

Week 5: Where Are We?



We've covered the basics of **document** representation and characterization.

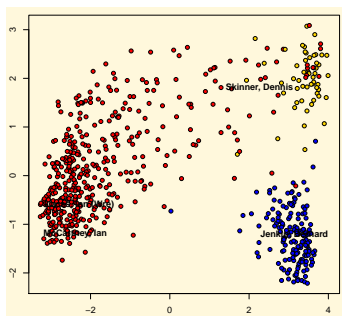
Now begin to think about documents as members of **categories** or **classes**

→ simple, fast **dictionary based** ways to classify/categorize
cover some 'major' dictionaries in **social science**
and demonstrate challenges that emerge in **constructing** and **using** dictionaries, especially for novel tasks.

Terminology

Unsupervised techniques: learning (hidden or latent) structure in **unlabeled** data.

e.g. PCA of legislators's votes: want to see how they are organized— by party? by ideology? by race?




Supervised techniques: learning relationship between inputs and a **labeled** set of outputs.


e.g. opinion mining: what makes a critic like or dislike a movie ($y \in \{0, 1\}$)?


CRITIC REVIEWS FOR STAR WARS: EPISODE VII - THE FORCE AWAKENS

All Critics (313) | Top Critics (48) | My Critics | Fresh (293) | Rotten (20)


 The new movie, as an act of pure storytelling, streams by with fluency and zip.


[Full Review...](#) | December 21, 2015

 **Anthony Lane**
New Yorker
★ Top Critic


 While Star Wars: The Force Awakens gets temporarily bogged down taking us back to the world that we left in 1983, it introduces us to the new and exciting torch-bearers of the franchise.


[Full Review...](#) | December 30, 2015

 **Blake Howard**
Graffiti With Punctuation

 At the end The Force Awakens looks more like a nostalgic film that will work as a transition to the new Star Wars' age. [Full Review in Spanish]

[Full Review...](#) | December 29, 2015

 **Salvador Franco Reyes**

 This film is a well-planned product that balances nostalgia with the capacity to attract new generations into the Star Wars universe. [Full Review in Spanish]

[Full Review...](#) | December 29, 2015

Overview: Supervised Learning

label some examples of each category

e.g. some reviews that were positive ($y = 1$), some that were negative ($y = 0$); some statements that were liberal, some that were conservative.

train a 'machine' on these examples (e.g. logistic regression), using the **features** (DTM, other stuff) as the 'independent' variables.

e.g. does the commentator use the word 'fetus' or 'baby' in discussing abortion law?

classify use the learned relationship—some $f(x)$ —to predict the outcomes of documents ($y \in \{0, 1\}$, review sentiment) not in the training set.

Dictionaries

Overview: Dictionary

idea: set of **pre-defined words** with specific connotations that allow us to **classify** documents automatically, quickly and accurately.

→ common in opinion mining/sentiment analysis, and in coding events or manifestos.

Classification with Dictionary Methods

Aim Typically we are trying to do one of two closely related things:

1 Categorize documents as belonging to a certain class

e.g. this review is 'positive', this speech is 'liberal'

2 Measure **extent to which** document is associated with given category

e.g. this review is generally 'positive', but has some negative elements.

We have a **pre-determined** list of words, the (weighted) presence of which helps us with (1) and (2).

More Specifically

We have a set of **key words**, with attendant scores,

e.g. for movie reviews: 'terrible' is scored as -1 ; 'fantastic' as $+1$

→ the **relative rate** of occurrence of these terms tells us about the overall **tone** or category that the document should be placed in.

i.e. for document i and words $m = 1, \dots, M$ in the dictionary,

$$\text{tone of document } i = \sum_{m=1}^M \frac{s_m w_{im}}{N_i}$$

where s_m is the score of word m

and w_{im} is the number of occurrences of the m th dictionary word in the document i

and N_i is the total number of all dictionary words in the document.

→ just add up the number of times the words appear and multiply by the score (normalizing by doc dictionary presence)

Example: Barnes' review of *The Big Short*

Director and co-screenwriter Adam McKay (Step Brothers) bungles a great opportunity to savage the architects of the 2008 financial crisis in The Big Short, wasting an A-list ensemble cast in the process. Steve Carell, Brad Pitt, Christian Bale and Ryan Gosling play various tenuously related members of the finance industry, men who made a killing by betting against the housing market, which at that point had superficially swelled to record highs. All of the elements are in place for a lacerating satire, but almost every aesthetic choice in the film is bad, from the U-Turn-era Oliver Stone visuals to Carell's sketch-comedy performance to the cheeky cutaways where Selena Gomez and Anthony Bourdain explain complex financial concepts. After a brutal opening half, it finally settles into a groove, and there's a queasy charge in watching a credit-drunk America walking towards that cliff's edge, but not enough to save the film.

Retain words in Hu & Liu Dictionary...

*Director and co-screenwriter Adam McKay (Step Brothers) bungles a **great** opportunity to **savage** the architects of the 2008 financial **crisis** in The Big Short, **wasting** an A-list ensemble cast in the process. Steve Carell, Brad Pitt, Christian Bale and Ryan Gosling play various **tenuously** related members of the finance industry, men who made made a **killing** by betting against the housing market, which at that point had **superficially swelled** to record highs. All of the elements are in place for a lacerating satire, but almost every aesthetic choice in the film is **bad**, from the U-Turn-era Oliver Stone visuals to Carell's sketch-comedy performance to the cheeky cutaways where Selena Gomez and Anthony Bourdain explain **complex** financial concepts. After a **brutal** opening half, it finally settles into a groove, and there's a queasy charge in watching a credit-**drunk** America walking towards that cliff's edge, but not **enough** to save the film.*

Retain words in Hu & Liu Dictionary...

great

savage
wasting

crisis

tenu-

ously

killing

superficially swelled

bad

complex

brutal

drunk

enough

Simple math...

negative 11

positive 2

total 13

$$\text{tone} = \frac{2-11}{13} = \frac{-9}{13}$$



Notes

Typically assume that each word in dictionary has **one of two values and sum totals** matter.

But no requirement that s_m be dichotomous or integer valued: could be **continuous**.

e.g. might want to differentiate 'good' from 'great' from 'best'. Hard to come up with rules!

NB Tone of the document can be presented as a continuous value, or used to put documents in categories via some **cutoff** rule.

e.g. all documents with $\text{tone} > 0$ are deemed 'positive'

NB Bag-of-words assn may be especially dubious for some dictionary tasks

e.g. context matters: "was **not** good" gets +1 !

Dictionaries I: General Inquirer

Stone (1965) begins efforts to automatically analyze **psychological states** of authors

'General Inquirer' combines several dictionaries to make total of 182 categories:

- ▶ Harvard IV-4 dictionary: psychology, themes, topics
- ▶ Lasswell dictionary: "commonsense categories of meaning", 8 basic value categories
- ▶ Semin and Fielder categories: interpersonal/psychological properties of words

General Inquirer (selected)

Entry	Source	Positiv	Negativ	Pstv	Affil	Ngtv	Hostile	Strong	Power
ABILITY	H4Lvd	Positiv						Strong	
ABJECT	H4		Negativ						
ABLE	H4Lvd	Positiv		Pstv				Strong	
ABNORMAL	H4Lvd		Negativ			Ngtv			
ABOARD	H4Lvd								
ABOLISH	H4Lvd		Negativ			Ngtv	Hostile	Strong	Power
ABOLITION	Lvd								
ABOMINABLE	H4		Negativ					Strong	
ABRASIVE	H4		Negativ				Hostile	Strong	
ABROAD	H4Lvd								
ABRUPT	H4Lvd		Negativ			Ngtv			
ABSCOND	H4		Negativ				Hostile		
ABSENCE	H4Lvd		Negativ						
ABSENT#1	H4Lvd		Negativ						
ABSENT#2	H4Lvd								
ABSENT-MINDED	H4		Negativ						
ABSENTEE	H4		Negativ				Hostile		
ABSOLUTE#1	H4Lvd							Strong	
ABSOLUTE#2	H4Lvd							Strong	

provides dictionaries and [software](#), which performs some stemming and [disambiguation](#) in terms of context

e.g. ADULT has two meanings: one is a 'virtue', one is a 'role'

Dictionaries II: Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al, <http://liwc.wpengine.com/>

LIWC2007 dictionary contains 2290 words and word stems (see also LIWC2015)

80 categories, organized **hierarchically** into 4 larger groups.

e.g. all anger words (e.g. hate) \subset negative emotion \subset affective processes \subset psychological processes

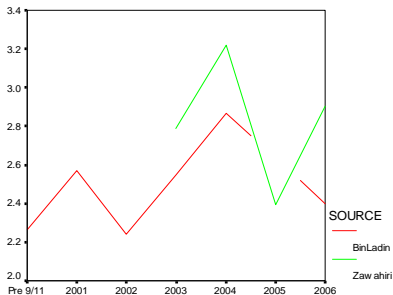
NB words can be in **multiple** categories, and each subdictionary score is incremented as such words appear.

Based on somewhat involved human coding/judgement and **proprietary**.

Pennebaker & Chung, 2007: Computerized Analysis of Al-Qaeda Transcripts

"The LIWC analyses suggest that Bin Ladin has been increasing in his cognitive complexity and emotionality since 9/11, as reflected by his increased use of exclusive, positive emotion, and negative emotion word use. "

C. Positive emotion (happy, love)



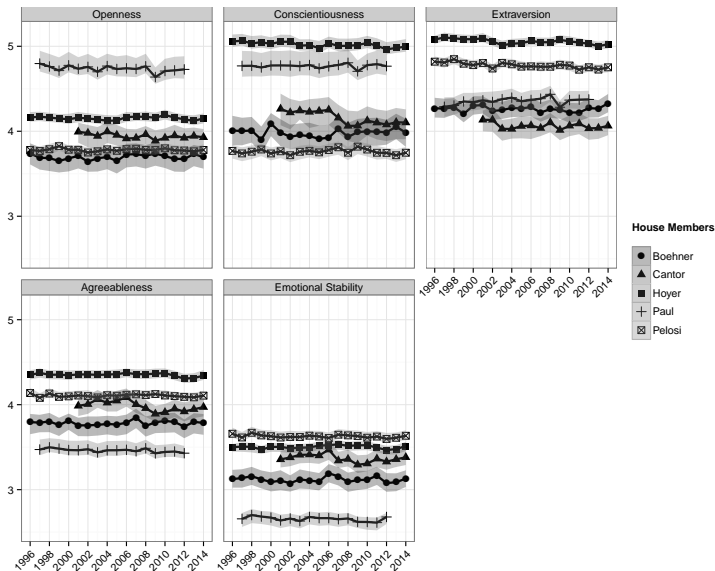
D. Negative emotion (hate, sad)



Application: Ramey, Klingler & Hollibaugh

Mairesse et al. (2007)
provide estimates of 'big 5'
personality traits from
LIWC categories

Ramey et al apply to
Congressional speech.



Dictionaries III: Young & Saroka's *Lexicoder Sentiment Dictionary*

Create dictionary specifically for **political communication**

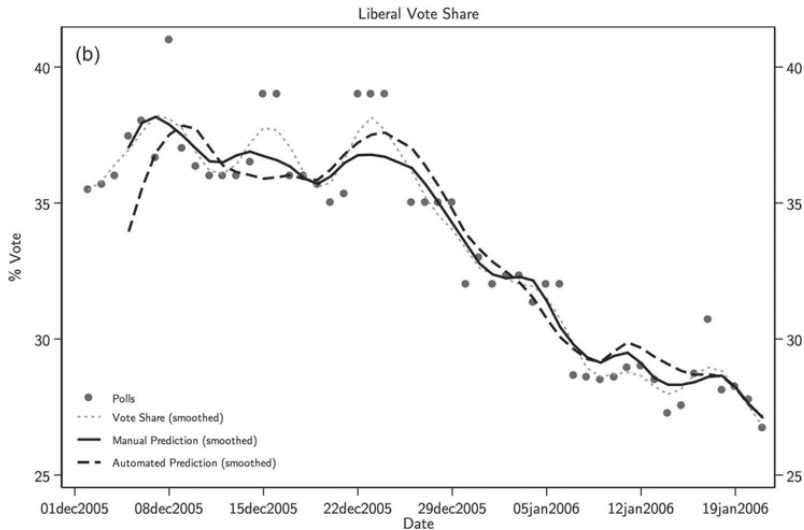
So combine General Inquirer and Roget's Thesaurus with...

RID Regressive Imagery Dictionary which "was designed to distinguish between primordial and conceptual thinking"

plus much hand coding and **validation** using KWIC (from 10k newspapers), plus some special **negation** handling.

NB high (0.75) correlation with LIWC, though outperforms it when compared to **manual coding** of NYT.

Predicting Liberal Poll Vote (2006) as function of media tone



Dictionaries IV: Laver & Garry

2000 Laver and Garry create dictionary for **manifestos** where basic unit is strings of ~ 10 words in length.

→ hierarchical, with topmost level pertaining to five policy domains:
economy, political system, social system, external relations, 'other' (waffle)

get good/valid results and high correlation with **expert surveys**.

1 1 1 ECONOMY/+State+/Budget
Budget

1 1 1 1 ECONOMY/+State+/Budget/Spending
Increase public spending

1 1 1 1 1 ECONOMY/+State+/Budget/Spending/Health

1 1 1 1 2 ECONOMY/+State+/Budget/Spending/Educ. and training

Dictionaries V: Hu & Liu

2004 Hu and Liu (“Mining and Summarizing Customer Reviews”) provide 6800 words which are **positive** and **negative** derived from amazon.com and others.



1,036 of 1,144 people found the following review helpful

★★★★★ **With Great Powers Comes Great Responsibility**

By [Tommy H.](#) on July 17, 2009

I admit it, I'm a ladies' man. And when you put this shirt on a ladies' man, it's like giving an AK-47 to a ninja. Sure it looks cool and probably would make for a good movie, but you know somebody is probably going to get hurt in the end (no pun intended). That's what almost happened to me, this is my story...

Be Careful...

In principle, it is straightforward to extend dictionary from one domain to another

→ matter of adding extra words in the various categories.

But much care is needed when a dictionary designed for one context is applied to another.

e.g. Loughran & MacDonald, 2011: common dictionaries fail badly when applied to financial texts

plus virtually impossible to validate dictionaries: very expensive, at least.

Making Dictionaries from Scratch

Not trivial, extending pre-existing is the norm.

Generally, need to ensure that we get **all relevant content** (no false negatives) and **only** that content (no false positives)

Works best when the contrasts are binary/obvious

e.g. obviously 'for' vs 'against'

NB Typically start with distinct **types** of documents (classified by hand), and learn which words are important for **discriminating** between them.

Word **embeddings** may offer automatic way forward here