Unsupervised Learning is Cool!
But how can we use this in our projects?

1.

2. Kickstarter
But first, let’s look at our dataset...
Data Dimensionality!

Cigarettes Consumed Per Day vs. Probability of Lung Cancer Developing
Data Dimensionality!

Cigarettes Consumed Per Day

Probability of Lung Cancer Developing
Data Dimensionality!

Cigarettes Consumed Per Day

Probability of Lung Cancer Developing
Data Dimensionality!

Probability of Lung Cancer Developing

Cigarettes Consumed Per Day
Data Dimensionality!

Cigarettes Consumed Per Day

Probability of Lung Cancer Developing
Data Dimensionality!

Cigarettes Consumed Per Day

Probability of Lung Cancer Developing

Difference from mean height?
Data Dimensionality!

Cigarettes Consumed Per Day

Probability of Lung Cancer Developing

1 Dimension!
Words and documents are the same way...
Words and documents are the same way...

\[
\begin{align*}
|V| & \quad X_{\text{tfidf}} \\
|D| & \\
\text{Pineapple} & \end{align*}
\]
Words and documents are the same way...

$X_{\text{tfidf}}$
Words and documents are the same way...

\[ \mathbf{X}_{\text{tfidf}} \]

\(|V|\)

\(|D|\)

Pineapple
Words and documents are the same way...

Pineapple

$|\mathbf{V}|$

$\bf{X}_{tfidf}$

$|\mathbf{D}|$

$\mathbb{R} | \mathbf{D} |$

(but, really -- a low dimensional subspace...)}
Words and documents are the same way...

|D|

|V|

$X_{tfd}$

“Pineapples were recalled..."
Words and documents are the same way...

$|D| \times_{tfidf} |V|$

"Pineapples were recalled..."
Words and documents are the same way...

X

$X_{\text{tfidf}}$

|D|

|V|

"Pineapples were recalled..."

(R) (but, really -- a low dimensional subspace...)

(x)
Key questions in unsupervised NLP:
Key questions in unsupervised NLP:

1. How many dimensions does our dataset actually live in?
Key questions in unsupervised NLP:

1. How many dimensions does our dataset actually live in?
2. How do we project our data down to those dimensions?
Key questions in unsupervised NLP:

1. How many dimensions does our dataset actually live in?
2. How do we project our data down to those dimensions?
3. Does any of this stuff actually do anything for our projects?
Key tool in Linear Algebra, NLP, Machine Learning, Data Science, Computer Vision, Algorithms, Matrix Computations, Optimization, Statistics...
Key tool in Linear Algebra, NLP, Machine Learning, Data Science, Computer Vision, Algorithms, Matrix Computations, Optimization, Statistics...
Key tool in Linear Algebra, NLP, Machine Learning, Data Science, Computer Vision, Algorithms, Matrix Computations, Optimization, Statistics...

\[
X = U \Sigma V^T
\]
Key tool in Linear Algebra, NLP, Machine Learning, Data Science, Computer Vision, Algorithms, Matrix Computations, Optimization, Statistics...

\[ X = UΣVT \]
$|V| X_{\text{tfidf}} = |D|$
$|D|$

$X_{\text{tfidf}} = |V|$
$|V| \times X_{tfidf} = k \times |V| \times |D|$
$X_{tfidf} = |V|$
$X_{tfidf} = |V|$
$X_{tfidf} = UV^T$
$X = U \Sigma V^T$
Key questions in unsupervised NLP:

1. How many dimensions does our dataset actually live in?

\[ X = U\Sigma V^T \]

2. How do we project our data down to those dimensions?

\[ X = U\Sigma V^T \]
Enough talk, time for some magic.
$\left| \mathcal{D} \right| \left| \mathcal{V} \right| X_{\text{tfidf}} + U \Sigma V^T = $
Latent Semantic Indexing (LSI)

\[ \mathbf{X}_{\text{tfidf}} + U \Sigma V^T = \]
The equation is:

\[ |D| X_{\text{tfidf}} + U \Sigma V^T = \]

Latent Semantic Indexing (LSI) =

Latent Semantic Analysis (LSA)
As a side note...

Great first NLP paper to read! Highly accessible :-)
What are topic models?
What are topic models?

Latent semantic indexing
(Deerwater et al. 1990)
What are topic models?

- Latent semantic indexing (Deerwater et al. 1990)
- Non-negative matrix factorization (Lee and Seung 1999)
What are topic models?

Latent dirichlet allocation (Blei et al. 2003)

Latent semantic indexing (Deerwater et al. 1990)

Non-negative matrix factorization (Lee and Seung 1999)
Why do we care??

$|V|$

$|D|$

$k$

$k$
Why do we care??

|V|   

|D|
Why do we care??

Interpretable, small number of features for text classification!
Why do we care??

- Interpretable, small number of features for text classification!
Document length *matters a lot*
Document length *matters a lot*

**Different regimes of supervised NLP**

(Jack’s opinions only! Lots of caveats!)
Document length *matters a lot*

**Different regimes of supervised NLP**

(Jack’s opinions only! Lots of caveats!)

- Less words
- 50-100 Words
- More words
Document length *matters a lot*

**Different regimes of supervised NLP**
(Jack’s opinions only! Lots of caveats!)

![Diagram showing the relationship between document length and topic model performance.](image-url)

- **50-100 Words**
  - Topic models fail
  - Topic models work

- **Less words**
- **More words**
Document length *matters a lot*

**Different regimes of supervised NLP**
(Jack’s opinions only! Lots of caveats!)

- Topic models fail
- Topic models work

- Less words
- 50-100 Words
- More words

Naive bayes, n-gram features + linear classifier are almost always pretty good in practice
:-)
So can we see if our Kickstarter will be successful?