DATA MINING HOMEWORK 1

VEDASRI UPPALA

500202008

uppala.1@osu.edu

TABLE OF CONTENTS

1. DATA SET DESCRIPTION

The dataset is a description of all the audio and video recordings of the talks that are available on the ted.com website as of 21 September 2017. Each row in the data set contains information about a specific talk. The data set has 2550 tuples and 17 attributes.

The following table describes the attributes in the dataset.

Name of the Attribute	Description
name	Name of the talk (includes title and speaker's name)
title	Title of the talk
description	Short description of what the talk is about
main_speaker	The first named speaker of the talk
speaker_occupation	Main speaker's occupation
num_speaker	Number of speakers involved in the talk
duration	Duration of the talk
event	Event where the talk happened
film_date	Date on which the talk is recorded
published_date	Date on which talk is uploaded on ted website
Comments	Number of comments made by viewers
tags	Themes associated with the talk
languages	Number of languages in which the talk is made available
ratings	Dictionary of several ratings given to the talk
Related_talks	List of recommended talks
url	URL of the talk
views	Number of views received by the talk

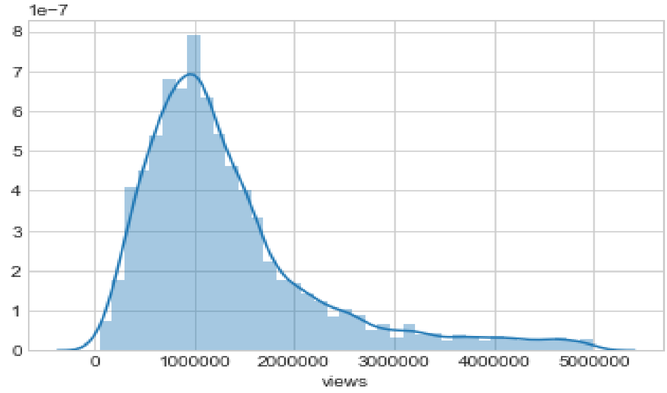
2. DATA PRE-PROCESSING

The data set has been loaded into a python data frame and the following transformations have been applied to ease the process of data exploration and analysis.

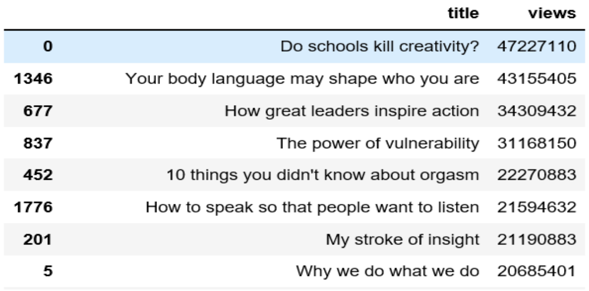
- The duration of each talk is given in minutes in the "duration". This has been converted into minutes.
- The values for the attributes "film_date" and "published_date" are given as unix time stamps. They have been converted to a readable dd-mm-yyyy format.
- The columns of the data have been logically re-ordered.
- The attribute 'name' doesnot give any useful information since it is a concatenation of title of the talk and the speaker's name. So this attribute has been eliminated.

3. ANALYSIS OF VIEWS

A histogram is plotted to visualize how the attribute ‘views’ is distributed.



The average(mean) number of views a ted talk garnered is 1698297 and the standard deviation is 2.498479e+06

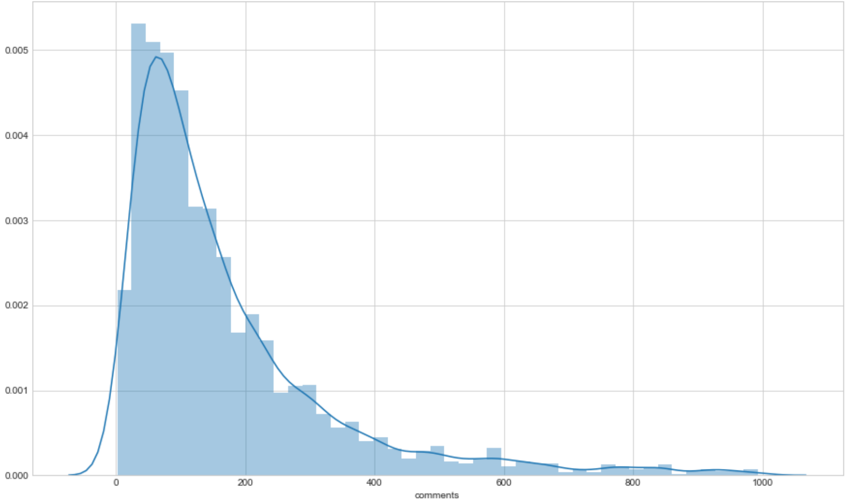


	title	views
0	Do schools kill creativity?	47227110
1346	Your body language may shape who you are	43155405
677	How great leaders inspire action	34309432
837	The power of vulnerability	31168150
452	10 things you didn't know about orgasm	22270883
1776	How to speak so that people want to listen	21594632
201	My stroke of insight	21190883
5	Why we do what we do	20685401

The data has been sorted in descending order on number of views each talk has garnered. It is observed that the talk titled “Do Schools kill creativity” by Ken Robinson is the most viewed talk in the history of ted.

4. ANALYSIS OF COMMENTS

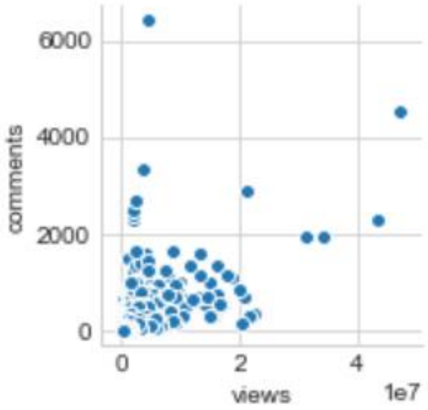
The distribution of comments is found to be the following.

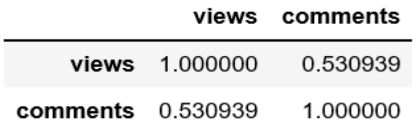


- *From the graph we can see that not many ted talks have comments greater than 1000.*
- *Also, it is found that each Ted talk on an average has 191 comments and the standard deviation is found to be 282.*

The talk titled “Militant Atheism” by Richard Dawkins has gathered highest number of comments (6404) in the history of Ted. This could be because Atheism is a controversial and debatable topic. I have also observed that the most commented talks in ted discuss contentious topics which lead to debates in the comments section.

Intuitively we expect the most viewed videos to have higher number of comments. So, I have tried to plot comments and views using pair plot in python. From the graph it seems that views and comments are correlated to each other.





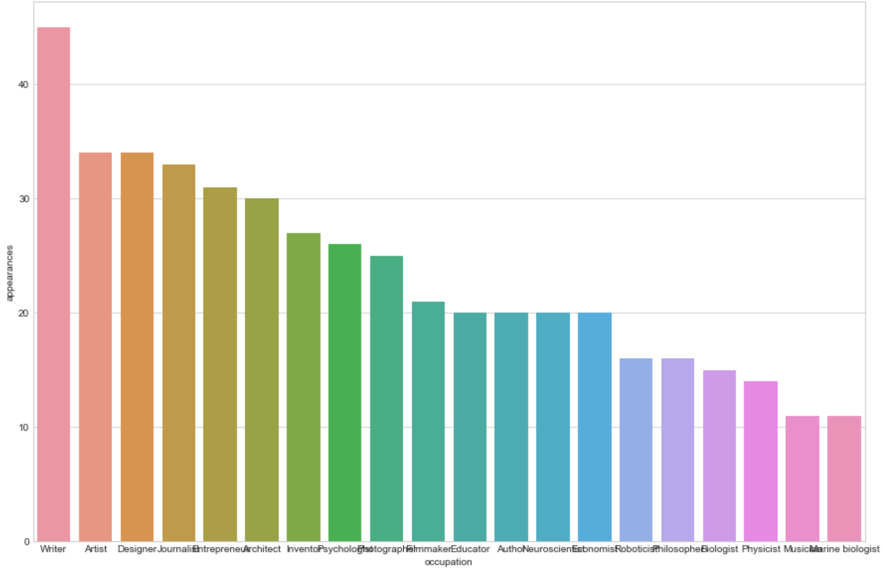
	views	comments
views	1.000000	0.530939
comments	0.530939	1.000000

Pearson coefficient = 0.530939. The obtained value of Pearson coefficient suggests medium to high positive correlation between views and comments of ted talks

5. ANALYSIS OF SPEAKERS OF TED TALKS

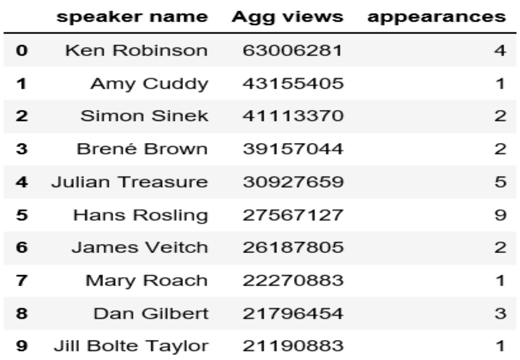
For each speaker, the number of appearances he/she has made was calculated. This data is analyzed and the mean was found to be around 1 suggesting that each ted speaker on an average gave 1 talk and the maximum number of talks given by a ted speaker is 9.

Next the relation between occupation of the speaker and the appearnces made by them is analyzed. The Bar graph below shows the results.



It can be observed from the bar graph that writers have made highest number of appearances as speakers in ted talks followed by Artists, Designers and Journalists.

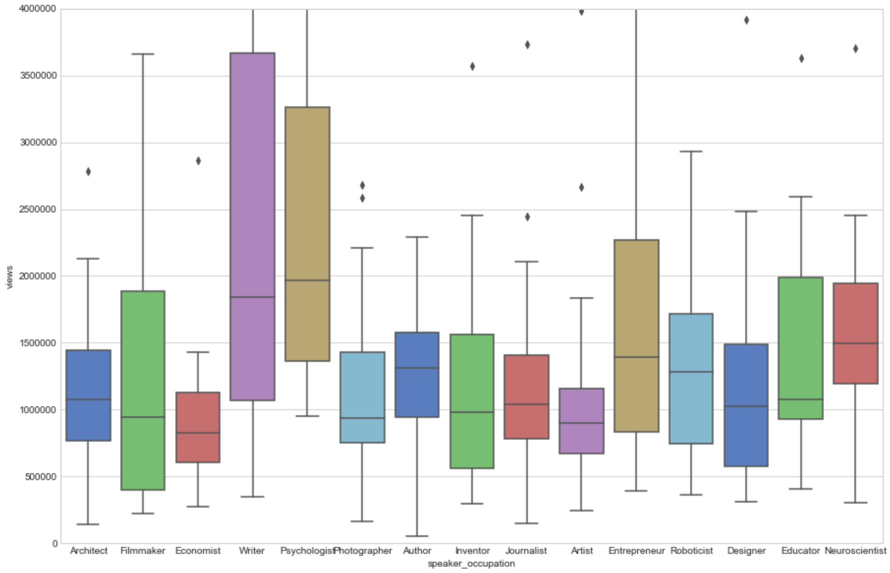
Next the aggregate views each speaker has received for all their videos has been calculated. Results are shown below.



	speaker name	Agg views	appearances
0	Ken Robinson	63006281	4
1	Amy Cuddy	43155405	1
2	Simon Sinek	41113370	2
3	Brené Brown	39157044	2
4	Julian Treasure	30927659	5
5	Hans Rosling	27567127	9
6	James Veitch	26187805	2
7	Mary Roach	22270883	1
8	Dan Gilbert	21796454	3
9	Jill Bolte Taylor	21190883	1

It can be seen that though Ken Robinson appeared only 4 times he has got the highest number of aggregate views. Hans Rosling who gave 9 ted talks stood in the fifth place when aggregate views are considered. This suggests that the viewers of ted talks consider content as more important than the familiarity with the speaker.

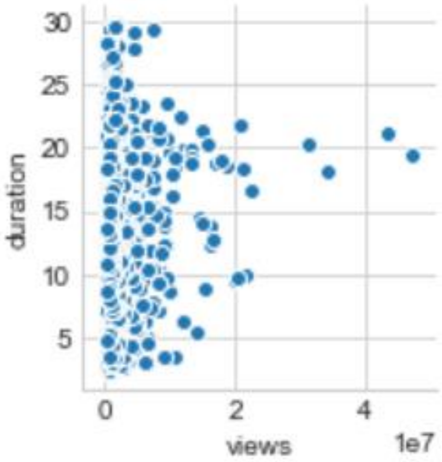
Next the relation between occupation of the speaker and number of views is explored. For each speaker occupation, the total number of views obtained is calculated and the box plot is shown below.

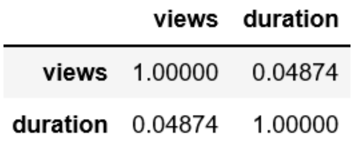


From the graph we can see that though writers have made more appearances, psychologists on average have garnered higher number of views compared to writers.

6. ANALYSIS ON DURATION OF TED TALKS

From the data given, the average duration of a ted talk is calculated and is found to be around 13 minutes. It has been observed that 75% of ted talks have duration which is approximately 17min. The longest talk is 87.6 minutes long which technically is not a ted talk. Intuitively, duration and views are expected to be negatively correlated because usually people tend to lose interest if the talk is long.



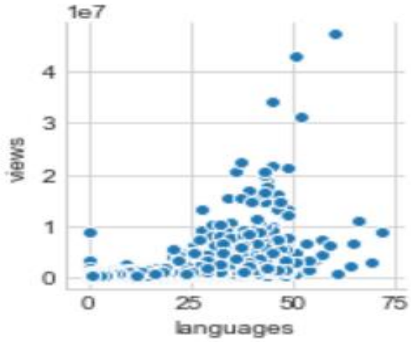


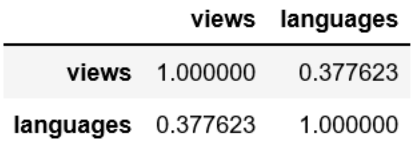
	views	duration
views	1.00000	0.04874
duration	0.04874	1.00000

The very low pearson coefficient and the plot for duration vs views indicate that duration of ted talks and views are actually not correlated to each other. So we can conclude that if the speaker is able to capture viewers' interest and if the content is worth watching then viewers don't have a problem with longer videos.

7. ANALYSIS ON LANGUAGES

The languages attribute describes the number of languages in which the talk is available. From my analysis, each ted talk on average is available in 27 languages. If a talk is available in more languages the number of viewers will also increase . To test this hypothesis the relation between number of languages in which the ted talk is available and the number of views is explored.

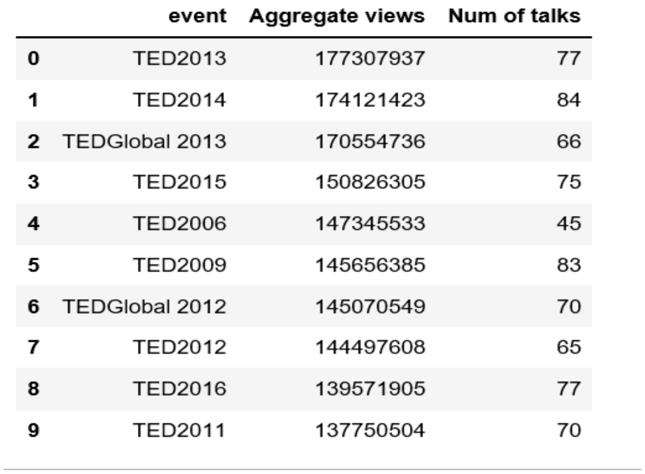




	views	languages
views	1.000000	0.377623
languages	0.377623	1.000000

Pearson coefficient is obtained as 0.377623 suggesting slight positive correlation which conforms to our expectation.

8. ANALYSIS ON EVENTS

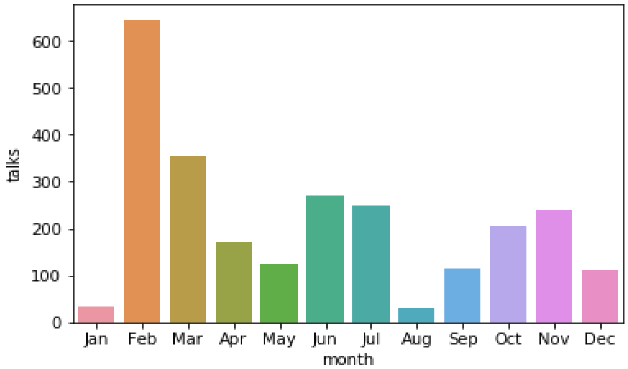


	event	Aggregate views	Num of talks
0	TED2013	177307937	77
1	TED2014	174121423	84
2	TEDGlobal 2013	170554736	66
3	TED2015	150826305	75
4	TED2006	147345533	45
5	TED2009	145656385	83
6	TEDGlobal 2012	145070549	70
7	TED2012	144497608	65
8	TED2016	139571905	77
9	TED2011	137750504	70

- *'TED 2014' has recorded the highest number of ted talks.*
- *The event TED2013 has more talks that are popular compared to other events since it has highest aggregate views inspite of not having highest number of talks*

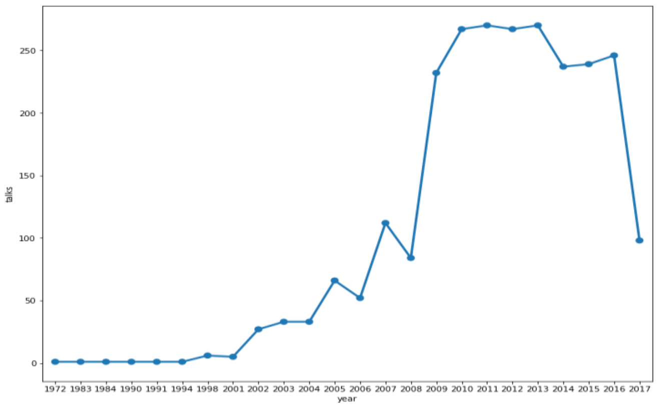
9. ANALYSIS ON DATE OF FILMING

Month is extracted from the 'filmed_date' attribute and the number of ted talks held in each month has been calculated. Results are shown in the bar graph below.



February is the month in which maximum number of ted talks have been recorded. On the other hand, in January and August least number of talks were recorded.

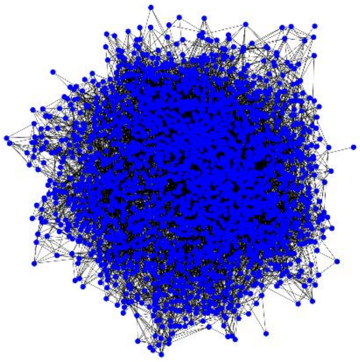
Next I have analysed how the number of ted talks filmed each year varied as the years passed by. For this the year information has been extracted from the ‘’filmed_date” attribute and the number of talks has been counted for each year. Point plot for the same is given below.

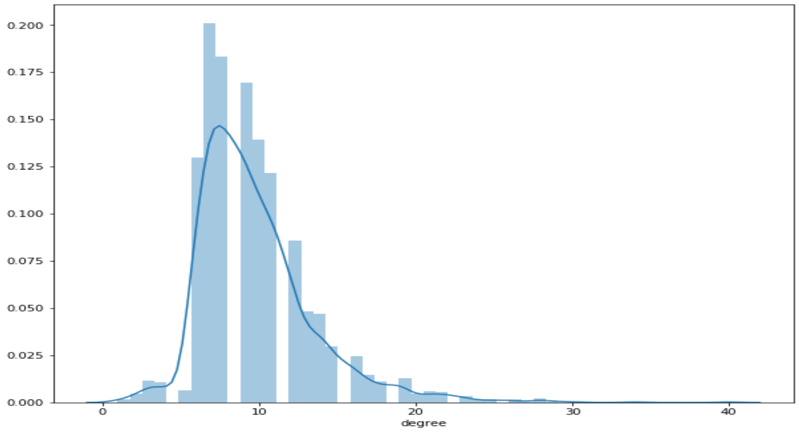


- *The number of talks have gradually increased till 2009 which tells us that popularity of ted has increased.*
- *In the year 2009 a steep increase is evident.*
- *From then more or less it has been constant and a drastic decrease has been observed in 2017.*

10.ANALYSIS ON RELATED TALKS

The related_talks attribute contains a list of related talks for each talk. Information is collected from this attribute to construct an undirected graph. The nodes in the graph are the talks in the given data set. The Graph contains edge A---B if A appears in ‘related_talks’ of B or B appears in related talks of A. This graph is used to analyze some important information. Degree of each vertex in the graph is calculated and is stored in a data frame. The plots showing the constructed graph and the distribution of degrees are given below.

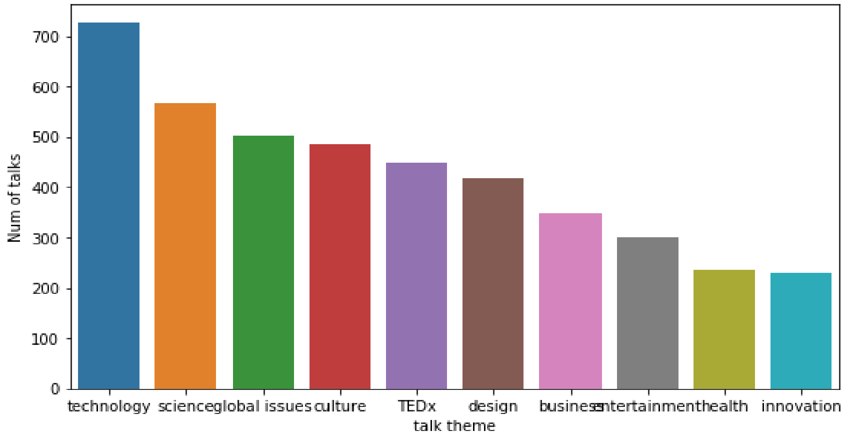




Each node’s degree in the graph has been calculated and the distribution has been plotted. *It has been found that on average each node in the graph has a degree 10. So every talk is related to 10 other talks. From this we can conclude that talks in ted are fairly related to each other.*

11. *ANALYSIS ON TAGS*

The “tags” attribute in the data set contains information about the themes associated with the particular talk. Number of talks held in each theme are calculated and results are plotted below**.**



➢ *Technology is the most popular theme*

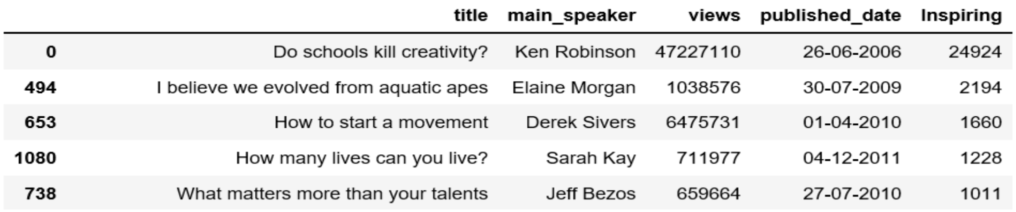
➢ *Ted website has the highest number of talks in Technology domain.*

➢ *Science and global issues follow technology in terms of popularity*

12.ANALYSIS ON RATINGS

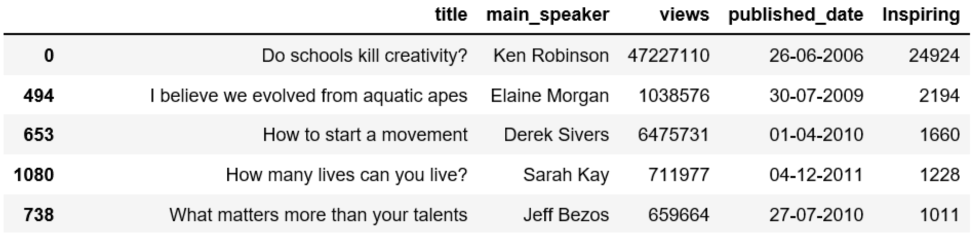
The attribute ratings contains the count of users who felt that the talk belongs to a particular category like funny, jaw-dropping, informative etc. For each of the talks, the count for ratings “funny”, “confusing”, “informative” and “inspiring” have been calculated and following conclusions are made.

- The talk which confused highest number of ted users (6600) is “How great leaders inspire action” by Simon Sinek.
- “The power of vulnerability” by Brene Brown was commented as “funny” by highest number of ted users(21444)
- As many as 21210 ted users have felt that the talk “ Your body language may shape who you are ” by Amy Cuddy is informative. Coincidentally, this talk is also the second most viewed talk on ted website.
- 24924 ted users have felt that the talk “Do schools kill creativity” by Ken Robinson is inspiring and this talk is the most viewed talk on ted website.



	title	main_speaker	views	published_date	Inspiring
0	Do schools kill creativity?	Ken Robinson	47227110	26-06-2006	24924
494	I believe we evolved from aquatic apes	Elaine Morgan	1038576	30-07-2009	2194
653	How to start a movement	Derek Sivers	6475731	01-04-2010	1660
1080	How many lives can you live?	Sarah Kay	711977	04-12-2011	1228
738	What matters more than your talents	Jeff Bezos	659664	27-07-2010	1011

Shown on the left are top 5 most inspiring talks



	title	main_speaker	views	published_date	Inspiring
0	Do schools kill creativity?	Ken Robinson	47227110	26-06-2006	24924
494	I believe we evolved from aquatic apes	Elaine Morgan	1038576	30-07-2009	2194
653	How to start a movement	Derek Sivers	6475731	01-04-2010	1660
1080	How many lives can you live?	Sarah Kay	711977	04-12-2011	1228
738	What matters more than your talents	Jeff Bezos	659664	27-07-2010	1011

Shown on the left are top 5 most informative talks