Springboard Data Science Capstone Project - Predicting news articles using topic modeling.

Ramaa Vissa

Apr 16, 2020

Contents

1. Introduction

If you are on your favorite news site browsing, wouldn't it be nice to view articles of interest. That is the main intention of this project.
There are several news websites that would like to show users favorite articles that they would like to browse and have a large user base. Every news site likes to have a huge subscription of users.
At the end of this prediction, Latent Semantic indexing technique will suggest a few topics based on content

similarity. The context of the words is taken into consideration for this technique.

Problem: “what are the favorite news article that a particular user will read” - The problem can also be stated as ‘***suggesting favorite content based news article recommendations to engage and captivate users***’

2. Data

The data set contains 6 months worth of data with various articles.
This is an NLP problem. This use case is a real life scenario, the acquired data sets are genuine and direct from our real world customer. The file itself was created by another technique that is outside the scope of the project. The file format is in the csv format. The video id name seems like a misnomer, it is simply an identifier id that is distinct for that record. The original data set is from a television channel in India. The channel runs a popular news website, and we were chosen to do predictions for the channel’s website. A test case measures the metrics of our technique.

	video_id	category_name	short_description	story_text	title
0	5653616771001	Entertainment News	The 34-year-old actor is not sure if he's goin...	Hack: \tThe 34-year-old actor is not sure if h...	Adam Driver still 'undecided' on seeing 'Star ...

Data dictionary

Video_id - identifier
Category Name - there are 10 categories
Short description - short description of the story text
Story_text - This is the article content. (Contains the text we are interested in)
Title - This is a short description of the title.

3. Model Overview

A brief overview of Latent Semantic Analysis.
This technique utilises the content based ***context*** for finding similar words and suggest similar topics. For example if the user has been browsing the Entertainment news category, which has news that is characteristic of that category, similar words are predicted.

Inner workings

Topic Model - This model can be defined as an Unsupervised technique to discover similar topics across documents. Topics have a certain

distribution in a document, and every topic has a proportion of different words it contains. It is used in NLP problems.

<u>Steps in involved in the Implementation</u>

M - number of text documents

N - number of total unique terms (words)

Our goal is to extract k topics. First generate a document term matrix of shape mXn having tf-idf scores.

Then reduce the mXn matrix to k dimensions using SVD. (Singular value decomposition)

SVD gives us vector representation of every document and term in our data. The length of each vector would be k. With these vectors find similar words and similar documents using cosine similarity.

(

SVD -

if A is a matrix and if SVD is used to decompose the matrix, we get U,S and a transpose matrix of V.

Uk - is the document - term matrix, each row in U is a vector representation of the corresponding document. k is the number of desired topics.

Vk - is the term topic matrix, gives the vector representation of terms in the story_text column.

S is a kXk matrix essentially a scalar value which equals the number of topics.
)

4. Data Preprocessing

- The story_text column will be cleaned and tokenized. First the stop words are removed and the words are sent through a lemmatizer to get base words. Unique words are filtered out after removing rare words. Then the corpus is created utilising doc2bow for the ‘story_text’, via a bag of words approach.

  Results from these techniques are discussed -
- Cosine similarity with Tf Idf vectorization.
- Gensim LDA model
- LSA

  The main objective of this project is to visualise and display the groups of topics that a particular user is interested in.

Model discussion -

- Cosine similarity - After tfidf vectorization of the cleaned text, a ‘similarity’ function determines if the word ‘news’ has other similar topics that the model can suggest. In our case the following are all the suggested topics, which suggests various news articles that are from the category ‘India news’ which is the largest category which aligns well when we visually inspect the articles.

['News Wrap: All the top headlines from across the world',
'Prime Minister directs withdrawal of fake news notice; onus put on Press Council of India',
"National Wrap: All That's Trending In The Country",
"National Wrap: All That's Trending In India",
'National Wrap: Latest Trending News In India On March 9',
'National Wrap: Latest Trending News In India On March 10',
'National Wrap: Latest Trending News In India On March 6',
'National Wrap: Latest Trending News In India On March 7',
'National Wrap: Latest Trending News In India On March 5',

'LIVE NOW: Republic World App - Light On Your Phone; Heavy On The News!']

With words ‘'post shared picture khan' our model suggests the following topics, which is correct.

["Shah Rukh Khan’s daughter Suhana shamed for wearing 'short dress' while meeting grandmother",
'Shah Rukh Khan’s 4-year-old son AbRam is furious with the paparazzi! Here’s why',
'Suhana Khan and her friends visit Taj Mahal; See first photos',
'Suhana Khan And Her BFF Hit The Pool; Beat The Summer In Style',
'Gauri Khan just made a big announcement about her daughter Suhana Khan',
'Shah Rukh Khan’s son Aryan mobbed in London?',
'This ‘Flawless’ Pic Of Suhana Khan Is Going Viral',
"Celebrations begin ahead of Shah Rukh Khan's 52nd birthday",
"Netizens shame Suhana Khan for 'pushing her butt out' while posing",
'This t-shirt of Kareena Kapoor Khan costs Rs 45,000?']

- Gensim LDA model
  After data pre processing steps the gensim lda model is used to train the corpus.
  Utilising the class ’similarities.Matrixsimilarity’ similar topics are predicted. Looking at the results the top has higher similarity scores than the ones at the bottom of the list.

(346, 0.99999994) Adam Driver still 'undecided' on seeing 'Star Wars: The Last Jedi'
(2129, 0.99999994) BEWARE: Akshay Kumar's look from 2.0 is out, and it is deadly
(4263, 0.99999994) Kathua-Rape-Murder Case: Protesters demand CBI probe
(4016, 0.9999998) Sridevi’s death: Here’s what Amar Singh claims about the ‘alcohol angle’
(1930, 0.99999976) Churni's next film on social media affecting personal lives
(968, 0.9999994) Bihar: Forced to marry widowed sister-in-law, 15-Year-Old boy ends life
(288, 0.9999989) Sridevi’s death: No conflict between two families of Boney Kapoor
(294, 0.99999887) Major fire breaks out in a factory in Rajasthan
(3746, 0.9999985) What is the link between Gagan Dhawan and Ahmed Patel?
(3766, 0.9999984) Chole Bhature was before "symbolic" fast, says Congress' Lovely. Logic gets panned on social media

- The topics from the LSA model will be interpreted. I have a sample output below and the interpretation from my end.

For reference our original data set has these categories, in the order of decreasing popularity which a typical user can browse any of these topics.

- India News
- Entertainment News

- Sports News
- World News
- R Bharat
- Technology News
- Business News
- Lifestyle
- Initiatives
- Karnataka Elections 2018

From the model analysis I can verify successfully that words from the topics from the model output are categorised as follows- topic 0 and topic 1 is ‘India news’. Topic 2 is Sports News. Topic 3 and 4 as ‘Entertainment News’.

Topic 0: film india minister actor khan congress post
Topic 1: Congress minister modi government gandhi prime party
Topic 2: pakistan india indian cricket match world team
Topic 3: court case salman police khan republic accused
Topic 4: sridevi kapoor dubai boney janhvi pakistan death

Conclusion/Interpretation:

The model suggestions align very well with our visual inspection of the results. All these are good techniques

that can be used. This is an unsupervised learning, the labels can be interpreted manually, and the model has predicted the topics correctly.

A sample of the data was taken to validate these results. A real life use case for testing showed that there was a promising 10% increase in user engagement while gauging the audience analytics.

Further enhancements:

An end to end project with Flask can be built. A service layer API can be built. A persistence layer such as Mongo DB can be used.