Instructions: <https://moodle2.cs.huji.ac.il/nu19/pluginfile.php/577268/mod_resource/content/1/project-clinic.pdf>. **Total should be 6 pages,**

Problem description & Research Questions

\*\*\*fill in from milestone\*\*

We assume that more energetic talks, funny, intellectual & inspirational talks would be popular, and boring slow ones would be unpopular

Explain why we chose transcript only talks, how we selected the data etc.

Data Brief

\*\*\*fill in from milestone\*\*

Our solution: Analysis process

**Data Collecting:**

For the project’s milestone, we built a crawler for Ted.com, saving the information to CSV format, sizing around 10 MB. We crawled in two batches:

Popular: ~500 most viewed talks, Unpopular: ~500 least viewed talks.

As we continued working, we encountered several issues:

* We didn’t have any information about talks that are neutral
* The Full Transcript for some ted talks had exceeded Excel & Spreadsheet cell’s characters limit (Elon Musk, we are talking about you), and broke our tables.

Thus, we decided to make some changes to our crawler:

First, we added a middle batch of neither most popular, nor least popular ted talks, in order to have a tiebreaker if needed and to improved our dataset accuracy.

Then, we doubled the size of the unpopular batch since we anticipated later-to-be-cleaned data. In total, we collected:

Popular: ~450 ted talks, Middle: ~450 ted talks ,Unpopular: ~800 ted talks

Last, we saved the collected data to Pickle format, to preserve our objects’ construction.

**Data Cleaning:**

In order to enable transcript analysis, we filtered out TedTalks from our collected dataset, if lacking any of the following fields: talk length, number of views, upload date and full transcript.

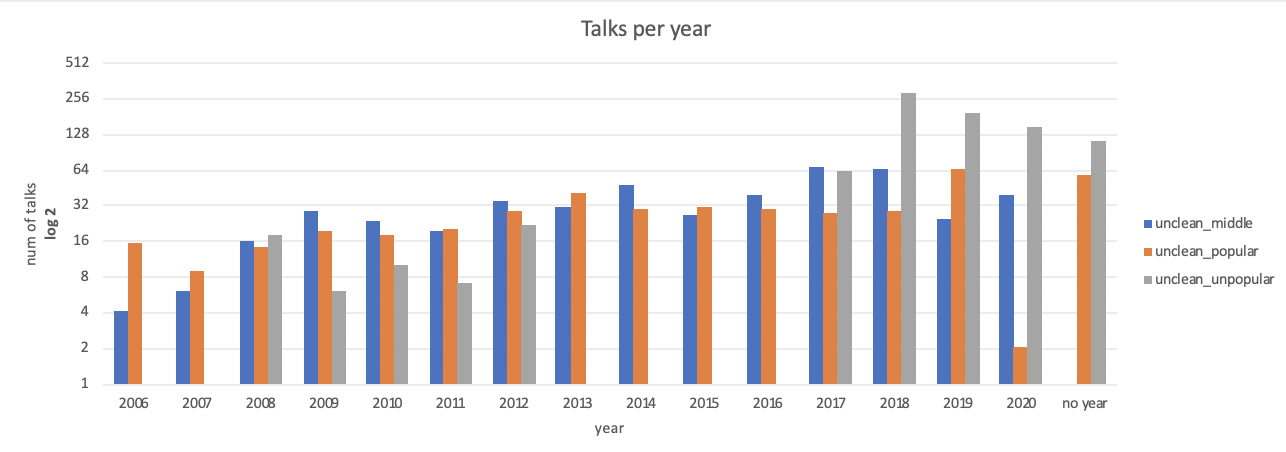
As we suspected, a significant amount of the data was filtered, and we remained with:

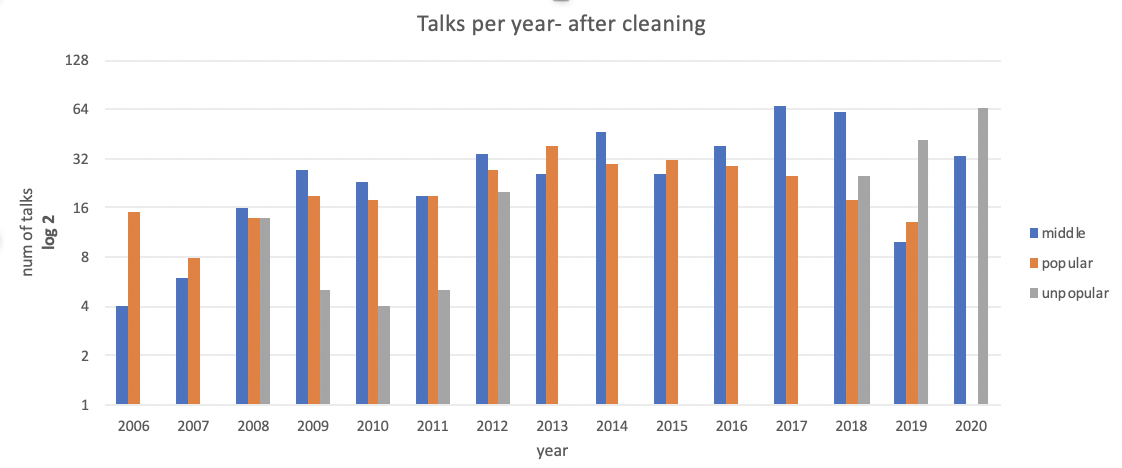
Popular: 305 ted talks, Middle: 437 ted talks, Unpopular: 183 ted talks

**Reality Check:**

In order to verify that the cleaned data set is accurate, and that no unwanted data was filtered out, we performed a reality check on both uncleaned & cleaned data.

First, we checked the number of ted talks per year:

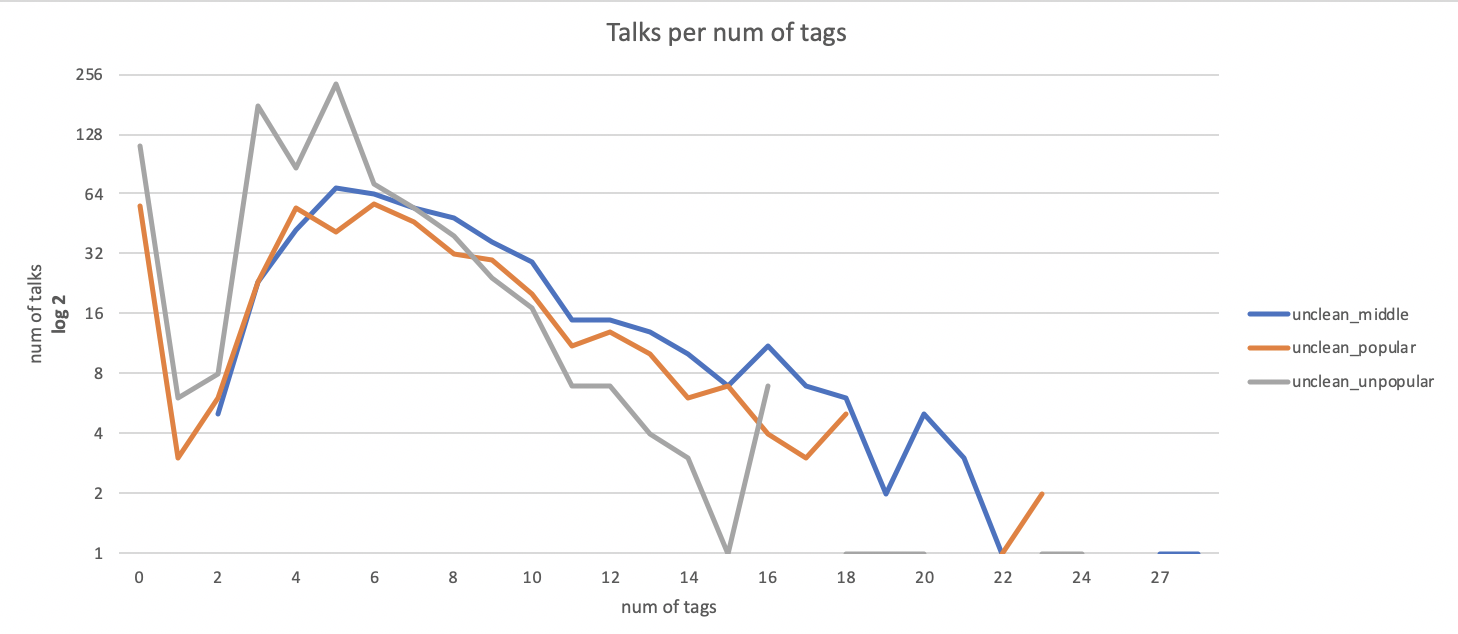


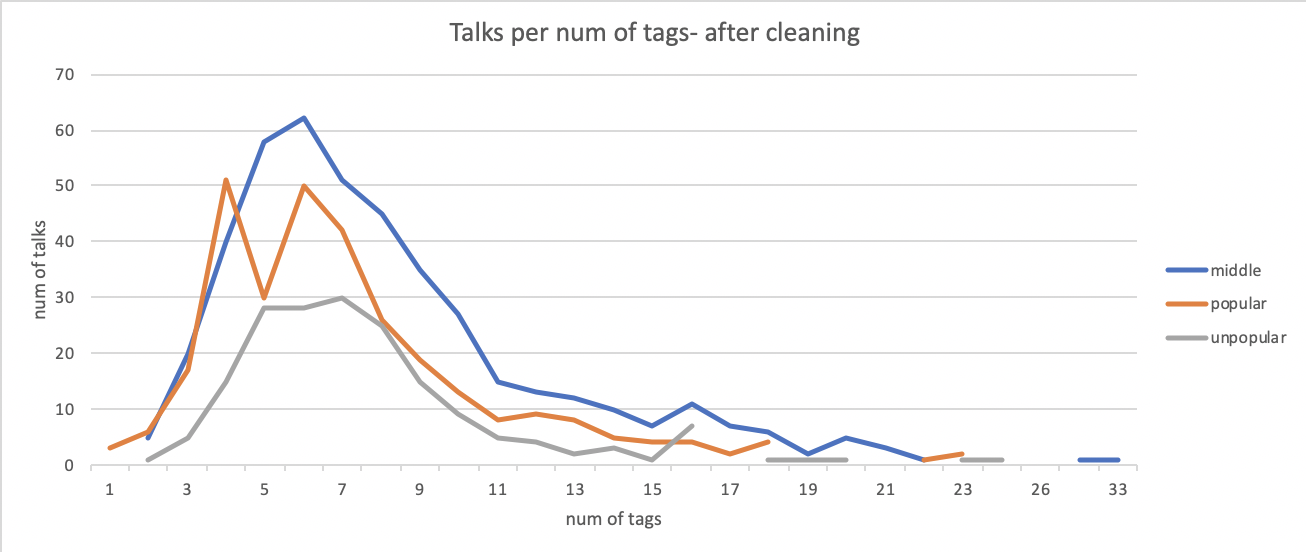


We can notice that there are popular and unpopular talks with no year, aka with no date of upload. We filtered those out upon cleaning, which can be seen in the second figure above.

Another issue that we see above is a lack of unpopular talks between the years 2013-2017. After investigation, we found out that some changes where made in Ted website HTML code on 2013 and 2017¹. By basing our crawler on the newest HTML version, we could not fully process the old HTML pages. About 100 talks from each of the unpopular and popular batches are lacking important information including upload date (year), tags, translation etc. It makes sense that Ted priorities adjusting the HTML format for popular and middle ted talks over the unpopular talks, which could be a reasonable explanation for the gap we found.

Second, we checked the number of ted talks per number of tags:

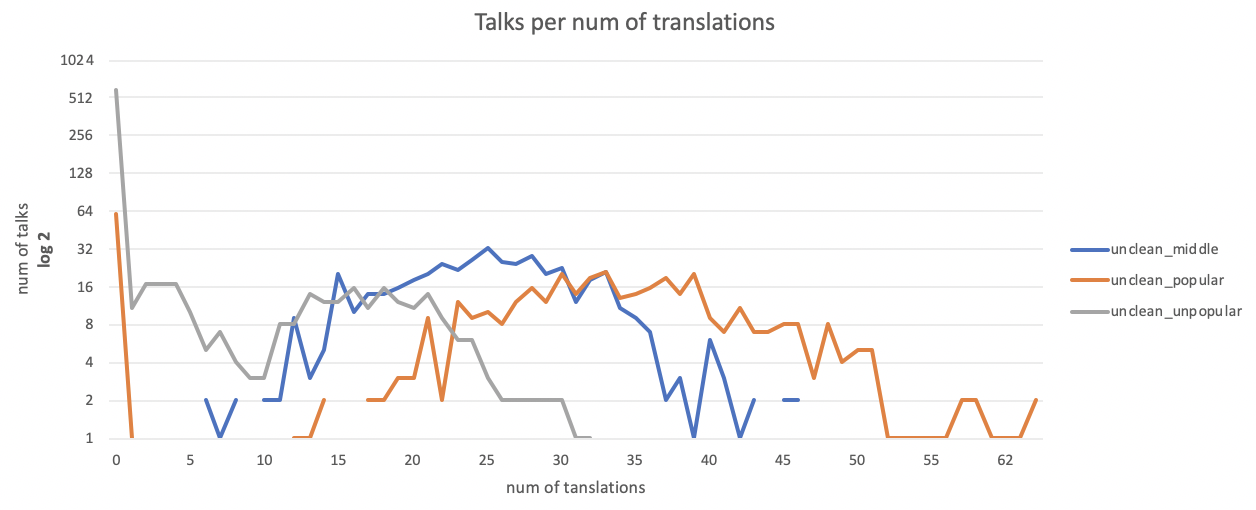


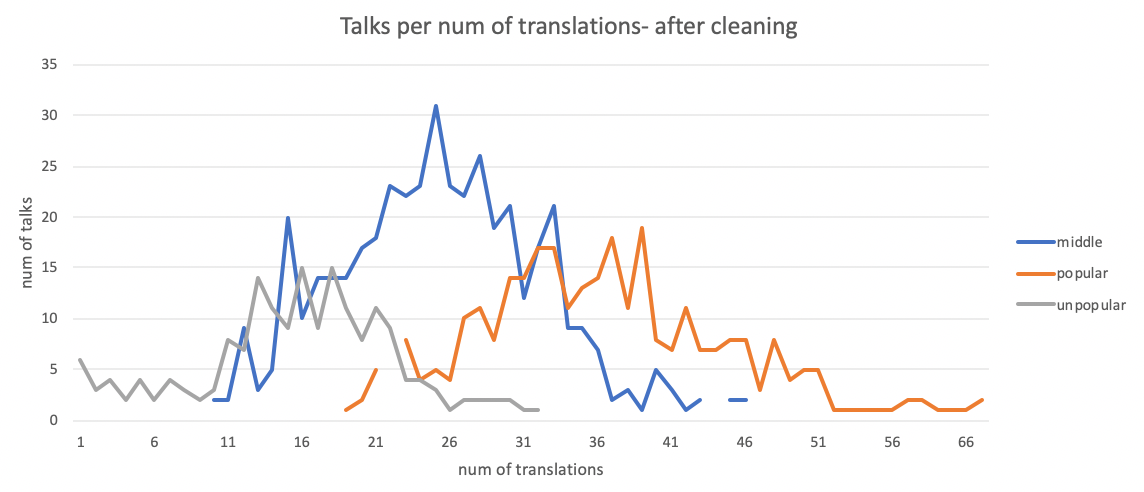


Here we notice many ted talks without tags, from both popular and unpopular ted talks. After the data cleaning, those talks were filtered out. Another difference is in the amount of unpopular talks that have a small amount of tags, which dropped significantly.

Both of these issues could be explained due to a change that was made in Ted website, between 2013 and 2017, as noted above.

Third, we checked the number of talks per number of translations:



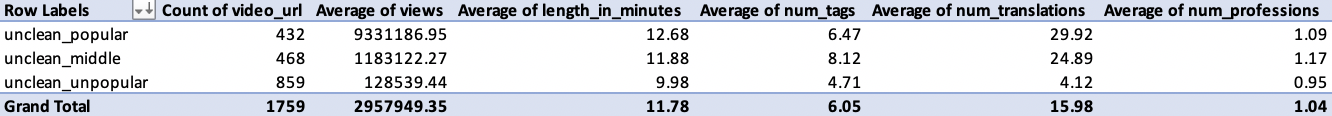


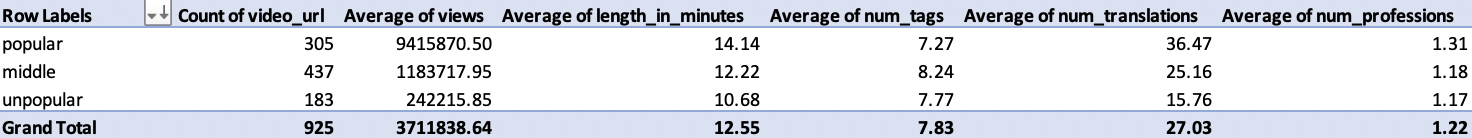
We notice again that there are unpopular and popular ted talks with no translation, which we predicted in the analysis above of the HTML changes made to the website.

We could also see three beautiful normal distributions in the figure above-

Popular: around 36 translations, Middle: around 25 translations, Unpopular: around 15 translations. Ted offers its community the option to help translating the talks, thus it makes sense that there’s a correlation between the views of a talk (popularity) to the number of translations the community offers.

Averages:





1. <https://thenextweb.com/insider/2013/10/30/ted-website-rebuilt-ground-new-generation-people-devices/>

+ 2017 ???

Evaluation

**Evaluation criteria:** (how we measure success)

\*\*\*fill in\*\*

**Setup:**

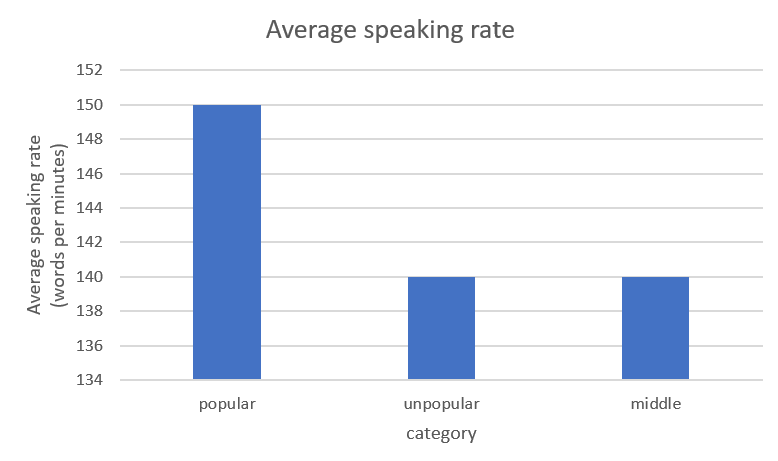
\*\*\*fill in\*\*

**Results & Visualization:**

First, to understand the lectures’ structure, we analyzed the following speech criteria:

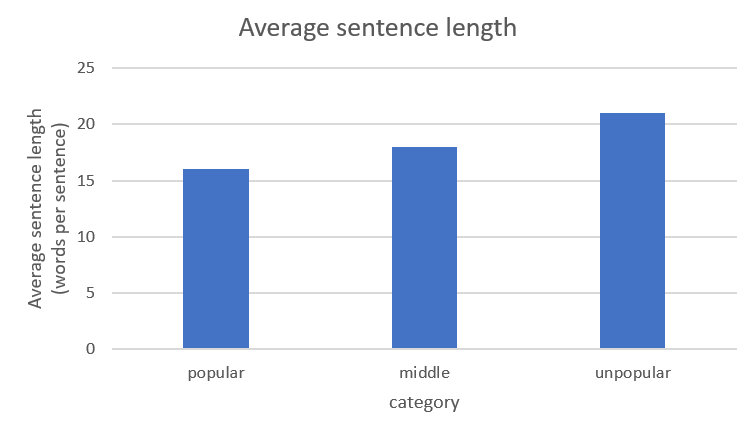
* Speaking rate-

the average speaking rate is \*\*\*\*. From the chart below we can see that popular talks’ speakers talk faster then unpopular & middle talks’ speakers. Higher speaking rate correlates with energized and knowledgeable talk \*\*\*FILL IN MISSING FROM RESEARCH. There’s an article with data that dows’nt match ours, they have a section about ted talks - <https://virtualspeech.com/blog/average-speaking-rate-words-per-minute#:~:text=The%20average%20speaking%20rate%20changes,podcasters%2C%20the%20wpm%20is%20higher.>\*\*\*



* Sentence length-

From the chart below, we can see that popular talks got shorter sentences compares to other talks (popular- 16, middle-18 and unpopular- 21).



* Top words (grouped by semantic meaning)-

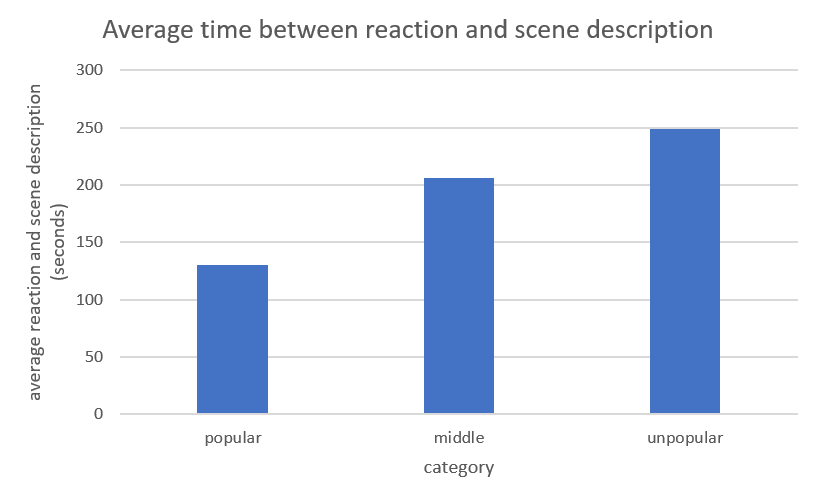
In the figure below we could see the top semantic groups in ted talks in general.

\*\*\*fill in, this is a cloud for all lectures (pop, mid, unpop). Not sure we can take a lot from it. I think we need to redo this with more words (top 50) without the semantic grouping we did. We could also do one for each category (bot, mid, pop)\*\*\*



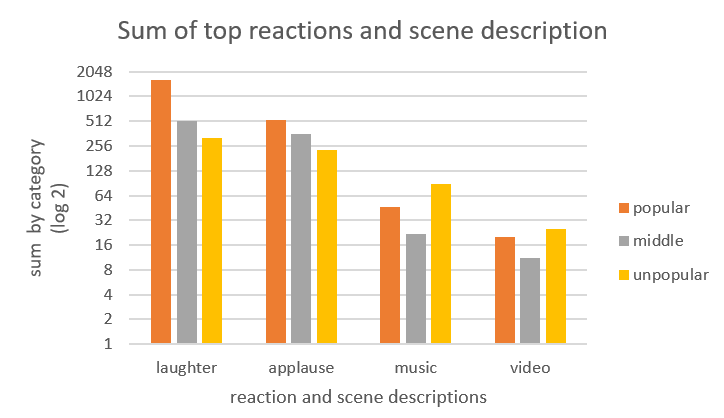
Second, we focused on the audience engagement of each category.

As can be seen in the bar chart below, there’s a correlation between a TedTalk’s popularity to the frequency of audience reactions and transcript’s scene description: popular talks- every ~2 minutes on average, middle talks- every ~3.5 minutes and unpopular talks- every ~4 minutes.



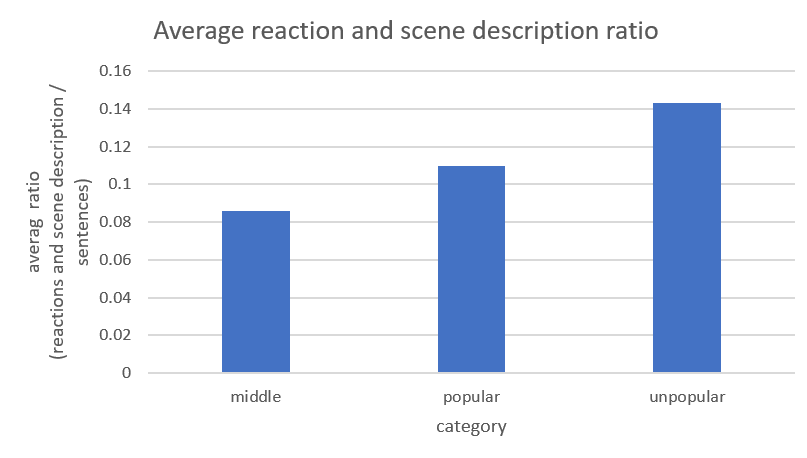
In the chart below we can see the top reactions and scene descriptions per category (in logarithmic scale).

* What we found interesting, is that popular talks have twice (!) the amount of laughter reactions then other talks. this correlates with our assumption (\*\*write more about it above\*\*).
* We can also see that there’s a correlation between popularity and applause.
* Another finding is that audio-visual aids (videos & music) don’t have a meaningful impact on the popularity of the talk. Unpopular talks use many aids as can be seen in the graph below, while popular talks use about half the amount. Checking with our control group (middle talks), we indeed find they got the least amount of aids, which reinforce the lack of correlation (???)

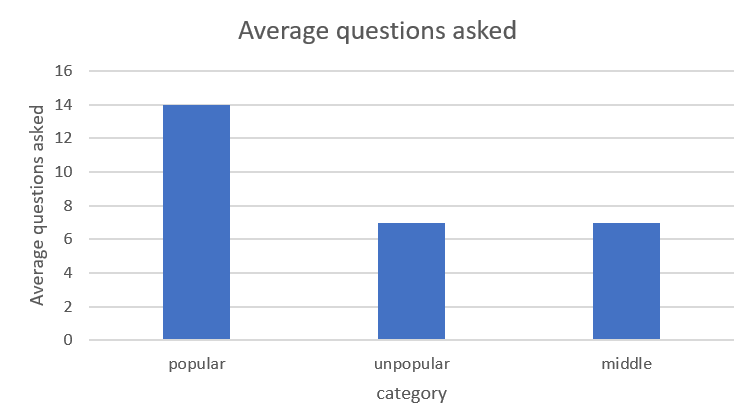


We dove a little deeper and were surprised by the ratio between reactions and scene descriptions / total sentences, as can be see in the chart below:

in general, less then 15% of sentences get a reaction from the audience. Middle talks have the lowest ratio of 0.08, then popular talks with 0.11, and unpopular talks with 0.143. This could be explained by the number of sentences on average of each category- \*\*\*\*\*FILL IN MISSING\*\*\*\*



Third, we zoomed in to our main project’s problem- including questions in the tedtalks as an engaging feature. In the figure below we could clearly see that asking question is a great way to interact with the audience. Popular talks have twice (!) as much asked questions per talk compared to other talks. in total, in the popular category there were 2584 question asked, middle had 1290 and unpopular got 1250.



Not All questions born the same. The English language differentiate between yes-no questions and WH-questions. The latter are more informative, and require a more detailed answer. By using WH-questions in a lecture, the audience’s curiosity is intrigued, and thus leads to attentiveness. In our research, we wanted to capture the use of those more interesting questions by category.

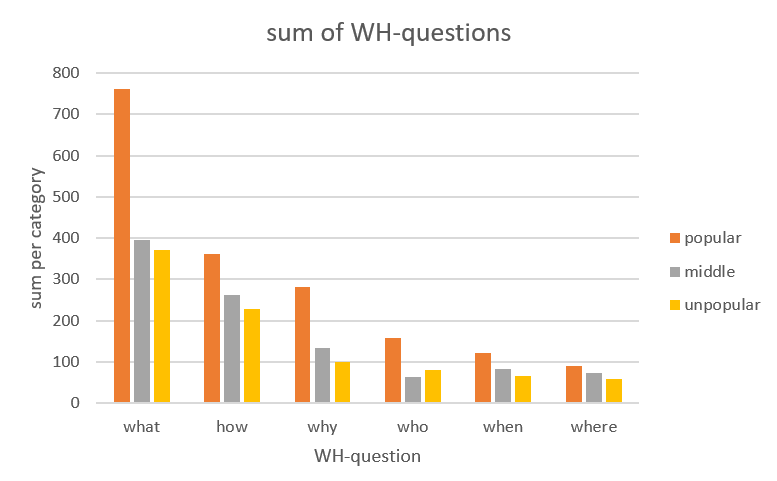
* In total, popular talks have 1773 WH-questions (0.74% of all questions), middle talks got 1013 (0.72%) and unpopular talks have only 906 (0.67%).
* On avarege, a single popular tedtalk got twice as many WH-questions then other talks, as can be seen in the table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **category** | **what** | **how** | **why** | **who** | **when** | **where** |
| popular | 4 | 2 | 2 | 1 | 1 | 0 |
| middle | 2 | 1 | 1 | 0 | 0 | 0 |
| unpopular | 2 | 1 | 1 | 0 | 0 | 0 |

As we can see in the figure below:

* The most popular question is ‘What’
* Top WH-questions (what, how, why, who) appear in popular talks twice (!) as much compared to other talks.

\*\*\*\*fill in why this is interesting\*\*



\*\*\*we should do the same analysis on 2 tags, or at least write it as a future work\*\*\*

Impediments (issues and how were handled)

\*\*\*\*\*see under Data Collecting, we could add more but not sure if necessary here\*\*

Future work

\*\*fill in\*\*

Conclusion

\*\*fill in\*\*