The Language of TEDTalks + Data Science

...

An Exploration of Natural Language Processing

Jay Kim

The Motivation: Why TEDTalks Data?

- **Personal Favorites**
  - Preference for certain topics (i.e., humanity-related)
  - Fertile ground for context, semantic analyses
- **Recent, Interesting Dataset- Linguistically, Topic-wise, etc.**
  - Part of ongoing study- potential to contribute (especially w/ cleaning, transcribing)

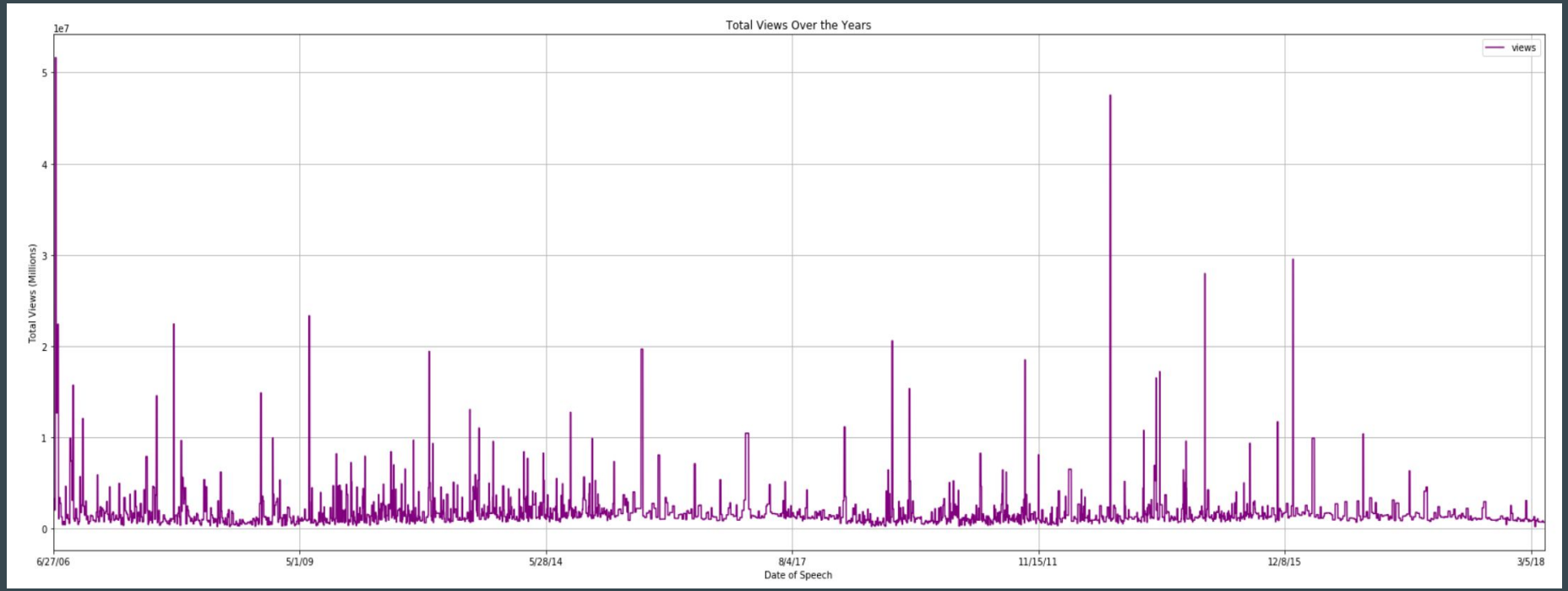
Sources & Tools

Exploratory Data Analysis

- Multiple Approaches of Pre-processing
  - Contractions, Accents
  - Stemming/Lemmatizing
  - Stop Words
  - Tokenization
  - Bag of Words
- Regular Expressions

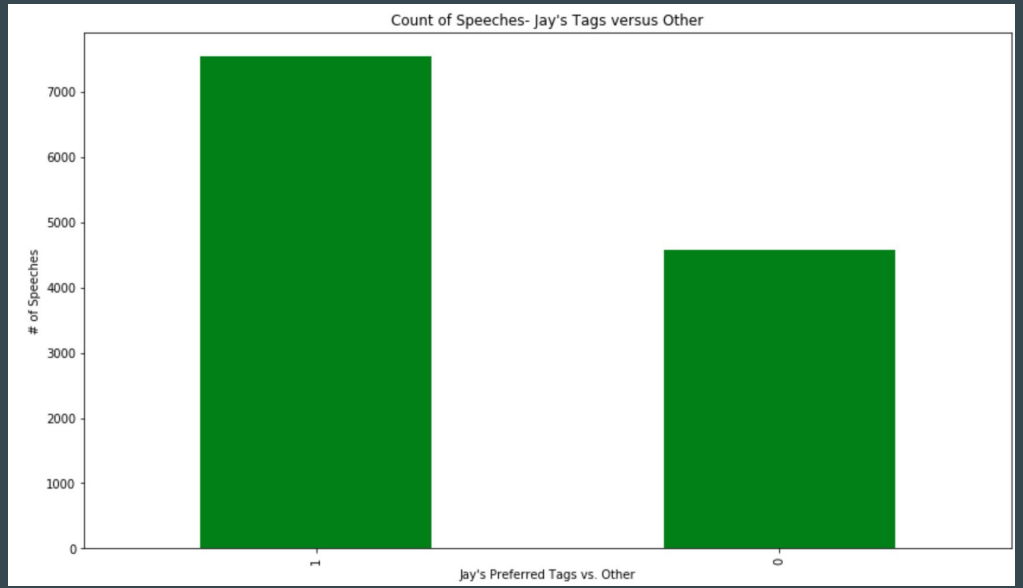
Exploratory Data Analysis

- Categorize & Classify by Total Views?



NLP Data Exploration is Nice and Necessary- But...

- **Where Am I Going? What's the End-Goal?**
  - Classification: Instead, decided to go by "Jay's Preferred Tags" - Humanity, Activism, Society, Social Good, Future, and Community
  - Similarity Component
  - Interactive Aspect



Topic Modeling

- LDA
- NMF
- Visuals

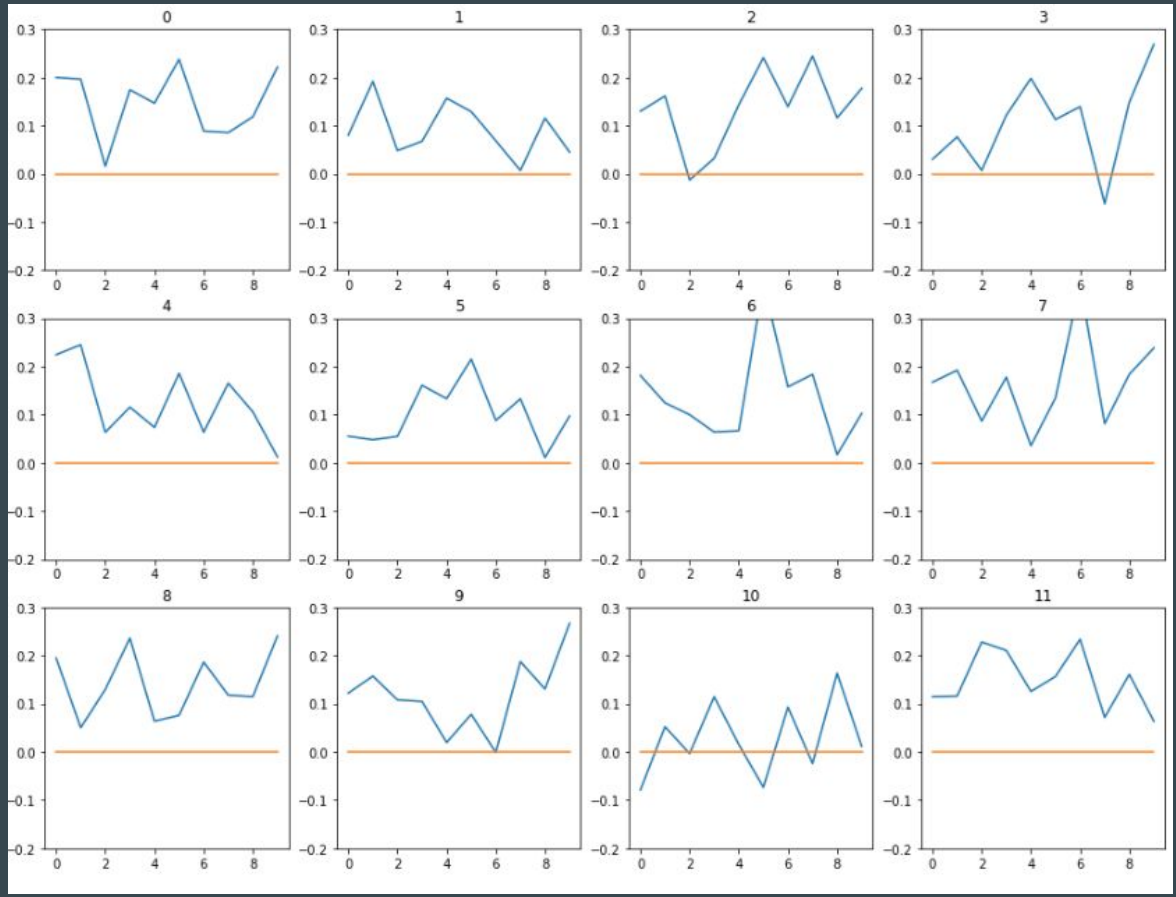
- Pattern Emerging

Topic Modeling + Sentiment Analysis

```
Out[200]: [(0,
  '0.009*"country" + 0.006*"problem" + 0.006*"new" + 0.006*"percent" + 0.005*"government" +
0.005*"city" + 0.005*"today" + 0.005*"good" + 0.005*"state" + 0.004*"life"'),
 (1,
  '0.009*"africa" + 0.007*"story" + 0.006*"african" + 0.005*"life" + 0.005*"day" + 0.004*"wa
r" + 0.004*"film" + 0.004*"woman" + 0.004*"new" + 0.004*"country"'),
 (2,
  '0.005*"good" + 0.005*"life" + 0.005*"little" + 0.005*"idea" + 0.005*"lot" + 0.005*"kind" +
0.005*"story" + 0.005*"day" + 0.004*"work" + 0.004*"new"'),
 (3,
  '0.008*"life" + 0.007*"human" + 0.007*"cancer" + 0.006*"planet" + 0.006*"new" + 0.005*"eart
h" + 0.005*"different" + 0.005*"cell" + 0.004*"lot" + 0.004*"specie"'),
 (4,
  '0.008*"kind" + 0.006*"new" + 0.006*"little" + 0.006*"idea" + 0.006*"lot" + 0.006*"day" +
0.005*"life" + 0.005*"different" + 0.004*"work" + 0.004*"space"'),
 (5,
  '0.016*"brain" + 0.008*"little" + 0.007*"robot" + 0.006*"car" + 0.006*"body" + 0.005*"cell"
+ 0.005*"kind" + 0.005*"different" + 0.005*"lot" + 0.005*"human"'),
 (6,
  '0.024*"woman" + 0.009*"life" + 0.008*"men" + 0.008*"day" + 0.007*"child" + 0.007*"girl" +
0.006*"school" + 0.006*"man" + 0.006*"family" + 0.005*"community"'),
 (7,
  '0.010*"life" + 0.007*"brain" + 0.006*"drug" + 0.005*"patient" + 0.005*"day" + 0.005*"lot"
+ 0.005*"disease" + 0.004*"human" + 0.004*"good" + 0.004*"story"')]
```

Sentiment Analysis

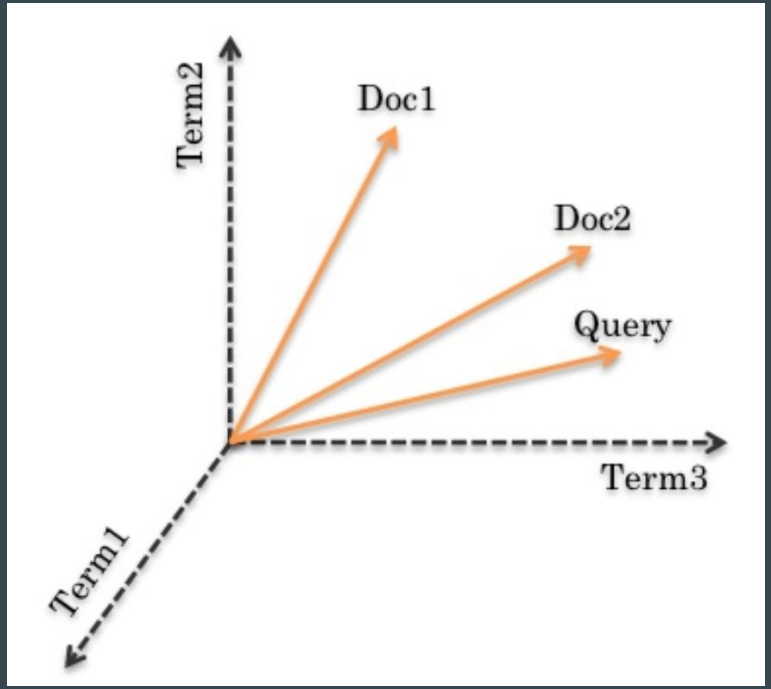
- Broke it up into 10-piece Text Clusters
- Can expand to a Full cross-comparison



Document Similarity + Classification

- TF-IDF Vectors (GenSim)

- Classification using Naive Bayes (+ SVM)



Document Similarity

Sample Search Results (right):

- ➔ Choice btw Searching Speech Transcript or Description
- ➔ Outputs a (transposed) dataframe

```
1 for TEDTalks Search
2 for TEDTalks Classifier: 1
Enter
1 to Search by Description
2 to Search by Speech Transcript: 1
Enter Search Term(s): calm
INDEX =  0
                                                                  0
public_url            https://www.ted.com/talks/al_gore_on_averting_...
headline                                      Averting the climate crisis
description           With the same humor and humanity he exuded in ...
event                                                          TED2006
duration                                                       0:16:17
published                                                      6/27/06
speaker_1                                                      Al Gore
speaker1_occupation                                    Climate advocate
speaker1_introduction  Nobel Laureate Al Gore focused the world's att...
speaker1_profile       Why you should listen\nFormer Vice President A...
speech_length                                                  12074
reached_threshold                                                  1
prefers                                                            1
polarity                                                    0.142703
subjectivity                                                0.451107
text                   avert climate crisis alternative humor humanit...

Process finished with exit code 0
```

Classification- Naive Bayes + SVM

- Run on various combinations of text categories:
  - Sub-par Precision Score
  - Ran using SVM, barely improved

```
Precision Score: 0.6742081447963801
Recall Score: 0.9141104294478528
Accuracy Score: 0.7304075235109718
F1 Score: 0.7760416666666667
```

- More models to run on, tune

```
1  predicted_svm = text_clf_svm.predict(x_test)
```

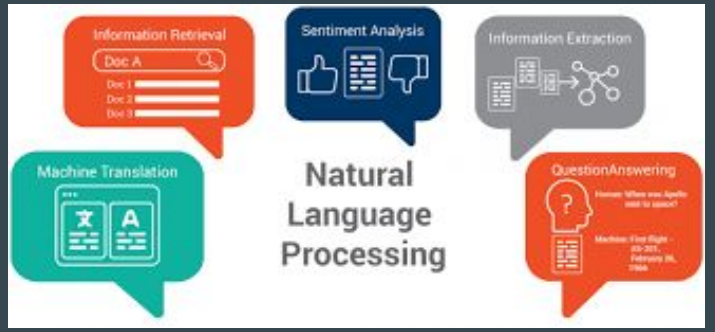
```
1  np.mean(predicted_svm == y_test)
```

```
0.7460815047021944
```

```
1  print_metrics(y_test, predicted_svm)
```

```
Precision Score: 0.6990291262135923
Recall Score: 0.8834355828220859
Accuracy Score: 0.7460815047021944
F1 Score: 0.7804878048780489
```

Conclusions



- Discovered some patterns,
  A start of teasing out factors
  - With that said- Challenges of NLP (and Rewards)
- Areas of expansion going forward
  - Will continue to be “fertile ground”
  - Transcript scroller + link to wikipedia profiles
  - More learning and applying pre-processing, other NLP techniques