMENT	R. Kelly Accused Of 'Knowingly And Intentionally' Infecting Woman With STD	3.316625
7	2018-01-29	ENTERTAINMENT	Here Are All The 2018 Grammy Winners	3.316625
8	2018-01-19	POLITICS	How We Arrived At A 'Shithole' Shutdown	3.316625
9	2018-01-03	POLITICS	Mitt Romney is Not The Answer	3.316625
10	2018-04-28	ENTERTAINMENT	All The Movies That Are Cool For The Summer	3.316625

Using BOW method, the Euclidean distance is very high, so we will investigate with the gensim Word2Vec embedding pre-trained by Google on a 100 billion-word Google News corpus.

Recommend 10 news articles based on the queried news headline by calculating the Euclidean distance and sort by closest similarity

```
News Headline:  Universities Tell Applicants That Protesting Gun Violence Won’t Affect Admissions

Recommended articles based on the above news headline:
```

	Publish_date	Category	Headline	euclidean similarity
1	2018-05-24	POLITICS	House Democrats Offer Internships For Students Affected By Gun Violence	0.890570
2	2018-02-21	EDUCATION	Texas District Says Students Protesting Gun Violence Will Get Suspended	0.935040
3	2018-04-20	POLITICS	Students From 2,600 Schools Plan Walk Outs To Protest Gun Violence	0.945961
4	2018-04-12	BLACK VOICES	Black Students Marched Against Gun Violence In Florida, But You Likely Didn’t Hear About It	0.967323
5	2018-03-14	POLITICS	Hundreds Of D.C.-Area Students Stage Gun Violence Protest At The White House	1.005005
6	2018-02-19	POLITICS	High School Students Lead Protest Against Gun Violence In Front Of White House	1.032745
7	2018-03-23	BLACK VOICES	Black Teens Affected By Gun Violence Speak Out Ahead Of March For Our Lives	1.044926
8	2018-03-14	MEDIA	Fox News All But Ignores Nationwide Student Walkouts To End Gun Violence	1.046540
9	2018-02-21	POLITICS	Florida School's Students And Parents Tearfully Ask Trump To Address Gun Violence	1.059892
10	2018-03-15	POLITICS	'Can't Buy A Kinder Egg, But I Can Buy An AR-15': NYC Students Protest Gun Laws On Walkout Day	1.065780

Recommend 10 news articles based on the queried news headline by calculating the Cosine distance and sort by closest similarity

```
News Headline:  Universities Tell Applicants That Protesting Gun Violence Won’t Affect Admissions

Recommended articles based on the above news headline:
```

	Publish_date	Category	Headline	cosine similarity
1	2018-05-24	POLITICS	House Democrats Offer Internships For Students Affected By Gun Violence	0.211393
2	2018-02-21	EDUCATION	Texas District Says Students Protesting Gun Violence Will Get Suspended	0.237304
3	2018-04-20	POLITICS	These Are The Students Walking Out Of School To Protest Gun Violence	0.251322
4	2018-04-20	POLITICS	Students From 2,600 Schools Plan Walk Outs To Protest Gun Violence	0.258947
5	2018-04-12	BLACK VOICES	Black Students Marched Against Gun Violence In Florida, But You Likely Didn’t Hear About It	0.274205
6	2018-03-13	POLITICS	Students Have The Right To Participate In Gun Violence Walkouts	0.286843
7	2018-03-15	POLITICS	School Walkouts Were Just The Beginning Of Students’ Activism On Gun Violence	0.292607
8	2018-03-14	POLITICS	These Photos Show The Strength Of Students As They Protest Gun Violence	0.298992
9	2018-03-14	MEDIA	Fox News All But Ignores Nationwide Student Walkouts To End Gun Violence	0.300997
10	2018-03-23	BLACK VOICES	Black Teens Affected By Gun Violence Speak Out Ahead Of March For Our Lives	0.301012

For the queried headline, comparing the recommended news articles by headline similarity between the Euclidean and Cosine distances, it can be noticed that 6 out of top 10 recommendations are similar. Key words in the headline used for recommending appears to be "Gun Violence" and "Applicants". Based on the context of the queried article, looks like the key recommendations are about Students, School and activities that are related to "Gun Violence".

Recommend 10 news articles based on the queried news headline and category by calculating the Euclidean distance and sort by closest similarity

```
# specify weights for headline and category
headline_category_model(528,10,0.1,0.8,'euclidean')
```

```
News Headline:  Universities Tell Applicants That Protesting Gun Violence Won't Affect Admissions
Category :  EDUCATION

Recommended articles based on the above news headline:
```

	Publish_date	Category	Headline	Weighted euclidean similarity	Word2Vec based euclidean similarity	Category based euclidean similarity
1	2018-02-21	EDUCATION	Texas District Says Students Protesting Gun Violence Will Get Suspended	0.103893	0.935040	0.0
2	2018-04-04	EDUCATION	Oklahoma Teachers Begin 110-Mile March To Protest Education Funding	0.130319	1.172869	0.0
3	2018-02-23	EDUCATION	West Virginia Teachers Are Making Sure Their Students Get Fed While They're On Strike	0.136202	1.225820	0.0
4	2018-02-06	EDUCATION	Homeless Students, Destroyed Campuses, 'Invisible Injuries': What California Schools Learned From Recent Disasters	0.141139	1.270249	0.0
5	2018-04-02	EDUCATION	Teachers Swarm Kentucky Capitol To Protest Pension Changes, School Budget Cuts	0.141732	1.275585	0.0
6	2018-04-06	EDUCATION	Puerto Rico To Shutter 283 More Schools This Summer As Education Crisis Deepens	0.141966	1.277695	0.0
7	2018-01-30	EDUCATION	Columbia University Refuses To Recognize Graduate Student Union	0.143250	1.289250	0.0
8	2018-02-07	EDUCATION	While Teachers Fight For Better Pay, West Virginia Lawmakers Discuss Opossums	0.144239	1.298151	0.0
9	2018-04-16	EDUCATION	Beyoncé Announces $100,000 In Scholarships For HBCU Students	0.144785	1.303064	0.0

Recommend 10 news articles based on the queried news headline and category by calculating the Cosine distance and sort by closest similarity

```
headline_category_model(528,10,0.1,0.8,'cosine')
```

```
News Headline:  Universities Tell Applicants That Protesting Gun Violence Won't Affect Admissions
Category :  EDUCATION

Recommended articles based on the above news headline:
```

	Publish_date	Category	Headline	Weighted cosine similarity	Word2Vec based cosine similarity	Category based cosine similarity
1	2018-02-21	EDUCATION	Texas District Says Students Protesting Gun Violence Will Get Suspended	0.026367	0.237304	0.0
2	2018-04-04	EDUCATION	Oklahoma Teachers Begin 110-Mile March To Protest Education Funding	0.043847	0.394620	0.0
3	2018-01-30	EDUCATION	Columbia University Refuses To Recognize Graduate Student Union	0.047527	0.427743	0.0
4	2018-04-02	EDUCATION	Teachers Swarm Kentucky Capitol To Protest Pension Changes, School Budget Cuts	0.048464	0.436180	0.0
5	2018-02-06	EDUCATION	Homeless Students, Destroyed Campuses, 'Invisible Injuries': What California Schools Learned From Recent Disasters	0.048697	0.438272	0.0
6	2018-02-23	EDUCATION	West Virginia Teachers Are Making Sure Their Students Get Fed While They're On Strike	0.049156	0.442404	0.0
7	2018-05-17	EDUCATION	The Controversial Way Some California Schools Are Handling Students' Misbehavior	0.050584	0.455252	0.0
8	2018-01-11	EDUCATION	Texas Schools Illegally Excluded Students With Disabilities: Federal Officials	0.051160	0.460438	0.0
9	2018-02-07	EDUCATION	While Teachers Fight For Better Pay, West Virginia Lawmakers Discuss Opossums	0.054208	0.487873	0.0

Here we are recommending articles based on category and headline. We specify the weights for the category and headline as parameters to the function based on which articles are selected. For the above recommendations we chose the weights to be 0.1 for the headline and 0.8 for category.

For the queried headline, comparing the recommended news articles by category and headline between Euclidean and Cosine distances, it can be noticed that 7 out of top 9 recommendations are similar. As the category was given more weight, all the recommended headlines are Education based and only one headline had the key word "Gun Violence".

Based on the context of the queried article, looks like the key recommendations are about Students, Schools, Teachers and Universities and no relevance to "Gun Violence".

Recommend 10 news articles based on the queried news headline, category and author by calculating the Euclidean distance and sort by closest similarity

```
headline_category_author_model(528,11,0.1,0.1,1,'euclidean')
```

```
News Headline :  Universities Tell Applicants That Protesting Gun Violence Won’t Affect Admissions
Category :  EDUCATION
Author :  Carla Herreria

Recommended articles based on the above news headline:
```

	Publish_date	Category	Authors	Headline	Weighted euclidean similarity	Word2Vec based euclidean similarity	Category based euclidean similarity	Author based euclidean similarity
1	2018-04-29	WORLD NEWS	Carla Herreria	Thousands Protest Across Spain After 5 Men Are Cleared Of Gang Rape	0.219107	1.215068	1.414214	0.0
2	2018-03-03	POLITICS	Carla Herreria	Steven Mnuchin Doesn't Want People To See Video Of His Heckled UCLA Talk	0.219438	1.219038	1.414214	0.0
3	2018-01-24	SPORTS	Carla Herreria	Trustee Defends MSU President, Dismissing Sex Abuse Reports As ‘Nassar Thing’	0.220176	1.227899	1.414214	0.0
4	2018-01-27	SPORTS	Carla Herreria	MSU Students Wear Teal To Show Support For Survivors Of Larry Nassar's Abuse	0.221702	1.246208	1.414214	0.0
5	2018-03-15	CRIME	Carla Herreria	Dylann Roof's Sister Accused Of Having Weapons At School During National Walkouts	0.221778	1.247125	1.414214	0.0
6	2018-04-15	QUEER VOICES	Carla Herreria	Prominent LGBTQ Lawyer Sets Self On Fire In 'Protest Suicide' Of Climate Change	0.222265	1.252965	1.414214	0.0
7	2018-03-28	BLACK VOICES	Carla Herreria	Stephon Clark's Brother Shuts Down City Hall Meeting As Protests Continue	0.222455	1.255244	1.414214	0.0
8	2018-03-30	BLACK VOICES	Carla Herreria	Howard Students Take Over Building To Protest University Embezzlement Scandal	0.223334	1.265800	1.414214	0.0
9	2018-02-01	POLITICS	Carla Herreria	Rep. Adam Schiff: GOP's FBI Memo Could Lead To 'Constitutional Crisis'	0.223996	1.273744	1.414214	0.0
10	2018-03-30	BUSINESS	Carla Herreria	MyFitnessPal Security Breach Affects 150 Million Users, Under Armour Reports	0.225209	1.288291	1.414214	0.0

Recommend 10 news articles based on the queried news headline, category and author by calculating the Cosine distance and sort by closest similarity

```
headline_category_author_model(528,11,0.1,0.1,1,'cosine')
```

```
News Headline :  Universities Tell Applicants That Protesting Gun Violence Won’t Affect Admissions
Category :  EDUCATION
Author :  Carla Herreria

Recommended articles based on the above news headline:
```

	Publish_date	Category	Authors	Headline	Weighted cosine similarity	Word2Vec based cosine similarity	Category based cosine similarity	Author based cosine similarity
1	2018-03-30	BLACK VOICES	Carla Herreria	Howard Students Take Over Building To Protest University Embezzlement Scandal	0.117567	0.410806	1.0	0.0
2	2018-04-29	WORLD NEWS	Carla Herreria	Thousands Protest Across Spain After 5 Men Are Cleared Of Gang Rape	0.119837	0.438043	1.0	0.0
3	2018-03-22	POLITICS	Carla Herreria	Hawaii Democrat Resigns In Response To Sexual Harassment Claims He Still Denies	0.121730	0.460760	1.0	0.0
4	2018-03-15	CRIME	Carla Herreria	Dylann Roof's Sister Accused Of Having Weapons At School During National Walkouts	0.121809	0.461703	1.0	0.0
5	2018-03-03	POLITICS	Carla Herreria	Steven Mnuchin Doesn't Want People To See Video Of His Heckled UCLA Talk	0.122080	0.464963	1.0	0.0
6	2018-01-24	SPORTS	Carla Herreria	Trustee Defends MSU President, Dismissing Sex Abuse Reports As 'Nassar Thing'	0.122857	0.474288	1.0	0.0
7	2018-03-02	POLITICS	Carla Herreria	Hawaii Democrat Defends Stance On Guns After Actress Questions Her Silence On Bill	0.123477	0.481730	1.0	0.0
8	2018-03-25	POLITICS	Carla Herreria	Sen. Marco Rubio Tells Students He Does Not Agree With The March For Our Lives	0.123645	0.483738	1.0	0.0
9	2018-04-15	QUEER VOICES	Carla Herreria	Prominent LGBTQ Lawyer Sets Self On Fire In 'Protest Suicide' Of Climate Change	0.123948	0.487372	1.0	0.0
10	2018-01-27	SPORTS	Carla Herreria	MSU Students Wear Teal To Show Support For Survivors Of Larry Nassar's Abuse	0.124005	0.488057	1.0	0.0

Here we are recommending articles based on author, category and headline. We specify the weights for the author, category and headline as parameters to the function based on which articles are selected. For the above recommendations we choose the weights to be 0.1 for the headline, 0.1 for category and 1.0 for the author.

For the queried headline, comparing the recommended news articles by author, category and headline between Euclidean and Cosine distances, it can be noticed that 7 out of top 10 recommendations are similar. As the author was given more weight, all the recommended headlines are from the author Carla Herreria with varying categories and shows no relevance to key word "Gun Violence" and very few articles have relevance to applicants or universities, which was expected due to the low weightage given.

Recommend 10 news articles based on the queried news headline, category, author and published day by calculating the Euclidean distance and sort by closest similarity

```
headline_category_author_pubday_model(528,10,0.1,0.1,0.1,1,'euclidean')
```

```
News Headline :  Universities Tell Applicants That Protesting Gun Violence Won't Affect Admissions
Category :  EDUCATION
Author :  Carla Herreria
Day-Month :  Sat_Feb

Recommended articles based on the above news headline:
```

	Publish_date	Categoty	Authors	Day and month	Headline	Weighted euclidean similarity with the queried article	Word2Vec based euclidean similarity	Category based euclidean similarity	Authors based euclidean similarity	Publishing day based euclidean similarity
1	2018-02-24	BLACK VOICES	Carla Herreria	Sat_Feb	Tribal Filipinos Were A Surprising Muse For 'Black Panther's' Dora Milaje	0.223684	1.493681	1.414214	0.000000	0.0
2	2018-02-17	POLITICS	Carla Herreria	Sat_Feb	Florida Gubernatorial Candidate Calls On Governor To Halt AR-15 Sales	0.227345	1.541267	1.414214	0.000000	0.0
3	2018-02-17	SPORTS	Carla Herreria	Sat_Feb	U.S. Figure Skater Nathan Chen Redeems Himself With Record-Setting Skate	0.242649	1.740228	1.414214	0.000000	0.0
4	2018-02-24	POLITICS	Jonathan Cohn	Sat_Feb	This Is What A Serious Gun Violence Policy Would Look Like	0.301676	1.093364	1.414214	1.414214	0.0
5	2018-02-03	POLITICS	Akbar Shahid Ahmed	Sat_Feb	Years Of U.S. Government Lies Could Soon Result In A Kurdish Massacre	0.310167	1.203742	1.414214	1.414214	0.0

	Publish_date	Categoty	Authors	Day and month	Headline	Weighted euclidean similarity with the queried article	Word2Vec based euclidean similarity	Category based euclidean similarity	Authors based euclidean similarity	Publishing day based euclidean similarity
6	2018-02-17	POLITICS	Sebastian Murdock	Sat_Feb	Floridians Tell Politicians Who Do The NRA's Bidding Their Time Is Up	0.311750	1.224327	1.414214	1.414214	0.0
7	2018-02-24	POLITICS	Mary Papenfuss	Sat_Feb	Trump's Defense Of Aide Accused Of Domestic Violence Is Cited In College Sexual Bias Lawsuit	0.313213	1.243348	1.414214	1.414214	0.0
8	2018-02-10	BLACK VOICES	Carol Kuruvilla	Sat_Feb	Students Walk Out After Princeton Professor Uses Racial Slur In Class On Hate Speech	0.314013	1.253738	1.414214	1.414214	0.0
9	2018-02-17	POLITICS	Mary Papenfuss	Sat_Feb	Dianne Feinstein Wants To Raise Minimum Age For Assault Weapon Purchases To 21	0.314396	1.258725	1.414214	1.414214	0.0

Recommend 10 news articles based on the queried news headline, category, author and published day by calculating the Cosine distance and sort by closest similarity

```
headline_category_author_pubday_model(528,10,0.1,0.1,0.1,1,'cosine')
```

```
News Headline :  Universities Tell Applicants That Protesting Gun Violence Won’t Affect Admissions
Category :  EDUCATION
Author :  Carla Herreria
Day-Month :  Sat_Feb

Recommended articles based on the above news headline:
```

	Publish_date	Categoty	Authors	Day and month	Headline	Weighted cosine similarity with the queried article	Word2Vec based cosine similarity	Category based cosine similarity	Authors based cosine similarity	Publishing day based cosine similarity
1	2018-02-17	POLITICS	Carla Herreria	Sat_Feb	Florida Gubernatorial Candidate Calls On Governor To Halt AR-15 Sales	0.124633	0.620228	1.0	0.0	0.0
2	2018-02-24	BLACK VOICES	Carla Herreria	Sat_Feb	Tribal Filipinos Were A Surprising Muse For ‘Black Panther’s’ Dora Milaje	0.129189	0.679462	1.0	0.0	0.0
3	2018-02-17	SPORTS	Carla Herreria	Sat_Feb	U.S. Figure Skater Nathan Chen Redeems Himself With Record-Setting Skate	0.137260	0.784383	1.0	0.0	0.0
4	2018-02-24	POLITICS	Jonathan Cohn	Sat_Feb	This Is What A Serious Gun Violence Policy Would Look Like	0.179825	0.337728	1.0	1.0	0.0
5	2018-02-24	POLITICS	Mary Papenfuss	Sat_Feb	Trump's Defense Of Aide Accused Of Domestic Violence Is Cited In College Sexual Bias Lawsuit	0.184988	0.404845	1.0	1.0	0.0
6	2018-02-10	BLACK VOICES	Carol Kuruvilla	Sat_Feb	Students Walk Out After Princeton Professor Uses Racial Slur In Class On Hate Speech	0.186303	0.421944	1.0	1.0	0.0

	Publish_date	Categoty	Authors	Day and month	Headline	Weighted cosine similarity with the queried article	Word2Vec based cosine similarity	Category based cosine similarity	Authors based cosine similarity	Publishing day based cosine similarity
7	2018-02-24	POLITICS	Carol Kuruvilla	Sat_Feb	Evangelical Leaders Say 'Pro-Life Ethic' Means Fighting For Gun Reform	0.187962	0.443500	1.0	1.0	0.0
8	2018-02-03	POLITICS	Akbar Shahid Ahmed	Sat_Feb	Years Of U.S. Government Lies Could Soon Result In A Kurdish Massacre	0.188847	0.455010	1.0	1.0	0.0
9	2018-02-17	POLITICS	Lee Moran	Sat_Feb	Former Mexican President: Mass Shootings Are Consequence Of Racism Like Trump's	0.190890	0.481564	1.0	1.0	0.0

Here we are recommending articles based on published day, author, category and headline. We specify the weights for the published day, author, category and headline as parameters to the function based on which articles are selected. For the above recommendations we choose the weights to be 0.1 for the headline, 0.1 for category, 0.1 for the author and 1.0 for the published day.

For the queried headline, comparing the recommended news articles by published day, author, category and headline between Euclidean and Cosine distances, it can be noticed that 7 out of top 9 recommendations are similar. As the published day was given more weight, all the recommended headlines are from the same days of the month with varying categories and authors, with some relevance to key words "Gun" and "Violence" and see no more references to key words "Students" or "Universities".

In general, between the Euclidean and Cosine distances there is not much difference in the recommended articles. It would be essential to have higher weightage to the headline so that similar articles of interest can be recommended to engage the users.