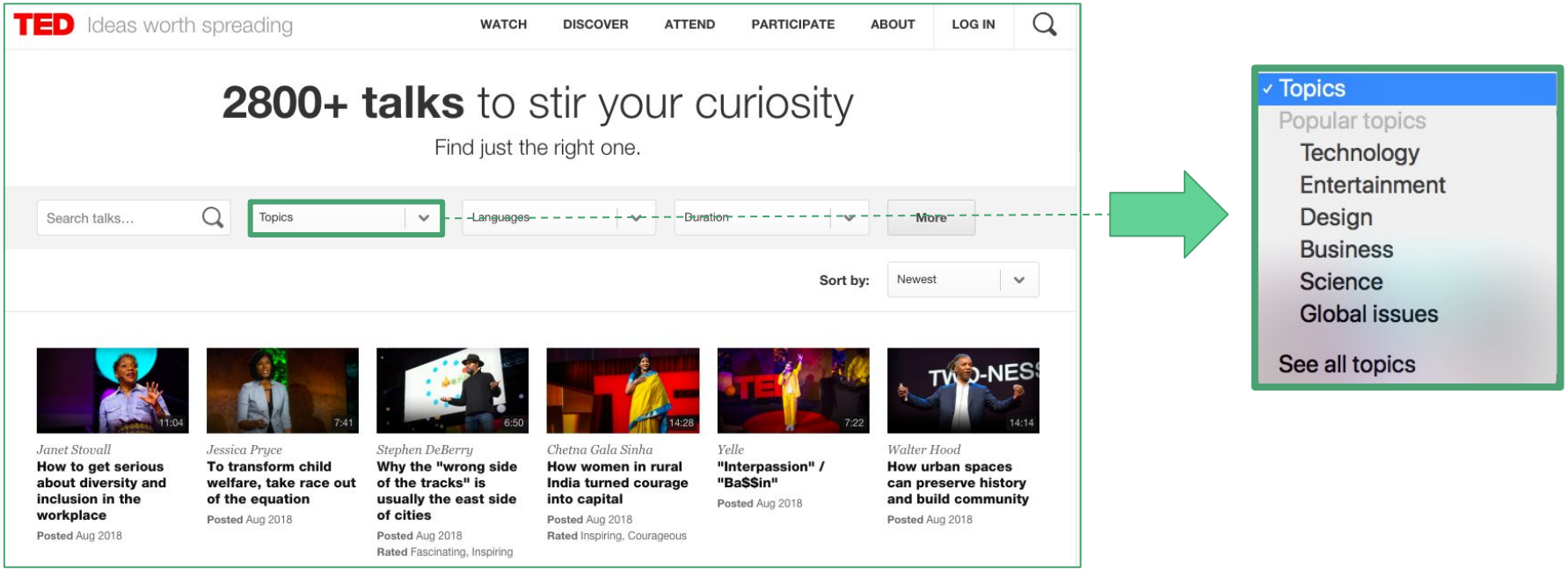
Predicting TED Talk Topics

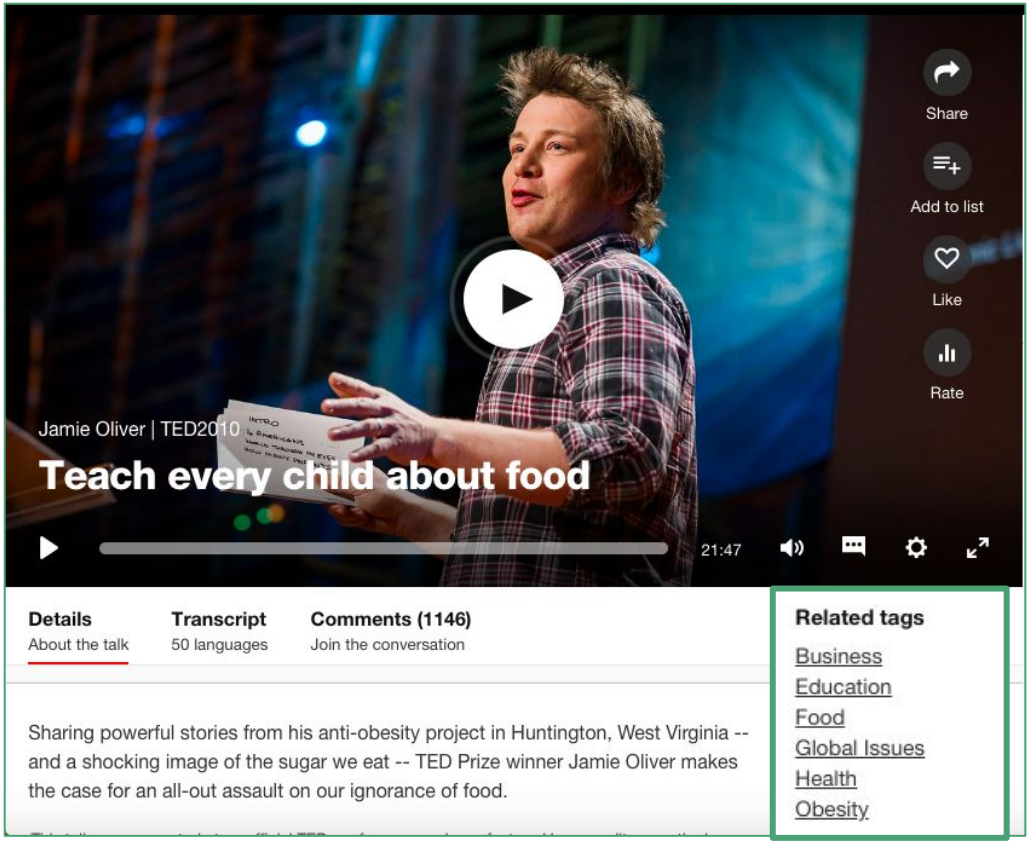
Laura Borton
Project Fletcher

What is the Purpose?



Group talks into 6 popular topics

Data



Years:
June 2006 - March 2018

Talks:
Over 2500

Relevant Features:

- **Title**
- **Summary**
- **Transcript**
- **Tags**

How well can a model predict?

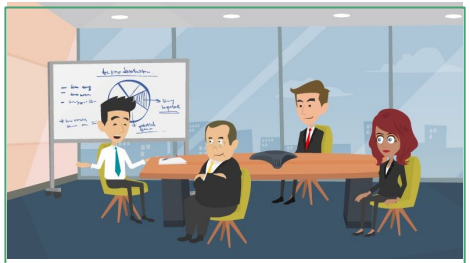
Predicted Talk Topic

Business

?

=

True Talk Topic



Business

Model

?

>

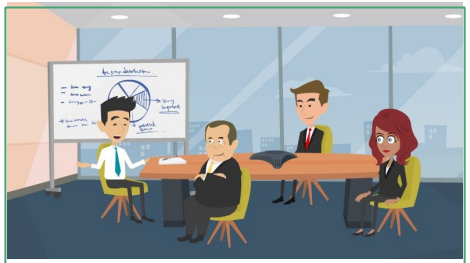
How well can a model predict?

Predicted Talk Topic

Business

55%

True Talk Topic



Business

Linear SVM:
55% Accuracy

>

17% Accuracy

Can the talks be classified differently?

Ideal number of clusters (20 - 25) gives:

1. **Topics already identified (but not the most popular)**

Cluster 0: Business / Economy

Cluster 10: Social Change / Activism

Can the talks be classified differently?

Ideal number of clusters (20 - 25) gives:

1. **Topics already identified (but not the most popular)**

Cluster 0: Business / Economy

Cluster 10: Social Change / Activism

2. **Nonsense: Same word, not topic**

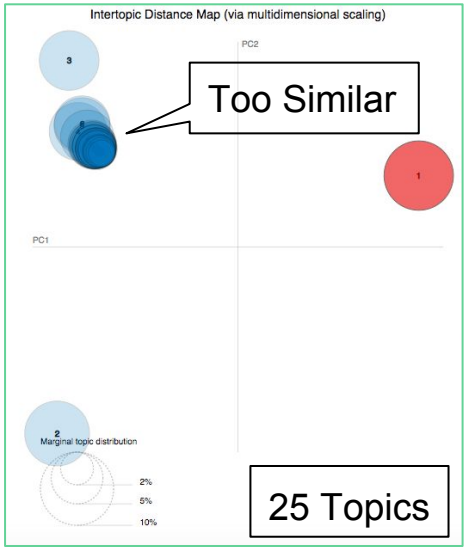
Warning: This talk might contain much more than you'd ever want to know about the way the world poops. But as sanitation activist (and TED Fellow) Francis de los Reyes asks ... doesn't everyone deserve a safe place to go?

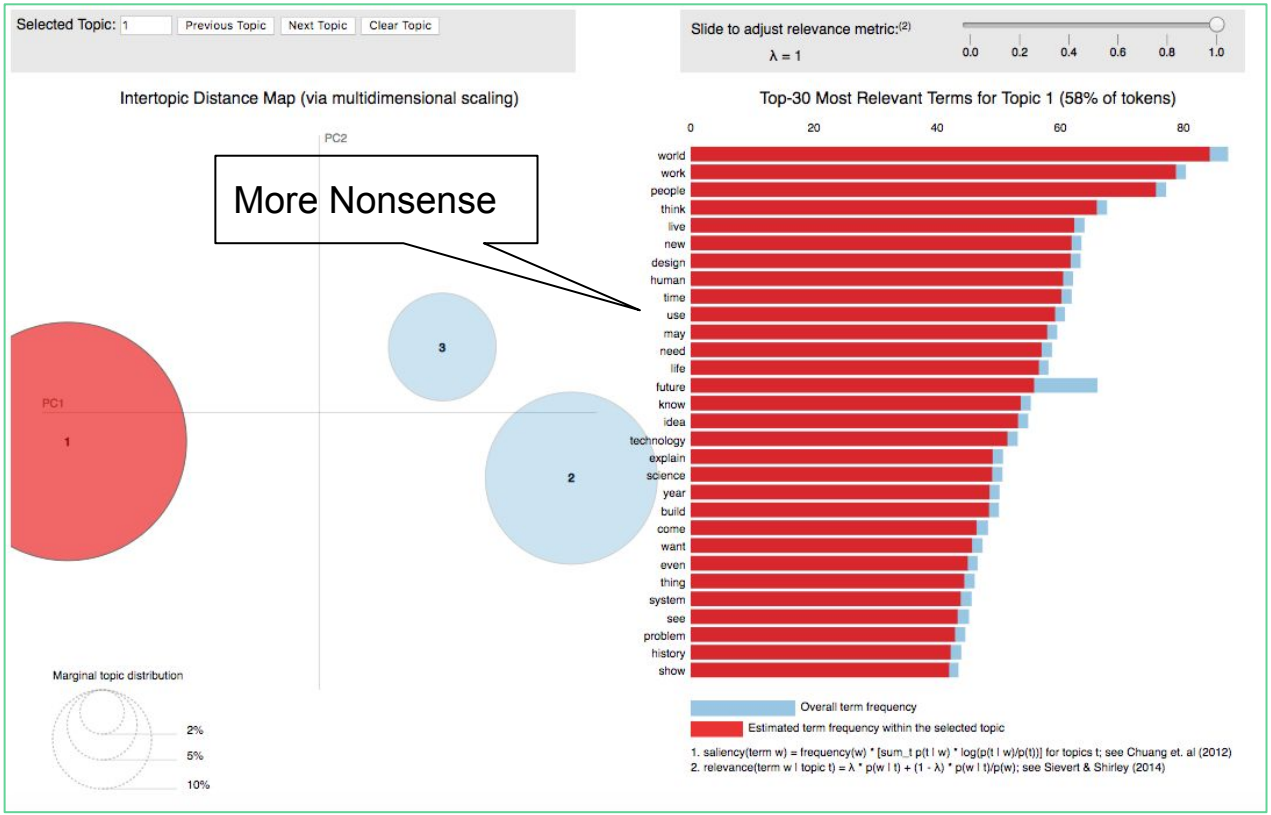
Let's admit it: aerial photo drones and UAVs are a little creepy, and they come with big regulatory and safety problems. But aerial photos can be a powerful way of telling the truth about the world: the size of a protest, the spread of an oil spill, the wildlife hidden in a delta. Sergei Lupashin demos Fotokite, a nifty new way to see the world from on high, safely and under control.

An epidemic of bad, inefficient, overcrowded meetings is plaguing the world's businesses ... and making workers miserable. David Grady has some ideas on how to stop it.

Cluster 4: World

Can the talks be classified differently?

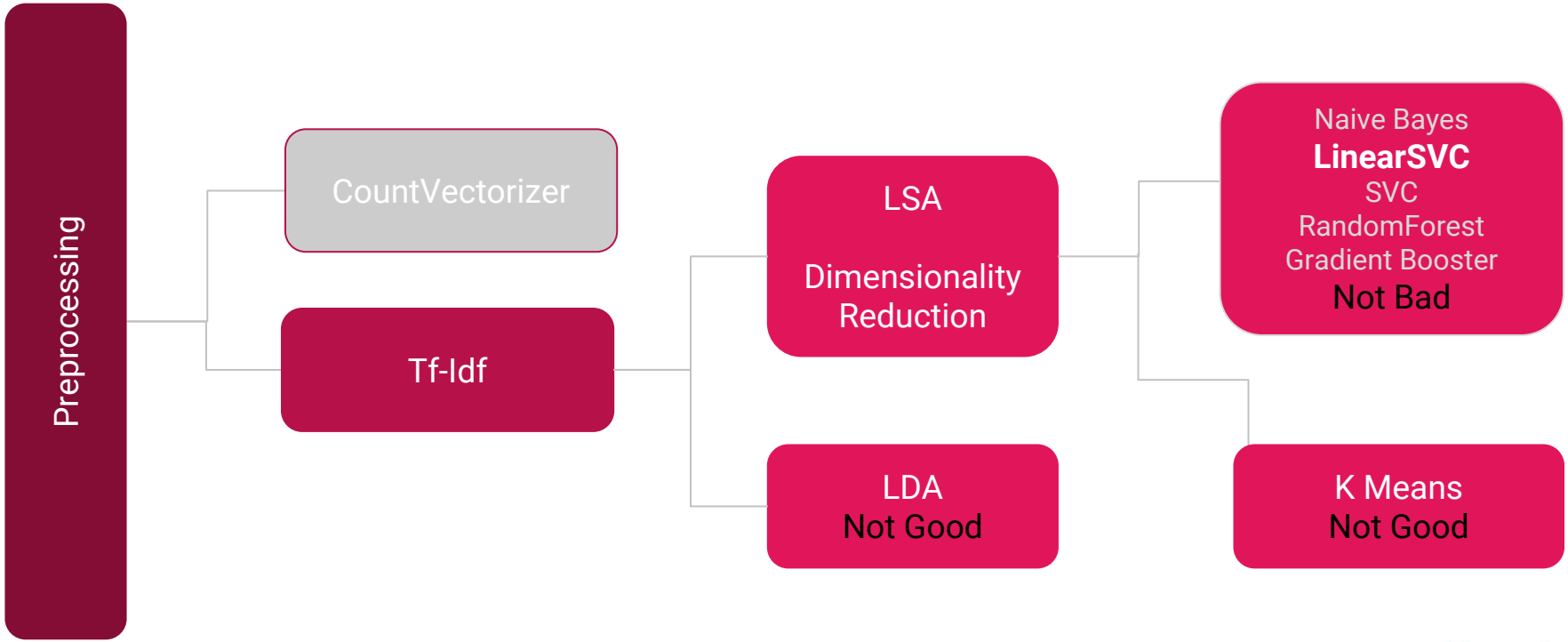




Going forward...

Use full talk transcripts for better category prediction

Summary



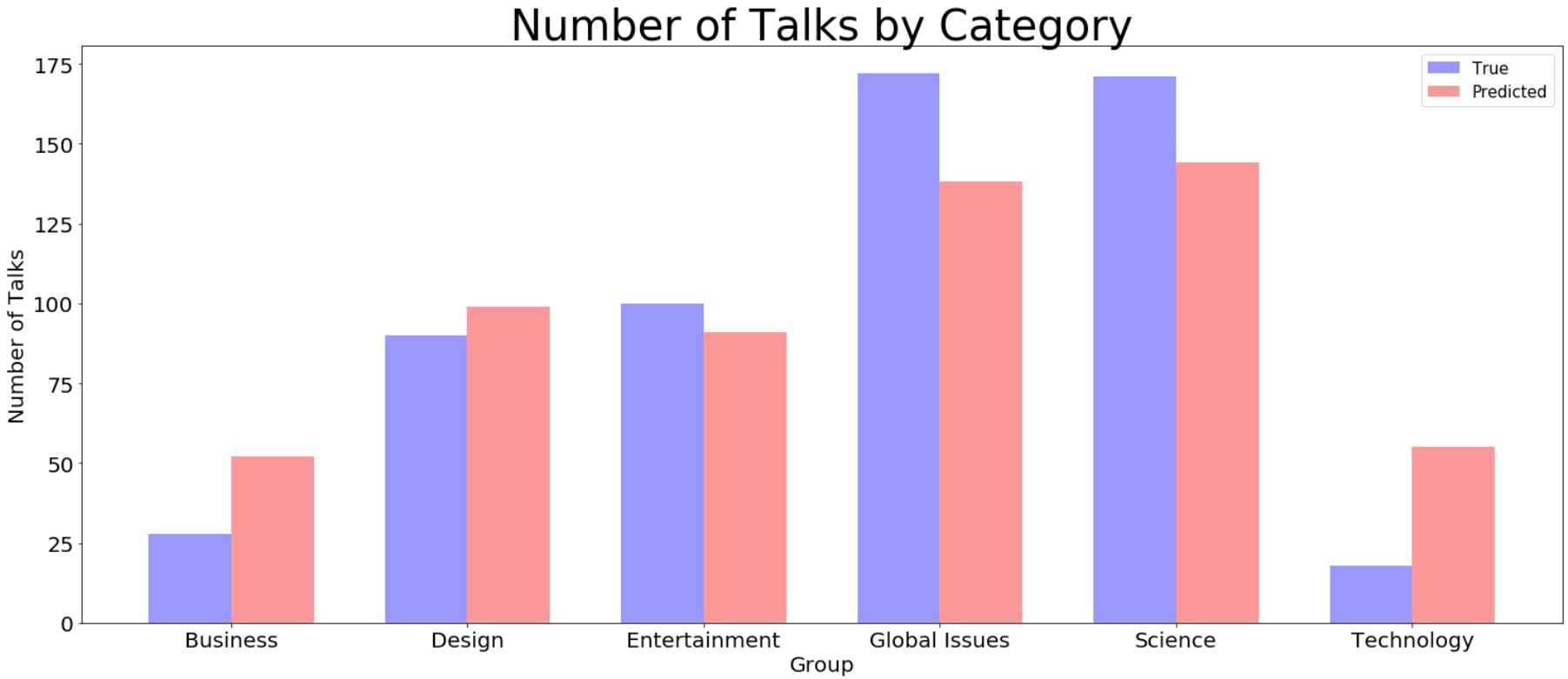
python™

Appendix

Supervised Learning

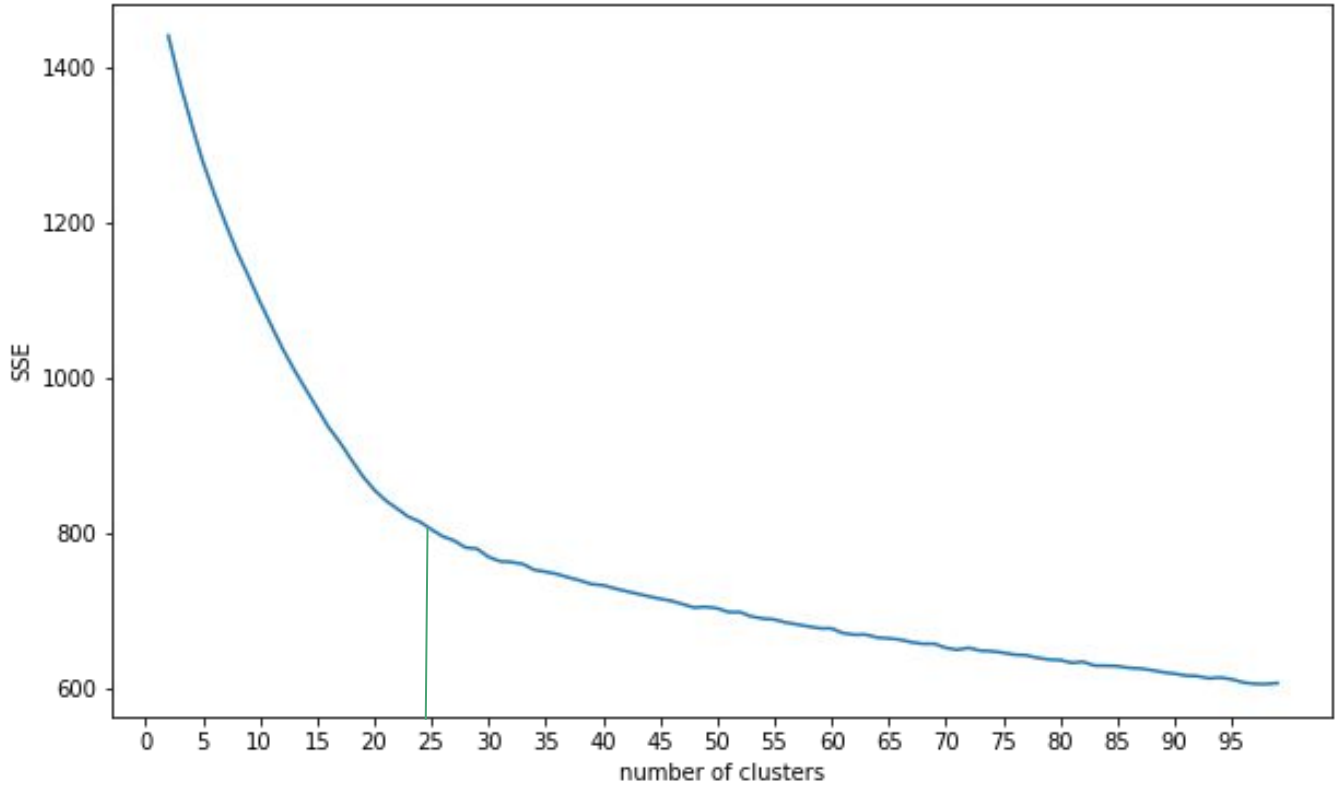
Algorithm	**Accuracy**
Naive Bayes (Gaussian) (CountVectorizer)	0.45
LinearSVC	**0.55**
SVC - linear	0.54
SVC - rbf	0.53
Random Forest	0.46
Gradient Boosting	0.49

Tags and Topics (Test Data Set)



Talks without one of the six popular topic tags were removed (2200 → 1950 talks)

The maximum number of clusters: 25



Tags and Topics

Talk Topics (Jamie Oliver Example)

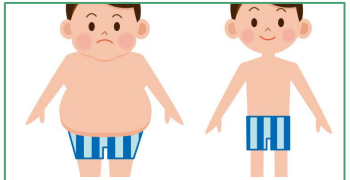

Business


Education

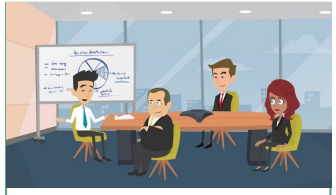

Global Issues


Food


Health


Obesity

Popular Topics


Business


Global Issues


Entertainment

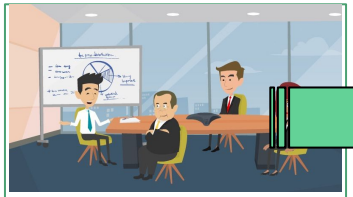

Design

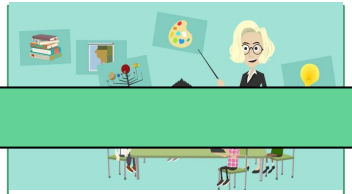

Technology

Science

Tags and Topics

Talk Topics (Jamie Oliver Example)

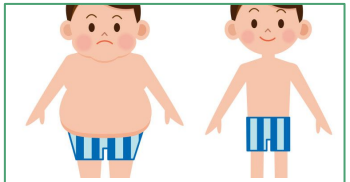

Business


Education

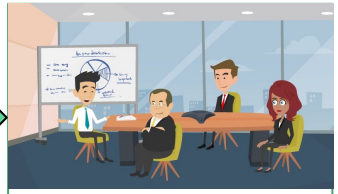

Global Issues


Food


Health


Obesity

Popular Topics


Business


Global Issues


Entertainment


Design


Technology

Science