Features and Errors

Sameer Singh and Conal Sathi

BANA 290: ADVANCED DATA ANALYTICS

SPRING 2018

April 24, 2018

Today's Overview

- Projects
- Case Study of NLP from Industry
- Features for Text Classification
- In-Class Activity 1
- Error Analysis of Classifiers
- In-Class Activity 2

Today's Overview

- Projects
- Case Study of NLP from Industry
- Features for Text Classification
- In-Class Activity 1
- Error Analysis of Classifiers
- In-Class Activity 2

Datasets: Text Classification

~5.2 million

yelp

10k to 9 million

amazon.com

Or, any other large-scale dataset
(we will provide a long list)

~2.3 million

Quora

~35,000

twitter

Goals: Train Classifiers

Algorithms

- Formulate as classification
- Train at least 3 classifiers (train/dev/test splits)
- Explore various features (more on this today)
- Tune all the hyper-parameters

Evaluation

- Identify the appropriate metrics/plots
- Present effect of hyper-parameters on metrics
- Compare the classifiers with each other
- Try different partitions of the data, e.g. easy vs hard

Analysis

- Compare features of classifiers (more on this today)
- Compare the errors made by the classifiers
- Other visualizations: word clouds, adversaries, etc.

Teams

2 per team

- Most teams should be this size
- Try to pick people with different backgrounds
- Discuss splitting of efforts early

3 per team

- Discouraged! (only if experienced w/ Python & ML)
- Need to do something beyond the requirements
  1. Combine with unsupervised learning
  2. Use sophisticated ML (like deep learning)
  3. Multiple datasets

Upcoming...

Homework

- Homework 2 will be up soon (so will HW1 grades)
- Will have most of the components of the project
- Due in ~two weeks, May 11th

Project

- Proposal: members (final) and dataset (planned)
- If 3 members, describe plan for extra work
- Instructions out soon, due May 15th

Today's Overview

Projects

Case Study of NLP from Industry

Features for Text Classification

In-Class Activity 1

Error Analysis of Classifiers

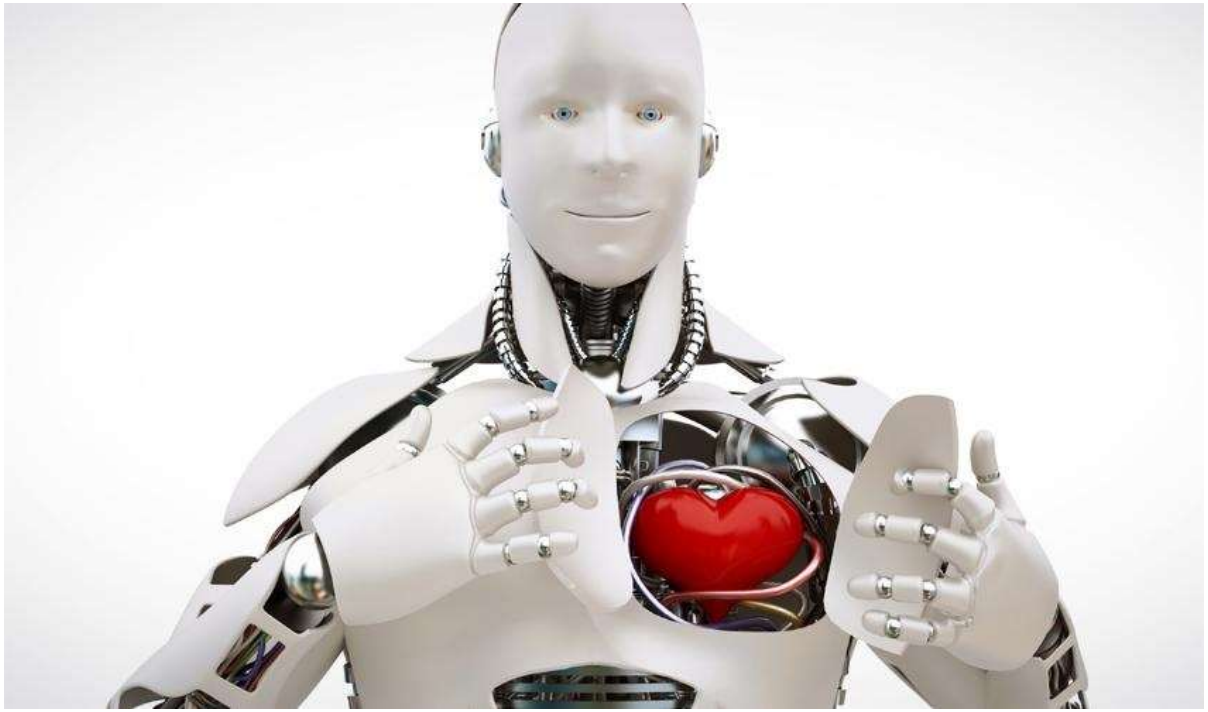
In-Class Activity 2

a human touch in machine learning

Conal Sathi

Data Alchemist at Slice Technologies

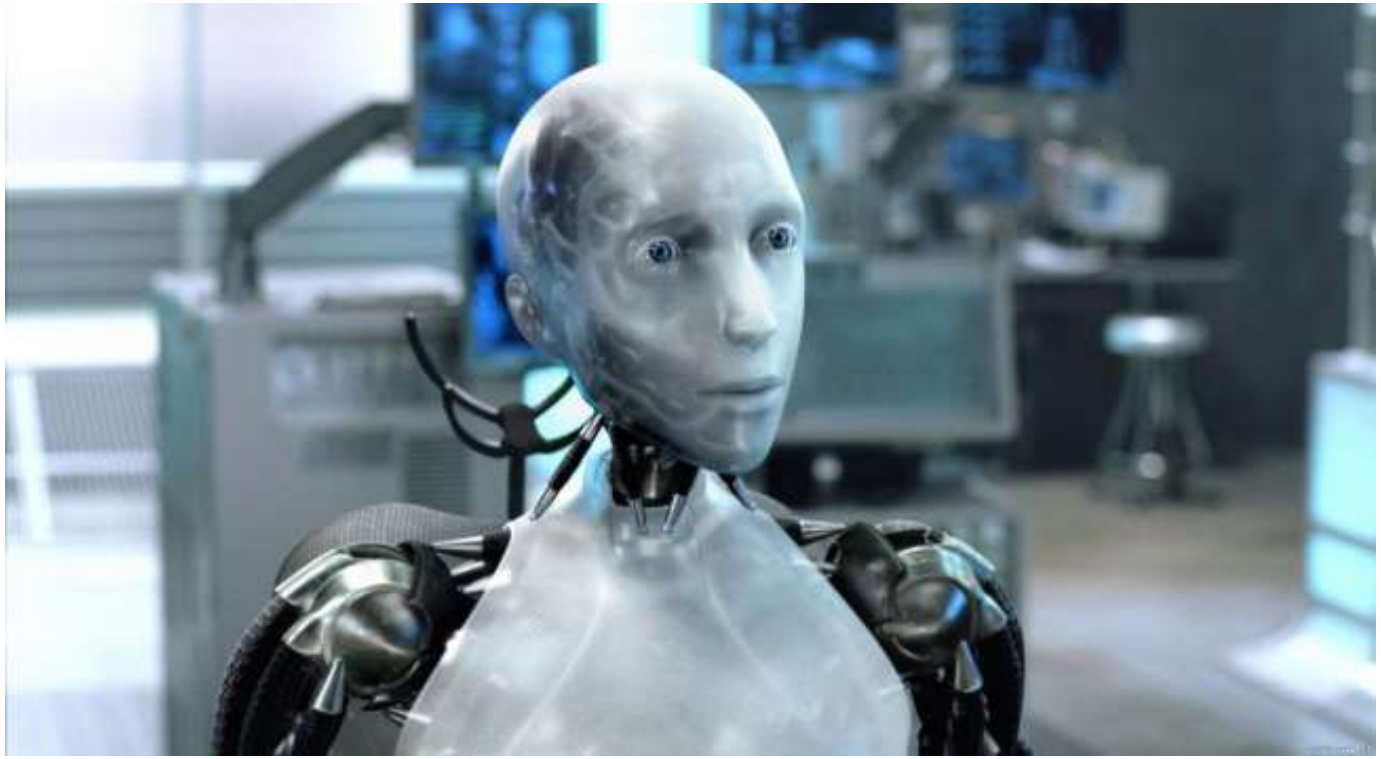
@aDataAlchemist



slice INTELLIGENCE

machine learning

computer systems that can learn from training data and make predictions for new (unlabeled) data



slice INTELLIGENCE

who do you trust?

who do you trust?

can they work together?

slice INTELLIGENCE

Slice: who we are

founded in 2010, Slice has amassed a 4.4-million person panel of US online shoppers to date

- data set: e-receipts in e-mail
- over 1 billion consumer purchases processed
- $100 billion in purchases

slice INTELLIGENCE

Slice: your smart shopping assistant

1. track shipments automatically
2. store shopping receipts
3. save money with price drop alerts
4. stay safe with recall alerts

TODAY CNN Forbes AARP lifehacker

Unroll.me: clean up your inbox

1. unsubscribe with one click

2. combine favorite subscriptions into one email
3. read what you want, when you want

CBS TIME The New York Times WSJ lifehacker

a treasure trove of data sits in e-receipts

the secret sauce to unique purchase insights

slice INTELLIGENCE

item-level consumer purchase data

across all merchants, channels, and time

NIKE

E-COMMERCE

package out for delivery

Nike Kobe X, 9.5, wide,
blue lagoon, $144.97

NORDSTROM

OFFLINE RETAIL

purchased Wed, March 9, 2016

cashmere blend scarf, $98.00

airbnb

TRAVEL

Los Angeles, CA

check in: Fri, April 15, 2016

check out: Sun, April 17, 2016

CIRQUE DU SOLEIL®

TICKETS

reservation for Varekai

Sat, March 12, 2016 @7:15pm

Square

PAYMENTS

payment confirmation

large non-fat mocha to La Stacione Cafe

slice INTELLIGENCE

machine learning for product semantics

Panasonic 35 - 100mm f / 2.8 X OIS

Samsung Galaxy S4 SPH - 16GB - Purple

slice INTELLIGENCE

machine learning for product semantics

Panasonic 35 - 100mm f / 2.8 X OIS

electronics & accessories
cameras & photo
camera lenses

brand: Panasonic
focal length: 35 - 100mm
aperture: 2.8

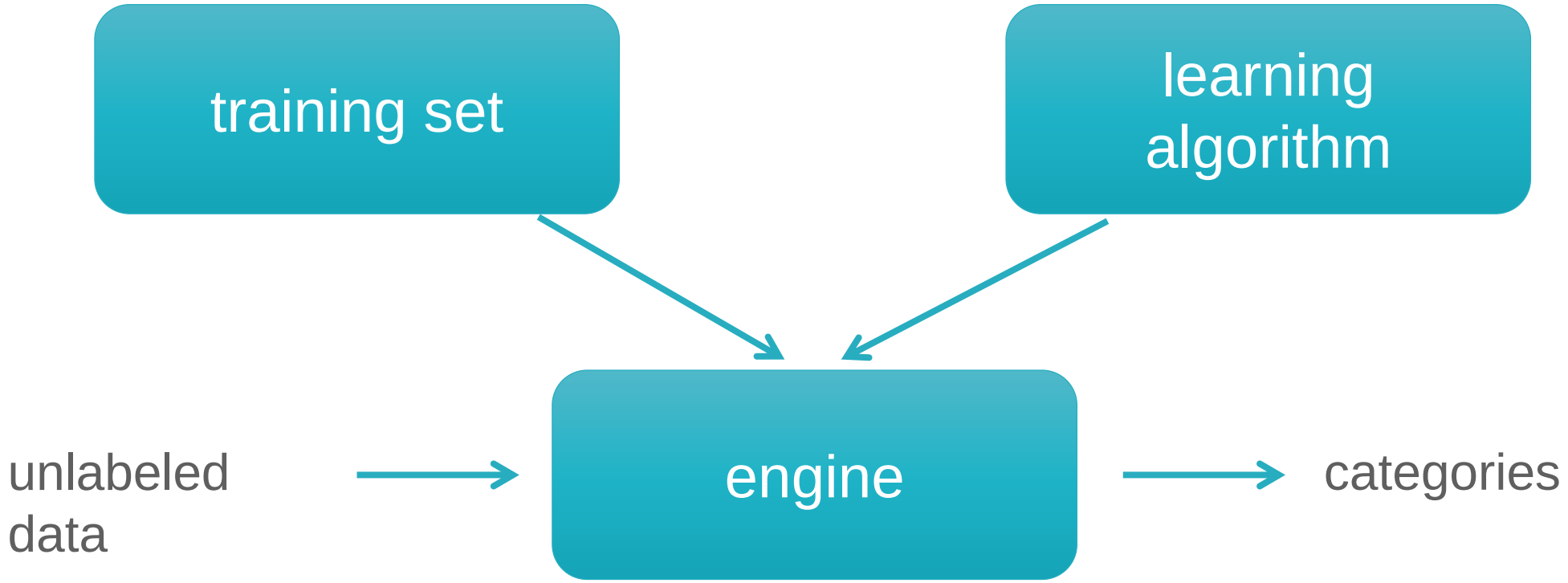
Samsung Galaxy S4 SPH - 16GB - Purple

electronics & accessories
telephones & mobiles
mobile phones

brand: Samsung
sub-brand: Galaxy
model: S4
memory: 16GB
color: purple

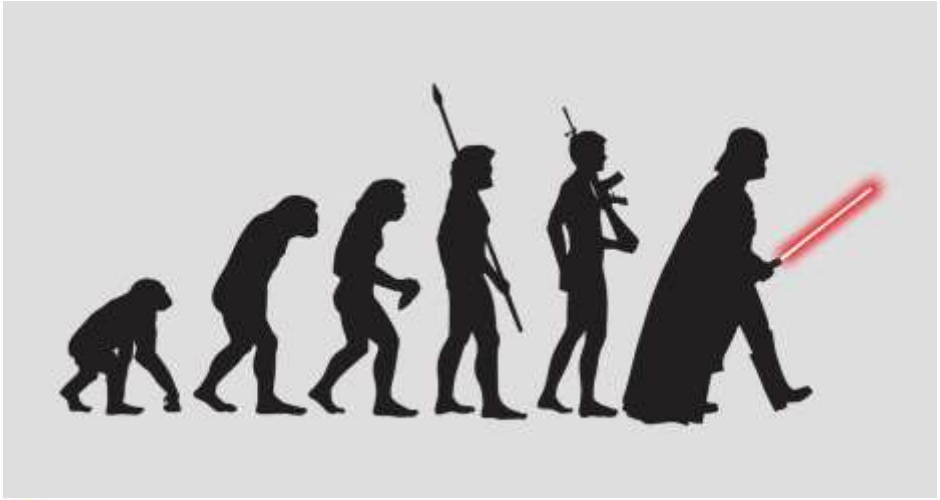
slice INTELLIGENCE

traditional approach to categorization



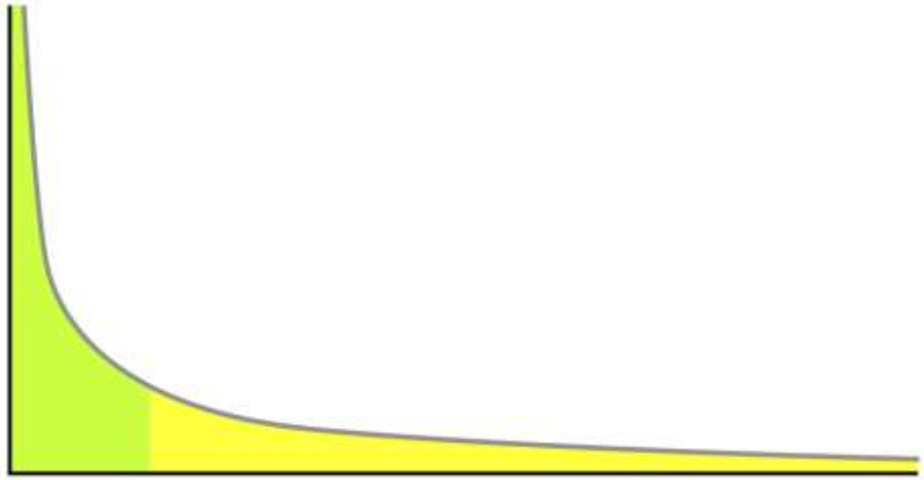
issues

need to continuously evolve



long tail of edge cases

- continuous improvement difficult

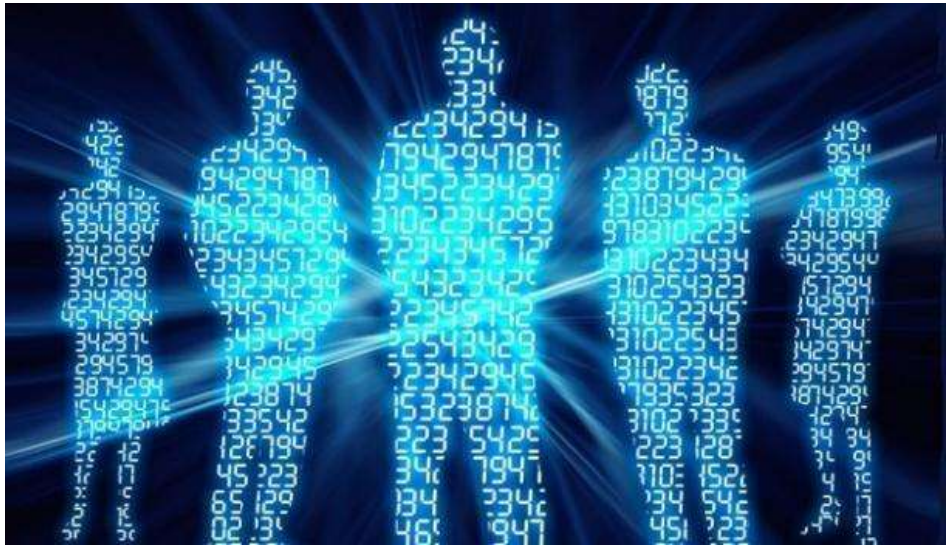


slice INTELLIGENCE

so what do we do?

use machines and humans together to create a continuously improving system

to do this, we need to scale humans to big data!



slice INTELLIGENCE

what do humans do?

humans can directly improve the engine by:

- adding training data
- creating rules

why rules?

- can quickly bring up the accuracy (and in a deterministic fashion)
- helpful when training data is sparse/not representative for a particular category
- effective and easy way to handle corner cases

slice INTELLIGENCE

types of workers

types of humans involved in our system:

- machine learning engineers
- analysts (domain experience)

add more human labor with:

- crowdsourced workers (very elastic supply, micro-tasks)
- outsourced workers (less elastic, more involved tasks)

slice INTELLIGENCE

crowdsourcing

get data/opinions/services from a large online community

Amazon mechanical turk

- initially invented in Amazon for internal data cleaning

employs over 500,000 people worldwide

can create tasks on demand

slice INTELLIGENCE

The goal is to categorize these 10 products that are typically bought online.

When you select a category from the drop down, the guidelines (if there are any for the chosen category) will appear below the drop downs.

Additional Rules:

- The product descriptions provided in the questions below are not adequate by themselves for determining the correct category. As a result we provide a link associated with each description to help you in the category selection process.
- **Clicking the link and reviewing the returned results to confirm your category selection is a required part of this task.** If you fail to complete this step, no matter how obvious the provided description is, you have not performed the task and you will not be paid.
- Randomly selecting a category is equivalent to not performing the task. We send each question to multiple people for comparison. We also compare against our internal results. We can easily detect users who are randomly selecting answers instead of carefully researching the correct categorization. **In such cases you will not be paid.**

Please note that not following the above rules may disqualify you from performing subsequent categorization tasks.

Question #1

Name of Product: Newclew Breathe Youre Home removable Vinyl Wall Decal Home Decor Large

Merchant: Amazon

Price: $9.00

Link to Search Results: **Click here for more information on the product**

Which category does the product best fit in?

Home & Kitchen

Choose a Category

Includes: furniture, small kitchen appliances, bakeware, glassware, serveware, laundry, cleaning supplies, décor, vacuums

Excludes: building supplies, harware, fixtures, light bulbs, food/drinks, big appliances, arts & crafts

outsourcing

the downside of crowdsourcing is that you have to give micro-tasks (not very involved tasks) and you do not work closely with the workers

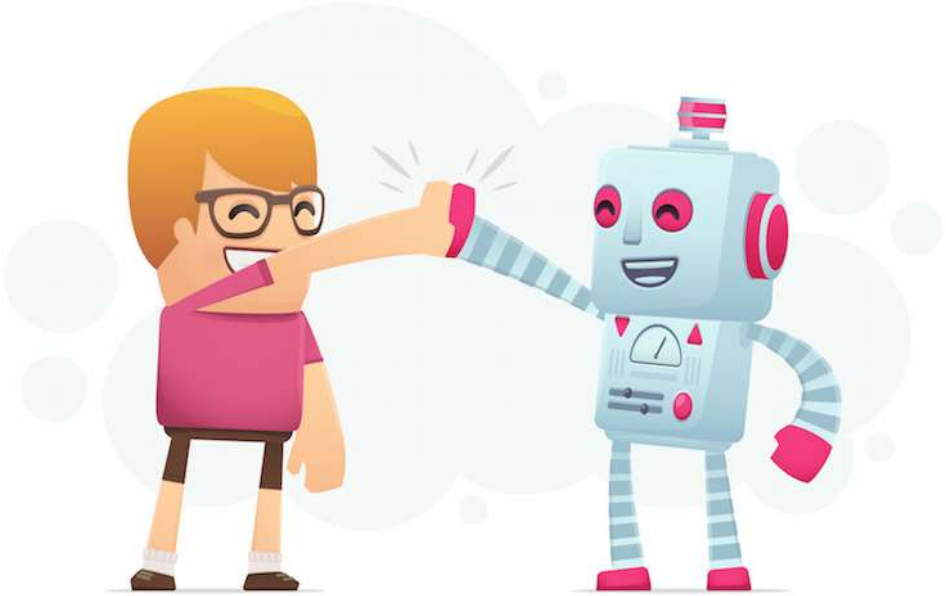
this is why we use outsourcing as well

- can give much more involved tasks (allow them to filter/search to find misclassifications)
- we can work closely with them to teach them the taxonomy and how to use our tools effectively
- they have much more practice on these tasks, so they can be very productive and eventually become domain experts

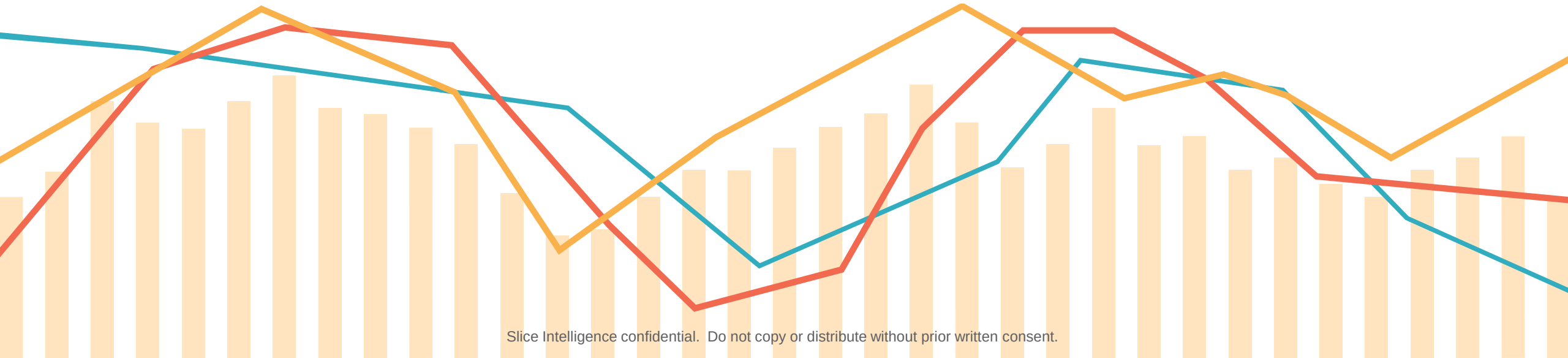
in conclusion….

with our system, we have created:

- continuous monitoring of the quality of our engine
- mechanism to quickly improve quality when issue is detected
- created large training data set (16 million items)
- extensive rules (12K)



slice INTELLIGENCE



questions?

visit sliceintelligence.com or email conal@slice.com

or follow us on Twitter: @aDataAlchemist | @sliceintel

Today's Overview

Projects

Case Study of NLP from Industry

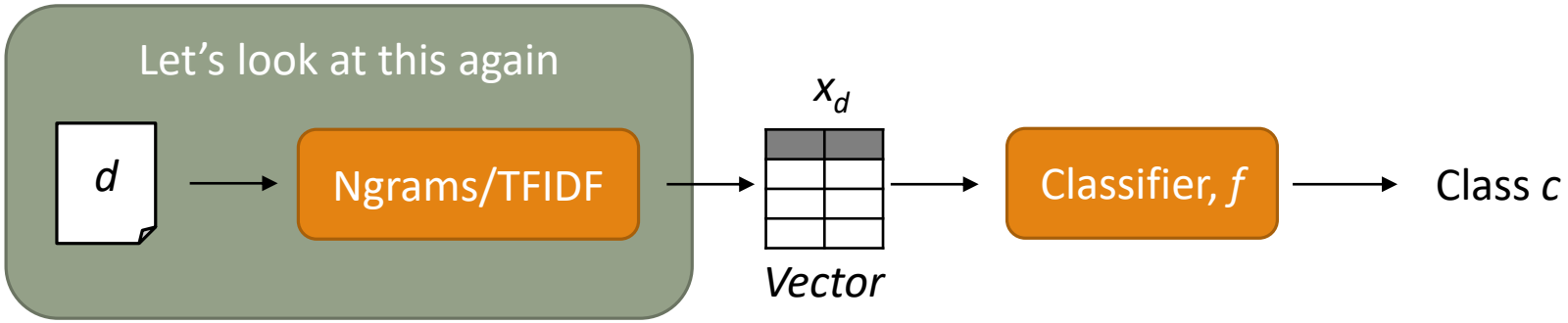
Features for Text Classification

In-Class Activity 1

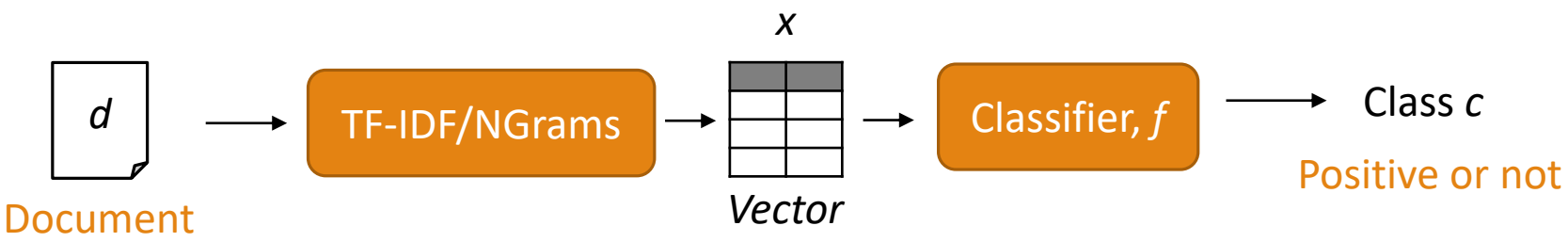
Error Analysis of Classifiers

In-Class Activity 2

Text Classification

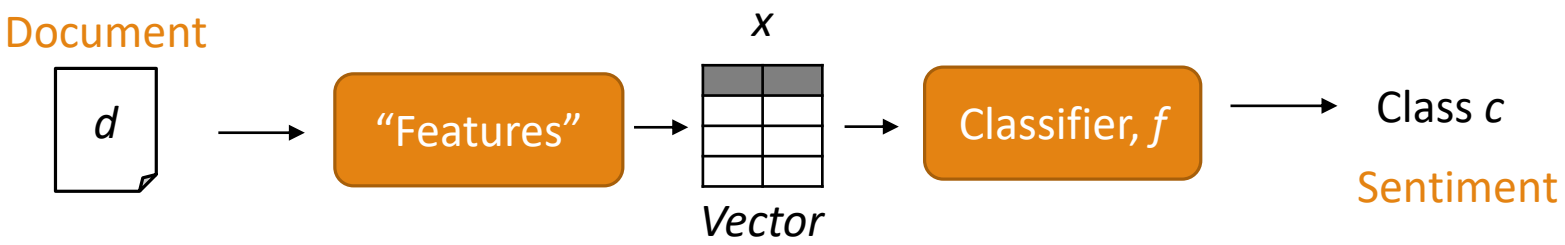


Sentiment Analysis



What about Lexicons? Regular Expressions? Their combinations?

Features Beyond Ngrams




```
def features(d):
    # implement code…
    return vector
```

Arbitrary vector that has everything you need!

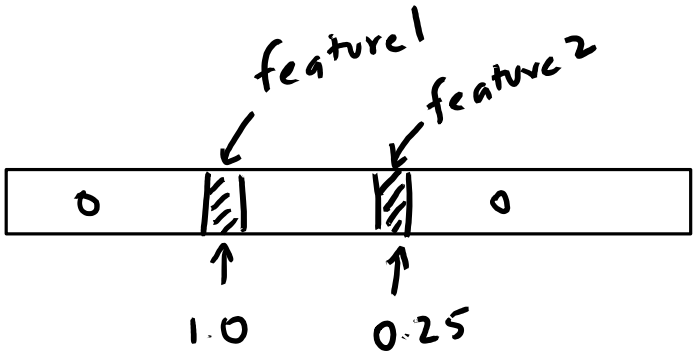
Feature Dictionary/Vector

```
def features(d):
    # implement code…
    return vector
```

Producing a vector is difficult:
- Usually, very sparse
- Indexing is difficult

Instead, return a dictionary!

```
{
  “feature1”: 1.0,
  “feature2”: 0.25,
}
```



Ngrams+TFIDF

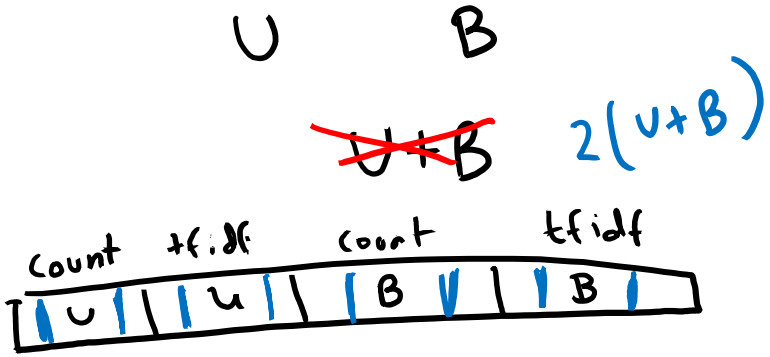
	Counts	TFIDF
Unigrams	{ "good": 1.0, "movie": 1.0, }	{ "good": 2.0, "movie": 0.1, }
Ngrams	{ "good movie": 1.0, "a good": 1.0, }	{ "good movie": 3.0, "a good": 1.0, }

Ngrams+TFIDF

Make sure features don't overlap by naming them

```
{
  "uni_count=good": 1.0,
  "uni_count=movie": 1.0,
  "uni_tfidf=good": 2.0,
  "uni_tfidf=movie": 0.1,
  "ng2_count=good movie": 1.0,
  "ng2_count=a good": 1.0,
  "ng2_tfidf=good movie": 3.0,
  "ng2_tfidf=a good": 1.0,
}
```

The vector will have unique position to each name



What about the lexicons?

```
{
  “inq_pos_count”: 5.0,
  “inq_neg_count”: 2.0,
  “inq_pos_more”: 1.0,
  ...
  “uni_count=movie”: 1.0,
  “uni_tfidf=good”: 2.0,
  “uni_tfidf=movie”: 0.1,
  “ng2_count=good movie”: 1.0,
  “ng2_count=a good”: 1.0,
  “ng2_tfidf=good movie”: 3.0,
  “ng2_tfidf=a good”: 1.0,
}
```

inq_pos_more than 5 : 0.0

Use lexicons to define new features!

Think of them as providing all the information you think the rules might need, but let ML do the rest.

2(U+B)+3

Any interesting information!

```
{
  “num_exclamations”: 3.0,
  “num_capitalized”: 2.0,
  “rev_length”: 25.0,
  ...
  “inq_pos_count”: 5.0,
  “inq_neg_count”: 2.0,
  “inq_pos_more”: 1.0,
  ...
  “uni_count=movie”: 1.0,
  “uni_tfidf=good”: 2.0,
  “uni_tfidf=movie”: 0.1,
  ...
}
```

`num_exclamations=1' : 0

`num.excla=2' : 0

`num.exclam=3' : 1

`num.exclam>3' : 0

Arbitrary vector that has everything you need!

Think of them as providing all the information you think the rules might need, but let ML do the rest.

Spam Filtering: SpamAssassin

- millions of (dollar), (dollar) NN,NNN,NNN.NN
- Phrases: impress .* girl
- One hundred percent guaranteed
- Generic Viagra
- Online Pharmacy
- From address: starts with numbers
- Subject is all capitals
- Claims you can be removed from the list
- 'Prestigious Non-Accredited Universities‘
- Non-Text: HTML has a low ratio of text to image area

Back to the definition

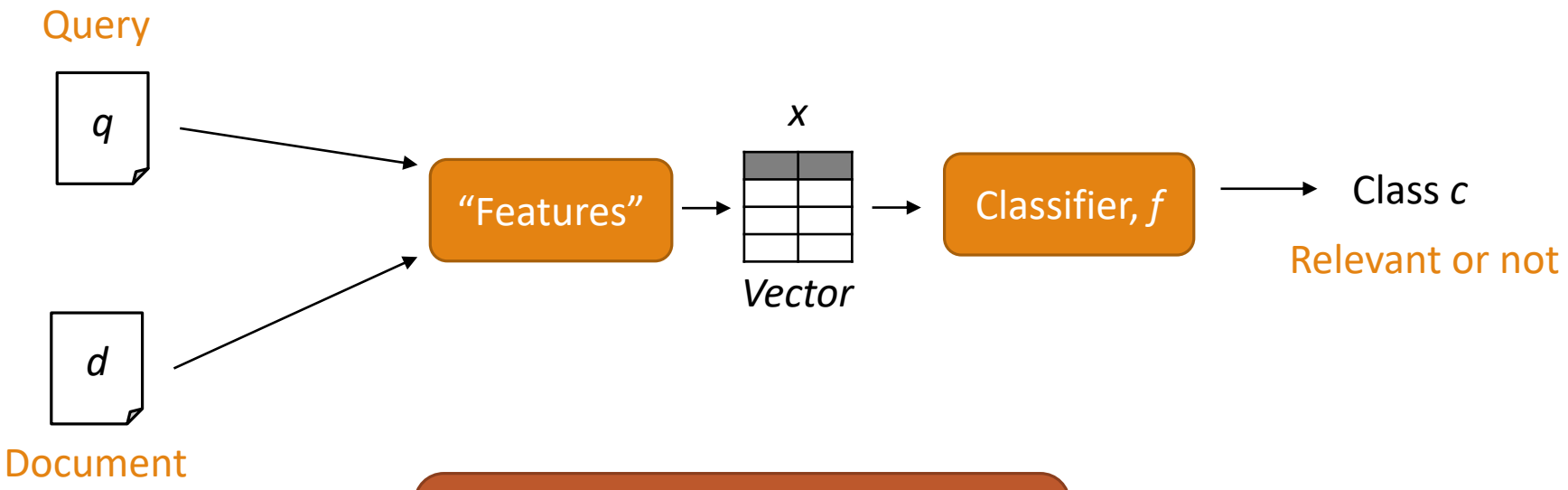
Input:

- a document d
- a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

Output: a predicted class $c \in C$

This is actually "Document Classification"

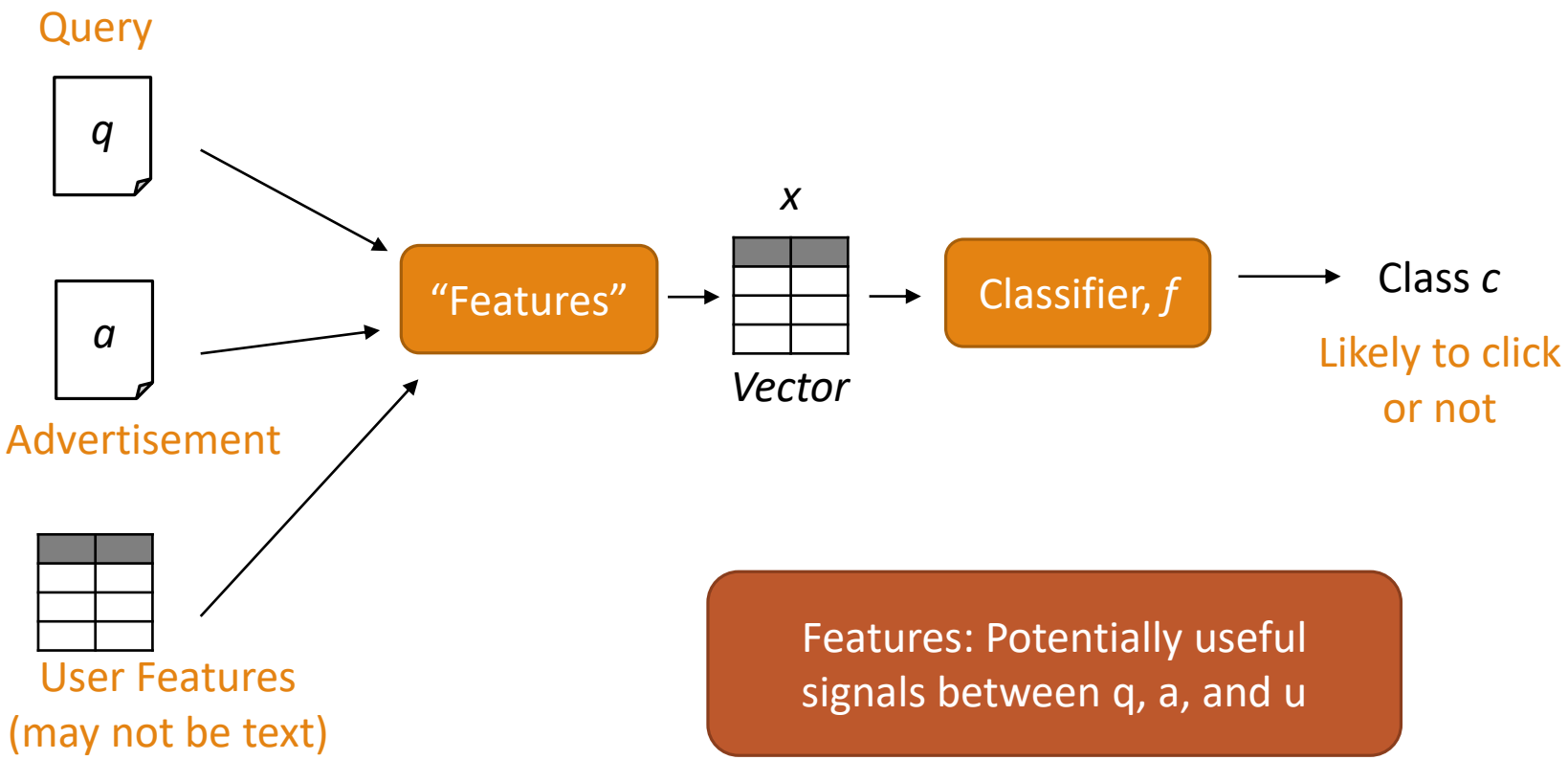
Information Retrieval



Features should describe how the query and document are similar

```
'WordOverlap': 5
```

Search Advertising



Named Entity Recognition

An important sub-task: find and classify names in text, for example:

- The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition

An important sub-task: find and classify names in text, for example:

- The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition

An important sub-task: find and classify names in text, for example:

- The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organi-
zation

NER: Entity Types

Stanford CoreNLP

3 class:	Location, Person, Organization
4 class:	Location, Person, Organization, Misc
7 class:	Location, Person, Organization, Money, Percent, Date, Time

spaCy.io

PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FACILITY	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LANGUAGE	Any named language.

From Stanford CoreNLP (http://nlp.stanford.edu/software/CRF-NER.shtml)

NER: Entity Types

Fine-grained Types

person	doctor	**organization**	terrorist_organization		
actor	engineer	airline	government_agency		
architect	monarch	company	government		
artist	musician	educational_institution	political_party		
athlete	politician	fraternity_sorority	educational_department		
author	religious_leader	sports_league	military		
coach	soldier	sports_team	news_agency		
director	terrorist				
location	body_of_water	**product**	camera	**art**	written_work
city	island	engine	mobile_phone	film	newspaper
country	mountain	airplane	computer	play	music
county	glacier	car	software	**event**	military_conflict
province	astral_body	ship	game	attack	natural_disaster
railway	cemetery	spacecraft	instrument	election	sports_event
road	park	train	weapon	protest	terrorist_attack
bridge					
building	time	chemical_thing	website		
airport	color	biological_thing	broadcast_network		
dam	award	medical_treatment	broadcast_program		
hospital	educational_degree	disease	tv_channel		
hotel	title	symptom	currency		
library	law	drug	stock_exchange		
power_station	ethnicity	body_part	algorithm		
restaurant	language	living_thing	programming_language		
sports_facility	religion	animal	transit_system		
theater	god	food	transit_line		

From Ling & Weld. AAAI 2012 (http://aiweb.cs.washington.edu/ai/pubs/ling-aaai12.pdf)

Is it Classification?

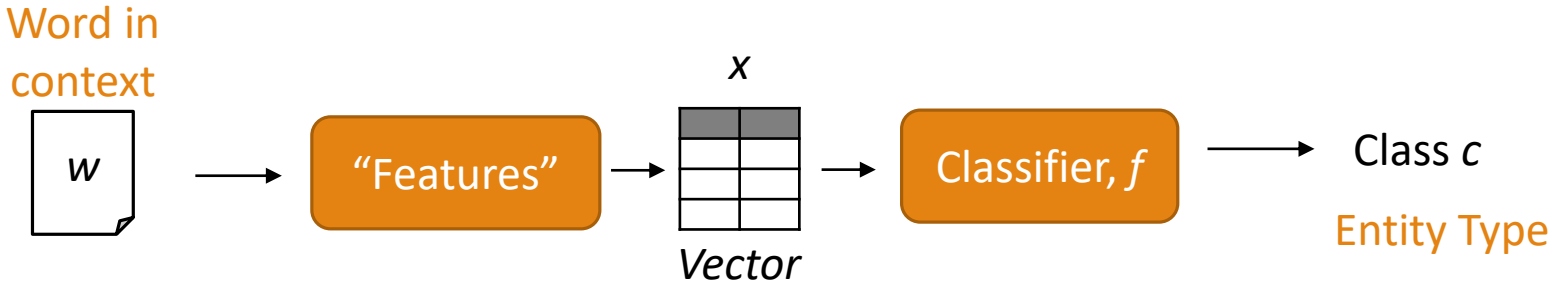
Barack Obama was born in Hawaii.

Yes!

Barack Obama was born in Hawaii.	PER
Barack Obama was born in Hawaii.	PER
Barack Obama was born in Hawaii.	OTH
Barack Obama was born in Hawaii.	OTH
Barack Obama was born in Hawaii.	OTH
Barack Obama was born in Hawaii.	LOC

Features can't just be BoWs/Ngrams

Let's talk about the features




```
def features(w, prev, next):
    # implement code…
    return dict
```

Barack Obama was born in Hawaii. → `features('Obama', 'Barack', 'was born in Hawaii.')`

Barack Obama was born in Hawaii. → `features('born', 'Barack Obama was', 'in Hawaii.')`

Features: Word Itself

Barack Obama was born in Hawaii.

Word

```
'WORD=Obama': 1.0
```

Cased

```
'LWORD=obama': 1.0,
'FirstCharCap': 1.0,
```

```
'AllCaps': 1.0
```

Not true for 'Obama'

Word Features: Lexicons

Barack Obama was born in Hawaii.

Lexicons

```
‘SomeLexicon’: 1.0,
‘Lexicon=FNames’: 1.0,
‘Lexicon=LNames’: 0.0,
‘PartOfName’: 1.0,
‘Lexicon=Places’: 0.0
‘NumOfLexicons’: 5.0
```

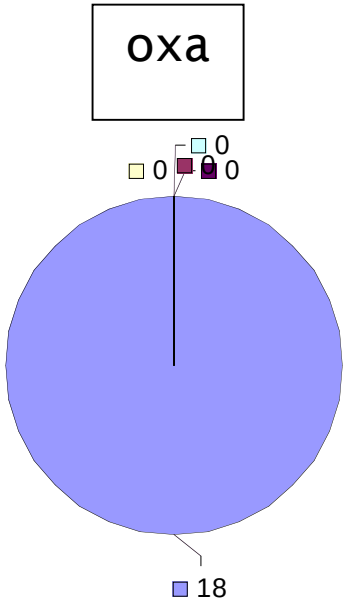
Word Features: Prefixes/Suffixes

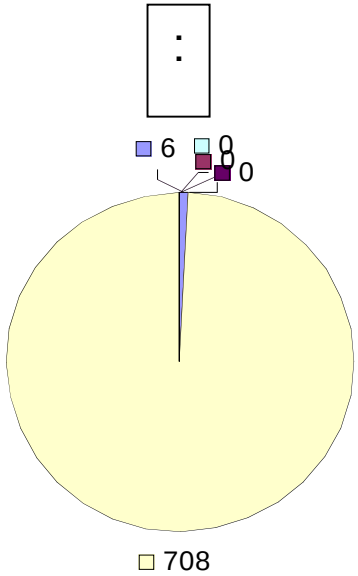
Barack Obama was born in Hawaii.

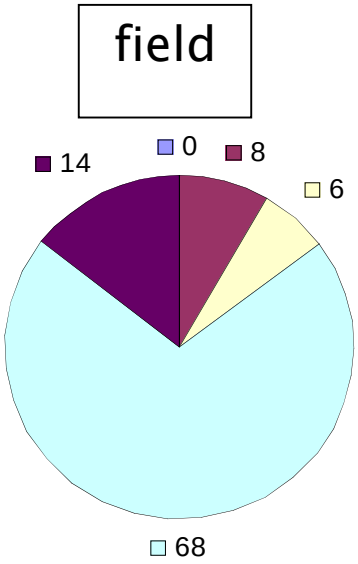
```
‘Prefix1=o’: 1.0,
‘Prefix2=ob’: 1.0,
‘Prefix3=oba’: 1.0,
 ...
‘Suffix1=a’: 1.0,
‘Suffix2=ma’: 1.0,
‘Suffix3=ama’: 1.0,
```

Why is this important?

Word Features: Substrings







Cotrimoxazole

Wethersfield

Alien Fury: Countdown to Invasion

```
'SubStr=oxa': 1.0,
'SubStr=field': 1.0,
```

Word Features: Shapes

Barack Obama was born in Hawaii.

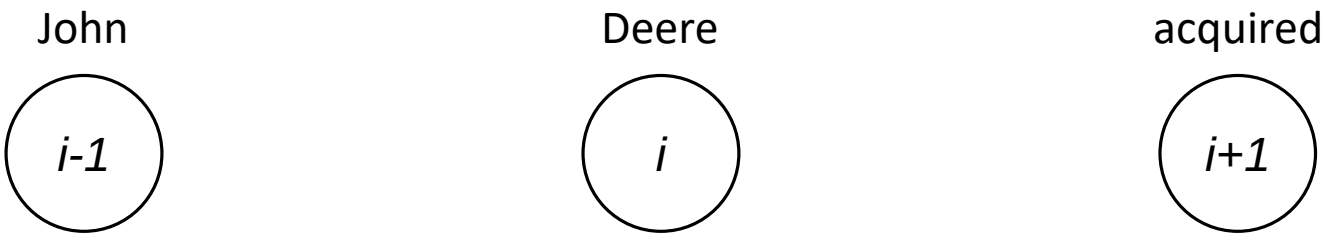
Some "regular expressions" might be useful for some classes

```
'WShape=Xxxxx': 1.0,
'Sshape=Xx': 1.0,
```

Shape(c)=

			John	DC-100	eBay
if A-Z	X				
if a-z	x	Word shapes	Xxxx	XX-ddd	xXxx
if 0-9	d				
o.w.	*c*	Short shapes	Xx	X-d	xXx

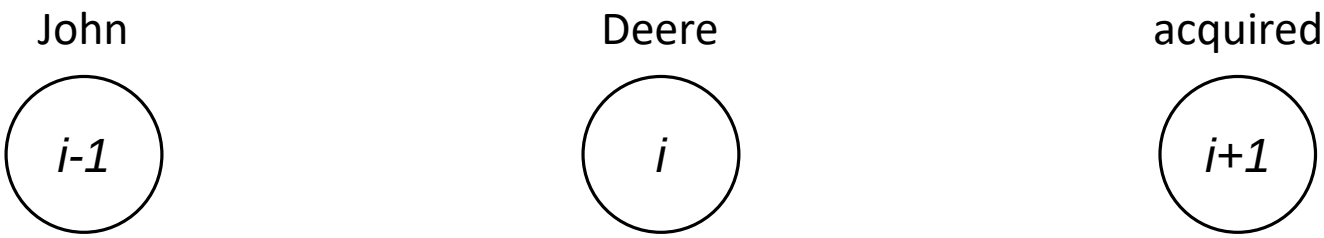
Features: Surrounding Context



WORD=Deere
LWORD=deere
FIRSTCAP=True
SSHAPE=Xx
LEXICON=company
…

NEXT_WORD=acquired

Features: Surrounding Context



PREV_WORD=John
PREV_LWORD=john
PREV_FIRSTCAP=True
PREV_SSHAPE=Xx
PREV_LEXICON=names
PREV_…

WORD=Deere
LWORD=deere
FIRSTCAP=True
SSHAPE=Xx
LEXICON=company
…

NEXT_WORD=acquired
NEXT_LWORD=acquired
NEXT_FIRSTCAP=False
NEXT_SSHAPE=x
NEXT_LEXICON=verb
NEXT_…

Similarly, Ngrams of Context, wider windows, etc.

Today's Overview

Projects

Case Study of NLP from Industry

Features for Text Classification

In-Class Activity 1

Error Analysis of Classifiers

In-Class Activity 2

Today's Overview

Projects

Case Study of NLP from Industry

Features for Text Classification

In-Class Activity 1

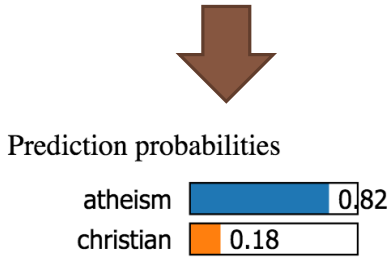
Error Analysis of Classifiers

In-Class Activity 2

Debugging Complex Classifiers

From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion.
If you'd like to know more, send me a note



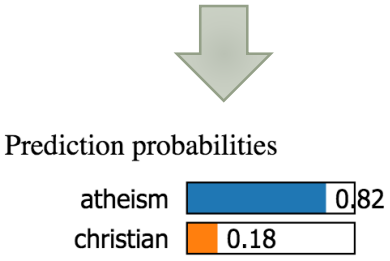
This model has 94% CV accuracy

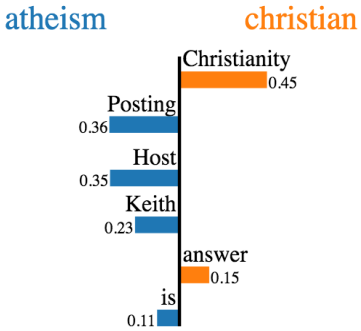
LIME Algorithm

Gives “weights” for **any** classifier!

From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion.
If you’d like to know more, send me a note





Today's Overview

Projects

Case Study of NLP from Industry

Features for Text Classification

In-Class Activity 1

Error Analysis of Classifiers

In-Class Activity 2