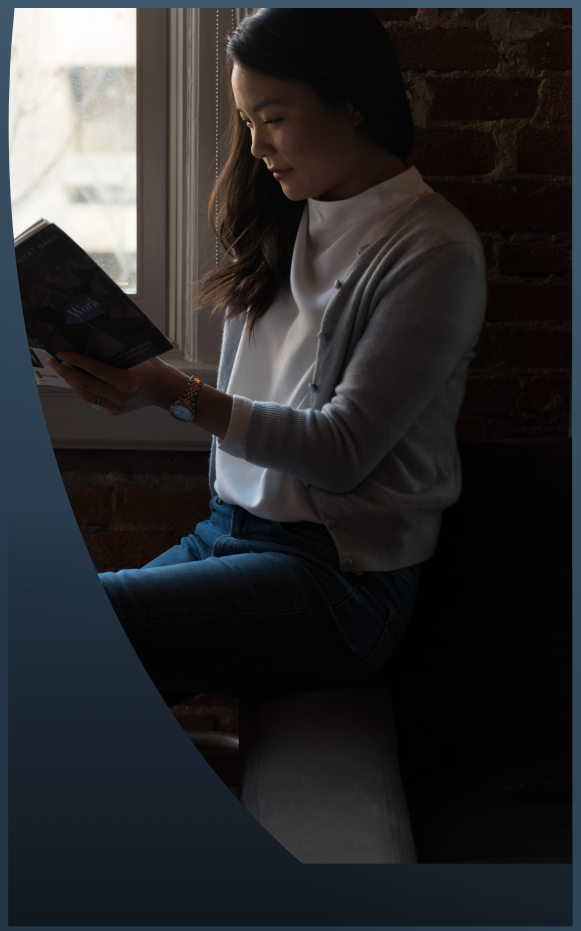
Booz | Allen | Hamilton®



INTRODUCTION TO NATURAL LANGUAGE PROCESSING

Summer Rankin, PhD
Kate Dowdy

JEAN BARTIK COMPUTING SYMPOSIUM
FEBRUARY 27-28, 2020
UNITED STATES NAVAL ACADEMY

AGENDA

INTRODUCTIONS

MACHINE LEARNING QUICK OVERVIEW

INSTALL SOFTWARE

OVERVIEW OF NLP (TOOLS & WORKFLOW)

CODING TUTORIAL

- CLEAN TEXT
- TOKENIZE
- VECTORIZE
- TOPIC MODELING
- VISUALIZATIONS

SUMMER RANKIN, PHD

Data Scientist @ Booz Allen Hamilton

PhD in Complex Systems & Brain Sciences

Neuroscientist @ Johns Hopkins University

Nonlinear signal processing & Stochastic Systems

FAU Florida Atlantic University

Johns Hopkins Medicine

Women in Data Science

Booz | Allen | Hamilton®

SummerRankin.com

rankin_summer@bah.com

KATE DOWDY

- Data Scientist @ Booz Allen
- BA Economics
- Former program manager at education/entrepreneurship nonprofits; former English teacher

WILLIAM & MARY
CHARTERED 1693

GA GENERAL ASSEMBLY

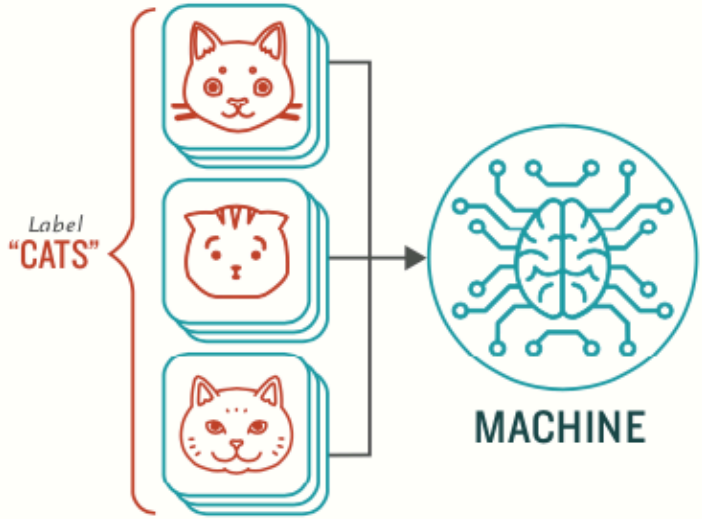
WOMEN IN DATA SCIENCE

Booz | Allen | Hamilton
dowdy_katherine@bah.com

How **Supervised** Machine Learning Works

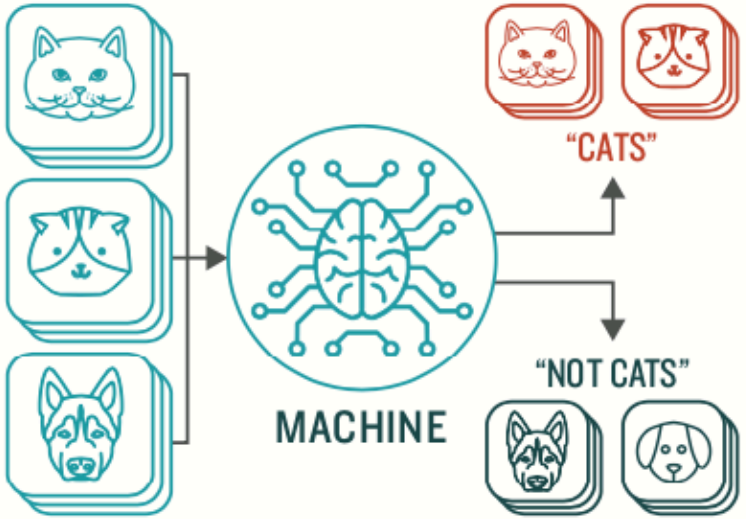
STEP 1

Provide the machine learning algorithm categorized or "labeled" input and output data from to learn

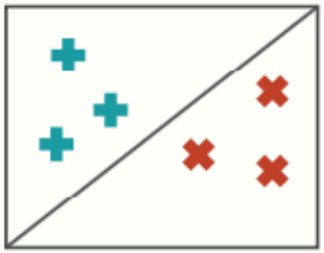


STEP 2

Feed the machine new, unlabeled information to see if it tags new data appropriately. If not, continue refining the algorithm

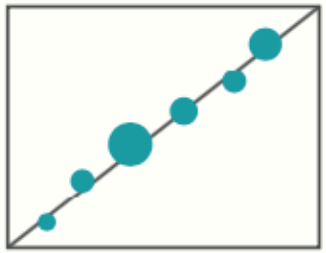


TYPES OF PROBLEMS TO WHICH IT'S SUITED



CLASSIFICATION

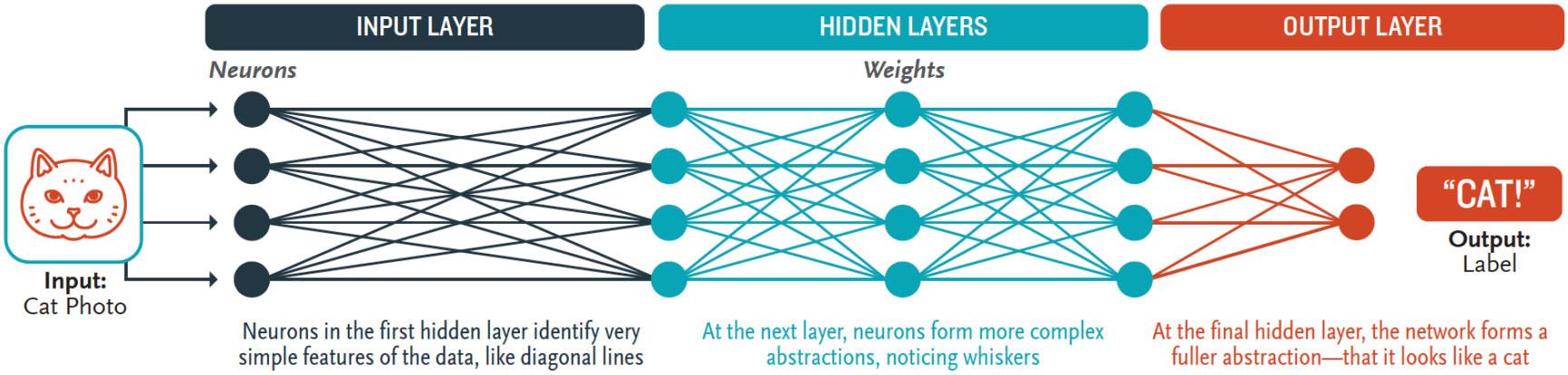
Sorting items into categories



REGRESSION

Identifying real values (dollars, weight, etc.)

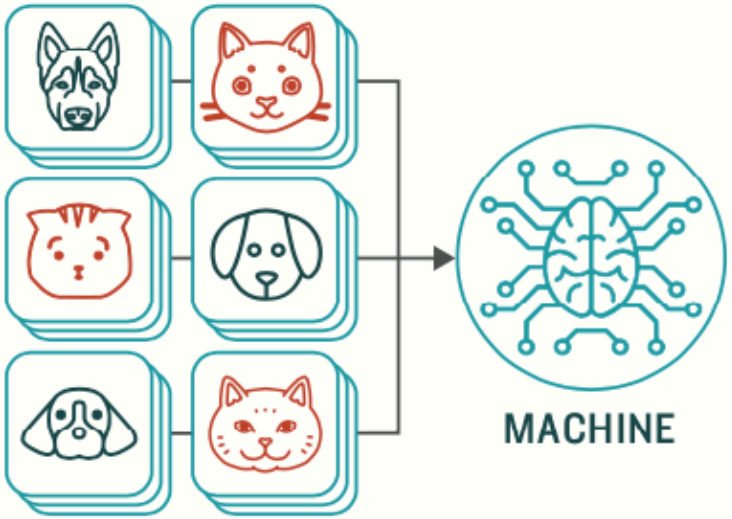
SUPERVISED LEARNING



How **Unsupervised** Machine Learning Works

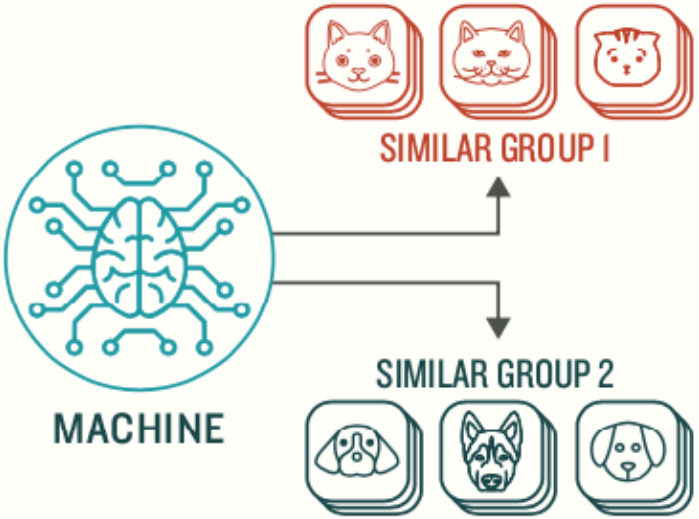
STEP 1

Provide the machine learning algorithm uncategorized, unlabeled input data to see what patterns it finds

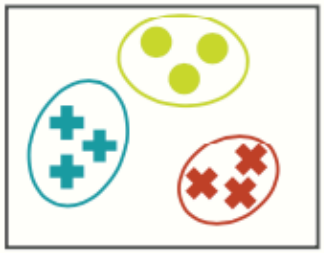


STEP 2

Observe and learn from the patterns the machine identifies



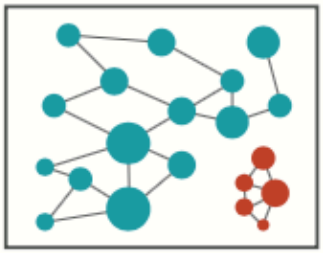
TYPES OF PROBLEMS TO WHICH IT'S SUITED



CLUSTERING

Identifying similarities in groups

For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment



ANOMALY DETECTION

Identifying abnormalities in data

For Example: Is a hacker intruding in our network?

INSTALLATION & SETUP

1. Install Miniconda

https://docs.conda.io/en/latest/miniconda.html#linux-installers

2. Pull the Github Repo

https://github.com/1fmusic/jean_bartik_computing_symposium_rankin

3. Open Anaconda Prompt

4. Navigate to Repo Folder & Execute:

```
conda env create –f environment.yml
conda activate jbcs2020
```

NATURAL LANGUAGE PROCESSING (NLP)

- Search and organize
- Recommenders
- Classification
- Chat bots
- Text creation

NLP TOOLS

Tools we'll use for this tutorial:

Natural Language Analyses with NLTK

NLTK (Natural Language Toolkit)
Text Processing

scikit learn

SciKit Learn
Modeling

pandas

$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$

Pandas
Data Manipulation

matplotlib

Matplotlib
Visualizations

Other tools for NLP using Python:

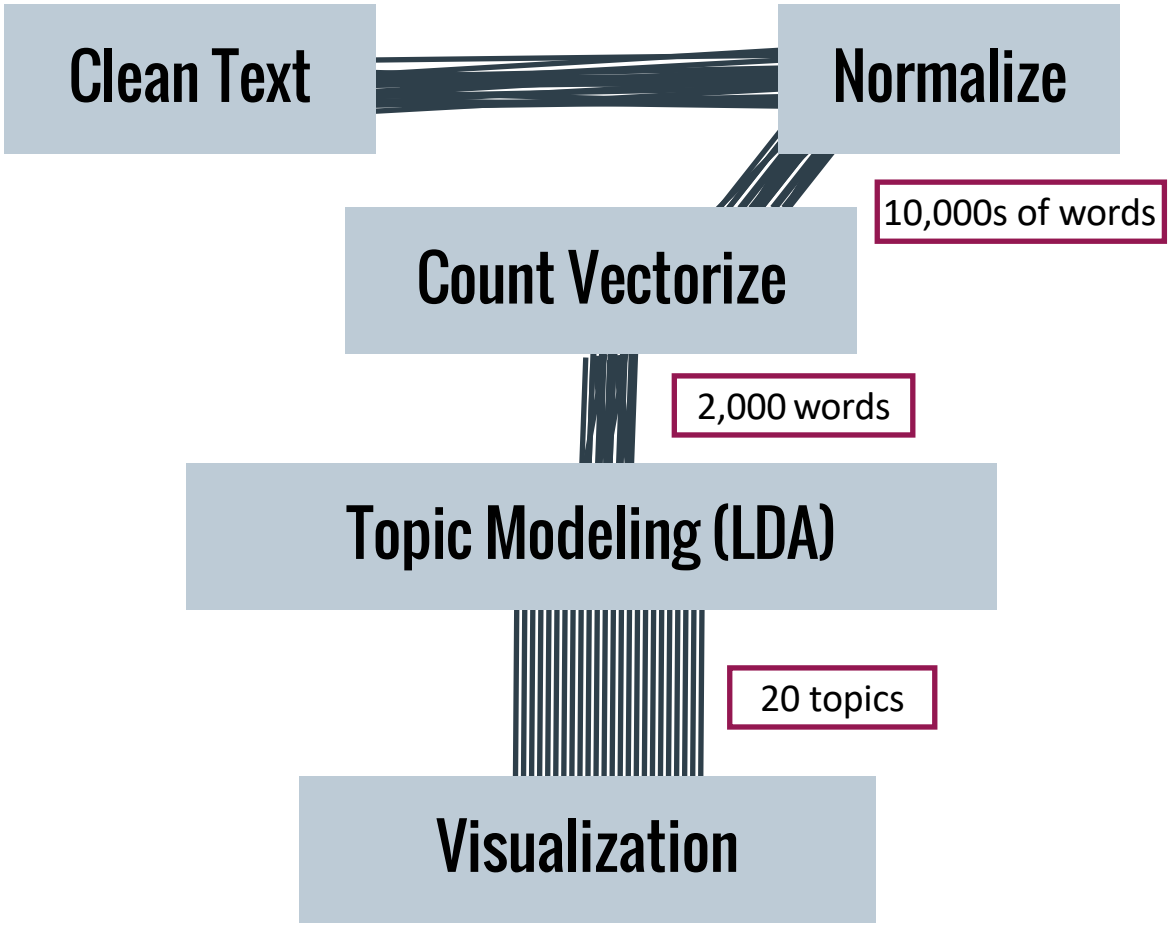
spaCy

spaCy
Text Processing / Entity Extraction

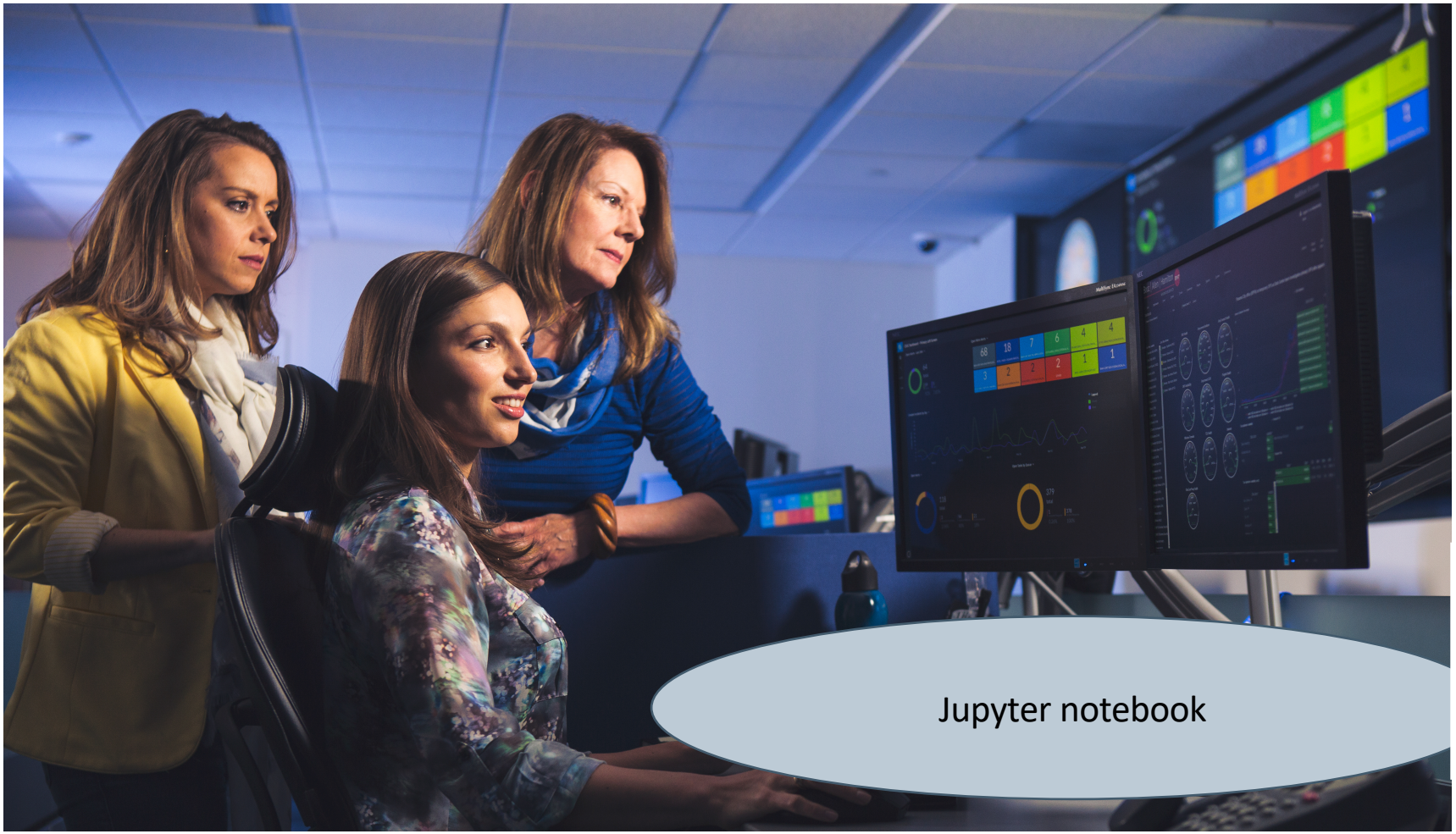
gensim
topic modelling for humans

gensim
Modeling

NLP WORKFLOW



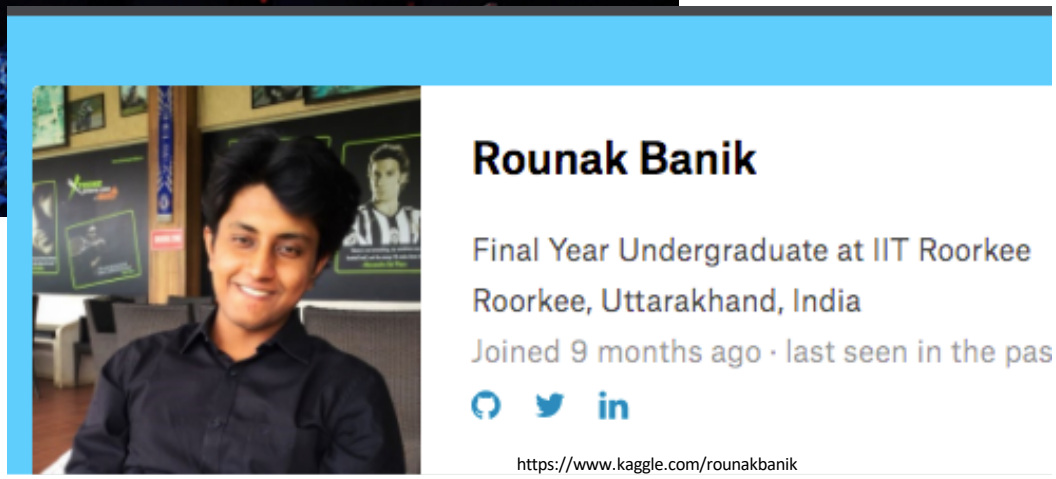
LET'S CODE!



CODING TUTORIAL: TED TALK TRANSCRIPTS

Data:

- 2,467 Transcripts



SAMPLE 1

Looking at the first ~520 characters of the second talk in our corpus:

In:

```
talks[1][:521]
```

Out:

```
Thank you so much, Chris. And it's truly a great honor to have the
opportunity to come to this stage twice; I'm extremely grateful. I
have been blown away by this conference, and I want to thank all of
you for the many nice comments about what I had to say the other
night. And I say that sincerely, partly because (Mock sob) I need
that.(Laughter)Put yourselves in my position.(Laughter)I flew on
Air Force Two for eight years.(Laughter)Now I have to take off my
shoes or boots to get on an airplane!(Laughter)(Applause)…
```

SAMPLE 2

Looking at the first ~520 characters of the fourteenth talk in our corpus:

In:

```
talks[15][:521]
```

Out:

```
(Music)(Music ends)(Applause)(Applause ends)Hi, everyone. I'm Sirena.
I'm 11 years old and from Connecticut. (Audience cheers)(Applause)Well,
I'm not really sure why I'm here.(Laughter)I mean, what does this have
to do with technology, entertainment and design? Well, I count my iPod,
cellphone and computer as technology, but this has nothing to do with
that.So I did a little research on it. Well, this is what I found. Of
course, I hope I can memorize it.(Clears throat)The violin is made of a
wood box and four metal "
```

CLEAN TEXT

What is useful? Remove unwanted items:

- Parentheticals, references
- Non-letters (emojis, symbols)
- Punctuation
- Stopwords
  - Common, but uninformative words (**stopwords**)
  - i.e. and, my, some

http://deckpropowerwash.com/graffiti-removal-2/

NON-SPEECH SOUNDS & EVENTS

Non-word behavior is transcribed. All of the speaker's non-speech sounds, audience sounds, videos, music, etc. are in parentheses. Examples:

```
(Applause)  (Applause ends)  (Pre-recorded applause) (Pre-recorded
applause and cheering)  (Audience cheers)  (Laughter) (Shouting)  (Mock
sob)  (Breathes in) (Baby cooing)  (Video)  (Singing)  (Heroic music)
(Loud music)  (Music)  (Music ends)  (Plays notes)  (Sighs)  (Clears
throat)  (Whispering)
```

Considerations:

- Is it safe to first go and take out everything that is in parentheses before we even tokenize so that we can just look at speech and lyrics?
- It would be interesting to collect these and keep a count in the main matrix, especially for things like 'laughter' or applause or multimedia (present/not present) in making recommendations or calculating the popularity of a talk

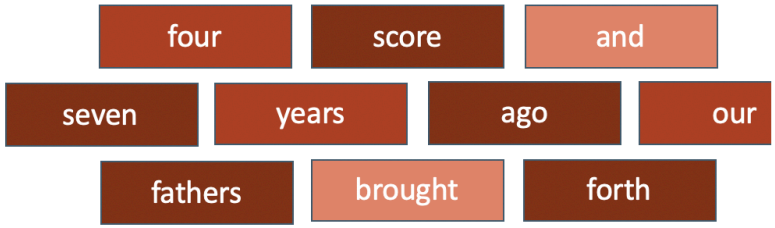
NORMALIZE

Make text less specific by removing:

- Pluralizations, endings (**lemmatization** or stemming to root word)
- Capital letters (**lower**case)

TOKENIZATION

Split text into small indivisible units for processing.



In:

```
import nltk

my_text = And it's truly a great honor to have the opportunity to
come to this stage twice; I'm extremely grateful.

print(nltk.wordpunct_tokenize(my_text))
```

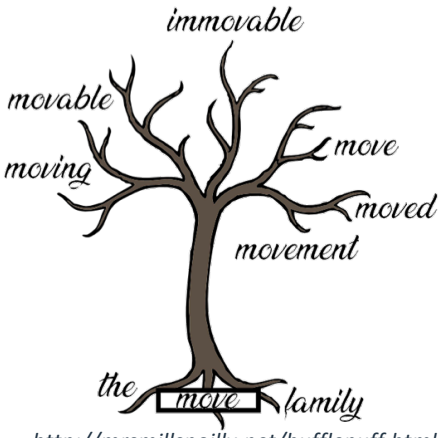
Out:

```
'And', 'it', "'", 's', 'truly', 'a', 'great', 'honor', 'to',
'have', 'the', 'opportunity', 'to', 'come', 'to', 'this', 'stage',
'twice', ';', 'I', "'", 'm', 'extremely', 'grateful', '.'
```

NORMALIZE: LEMMATIZATION

A method for getting the word root using vocabulary and morphological analysis.

- Reduces number of unique words.
- It will replace the ending with the correct letters instead of chopping it off like some of the stemming functions. This still leaves us will a few non-stemmed words.



http://mrsmillspoilly.net/hufflepuff.html

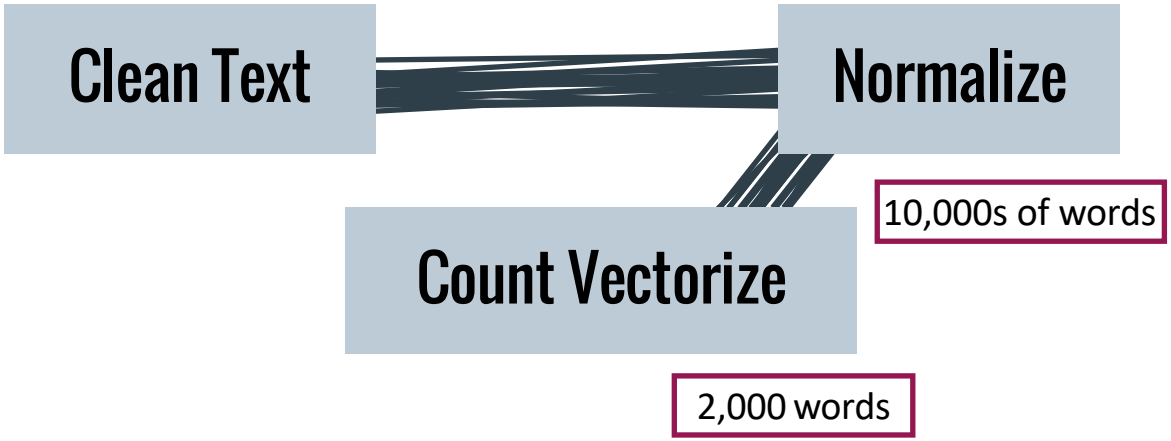
In:

```
import nltk
lemmizer = nltk.WordNetLemmatizer()

my_text = With capabilities educate the children

for word in nltk.wordpunct_tokenize(my_text):

    print(word, lemmizer.lemmatize(word.lower()))
```

Out:

```
With with
capabilities capability
educate educate
the the
children child
```

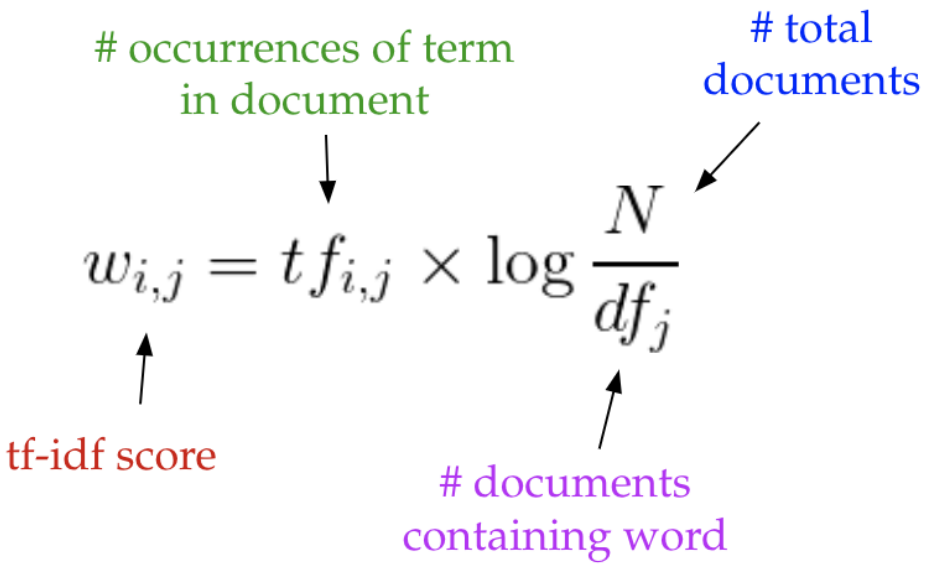
COUNT VECTORIZE



- Begin by turning terms into numbers by creating a document-term matrix.
- Each row is a document, each column is a term, and the data is (count) the number of times each term appears in that document.

	Feel	Like	Waste	Time	never	…
Doc I	1	10	2	5	20	
Doc 2	0	0	1	4	2	

TF-IDF (TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY)



https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

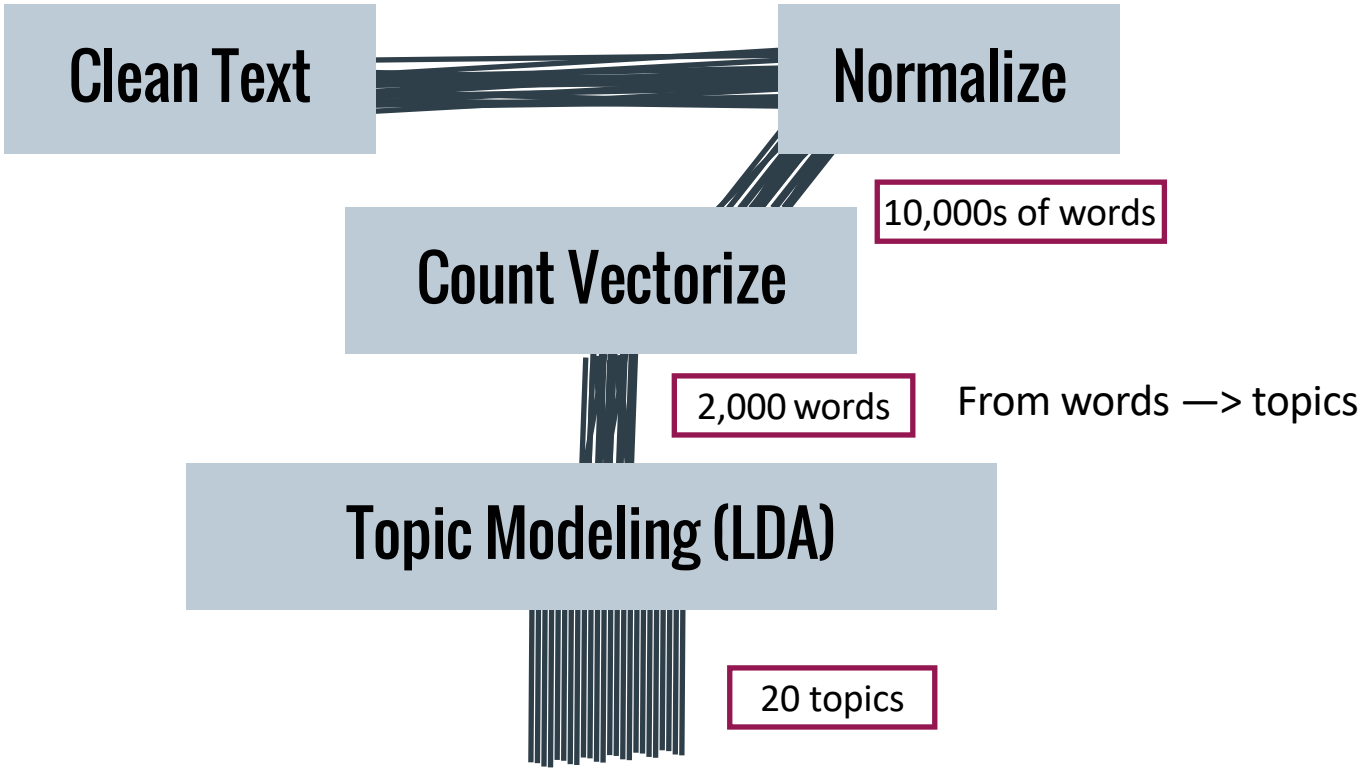
DIMENSIONALITY REDUCTION

Now, we want to take our sparse document-term matrix and perform dimensionality reduction so that we can find topics (collections of words) that describe the relationship between documents.

abandoned	abandoned dog	abandoned dog brought	abandoned puppy	abandoned puppy rain	abide	abide zoning
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	0.026146	0.013073	0.013073	0.013073	0.013073	0.000000
3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

- Dimensionality reduction
  - Reduces noise
- From words ——> topics
- Strength of topics for a talk
  - Max = talk topic

TOPIC MODELING



War/Politics: War government political power law
Climate: Energy water climate oil carbon fuel
Education: Kid school music play student teacher

HOW TOPIC MODELING WORKS

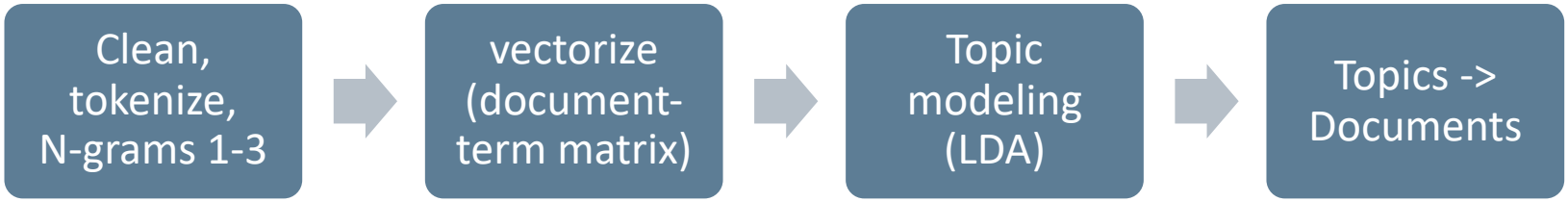
A topic is basically a probability distribution over all possible words.

- Some words are more likely to appear in a particular topic than others.

pyLDAvis

Topic = Dictionary of broad concept

	Topic 1	Topic 2	Topic 3
aortic	.02	.89	0.0302



	dialysis	kidney	aortic	valve	fever	…
Doc I	1	10	2	5	20	
Doc 2	0	0	1	4	2	

	Topic 1	Topic 2	Topic 3
Doc I	.001	.660	0.0302

Assign Max Topic

tSNE

HOW TOPIC MODELING WORKS (CONT.)

Generative process for each document.

1. Randomly assign every term to a topic
2. Term-by-term, assume all other terms are correctly identified (regarding topic) and assign term to topic that is most likely (each term gets a score for each topic)

For a given topic, what is the likelihood that it is present in this document?

$$P_{(Z|W,D)} = \frac{(\#\ of\ terms\ in\ Z)+\beta_u}{(\#\ of\ terms\ in\ Z) for\ all\ D*\beta} (\#\ of\ terms\ in\ Z\ for\ D + \alpha)$$

topics

Z = topic
W = term
D = document
β = smoothing
$\propto$ = learning rate

Get a percentage breakdown of the amount of each topic that is present in a document. (topic proportion)

LATENT DIRICHLET ALLOCATION (LDA)

Allows us to convert from **terms to topics**.

LDA creates a latent (topic) space where documents that contain the same topic will be closer

- b/c documents from the same topic tend to share the same words.

Clustering can be used to find **similar** documents

Latent space = less dimensions than term space

Assume some topic distribution in the corpus.

Look at corpus and find the topic and word distributions that would most likely generate this data.

A Bayesian process needs a prior, this model uses Dirichlet for the probability distribution (over all possible words).

The sparse Dirichlet distribution assumes each topic will only be made from a small subset of the total available space.

Dirichlet distribution performs better than other choices.

LDA: CODE

INPUT:

Number of topics

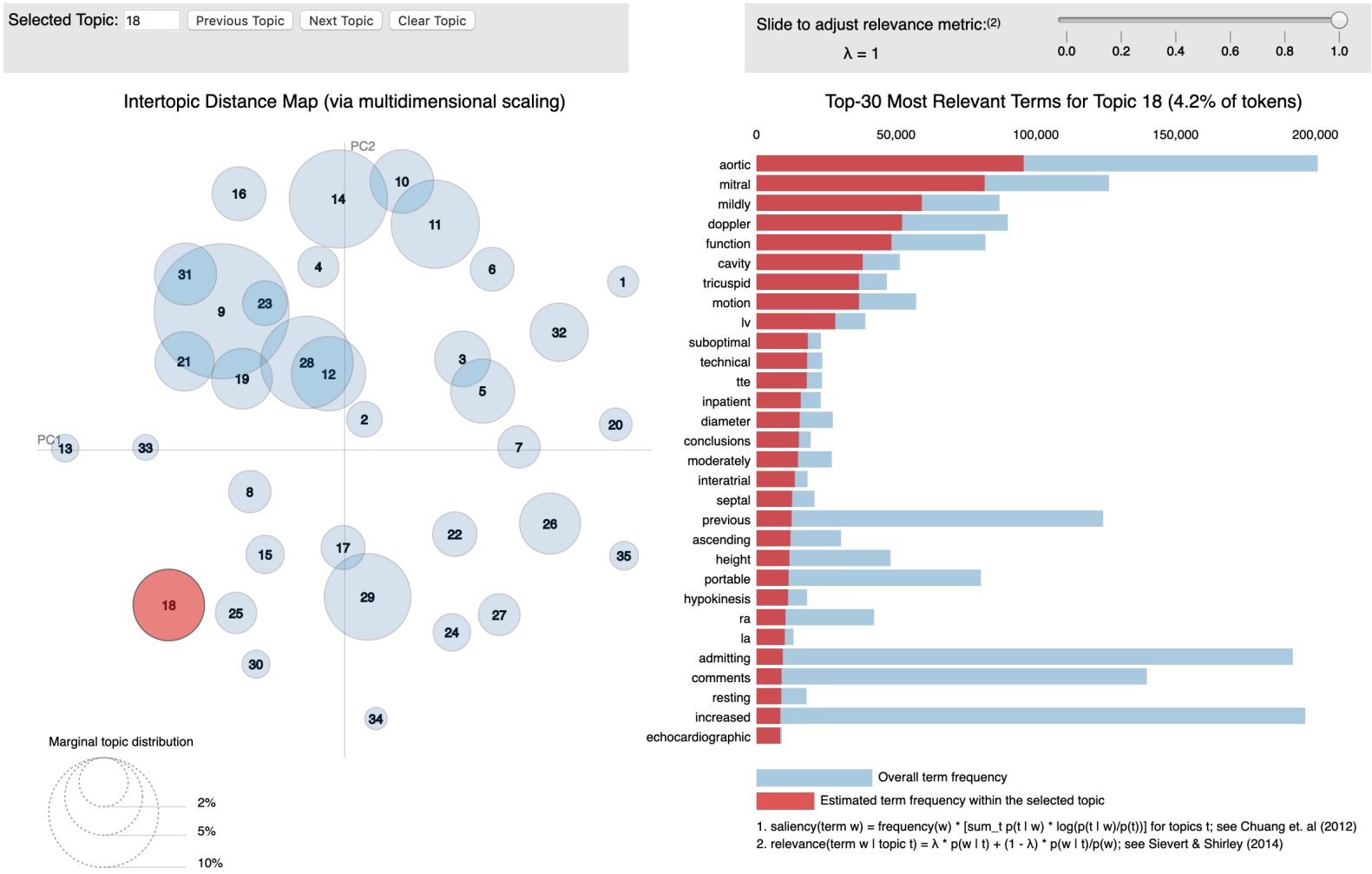
```
lda = LatentDirichletAllocation(n_components=10, max_iter=100)

lda_dat = lda.fit_transform(vectorized_corpus)

Print(lda_dat[0])
```

Show topic proportions for document 0

OUTPUT:

```
0.00131606, 0.00131646, 0.00131599, 0.10262705, 0.001316 , 0.0013159 ,
0.88684465, 0.00131593, 0.00131609, 0.00131587
```

Document is 88% topic 7


Selected Topic: 18
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 18 (4.2% of tokens)
0
50,000
100,000
150,000
200,000
aortic
mitral
mildly
doppler
function
cavity
tricuspid
motion
lv
suboptimal
technical
tte
inpatient
diameter
conclusions
moderately
interatrial
septal
previous
ascending
height
portable
hypokinesis
ra
la
admitting
comments
resting
increased
echocardiographic
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

THANK YOU

rankin_summer@bah.com

www.SummerRankin.com

dowdy_katherine@bah.com

TOPICS

- Art design sort image piece film
- Car hour road mile drive vehicle
- War government political power law
- DNA gene specie animal food plant
- Internet data information phone
- Bird bee flu iran expert early guy
- Patient health care doctor medical
- Social china society India growth
- Computer technology robot machine
- Energy water climate oil carbon fuel

- Visual Art
- Driving
- War
- DNA Biology
- Internet
- Junk
- Health care
- Developing countries' economy
- Artificial intelligence/ computing
- Climate change

- Education
- Economics, stock market
- Marine biology
- Microbiology of disease
- Family
- Earth science/ origin of universe
- Architecture
- Food security
- neuroscience/ consciousness

TOPICS

- Art design sort image piece film
- Car hour road mile drive vehicle
- War government political power law
- DNA gene specie animal food plant
- Internet data information phone
- Bird bee flu iran expert early guy
- Patient health care doctor medical
- Social china society India growth
- Computer technology robot machine
- Energy water climate oil carbon fuel

- Visual Art
- Driving
- War
- DNA Biology
- Internet
- Junk
- Health care
- Developing countries' economy
- Artificial intelligence/ computing
- Climate change

TOPICS

- Kid school music play student teacher
- Dollar company money business cost
- Water ocean fish animal sea ice
- Cell cancer disease drug body blood
- Woman man family girl love man child
- Earth universe planet space light star
- City building space community project
- Africa child food community family poor
- Love experience self god fear moment
- Brain body neuron memory mind
- Education
- Economics, stock market
- Marine biology
- Microbiology of disease
- Family
- Earth science/ origin of universe
- Architecture
- Food security
- Writing? junk?
- neuroscience/ consciousness

BI-GRAMS

year ago	2074	can not 877
little bit	1607	first time 751
year old	1365	every day 692
united states	1103	many people 656
one thing	1041	last year 604
around world	937	every single 573
thank much	931	one day 559
new york	894	10 year 541