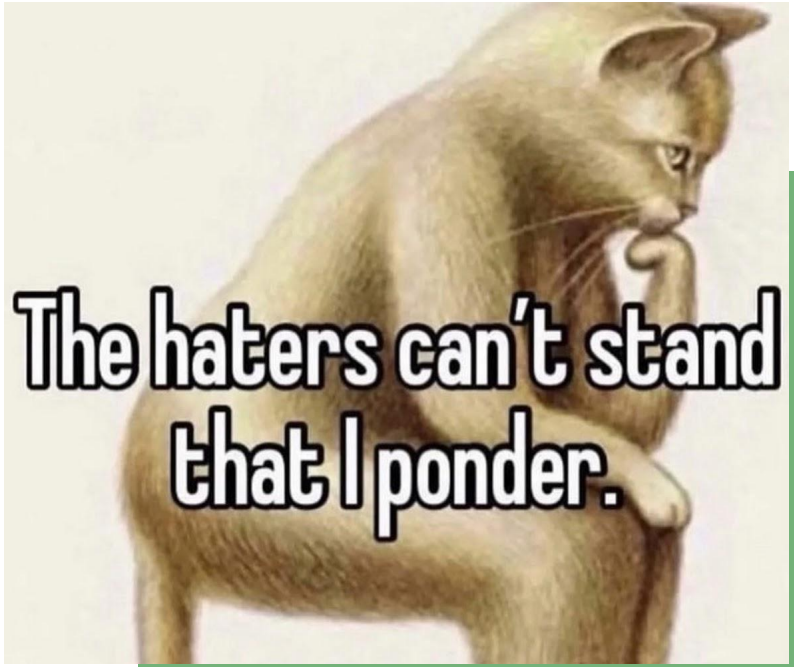
Week 8 lab

COGS 108, 9:00-9:50AM (B01)

https://github.com/alexavndra/cogs108-b01

Reminders!!

- A3 due TODAY at 11:59PM
  - SUBMIT ON DATAHUB
- D7 is due MONDAY, November 27th at 11:59PM
- Make some OH appointments!
  - https://calendly.com/alexandrarh/office-hours



Nonparametrics

An overview (sort of)!

Nonparametrics: what is it?

Nonparametrics is when I take the MATH 20E midterm and believing I'll get a 90%...without taking the practice midterm and studying, so I don't know the distribution of questions/topics (I lied btw).

Nonparametrics (actual, not funny): a statistical method in which the data are not assumed to come from prescribed models that are determined by a small number of parameters (ibm.com)

- ➢ Data distribution is usually unknown
- ➢ Best to use when data can't be parameterized (e.g. ordinal/ranked data)

Still uses parameters, just doesn't assume/parameterize UNDERLYING distribution.

D7: Text analysis (and TF-IDF)

The topic + what you'll be using!

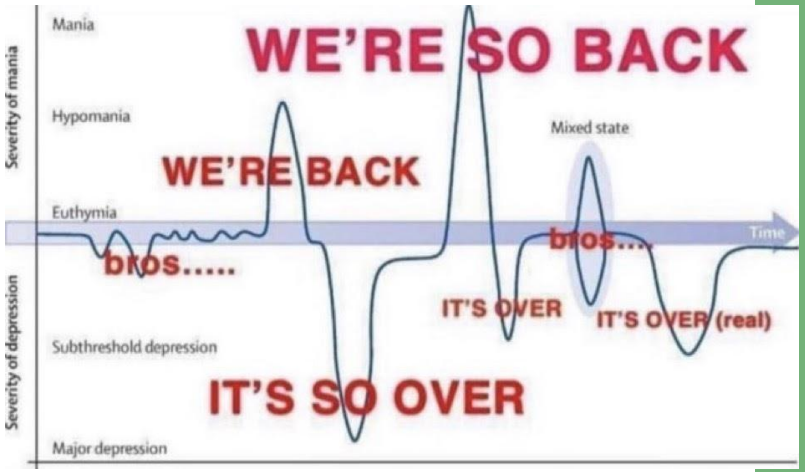
Here's a rundown of what you'll be focusing on...

Line graphs: helps analyze the relevance of certain words within a text dataset or document

- ➢ Sentiment analysis?

TF-IDF analyzers: statistical measure that evaluates how relevant a word is to a document in a collection of documents.

- ➢ Utilizes TfIdfVectorizer
  - Sublinear_tf: Apply sublinear tf scaling, i.e. replace tf with 1 + log(tf).
  - max_features: Builds a vocabulary that only consider the top max_features ordered by term frequency across the corpus



How TF-IDF works (mathematically)

$$\text{TF-IDF}(t, d, D) = \log(1 + \text{frequency of } t \text{ in } d) \quad \log\left(\frac{\text{Total \# of documents}}{\text{\# of documents that contain the word } t}\right)$$

Term frequency

Inverse document frequency

Next time!

D8, EDA, etc...

D7 Demo