**Ted-Talks: How to talk your way into their most viewed list?**

Razi Haj, [Razi.Haj@mail.huji.ac.il](mailto:Razi.Haj@mail.huji.ac.il), 205739386

Noa Levitzky, [noa.levitzky@mail.huji.ac.il](mailto:noa.levitzky@mail.huji.ac.il), 205970783

1. **Project Description & Research Questions**

Talk slow. Keep your crowd engaged. Cut to the chase. We've all heard many techniques on how to craft a powerful speech. In this project, we will attempt to validify some of them, by analyzing how they were employed in different Ted talks and using the top viewed ones as a benchmark. We'll examine the engagement rate of the audience, the speaking rate of the presenters, and their use of specific rhetoric techniques to see if they're the most common in the poplar talks.

1. **Data Description**
   1. **Talk Tiers**

Using a web crawler we built for Ted.com, we gathered data on 1,600 talks from a list with all the talks ever uploaded, sorted by view numbers. These talks were selected according to on three tiers, based on the number of views: most viewed (will also be referred to as 'top' or 'popular'), least viewed ('bot' or 'unpopular') and the control group ('mid'). The latter was extracted from the middle of the full talk list. Since the tiers had to be of equal size, we had to select it according to the smallest viable group. After multiple crawls and data cleans, we settled on 183 talks per tier.

* 1. **Data Cleaning**

This process was challenging, to say the least. When we worked on our milestone this stage was not required, and we didn't assume so many talks would not be suitable. Since we rely on the transcript provided to analyze many techniques, every talk without one was removed, as well as talks without any timestamps. On top of that, some talks were missing the view count, or other relevant information, and got removed as well. Eventually, we remained with 305 top talks, 437 mid talks, and 183 bot talks. Unfortunately, the sample size was smaller than we would have wanted (549 in total).

* 1. **General Fields Gathered**

Speaker's name, description, and profession; Talk's views number, full transcript, and its available translations; video title, length, and upload date; Related tags.

1. **Reality Checks**

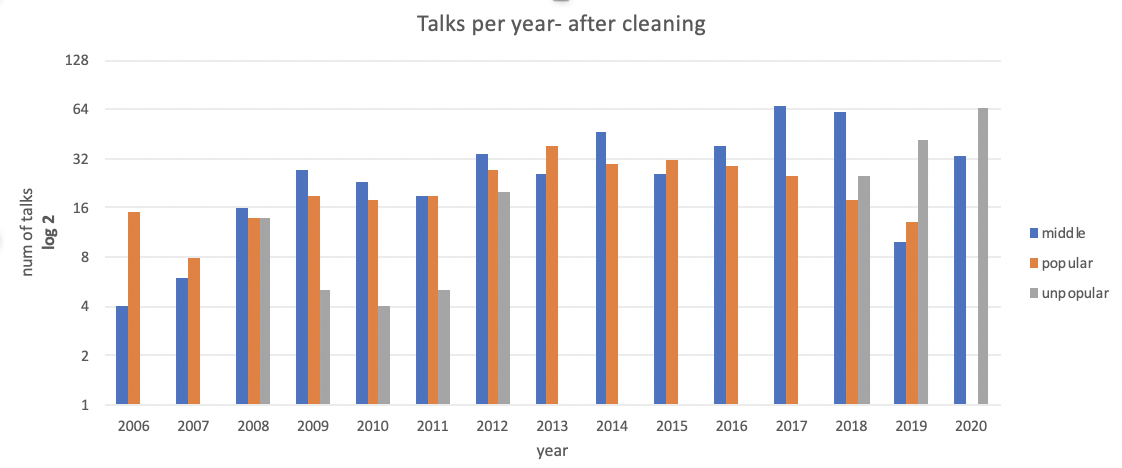


Fig 1: Tier distribution of the talks per year (after cleaning)

1. **Talk Distribution Per Year**

As we would have expected, figure #1 shows a similar distribution of popular talks throughout the years, except 2020 – where talks didn't have much time to accumulate millions of views to become popular. But a recurring problem since the milestone, in which the years 2013-2017 didn't have any unpopular talks, wasn’t solved even after updating the crawler multiple times. Researching this matter, we've discovered that during 2013 Ted.com had started to experiment with crowd-based inputs, allowing users to add and translate transcripts.[[1]](#footnote-1) This probably resulted in a change to the HTML that our crawler couldn’t handle.

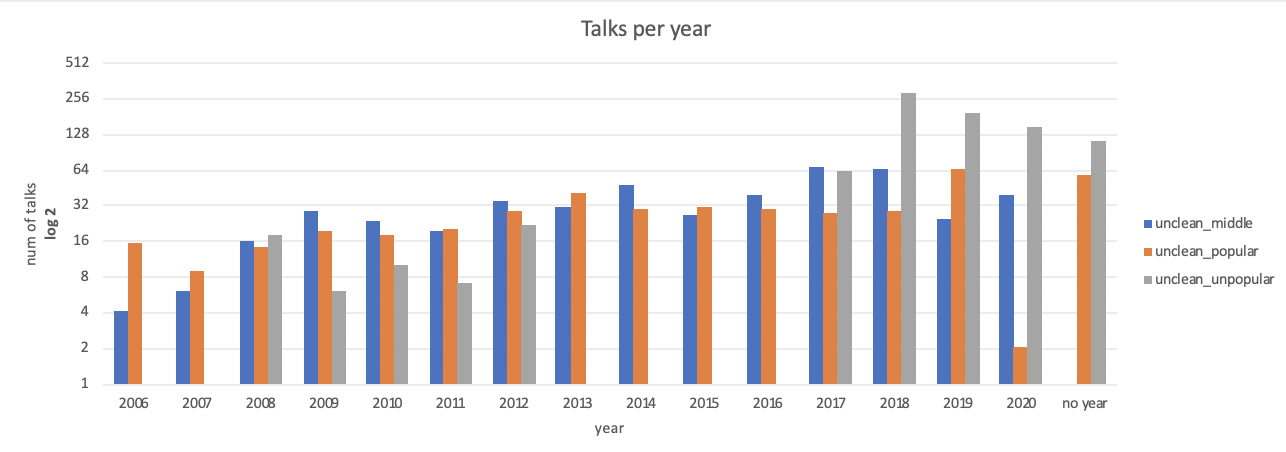


Fig 2: Tier distribution of the talks per year (before cleaning)

Figure #2 visualize another major problem, in which many of the unpopular talks had missing transcripts, to begin with. All of the 2017 unpopular talks were deleted because of that reason. Although these explanations are somewhat lacking, we couldn't invest more time into this crawler.

1. **Number Of translations**

Another check we did was on the number of translations. We would have expected to find a higher amount of translations for the popular talks, whether because they increase the availability and accessibility of the talks, or, most likely, are a by-product of such popularity – since viewers can help translate the talks, it makes sense that they will put more effort into successful ones, rather than the rest.

Fig 3: Number of translations for each tier (after cleaning)

And so, as can be seen from the figure, the distribution is indeed matching our hypothesis: the more popular the talk is, the more translations it has.

1. **Evaluation Criteria**

We would have wanted to find some sort of consistent correlation between the techniques we tested and the tiers. If a certain technique would be effective in creating a successful talk, we would expect to see it being implemented the most in the top-tier, and the least in the bot-tier. The implementation of the mid-tier should, in the same fashion, be between the other two. If a certain technique would be ineffective, we would have expected to see it similar in all levels. similar results in the top and bot tiers, and a different one with the control group would not demonstrate a successful correlation.

1. **Analysis**
   1. **Average Speaking Rate (Words Per Minute – wpm)**

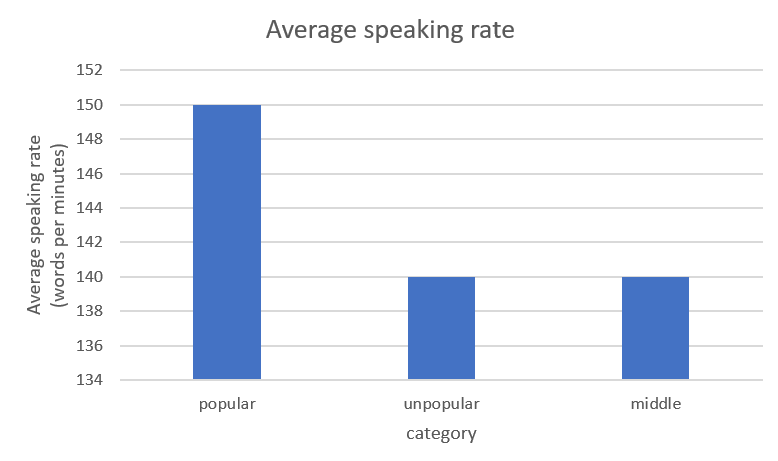
This is calculated by dividing total words by the length of the talk. We used an NLTK module to parse the transcript into words.

Fig 4: Average speaking rate per tier

The presenters of the popular talks had, on average, a faster-talking pace. Although there was no definitive answer, one article placed 150 wpm at the higher end of the "presentational"/ "conversational" talk types. It is said that fast-paced talks correlate to high energy, passion, urgency, and excitement, while slower-paced indicate a matter of importance, sadness, and the seriousness of a point.[[2]](#footnote-2)

* 1. **Average sentence length**

This was calculated in a similar way to the speaking rate.

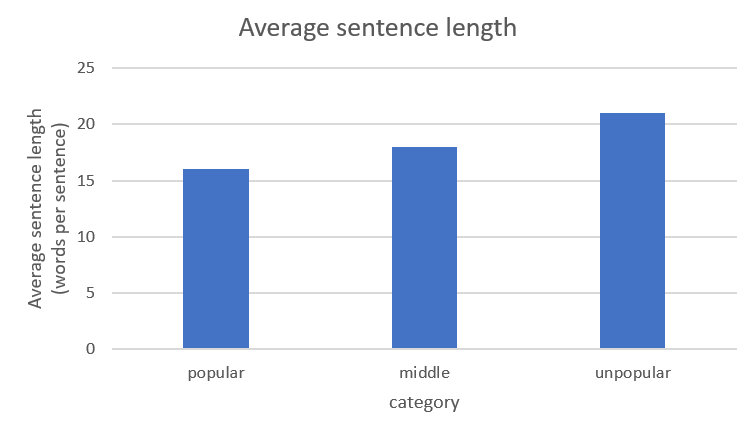


Fig 5: Average sentence length per tier

The sentences of popular talks had 16 words on average, and the control group had 18. The unpopular talks had a staggering 21 words on average. This is no small matter since each talk has tens of sentences, so overall even a small gap can be translated into a big change.

* 1. **Crowd Reaction**

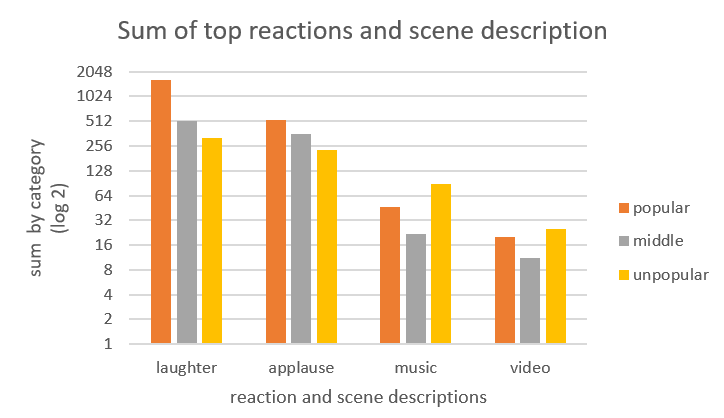


Fig 6: Sum of the top different scene description for all the talks of a tier

Each transcript also included the reaction of the crowd, as a description of the scene. Using NLTK, we could count how many instances of laughter and applause each talk had. Looking at the graph, one thing is bluntly clear: popular talks have twice as much laughter and applauses than the control group and 5 times as much laughter than the unpopular talks. We can also see that the use of video clips in talks is similar between the top and bot tiers, and twice as common in the unpopular talks.

* 1. **The spread of the crowd's reaction**

Part of getting the crowd engaged in a talk, is to break the monotonous of it. Thus, we wanted to calculate the average "spread" of these reactions. Simply put, we wanted to know how many seconds it takes until something interesting happens in a talk, so interesting that people start laughing, or applauding.

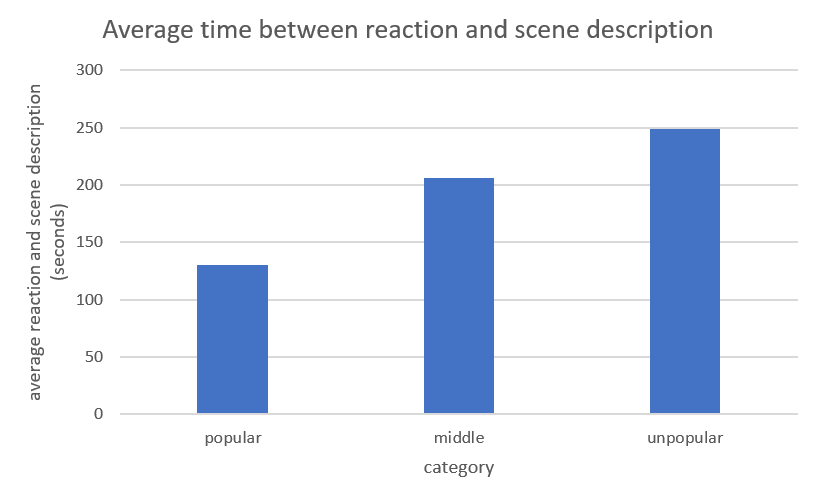


Fig 7: Average time in seconds until the crowd reacts to the talk

As expected, it takes around 2 minutes for something "exciting" to happen in popular talks, and as we go down the tiers, that time is increasing consistently.

* 1. **Asking questions**

Asking the crowd questions, many claims, draws the listeners in and keeping them concentrated on the subject at hand. To test that matter we first counted the total questions each tier had. The top-tier took first place with a total of 2584 questions asked – twice as much as the other two. Coming in close were the mid (1290) and bot (1250). Calculating the average number of questions asked can further support this point.

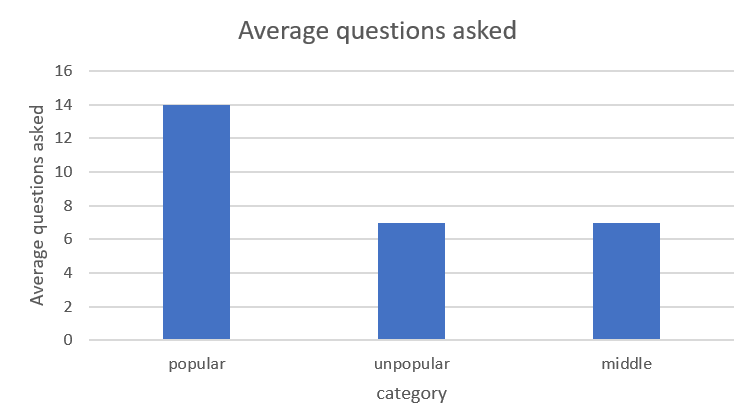


Fig 9: Average amount of questions asked in a talk, per tier

* 1. **Quality of Questions Asked**

Some copywriters claim that not all questions are created equal. If one's looking to encourage thinking, questions you could answer with a simple "yes" or "no", are probably not going to achieve that goal.[[3]](#footnote-3) Some claim that generally, "WH" questions are of a higher quality. To test this theory, we counted the number of questions containing "WH" questions and divided them by the total number of questions. We called this "Quality Questions Ratio" (QQR).

Fig 9: Quality Questions Ratio, per tier

Yet again, top talks were in the lead, with 74% quality questions (0.74), followed by the mid-tier with 72% (0.72), and lastly the bottom, with 0.67 QQR. Although the QQR seems similar to all tiers, taking into account the major difference in the total questions that were asked demonstrate a different story: top presenters ask significantly more questions, and most of them are of "higher" quality. Alternatively, if this difference is minor indeed, then we could argue that asking more questions, of any type, is helpful in keeping the crowd engaged.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question Types** | **What** | **How** | **Why** | **Who** | **When** | **Where** |
| Top | 4 | 2 | 2 | 1 | 1 | 0 |
| Mid | 2 | 1 | 1 | 0 | 0 | 0 |
| Bot | 2 | 1 | 1 | 0 | 0 | 0 |

Fig 10: Average question types per talk

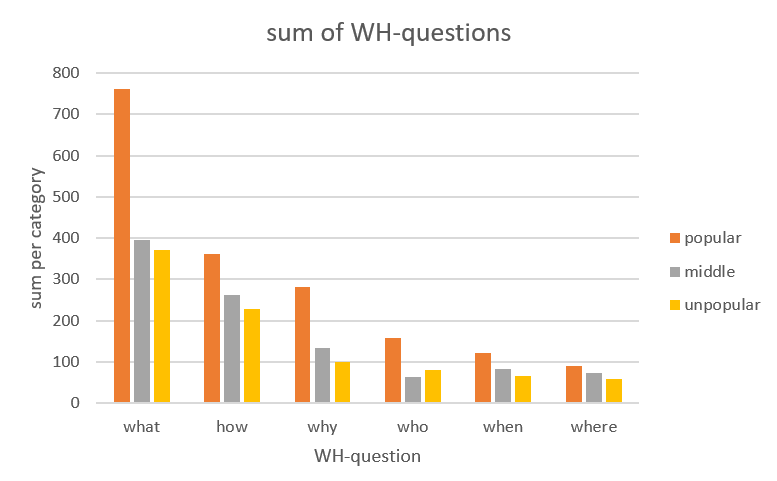


Fig 11: Sum of question types per tier

We wanted to examine, with the help of figures 10 and 11, if there's something to infer regarding the use of questions. We could see that on average, each successful talk is using a larger variety of questions, and the most common question type is "what". Alas, we couldn’t think of an interesting explanation for this observation.

* 1. **Day of Publication**

Does the day of publication have an influence on views? Many researches mention that Friday is the least productive date, since people are more exited for the upcoming weekend.[[4]](#footnote-4) This results in procrastination, and, probably, more time wasted on watching videos.

Fig 12: Views accumulated per weekday

Fig 13: Number of talks published per weekday

These two graphs demonstrate that point, assuming the views are mostly from countries where the weekday is on Saturday and Sunday, like the States. From Figure # 12, We can see that Friday accumulated the most views – 30% more than Wednesday, the second-best day. To further support this point, figure #13 shows that top-viewed Talks were published the most by Friday, and the least viewed were published on Monday, which is the most productive day, according to that same article. Alas, we're not sure if this is an explanation or simply a correlation, since views are accumulated over many weeks. Posting on Friday might give a talk a temporary boost to the views, but viewers could stumble upon any other talk after it have been upload, regales to the initial upload day. A possible explanation for that might be, that high traffic videos get recommended more easily to other viewers. That way, the initial "boost" of a video published on Friday gets it recommended more, which in turn increases the popularity, and so forth.

1. **Impediments**

Firstly, building a crawler is susceptible to changes, and at times is not reliable. During our work, we had to tweak it many times, and it was still not perfect. These changes were mainly on the HTML identifiers, of how to locate the information, and rebuilding the crawler with Selenium library, since beautiful soup doesn't support dynamic pages.

Secondly, data cleaning should be an important factor, and adding the task of NLP on top of that meant that our initial sample size should have been a lot bigger than it was. Given enough time, we should have crawled the entire website.

Thirdly, NLP takes a lot of processing power. We had to learn how to "pickle" files, which means saving processed objects into the hard-drive. This way loading the files into python resulted in the post-processed data, and saved us many hours of waiting.

Lastly, many of the code for future work was already implemented in the code, but since we had time and page restrictions, we decided to not use all of them in this paper.

1. **Future work**

There were so many paths we could have taken with this project. We would have wanted to run all these tests only on the first 3 minutes of each talk, to see if the opening of the talk is indeed as important as people claim it is. Continuing that methodology, we can also analyze and compare specific tags, and see how these tactics influence different fields. We could add more criteria, such as sentiment analysis (to see if optimistic/ pessimistic matters have an influence on popularity). We could also attempt to identify "sleeping beauties", which describes a sudden and surprising spike of views of a talk, several years after it was initially published.

1. **Conclusion**

In this paper, we have found multiple correlations between certain speech tactics, and the success rate of a Ted talk. We could clearly see that in top viewed talks the crowed is more engaged, by laughing and clapping. These instances were shorter in-between, so this higher rate was not caused only by longer talks. We noticed that the crowd was being asked significantly more questions by the top tier than the other two and that these questions were usually composed as open, rather than requiring a binary answer. Furthermore, one should try to get his video published on Friday, since people are more likely to watch it – utilizing human escapism at its finest.

1. https://thenextweb.com/insider/2013/10/30/ted-website-rebuilt-ground-new-generation-people-devices/ [↑](#footnote-ref-1)
2. https://virtualspeech.com/blog/average-speaking-rate-words-per-minute#:~:text=The%20average%20speaking%20rate%20changes,podcasters%2C%20the%20wpm%20is%20higher [↑](#footnote-ref-2)
3. https://www.forbes.com/sites/nickmorgan/2015/11/24/whats-the-right-way-to-ask-audiences-questions/#c5f0cc4ed157 [↑](#footnote-ref-3)
4. https://www.getflow.com/blog/productive-day [↑](#footnote-ref-4)