Springboard Data Science Capstone Project - Predicting User preferred news articles

Ramaa Vissa

apr 14, 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.
This is an NLP problem.

Problem: “what are the favorite news article that a particular user will read”

2. Data

The data has text columns, video_id, story_text,category_name,short_description.

The acquired data sets were created by logs and is already in the csv format.

	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 ...

3. Data Cleaning and Model discussion

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’. Now utilise the gensim lda model and train the corpus. A list of topics is created using the ldamodel.

I chose a few parameters to tune the model and saved the model to disk. (LDA is an unsupervised generative model that assigns topic distribution to documents.)
To find the similar topics, cosine similarity is used to find the articles that might interest users.
LSA has been utilized to find similar topics. Code implementation is in the Github folder.

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.