Lecture 06:
Texto como Datos

Aprendizaje y Minería de Datos para los Negocios

Ignacio Sarmiento-Barbieri

Universidad de los Andes

November 3, 2021

Agenda

1. Text as Data
2. Aspectos Operativos
   - Tokenization
3. Modelos de Tópicos
4. Modelos de Tópicos
   - Latent Dirichlet Allocation
5. Word Embeddings
6. Word Embedding
7. Break
8. `R para ML`

Text as Data: Motivation

Comencemos con una historia: Sesgo en el discurso partidario

Econometrica, Vol. 78, No. 1 (January, 2010), 35–71

WHAT DRIVES MEDIA SLANT? EVIDENCE FROM U.S. DAILY NEWSPAPERS

BY MATTHEW GENTZKOW AND JESSE M. SHAPIRO[1]

We construct a new index of media slant that measures the similarity of a news outlet's language to that of a congressional Republican or Democrat. We estimate a model of newspaper demand that incorporates slant explicitly, estimate the slant that would be chosen if newspapers independently maximized their own profits, and compare these profit-maximizing points with firms' actual choices. We find that readers have an economically significant preference for like-minded news. Firms respond strongly to consumer preferences, which account for roughly 20 percent of the variation in measured slant in our sample. By contrast, the identity of a newspaper's owner explains far less of the variation in slant.

KEYWORDS: Bias, text categorization, media ownership.

Text as Data: Motivation

Gentzkow and Shapiro: What drives media slant? Evidence from U.S. daily newspapers (*Econometrica*, 2010)

- Construir un modelo económico para la demanda de los periódicos que incorpore el partidismo político (republicano vs demócrata)
  - ¿Cuál sería la "inclinación" independiente de maximización de beneficios?
  - Compare esto con el sesgo estimado a partir del texto del periódico.

Text as Data: Motivation

- Jerry Moran, R-KS, dice "death tax" a menudo y su distrito votó 73% por George W. Bush en 2004.

$$\mathbf{X}_{\text{text}} = f(\text{ideology}) \approx g(Y_{Bush})$$

$$\Rightarrow \text{"death tax" es republicano}$$

$$\Rightarrow \text{el Wall Street Journal se inclina a la derecha.}$$

- William Jefferson, D-LA, dice "estate tax" a menudo y su distrito votó 24% por George W. Bush en 2004.

Text as Data: The Big Picture

- **Text is a vast source of data for business**
- It comes connected to interesting "author" variables
  - What you buy, what you watch, your reviews
  - Group membership, who you represent, who you email
  - Market behavior, macro trends, the weather
- Opinion, subjectivity, etc.
- Sentiment is *very* loosely defined: Observables linked to the variables motivating language choice

Text as Data: The Big Picture

- **Text is also super high dimensional**
- And it gets higher dimensional as you observe more speech.
- Analysis of phrase counts is the state of the art (hard to beat).

Aspectos Operativos

Information Retrieval and Tokenization

- A passage in '*As You Like It*' from Shakepeare:

  All the world's a stage,
  and all the men and women merely players:
  they have their exits and their entrances;
  and one man in his time plays many parts...

- What the econometrian sees:

world	stage	men	women	play	exit	entrance	time
1	1	2	1	2	1	1	1

- This is the Bag-of-Words representation of text.

Possible tokenization steps

- Remove words that are super rare (in say $< \frac{1}{2}\%$, or $< 15\%$ of docs; this is application specific). For example, if Argentine occurs only once, it's useless for comparing documents.
- Stemming: 'tax' $\leftarrow$ taxing, taxes, taxation, taxable, ...
  A stemmer cuts words to their root with a mix of rules and estimation.'Porter' is standard for English.
- Remove a list of stop words containing irrelevant tokens.
  If, and, but, who, what, the, they, their, a, or, ...
  Be careful: one person's stopword is another's key term.
- Convert to lowercase, drop numbers, punctuation, etc ...
  Always application specific: e.g., don't drop :-) from tweets.

The n-gram language model

- An n-gram language model is one that describes a dialect through transition probabilities on n consecutive words.
- An n-gram tokenization counts length-n sequences of words.
  A unigram is a word, bigrams are transitions between words.
  e.g., `world.stage`, `stage.men`, `men.women`, `women.play`, ...
- This can give you rich language data, but be careful: n-gram token vocabularies are very high dimensional (p^n)
- More generally, you may have domain specific 'clauses' that you wish to tokenize.
- There is always a trade-off between complexity and generality.
- Often best to just count words.

Text as Data: Wordle

- Often best to just count words.
- For example, occurrences by party for some partisan terms

Congress	State	Party	America	Death Tax	Estate Tax	· · ·
63	NM	dem	108	30	140	
		gop	100	220	12	

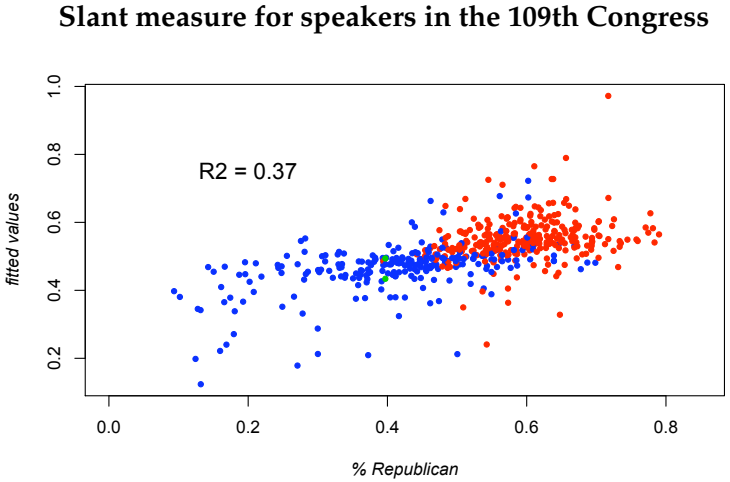
Text Regression

- Once you have text in a numeric format, we can use all the tools we learned so far

$$y = f(word\ counts) + u \tag{1}$$

- where you can use lasso, PCA, etc. to do dimentionality reduction

Text Regression

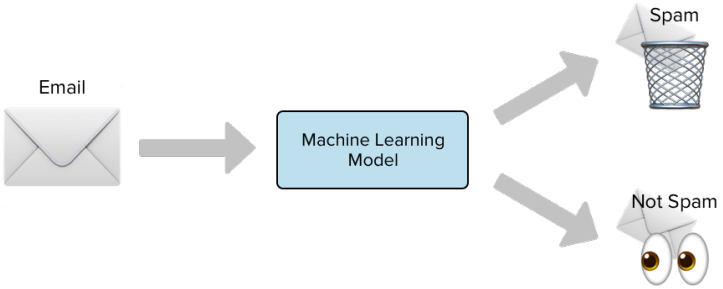


Democrats get low z_{slant} and Republicans get high z_{slant}.

Do this for newspaper text and you'll get a similar picture

Text Regression

- Another example: Classify emails into spam



$$\text{logit}\left[\texttt{spam}\right] = \alpha + f\beta \tag{2}$$

- where $f_i = \frac{x_i}{\sum_j x_{ij}}$ are the normalized text counts

Modelos de Tópicos

Topic Models

- Text is super high dimensional
- there is often abundant *unlabeled* text
- Some times unsupervized factor model is a popular and useful strategy with text data
- You can first fit a factor model to a giant corpus and use these factors for supervised learning on a subset of labeled documents.
- The unsupervised dimension reduction facilitates the supervised learning

Topic Models: Example

- We have 6166 reviews, with an average length of 90 words per review, we8there.com.
- A useful feature of these reviews is that they contain both text and a multidimensional rating on overall experience, atmosphere, food, service, and value.
- For example, one user submitted a glowing review for Waffle House #1258 in Bossier City, Louisiana: *I normally would not revue a Waffle House but this one deserves it. The workers, Amanda, Amy, Cherry, James and J.D. were the most pleasant crew I have seen. While it was only lunch, B.L.T. and chili, it was great. The best thing was the 50' s rock and roll music, not to loud not to soft. This is a rare exception to what you all think a Waffle House is. Keep up the good work.*
  Overall: 5, Atmosphere: 5, Food: 5, Service: 5, Value: 5.

Topic Models: Example

- Puedo resumir estos textos a algunos simples temas?
- Puedo usar PCA
- Tambien podemos usar LDA (más apropiado)

Principal Component Regression

- The concept is very simple: instead of regressing onto X, use a lower dimension set of principal components F as covariates.
- This works well for a few reasons:
  - PCA reduces dimension, which is always good.
  - Higher variance covariates are good in regression, and we choose the top PCs to have highest variance.
  - The PCs are independent: no multicollinearity.
- The 2-stage algorithm is straightforward. For example,

```
mypca = prcomp(X, scale=TRUE)
z = predict(mypca)[,1:K]
reg = glm(y~., data=as.data.frame(z))
```

Latent Dirichlet Allocation

- The approach of using PCA to factorize text was common before the 2000s.
- Versions of this algorithm were referred to under the label latent semantic analysis.
- However, this changed with the introduction of topic modeling, also known as Latent Dirichlet Allocation (LDA), by Blei et al. in 2003.
- These authors pointed out that the squared error loss (i.e., Gaussian model) implied by PCA is inappropriate for analysis of sparse word-count data.

Latent Dirichlet Allocation

TRANSPARENCY AND DELIBERATION WITHIN THE FOMC: A COMPUTATIONAL LINGUISTICS APPROACH*

STEPHEN HANSEN
MICHAEL MCMAHON
ANDREA PRAT

How does transparency, a key feature of central bank design, affect monetary policy makers' deliberations? Theory predicts a positive discipline effect and negative conformity effect. We empirically explore these effects using a natural experiment in the Federal Open Market Committee in 1993 and computational linguistics algorithms. We first find large changes in communication patterns after transparency. We then propose a difference-in-differences approach inspired by the career concerns literature, and find evidence for both effects. Finally, we construct an influence measure that suggests the discipline effect dominates. *JEL Codes*: E52, E58, D78.

Latent Dirichlet Allocation

THE PARTICIPATION DIVIDEND OF TAXATION: HOW CITIZENS IN CONGO ENGAGE MORE WITH THE STATE WHEN IT TRIES TO TAX THEM*

JONATHAN L. WEIGEL

This article provides evidence from a fragile state that citizens demand more of a voice in the government when it tries to tax them. I examine a field experiment randomizing property tax collection across 356 neighborhoods of a large Congolese city. The tax campaign was the first time most citizens had been registered by the state or asked to pay formal taxes. It raised property tax compliance from 0.1% in control to 11.6% in treatment. It also increased political participation by about 5 percentage points (31%): citizens in taxed neighborhoods were more likely to attend town hall meetings hosted by the government or submit evaluations of its performance. To participate in these ways, the average citizen incurred costs equal to their daily household income, and treated citizens spent 43% more than control. Treated citizens also positively updated about the provincial government, perceiving more revenue, less leakage, and a greater responsibility to provide public goods. The results suggest that broadening the tax base has a "participation dividend," a key idea in historical accounts of the emergence of inclusive governance in early modern Europe and a common justification for donor support of tax programs in weak states. *JEL* Codes: H20, P48, D73.

Latent Dirichlet Allocation

TABLE VII

TOPICS OF CITIZEN COMMENTS AT TOWN HALLS AND WRITTEN-IN COMMENTS ON SUBMITTED EVALUATIONS

Order	(1)	(2)	(3)	(4)	(5)
Panel A: Topics of citizen comments at town hall meetings					
1	pay	tax	necessary	pay	pay
2	necessary	population	population	take	must
3	population	necessary	collectors	without	population
4	tax	pay	pay	decision	why
5	why	know	know	why	others
6	agents	do	see	necessary	collectors
7	time	collectors	tax	participation	agents
8	collectors	why	without	tax	nothing
9	communes	nothing	information campaign	others	participation
10	manager	schools	transparency	agents	tax
Panel B: Topics of written-in comments on submitted evaluations					
1	government	government	government	government	government
2	water	provincial	provincial	provincial	province
3	ask	should	should	work	country
4	roads	more	population	province	leaders
5	electricity	work	especially	do	population
6	improve	public	erosion	better	good
7	jobs	goods	needs	ask	ask
8	people	concerning	people	would	development
9	more	ask	security	central	love
10	who	because	take	Kasaï	could

Notes. This table reports the first ten words in each of the five main topics identified by latent Dirichlet allocation (Blei, Ng, and Jordan 2003) applied to two sources of text that offer insight into citizens' reasons

Latent Dirichlet Allocation

- Blei et al. proposed you take the bag-of-words representation seriously and model token counts as realizations from a multinomial distribution.
- Topic models are built on a simple document generation process:
  - For each word, pick a "topic" k. This topic is defined through a probability vector over words, say, θ_k with probability θ_{kj} for each word j.
  - Then draw the word according to the probabilities encoded in θ_k .
- After doing this over and over for each word in the document, you have proportion ω_{i1} from topic 1, ω_{i2} from topic 2, and so on.

Latent Dirichlet Allocation

- This basic generation process implies that the full vector of word counts, x_i, has a multinomial distribution:

$$x_i \sim MN(\omega_{i1}\theta_1 + \cdots + \omega_{iK}\theta_K, m_i) \tag{3}$$

- where $m_i = \sum_j x_{ij}$ is the total document length and, for example,
- the probability of word j in document i will be $\sum_k \omega_{ik}\theta_{kj}$

Latent Dirichlet Allocation vs PCA

- Recall our PC model:

$$E(x_i) = \delta_{i1} F_1 + \cdots + \delta_{iK} F_K \tag{4}$$

- The analogous topic model representation, implied by the above equation, is

$$E(x_i) = \omega_{i1} \theta_1 + \cdots + \omega_{iK} \theta_K \tag{5}$$

- such that topic score ω_{ik} is like PC score δ_{ik} and
- θ_k topic probabilities are like rotations F_k.
- The distinction is that the multinomial in implies a different loss function (from a multinomial) rather than the sums of squared errors that PCA minimizes.
- Note that we condition on document length here so that topics are driven by relative rather than absolute term usage.

Word Embeddings

Word Embedding

- This is a new method that have come out of work in deep learning.
- Word embedding was originally motivated as a technique for dimension reduction on the inputs to a deep neural network.
- However, word embedding turns out to be valuable in its own right: it imposes a spatial structure on words, making it possible for those studying language to reason about distances between meanings and consider the algebra behind combinations of words in documents.

Word Embedding

- In the original deep learning context, embedding layers replace each word with a vector value, such that, for example, hotdog becomes the location [1,–5, 0.25] in a three-dimensional embedding space
- Compare this to the standard bag-of-words representation, where hotdog would be represented as a binary vector that is as long as there are words in the vocabulary, say, p.
- This binary vector will have p–1 zeros and a one in the hotdog dimension.
- The word embedding has translated the language representation from a large binary space to a smaller real-valued (and much richer) space.

Word Embedding

- There are a variety of different embedding algorithms—as many as there are different architectures for deep neural networks.
- The most common and general embeddings are built around word co-occurrence matrices.
- This includes the popular `Glove` and `Word2Vec` frameworks.
- What is co-occurrence?
  - Two words co-occur if they appear within the same sentence and within b words of each other. Where b is the "window size"
  - For a vocabulary size p, this leads to a sparse $p \times p$ co-occurrence matrix where each [i, j] entry is the number of times that words i and j co-occur. Call this matrix C.
  - A word embedding algorithm seeks to approximate C as the product of two lower-dimensional matrices

Word Embedding

- A word embedding algorithm seeks to approximate C as the product of two lower-dimensional matrices

$$C \approx UV' \tag{6}$$

- Here, U and V are each $p \times K$ dimensional dense and real valued matrices.
- K is the dimension of the embedding space; hence, $K << p$ and both U and V are very tall and thin matrices.
- Each row of U and of V, u_j and v_j is then a K-dimensional embedding of the jth word.
- The implication is that these embeddings summarize the meaning of words as their inner product defines how much you expect them to co-occur.

  Recall that the inner product is a standard measure of distance in linear algebra (e.g. $e'e$)

Word Embedding

- One way to find U and V is to solve $C \approx UV'$ through the singular value decomposition (SVD).
- SVD is a factorization of a real or complex matrix, that serves for example, to find the eigenvalues and eigenvectors of square symmetric matrices (and hence in calculating principal components).
- In practice, most of the software embedding solutions use alternatives to SVD that are designed to deal with the high amount of sparsity in C (since most words never co-occur in limited windows for standard corpora).
- Under many algorithms, especially when co-occurrence is symmetric, U and V will be mirror images of each other.
- Thus, it is standard to take one of these vectors u_j as the single embedding location for word j.

Word Embedding

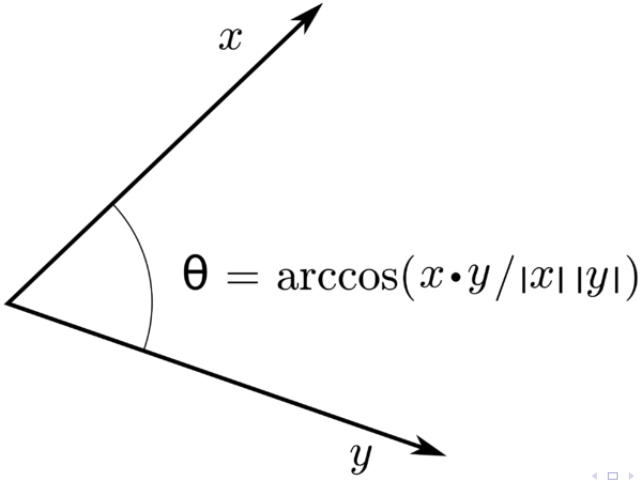
- These locations were originally viewed as an intermediate output—as a processing step for inputs to a deep neural network.
- However, social scientists and linguists have discovered that the space of word locations contains rich information about the language of the documents used to train the embedding.
- Word embeddings preserve semantic relationships.
  - Words with similar meaning have similar representations.
  - Dimensions induced by word differences can be used to identify cultural concepts
- For example, the vector difference `man - woman` isolates a gender dimension in the space.

Word Embedding

- For example, the vector difference `man - woman` isolates a gender dimension in the space.
  - The dimensions are useful because they produce quantitative measures of similarity between the associated concepts and specific words in the corpus.
  - In this case, we can understand the gender connotation of a given word by taking the cosine of the angle between the vector representation of the word and the differenced vector representing the gender dimension (why?)

Word Embedding

- Recall the geometric interpretation of the angle between two vectors defined using an inner product



Word Embedding

- Words with male connotations – e.g. male first names – are going to be positively correlated with `man - woman`.
- Female words, in turn, will be negatively correlated with the dimension.
- This framework provides an intuitive approach to measuring stereotypical associations in a given corpus.
- Bolukbasi et al (2016) is a nice example

Word Embedding: Example 1

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words *receptionist* and *female*, while maintaining desired associations such as between the words *queen*

Word Embedding: Example 1

- They trained a standard word2vec embedding algorithm on the Google News corpora of news articles.
- Then look at the differences between established gender words (for example, the vector for man minus the vector for woman, or father minus mother) to establish an axis in the embedding space that spans from masculinity to femininity.
- They then calculate the location along this axis for a large number of terms that should be gender-neutral.
- The embedding space has learned—from how the words are used in news articles—that these professions are stereotypically viewed as female and male occupations.

Word Embedding: Example 1

Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 1: **Left** The most extreme occupations as projected on to the *she*−*he* gender direction on w2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded. **Right** Automatically generated analogies for the pair *she-he* using the procedure described in text. Each automatically generated analogy is evaluated by 10 crowd-workers to whether or not it reflects gender stereotype.

Word Embedding: Example 2

Language from police body camera footage shows racial disparities in officer respect

Rob Voigt[a,1], Nicholas P. Camp[b], Vinodkumar Prabhakaran[c], William L. Hamilton[c], Rebecca C. Hetey[b], Camilla M. Griffiths[b], David Jurgens[c], Dan Jurafsky[a,c], and Jennifer L. Eberhardt[b,1]

[a]Department of Linguistics, Stanford University, Stanford, CA 94305; [b]Department of Psychology, Stanford University, Stanford, CA 94305; and [c]Department of Computer Science, Stanford University, Stanford, CA 94305

Contributed by Jennifer L. Eberhardt, March 26, 2017 (sent for review February 14, 2017; reviewed by James Pennebaker and Tom Tyler)

Using footage from body-worn cameras, we analyze the respectfulness of police officer language toward white and black community members during routine traffic stops. We develop computational linguistic methods that extract levels of respect automatically from transcripts, informed by a thin-slicing study of participant ratings of officer utterances. We find that officers speak with consistently less respect toward black versus white community members, even after controlling for the race of the officer, the severity of the infraction, the location of the stop, and the outcome of the stop. Such disparities in common, everyday interactions between police and the communities they serve have important implications for procedural justice and the building of police–community trust.

racial disparities | natural language processing | procedural justice | traffic stops | policing

some have argued that racial disparities in perceived treatment during routine encounters help fuel the mistrust of police in the controversial officer-involved shootings that have received such great attention. However, do officers treat white community members with a greater degree of respect than they afford to blacks?

We address this question by analyzing officers' language during vehicle stops of white and black community members. Although many factors may shape these interactions, an officer's words are undoubtedly critical: Through them, the officer can communicate respect and understanding of a citizen's perspective, or contempt and disregard for their voice. Furthermore, the language of those in positions of institutional power (police officers, judges, work superiors) has greater influence over the course of the interaction than the language used by those with less power (12–16). Measuring officer language thus provides a quantitative lens on one key aspect of the quality or tone of

Word Embedding: Example 3

Stereotypes in High-Stakes Decisions:

Evidence from U.S. Circuit Courts

Elliott Ash, ETH Zurich

Daniel L. Chen, Toulouse School of Economics

Arianna Ornaghi, University of Warwick*

March 12, 2020

Abstract

Stereotypes are thought to be an important determinant of decision making, but they are hard to systematically measure, especially for individuals in policy-making roles. In this paper, we propose and implement a novel language-based measure of gender stereotypes for the high-stakes context of U.S. Appellate Courts. We construct a judge-specific measure of gender-stereotyped language use – *gender slant* – by looking at the linguistic association of words identifying gender (male versus female) and words identifying gender stereotypes (career versus family) in the judge's authored opinions. Exploiting quasi-random assignment of judges to cases and conditioning on detailed biographical characteristics of judges, we study how gender stereotypes influence judicial behavior. We find that judges with higher slant vote more conservatively on women's rights' issues (e.g. reproductive rights, sexual harassment, and gender discrimination). These more slanted judges also influence workplace outcomes for female colleagues: they are less likely to assign opinions to female judges, they are more likely to reverse lower-court decisions if the lower-court judge is a woman, and they cite fewer female-authored opinions.

Volvemos en 15 mins con R

R para ML

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Text as Data: What is slant?

- Text: phrase-counts by speaker in 109^{th} US Congress (05-06)
- Sentiment: two-party constituent vote-share for Bush in 2004.
- Use covariance between phrase frequencies (f_{ij}) and 'Bush' sentiment (y_i) to build an index of partisanship for text.

$$z_i^{slant} = \sum_j \text{cov}(f_j, y) f_{ij}$$

- For example, if phrase j forms a high proportion of what you say, and usage of phrase
- j is correlated with Bush vote-share, then this contributes a positive amount to your slant score.
- This is a type of *marginal regression*.