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.

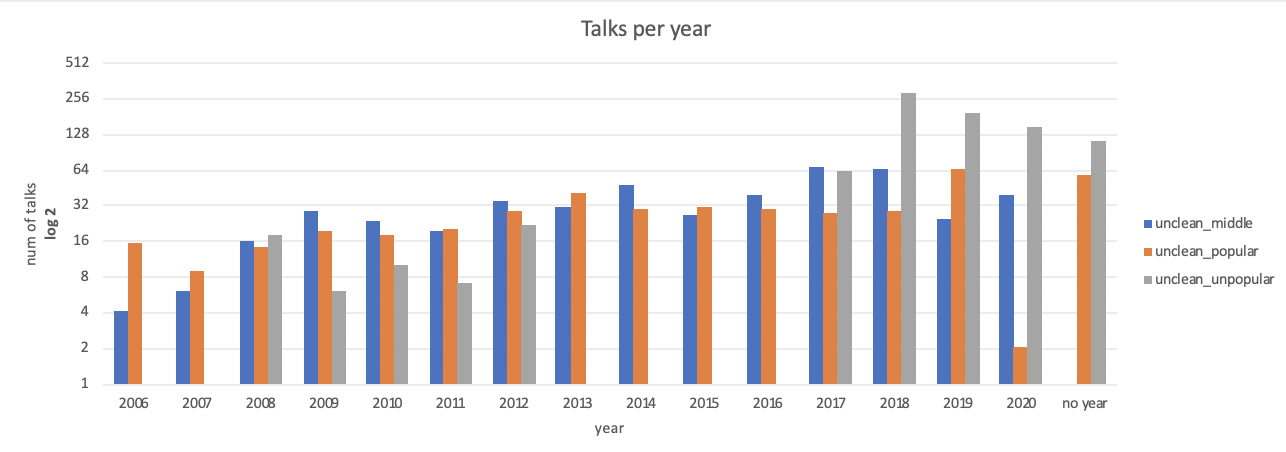
As we suspected, a significant amount of the data was filtered, and we remained with:

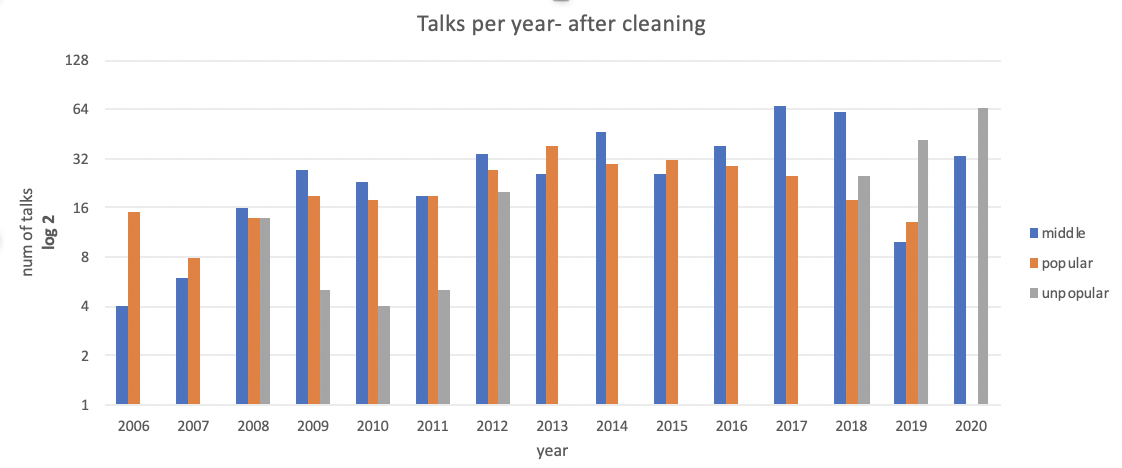
Popular: 305 ted talks, Middle: 437 ted talks, Unpopular: 183 ted talks

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**:

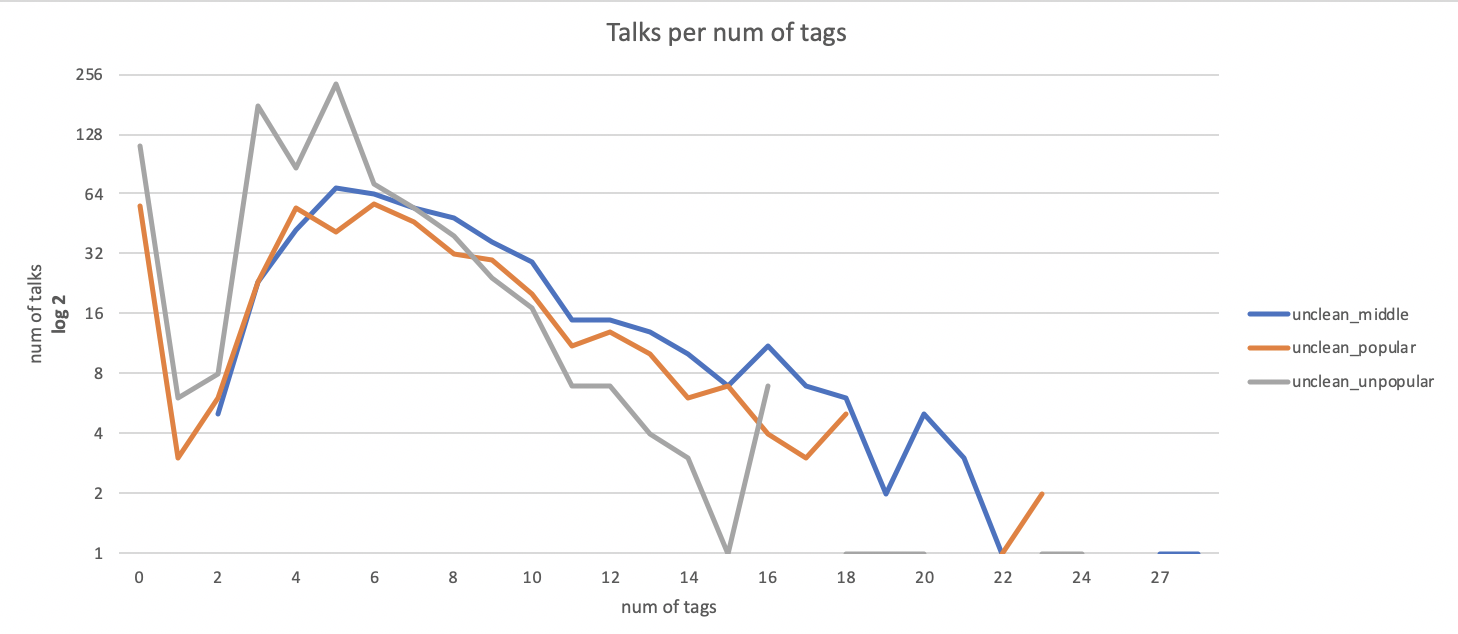


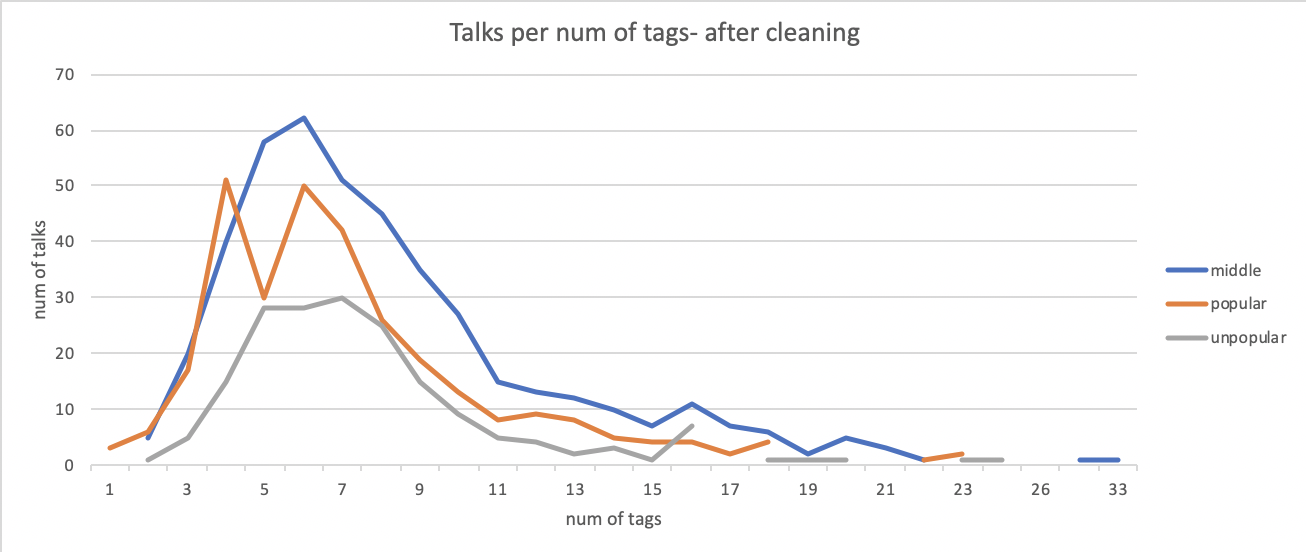


We can notice that there are popular and unpopular talks with no year, aka with no date of upload. We filtered those out upon cleaning, which can be seen in the second figure above.

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**:

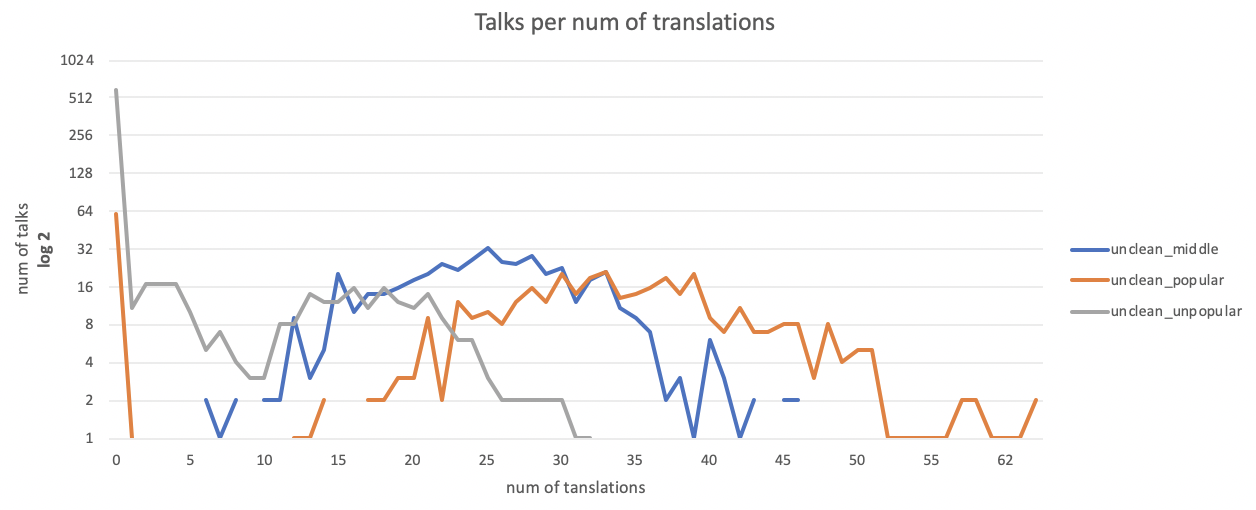


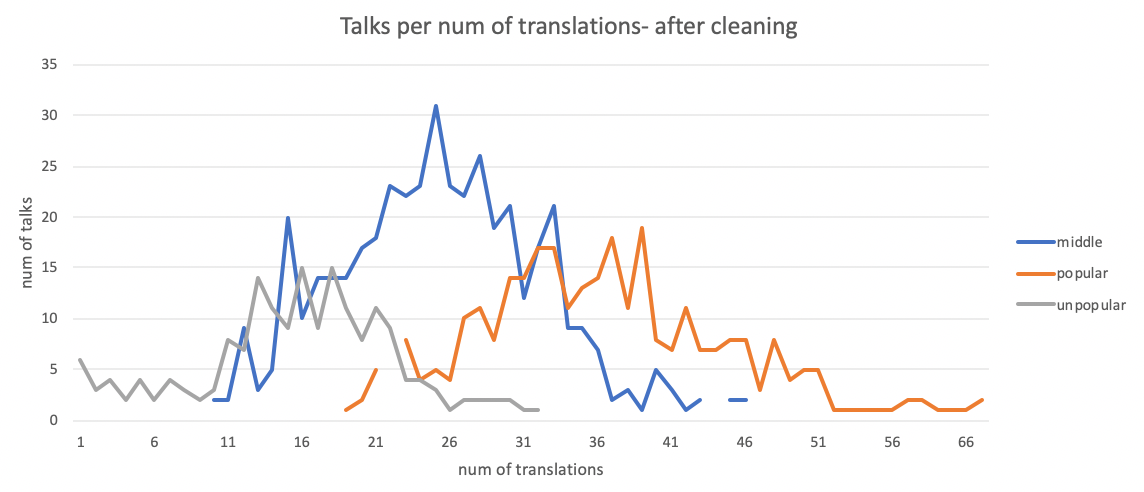


Here we notice many ted talks without tags, from both popular and unpopular ted talks. After the data cleaning, those talks were filtered out. Another difference is in the amount of unpopular talks that have a small amount of tags, which dropped significantly.

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**:





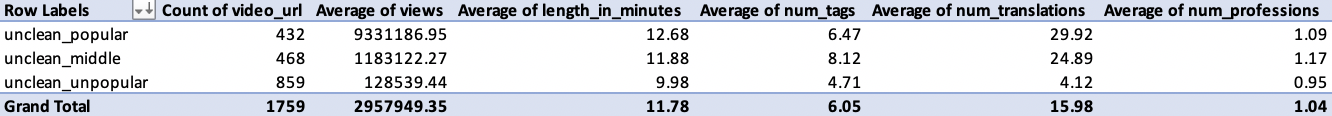
We notice again that there are unpopular and popular ted talks with no translation, which we predicted in the analysis above of the HTML changes made to the website.

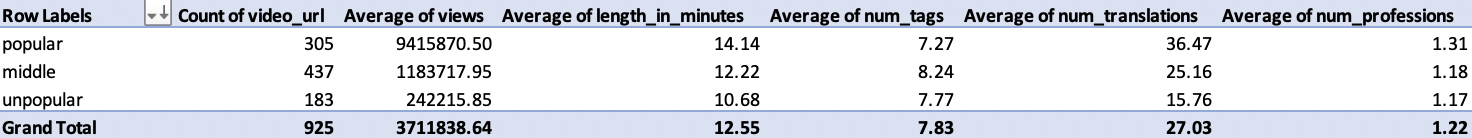
We could also see three beautiful normal distributions in the figure above-

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 ???