Disaster Awareness Notifications

Sarah Berner
Kenneth Lee
True McKee
Anna Pavlova

Background

- At any given time there are a lot of incidents, reported by a wide range of sources, that need to be monitored. Additionally, these incidents have a wide range of impact - from relatively routine to disastrous. To monitor this manually is unrealistic

- Relevance is important. A Google search will not necessarily return the appropriate context or immediacy required for the purpose of dispersing resources. Example: Search “flood disasters today.”
  - Are Recent Flood Disasters the Result of Climate Change
  - Flooding + Natural disasters and extreme weather | Environment

Objectives

1. Create a model that is able to classify articles from a wide variety of sources as relevant to an impending or currently occurring disaster

2. Integrate that model into a function - eventually a background-running application/notification system - that alerts the appropriate disaster relief employee(s) to the disaster and its location in order to begin the process of providing disaster relief services to the affected area

Data Collection and Cleaning

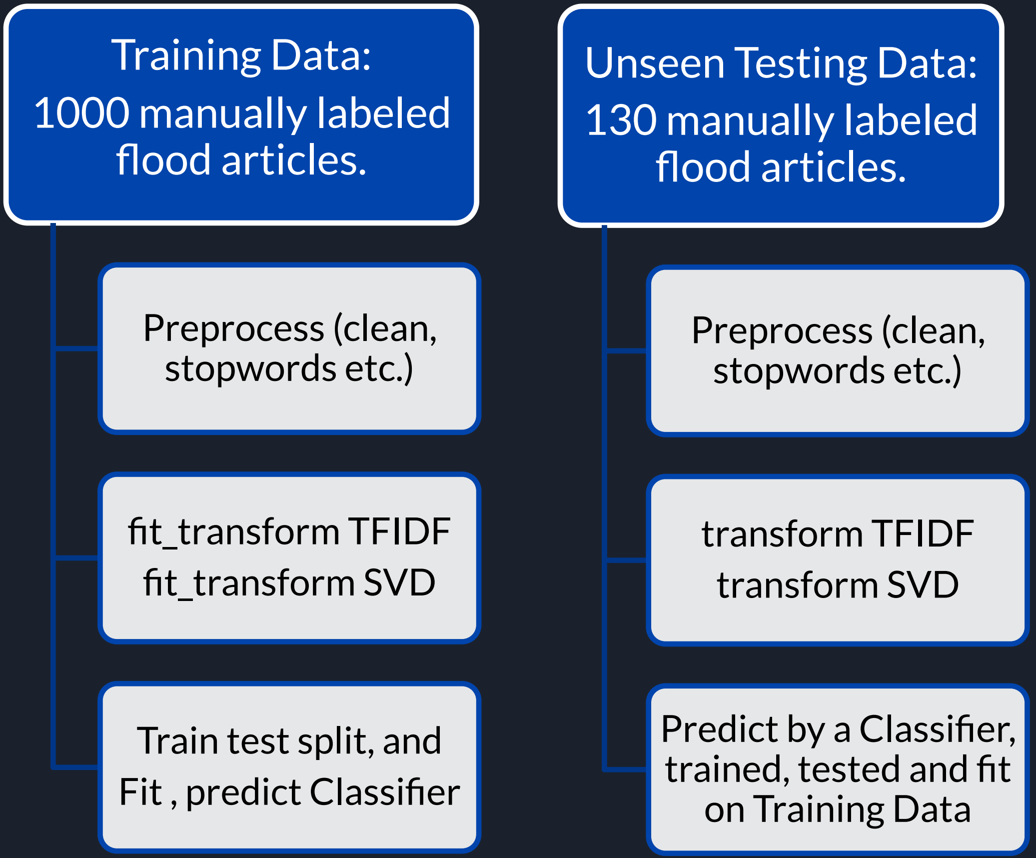
- NewsAPI
    - Thirteen (primarily US-based) sources
    - Articles returned searching for “flood”
    - 6 months back
    - All articles (as opposed to “Top Headlines”)

- Read in articles to a dataframe and proceeded to manually classify to flood_disaster_relevant or not

Data Collection and Cleaning

content	description	publishedAt	title	source_name	flood_relevance
Chat with us in Facebook Messenger. Find out w...	A pair of environmental reports reveal the wor...	2019-01-16T22:52:24Z	Melting ice could flood Brooklyn Bridge	CNN	0
The amount the ground can soak up is limited, ...	Flooding in the town of Hamburg, Iowa, on Marc...	2019-04-08T20:58:04Z	Powerful Storm Threatens More Misery in Flood-...	The New York Times	1
Many of the works in Programmed: Rules, Codes ...	The artist's monumental video wall, featuring ...	2019-04-04T17:45:31Z	Last Chance: Nam June Paik at the Whitney: A W...	The New York Times	0
This means that electric utilities, in particu...	A firefighter checked out burned vehicles and ...	2019-01-29T19:09:38Z	The Very High Costs of Climate Risk	The New York Times	0
A major storm is now moving out of the Northwe...	A tornado was observed in Port Orchard, Washin...	2018-12-19T12:20:27Z	Storm that spawned tornado in Washington now m...	ABC News	1

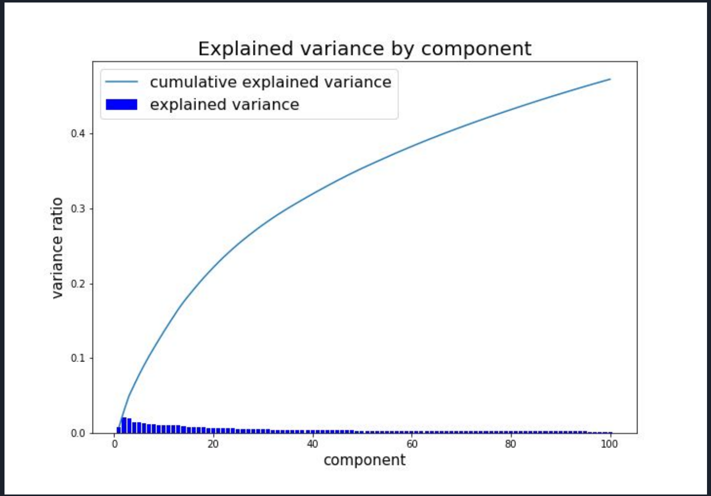
Data Transformation

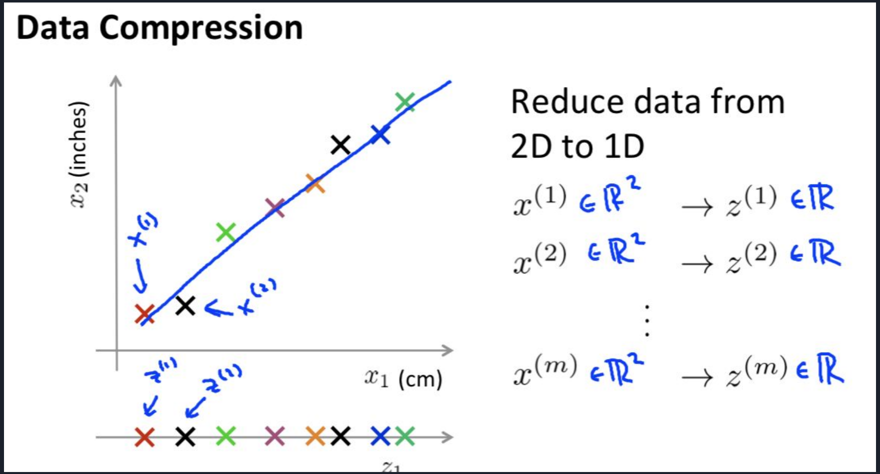


SVD Data Compression

- Reduce computing power by reducing data dimensionality from n to k

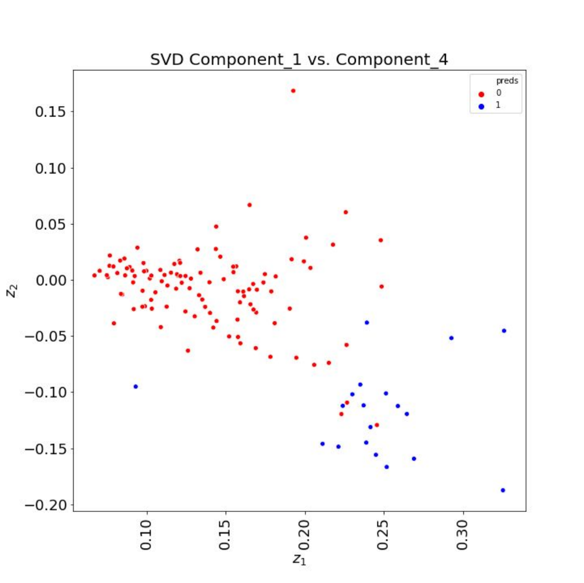
- Choose number principal components k = 100





Ref: Coursera “Machine Learning”, Andrew Ng.

SVD Visualization



Component 1 weights

rain	0.178262
snow	0.162953
storm	0.161420
flooding	0.127096
weather	0.126887
inches	0.118378
people	0.110871
river	0.101504
heavy	0.100614
new	0.097735
trump	0.097261
water	0.095494
california	0.094935
flood	0.087606
nebraska	0.083761
thursday	0.083737
morning	0.082803
areas	0.081431
year	0.081242
friday	0.079965

Component 4 weights

woman	-0.054716
police	-0.060450
murder	-0.036421
year old	-0.056000
teen	-0.038144
crash	-0.036373
california	0.014928
old	-0.059672
death	-0.046452
killing	-0.037224
news headlines	-0.028764
headlines today	-0.028764
man	-0.048334
suspect	-0.023583
headlines	-0.028778
car	-0.025896
mom	-0.032635
arrested	-0.030022
officer	-0.032467
storm	0.173950

Evaluating Model Performance: Training Data

Model	Train Score	Test Score	Precision	Sensitivity
kNN (k = 7)	0.862	0.833	0.816	0.775
Bagged Decision Trees	0.873	0.846	0.853	0.764
Random Forest	0.893	0.854	0.843	0.833

Goal:

- Minimize False Negatives

Conclusion:
Random Forest has best balance of:

- Train/Test Score
- Precision
- Recall

Evaluating Model Performance: Unseen Test Data

- Zero False Negatives
- False Positives may actually be True

'Parts of southern Africa have been left devastated after Cyclone Idai swept through Mozambique, Malawi and Zimbabwe, destroying towns and villages in its path.'

	Predicted	
Random Forest	Negative	Positive
Actual Neg	104	14
Actual Pos	0	9

	Predicted	
kNN	Negative	Positive
Actual Neg	108	10
Actual Pos	0	9

	Predicted	
BDTrees	Negative	Positive
Actual Neg	105	13
Actual Pos	0	9

Limitations

- Limited training data - NewsAPI only goes back 6 months
  - Additionally, a published list of sources was removed mid-project so we were stuck with a specific, non-expandable set of sources
- Indexing local news sources a massive undertaking
- Some disasters are relatively infrequent (e.g. earthquakes, tornados) so not many articles to train on

Next Steps

- Turn function into production application that runs automatically in background of user's workflow

- Expand disaster recognition beyond floods

- Expand source info beyond NewsAPI
    - E.g. Local news sources

Thank you!

Questions? Comments?