TED TALK ANALYSIS

- FOCUSING ON TRANSCRIPT AND POPULARITY

Soobin Choi
LING 2234
Dr. Narae Han

1. Motivation / Research Question

- *What make a Popular Ted Talk?*
- Two biggest part of the talks:
  verbal aspect / nonverbal aspect
- RQ: Can the popularity of a talk be predicted solely based off of the content of it?
  - Popularity here: each talk's share of positive / negative review in *rating* column

TED Ideas worth spreading

2. Data Processing

Before Merging/sorting:

```
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   comments            2544 non-null   int64
 1   description         2544 non-null   object
 2   duration            2544 non-null   int64
 3   event               2544 non-null   object
 4   film_date           2544 non-null   int64
 5   languages           2544 non-null   int64
 6   main_speaker        2544 non-null   object
 7   name                2544 non-null   object
 8   num_speaker         2544 non-null   int64
 9   published_date      2544 non-null   int64
 10  ratings             2544 non-null   object
 11  related_talks       2544 non-null   object
 12  speaker_occupation  2544 non-null   object
 13  tags                2544 non-null   object
 14  title               2544 non-null   object
 15  url                 2544 non-null   object
 16  views               2544 non-null   int64
dtypes: int64(7), object(10)
memory usage: 357.8+ KB
None
-----------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2467 entries, 0 to 2466
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   transcript  2467 non-null   object
 1   url         2467 non-null   object
dtypes: object(2)
memory usage: 38.7+ KB
None
```

After Merging (and a bit of sorting):

```
# combine two files
ted_clean = pd.merge(ted_main, tran, on = 'url')
ted_clean.info()
```

[5] ✓ 0.1s

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 972 entries, 0 to 971
Data columns (total 18 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   comments            972 non-null    int64
 1   description         972 non-null    object
 2   duration            972 non-null    int64
 3   event               972 non-null    object
 4   film_date           972 non-null    int64
 5   languages           972 non-null    int64
 6   main_speaker        972 non-null    object
 7   name                972 non-null    object
 8   num_speaker         972 non-null    int64
 9   published_date      972 non-null    int64
 10  ratings             972 non-null    object
 11  related_talks       972 non-null    object
 12  speaker_occupation  972 non-null    object
 13  tags                972 non-null    object
 14  title               972 non-null    object
 15  url                 972 non-null    object
 16  views               972 non-null    int64
 17  transcript          972 non-null    object
dtypes: int64(7), object(11)
memory usage: 144.3+ KB
```

2. Data Processing

- Why only *TED original?*
  - TEDx, TEDMED, TED India, … : programs supported by TED organization when a specific group wants to hold conferences similar to TED talk.
  - Topics depend on the nature of the group.

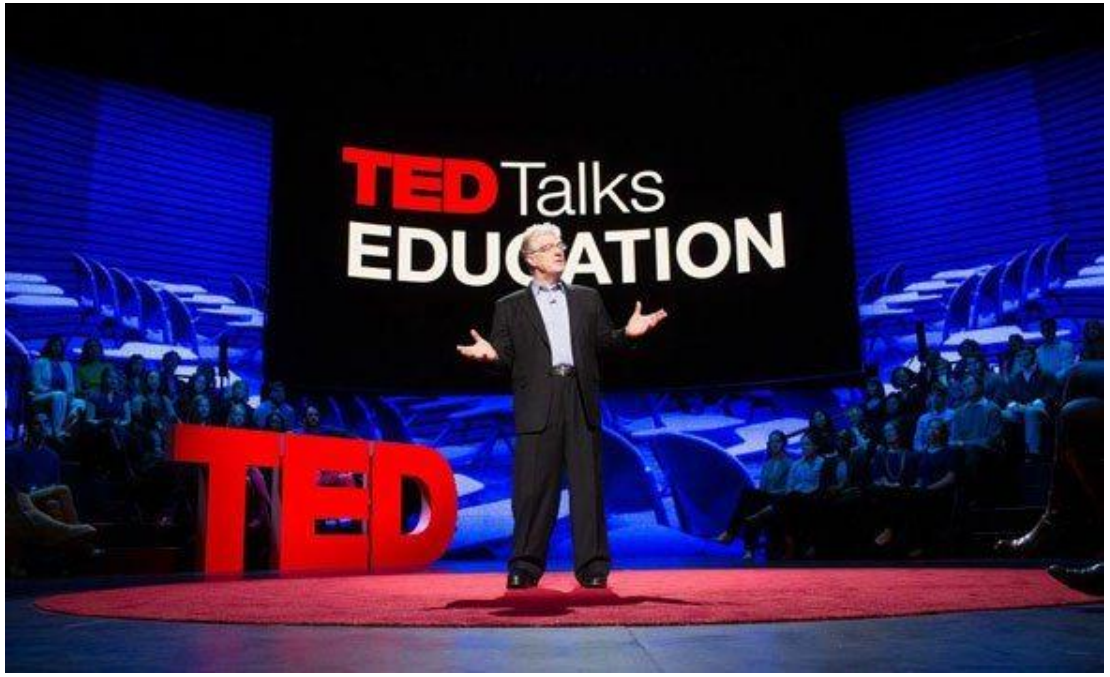


Image from: http://www.findaspark.co.uk/resource/schools-kill-creativity/

2. Data Cleaning

- *'ratings'* column

```
ted_clean[['ratings']][:1]
```

[36] ✓ 0.0s Python

	ratings
0	[{'id': 7, 'name': 'Funny', 'count': 19645}, {'id': 1, 'name': 'Beautiful', 'count': 4573}, {'id': 9, 'name': 'Ingenious', 'count': 6073}, {'id': 3, 'name': 'Courageous', 'count': 3253}, {'id': 11, 'name': 'Longwinded', 'count': 387}, {'id': 2, 'name': 'Confusing', 'count': 242}, {'id': 8, 'name': 'Informative', 'count': 7346}, {'id': 22, 'name': 'Fascinating', 'count': 10581}, {'id': 21, 'name': 'Unconvincing', 'count': 300}, {'id': 24, 'name': 'Persuasive', 'count': 10704}, {'id': 23, 'name': 'Jaw-dropping', 'count': 4439}, {'id': 25, 'name': 'OK', 'count': 1174}, {'id': 26, 'name': 'Obnoxious', 'count': 209}, {'id': 10, 'name': 'Inspiring', 'count': 24924}]

```
ted_clean[['ratings_tuple']][:1]
```

[47] ✓ 0.0s Python

	ratings_tuple
0	((Funny, 19645), (Beautiful, 4573), (Ingenious, 6073), (Courageous, 3253), (Longwinded, 387), (Confusing, 242), (Informative, 7346), (Fascinating, 10581), (Unconvincing, 300), (Persuasive, 10704), (Jaw-dropping, 4439), (OK, 1174), (Obnoxious, 209), (Inspiring, 24924))

```
'pos_neg', 'label']].head(10)
```

[42] ✓ 0.0s Python

	pos_neg	label
0	(98.7721, 1.2279)	pos
1	(79.3633, 20.6367)	neg
2	(86.8932, 13.1068)	neg
3	(95.7178, 4.2822)	pos
4	(98.7782, 1.2218)	pos
5	(91.2974, 8.7026)	pos
6	(86.2917, 13.7083)	neg
7	(91.4672, 8.5328)	pos
8	(83.8344, 16.1656)	neg
9	(79.8812, 20.1188)	neg

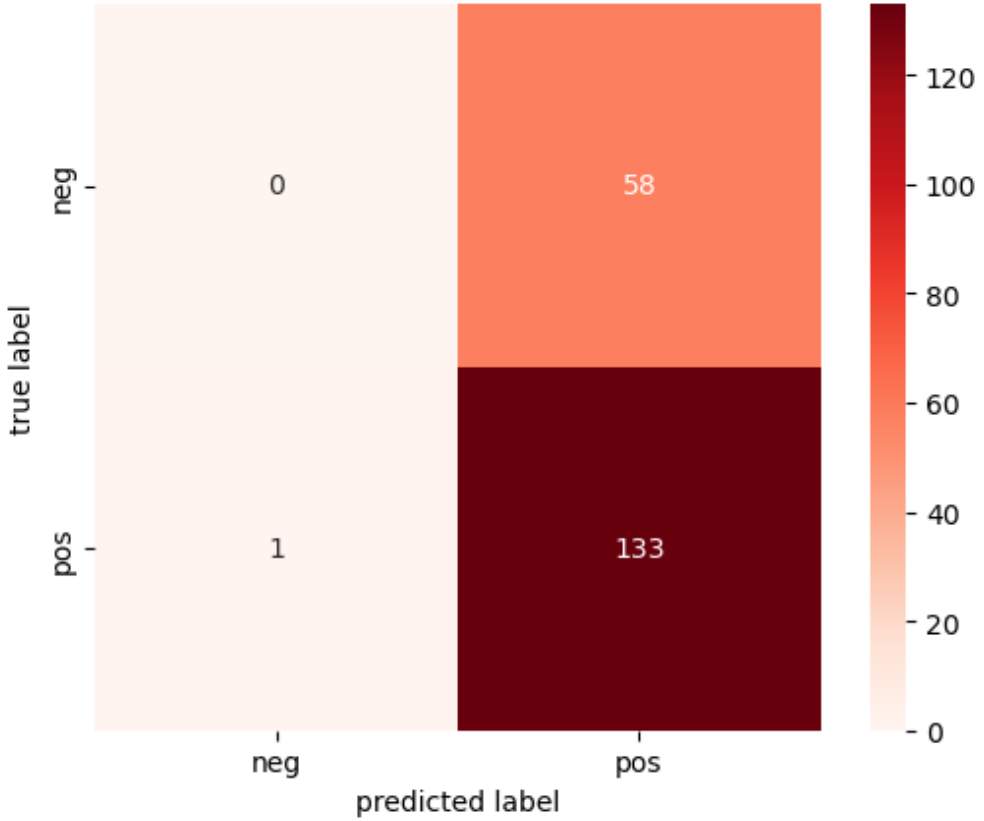
3. Analysis

- Features I am using:
  Tf-idf (unigram, bigrams, trigrams), Mean K-band, Mean Sentence Length
- Models I am using:
  Multinomial NB, SVM

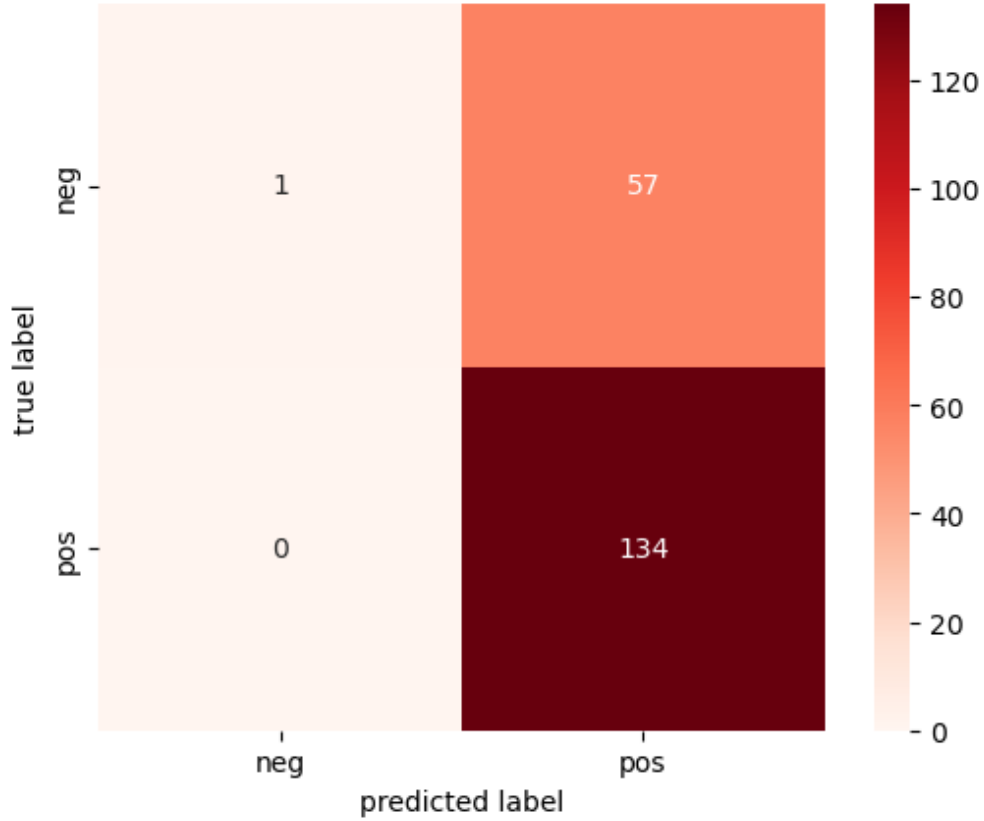
- Hypothesis testing here:
  1. It is possible to predict the popularity of talks based on their transcript
  2. There is a positive correlation between the rating `obnoxious` and mean k-band
  3. There is a positive correlation between the rating `longwinded` and mean sentence length

3. Analysis 1 – Multinomial NB (unigram)

1. Max Feature = 1500, Accuracy Score = 69%

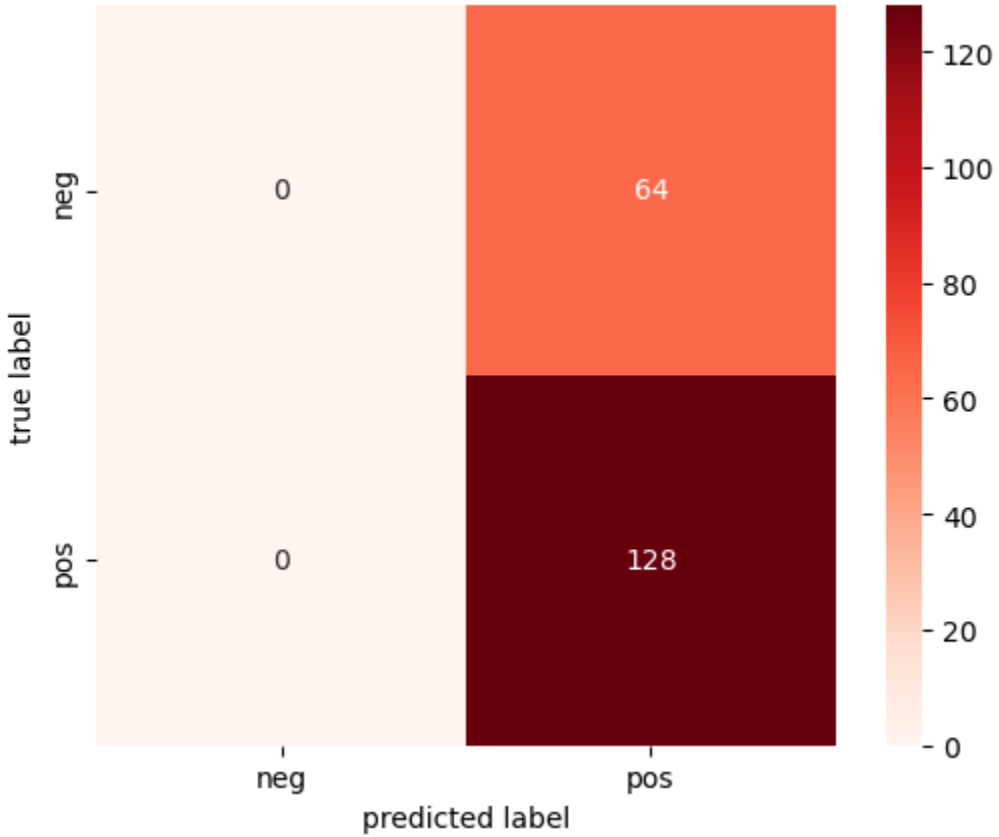


2. Max Feature = 3000, Accuracy Score = 70%



3. Analysis 1 – Multinomial NB (bigrams)

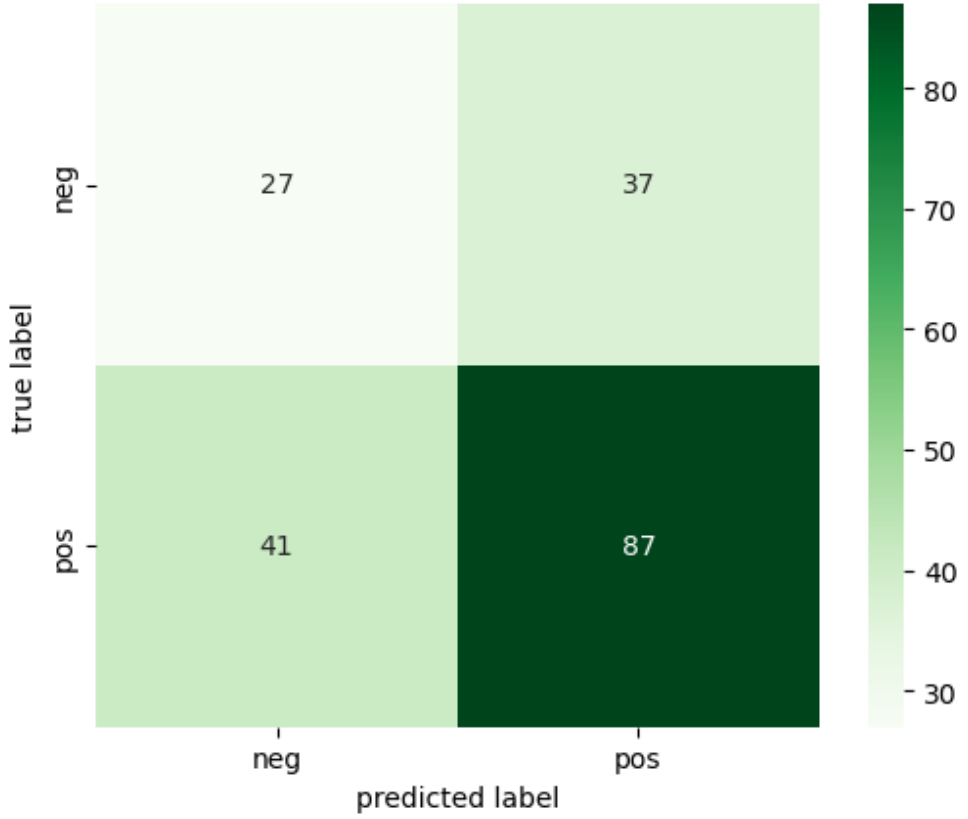
Max feature = 20000, Accuracy Score = 66%



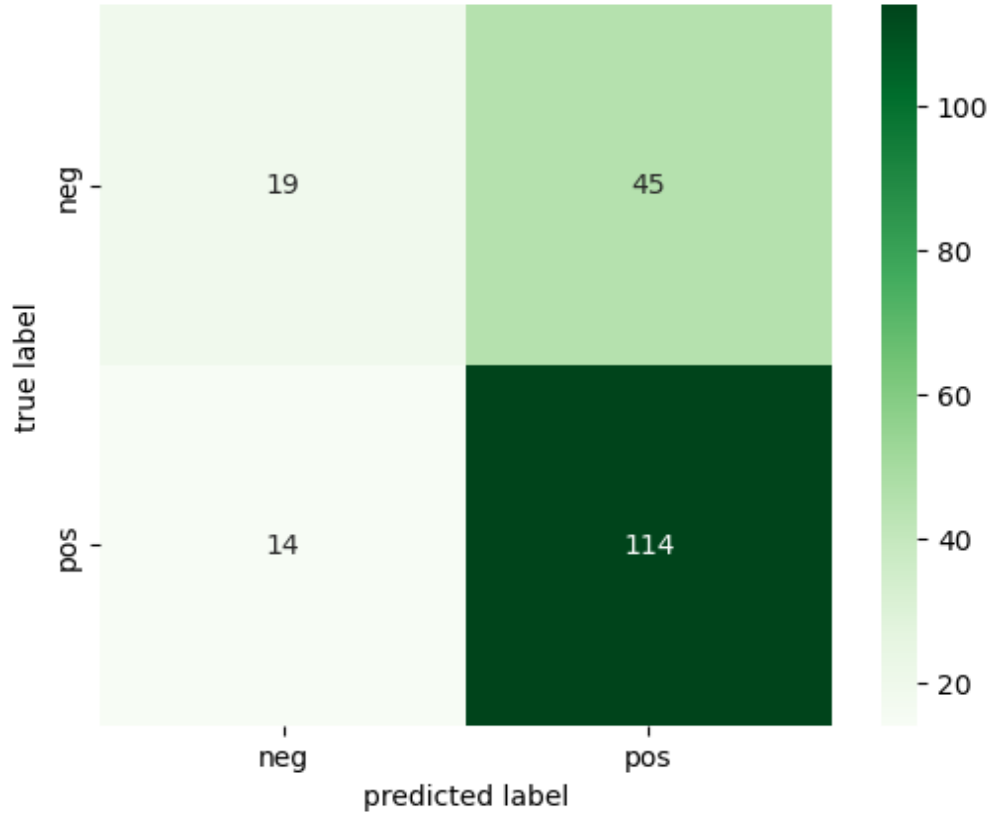
- The model classified all of the transcripts as positive.
- This indicates that there is no difference between the talks with more positive rating and those with less positive rating when it comes to tf-idf feature.

3. Analysis 1 – SVM (unigram)

Max Feature = 15000, Accuracy Score = 59%, C = 1E5

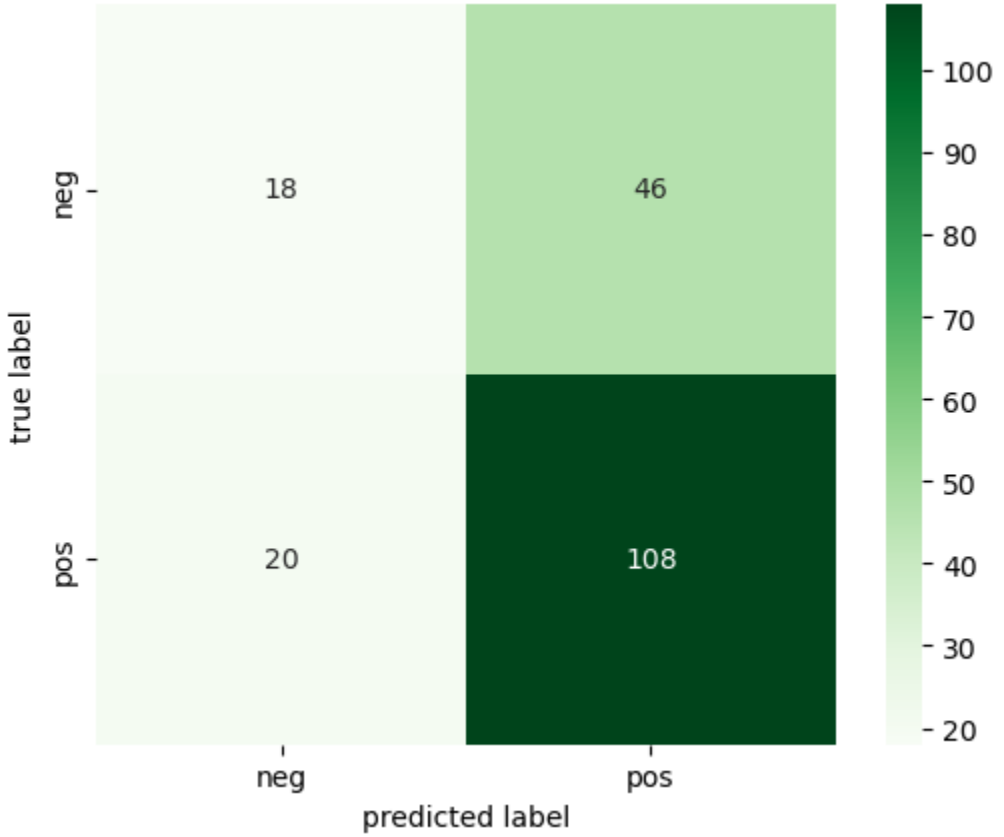


Max Feature = 15000, Accuracy Score = 65%, C = 1



3. Analysis 1 – SVM (ngrams)

Max Feature = 15000, Accuracy Score = 65%, C = 1E5



- Overall, the accuracy was lower than that of NB model
- However, SVM models were more successful in classifying true negatives. – SVM more sophisticated than NB

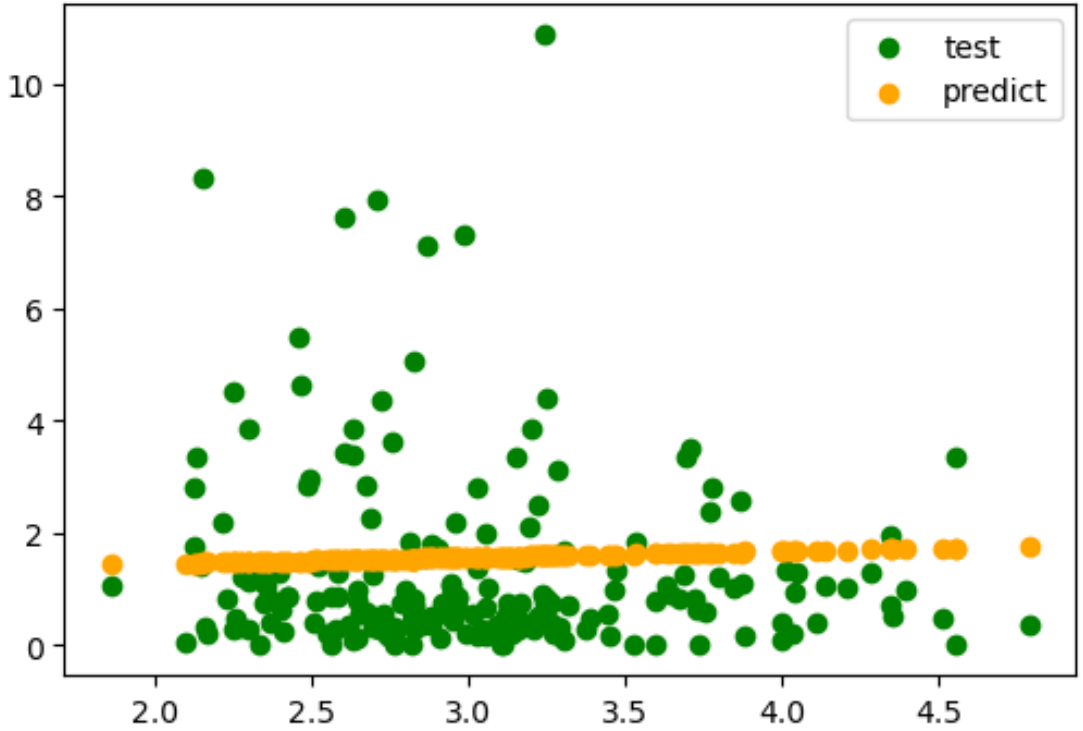
- Hypothesis – rejected.

3.

	transcript
361	I'll start with my favorite muse, Emily Dickinson, who said that wonder is not knowledge, neither is it ignorance. It's something which is suspended between what we believe we can be, and a tradition we may have forgotten. And I think, when I listen to these incredible people here, I've been so inspired — so many incredible ideas, so many visions. And yet, when I look at the environment outside, you see how resistant architecture is to change. You see how resistant it is to those very ideas. We can think them out. We can create incredible things. And yet, at the end, it's so hard to change a wall. We applaud the well-mannered box. But to create a space that never existed is what interests me; to create something that has never been, a space that we have never entered except in our minds and our spirits. And I think that's really what architecture is based on.Architecture is not based on concrete and steel and the elements of the soil. It's based on wonder. And that wonder is really what has created the greatest cities, the greatest spaces that we have had. And I think that is indeed what architecture is. It is a story. By the way, it is a story that is told through its hard materials. But it is a story of effort and struggle against improbabilities. If you think of the great buildings, of the cathedrals, of the temples, of the pyramids, of pagodas, of cities in India and beyond, you think of how incredible this is that that was realized not by some abstract idea, but by people.So, anything that has been made can be unmade. Anything that has been made can be made better. There it is: the things that I really believe are of important architecture. These are the dimensions that I like to work with. It's something very personal. It's not, perhaps, the dimensions appreciated by art critics or architecture critics or city planners. But I think these are the necessary oxygen for us to live in buildings, to live in cities, to connect ourselves in a social space.And I th...

3. Analysis 2 – K-band & obnoxious

- Model: Regression
- Result: very low correlation
  - Coef: 0.11
  - Mean Absolute Error: 1.24
- Conclusion: there are other elements that induce ‘obnoxious’ rating more than the level of the words in the transcript.

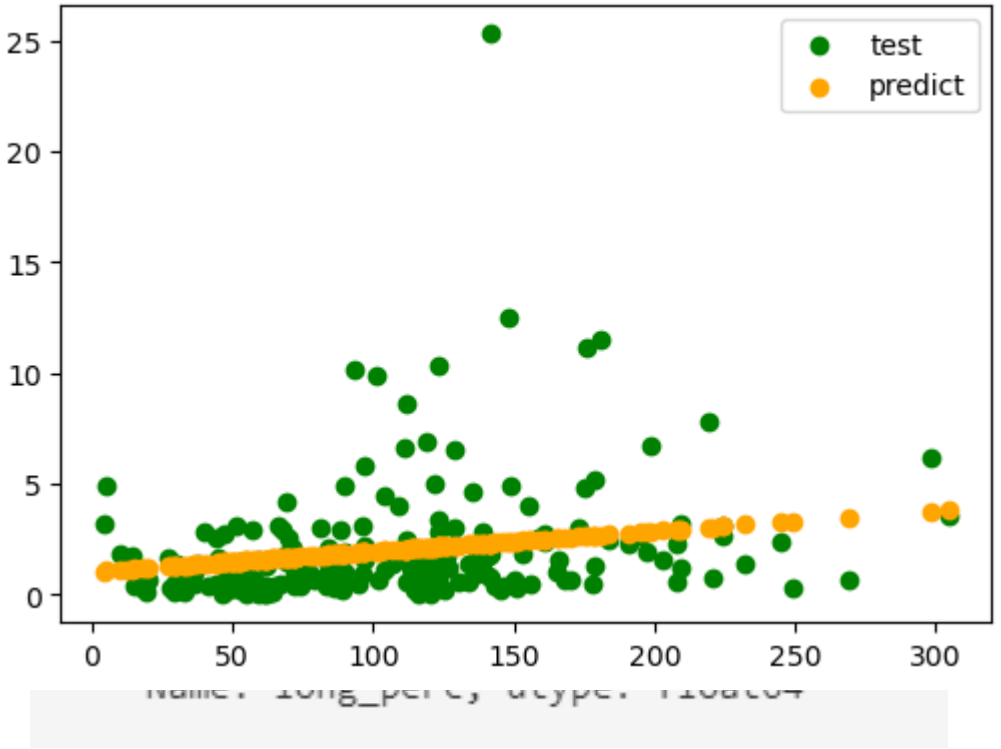


3. A

	transcript
318	When I was five years old I fell in love with airplanes. Now I'm talking about the '30s. In the '30s an airplane had two wings and a round motor, and was always flown by a guy who looked like Cary Grant. He had high leather boots, jodhpurs, an old leather jacket, a wonderful helmet and those marvelous goggles — and, inevitably, a white scarf, to flow in the wind. He'd always walk up to his airplane in a kind of saunter, devil-may-care saunter, flick the cigarette away, grab the girl waiting here, give her a kiss. (Laughter) And then mount his airplane, maybe for the last time. Of course I always wondered what would happen if he'd kissed the airplane first. (Laughter)But this was real romance to me. Everything about flying in those years, which was — you have to stop and think for a moment — was probably the most advanced technological thing going on at the time. So as a youngster, I tried to get close to this by drawing airplanes, constantly drawing airplanes. It's the way I got a part of this romance. And of course, in a way, when I say romance, I mean in part the aesthetics of that whole situation. I think the word is the holistic experience revolving around a product. The product was that airplane. But it built a romance. Even the parts of the airplane had French names. Ze fuselage, ze empanage, ze nessal. You know, from a romance language. So that it was something that just got into your spirit. It did mine.And I decided I had to get closer than just drawing fantasy airplanes. I wanted to build airplanes. So I built model airplanes. And I found that in doing the model airplanes the appearance drawings were not enough. You couldn't transfer those to the model itself. If you wanted it to fly you had to learn the discipline of flying. You had to learn about aeronautics. You had to learn what made an airplane stay in the air. And of course, as a model in those years, you couldn't control it. So it had to be self-righting, and stay up without crashing. So I had t...

3. Analysis 3 – Sentence Length & `longwinded`

- Model: Regression
- Result: No correlation.
  - Coef: 0.009
  - Mean Absolute Error: 1.59
- Conclusion: The length of the sentence has no significant effect on being rated as longwinded.



4. Summary

- The textual features of transcript (Tf-idf (ngrams), K-band, sentence length) does *not* affect the talk's rating / popularity in a meaningful way.
- My best guess: nonverbal aspects of the talk (tone, speed rate, body language, …) matters more.

THANK YOU ☺

Any Questions?