

CSC401/2511
ONE OF THE FIRST TUTORIALS EVER,
EVERYBODY SAYS SO

Frank Rudzicz

THE WORLD WE LIVE IN

Top stories



Man Charged in
Gunfire at Pizzeria
Cites Fake News of
'Child Sex Slaves'

The New York Times · 36



Comet Ping Pong
Gunman Facing 4
Charges

Washington Post · 55 mins



N.C. man told
police he was
armed to save
children and left Com...

Washington Post · 3 hours

→ More news for comet



THE WORLD WE LIVE IN



The_Donald before learning the Vegas shooter was an old white man

100 points 14 hours ago
+ This is why we do the travel ban
permalink source upvoted down score-953 percent report give gold reply hide child comments

100 points 14 hours ago
+ Fuck the travel ban. Forced deportation. 1 week or 10 years in jail. GTFO
permalink source upvoted down score-953 percent report give gold reply hide child comments
hide more comments (4 replies)

14 points 14 hours ago
I know what we are all thinking there is no doubt that this is an Islamic terror attack.

14 points 14 hours ago
Well, here's the culture war the left wants. Time to pick a side. Time to be as offensive as possible. Time to shantelope. Time to profile. Time to pack heat whenever you go. Time to shoot back.
permalink source down score-953 percent report give gold reply hide child comments

40 points 14 hours ago
Travel ban is useless, we need a Muslim Ban, and we need it yesterday!
permalink source upvoted down score-953 percent report give gold reply hide child comments

18 points 14 hours ago
+ Not only that, we need a Muslim Deportation. All visas revoked. All "refugees" need to leave now. That 9/11 does not belong in Western civilization.
permalink source down score-953 percent report give gold

18 points 14 hours ago
Terror attacks around the globe and now in the most harem city in the world...this is pretty obvious Islam is involved.
permalink source down score-953 percent report give gold reply hide child comments

14 points 14 hours ago
+ Is there any confirmation on the shooters being muslim?
permalink source upvoted down score-953 percent report give gold reply

18 points 14 hours ago
+ Mayor B Bell says, "Signs point to yeh."
permalink source upvoted down score-953 percent report give gold

15 points 14 hours ago
+ Outlook good
permalink source upvoted down score-953 percent report give

14 points 14 hours ago
RELIGION OF PEACE TRO
permalink source down score-953 percent report give gold reply hide child comments

14 points 14 hours ago
ISIS or ANTIFA? We will soon find out
permalink source upvoted down score-953 percent report give gold reply

After

1/11 88 points 1 hour ago

Something is not right here. This narrative doesn't make any sense, there is something big missing that we'll probably find out in the next few days.

1/11 4 points 1 hour ago

Don't you find it odd that the shooter had no motivation? Have you seen the born identity super soldier concept? They basically program shooters to act when given a command.

1/11 33 points 1 hour ago

This feels fishy. A 60 year old retiree manages to get a fully automatic weapon and several hundred rounds of ammo. He decides to go on a rampage and a country music concert? What are we missing here?

1/11 30 points 1 hour ago

Justin?

possible source added name name 100 points 1 hour ago

1/11 3 points 1 hour ago

This smells like a false flag, research this.

possible source added name name 100 points 1 hour ago

1/11 55 points 1 hour ago

Yeah a white supremacist is gonna shoot up a country music concert. Uh huh. Either false flag or demo-scare tactic. possible source added name name 100 points 1 hour ago

1/11 53 points 1 hour ago

Oh fuck, didn't Alex Jones very recently warn of an impending false flag?

possible source added name name 100 points 1 hour ago

1/11 43 points 1 hour ago

Is this shooter may have been FDS. Why did the FBI already blame this on terrorism... something smells. possible source added name name 100 points 1 hour ago

1/11 36 points 1 hour ago

Why is EVERY news outlet pushing the SDS claim?? This is really smelling like a false flag up to me now. Some is my main suspect.

possible source added name name 100 points 1 hour ago

1/11 40 points 1 hour ago

Both was shifty using this to attack suppressors and the ebull rns. Fuck that stupid cunt. This has false flag written all over it.

possible source added name name 100 points 1 hour ago

1/11 24 points 1 hour ago

Something about the known facts is very fishy right now.

We don't even know for sure at the moment. Paddock was the killer. They found him dead of an apparent suicide in the hotel room, but he might have been killed by whoever actually shot the guns and those people left before the police arrived. I'm not calling conspiracy yet but its probably way too early and the pieces we have so far don't really fit yet.



Silence is Consent

1 hr · 🌐

...

REPORT: Was Vegas Shooter Part of Antifa? Here's What The Media Isn't Saying (VIDEO)



REPORT: Was Vegas Shooter Part of Antifa? Here's What The Media Isn't Saying (VIDEO) · Silence is Consent

SILENCEISCONSENT.NET



Like



Comment



Share



👍👎😱 305

Top Comments ▾

243 Shares

54 Comments



WorldTruth.TV

22 hrs · 🌐



...

Five Things That Just Don't Add Up About The Las Vegas Mass Shooting



Five Things That Just Don't Add Up About The Las Vegas Mass Shooting

Our hearts and prayers go out to all those killed or injured in the Las Vegas shooting, and in a nation where so many anti-Americans are kneeling in...

WORLDTRUTH.TV | BY [EDDIE LEVIN](#)



Like



Comment



Share



👍😱😞 1.1K

[Top Comments](#) ▼

7,254 Shares

257 Comments



The People's Voice

17 hrs · 🌐

...

The mainstream media narrative about the Las Vegas shooting has been debunked by two videos proving there were multiple gunmen involved in an orchestrated attack.



Las Vegas: Video Confirms Multiple Shooters, Co-ordinated Attack

The mainstream media narrative about the shooting has been debunked by two videos proving there were at least two gunmen.

YOURNEWswire.COM | BY BAXTER DMITRY



Like



Comment



Share



2.1K

Top Comments ▾

6,072 Shares

541 Comments

LOGICAL FALLACIES

 strawman Misrepresenting someone's argument to make it easier to attack.	 false cause Presuming that a real or perceived relationship between things means that one is the cause of the other.	 appeal to emotion Handicapping an emotional response in place of a valid or compelling argument.	 the fallacy fallacy Presuming that because a claim has been poorly argued, or a fallacy has been made, that it is necessarily wrong.
 slippery slope Asserting that if we allow A to happen, then Z will consequently happen too, therefore A should not happen.	 ad hominem Attacking your opponent's character or personal traits instead of engaging with their argument.	 tu quoque Avoiding having to engage with criticism by turning it back on the criticiser - answering criticism with criticism.	 personal incredulity Saying that because one finds something difficult to understand that it's therefore not true.
 special pleading Moving the goalposts or making up exceptions when a claim is shown to be false.	 loaded question Asking a question that has an assumption built into it so that it can't be answered without appearing guilty.	 burden of proof Saying that the burden of proof lies not with the person making the claim, but with someone else to disprove.	 ambiguity Using double meanings or ambiguous language to mislead or misrepresent the truth.
 the gambler's fallacy Believing that 'runs' occur to statistically independent phenomena such as roulette wheel spins.	 bandwagon Appealing to popularity or the fact that many people do something as an attempted form of validation.	 no true scotsman Making what could be called an appeal to purity as a way to dismiss relevant criticisms or flaws of an argument.	 genetic Judging something good or bad on the basis of where it comes from, or from whom it comes.
 black-or-white Where two alternative states are presented as the only possibilities, when in fact more possibilities exist.	 begging the question A circular argument in which the conclusion is included in the premise.	 the texas sharpshooter Cherry-picking data clusters to suit an argument, or finding a pattern to fit a presumption.	 middle ground Saying that a compromise, or middle point, between two extremes is the truth.
 appeal to authority Using the opinion or position of an authority figure or institution of authority in place of an actual argument.	 composition/division Assuming that what's true about one part of something has to be applied to all, or other, parts of it.	 appeal to nature Making the argument that because something is 'natural' it is therefore valid, justifies, inevitable, or ideal.	 anecdotal Using personal experience or an isolated example instead of a valid argument, especially to dismiss statistics.

WHAT CAN BE DONE?

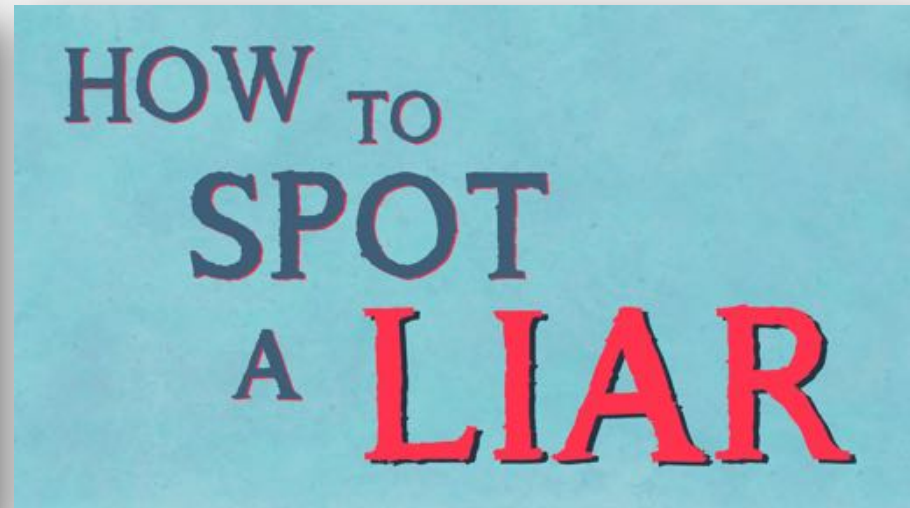
- There are probably many solutions, including better education and a ramping down of political zealotry from Our Glorious Leaders.
- But this is a class on natural language processing.
- *Can we detect bias automatically from online texts?*

LANGUAGE ANALYSIS AND LYING

“Don’t use a big word when a diminutive one
would suffice.”



<https://youtu.be/4ab2ZeZ-krY>



<https://youtu.be/H0-VVkpmTPrM>

REDDIT CORPUS

- We have curated data from Reddit by scraping subreddits, using Pushshift, by perceived political affiliation.

Left (598,944)	Center (599,872)	Right (600,002)	Alt (200,272)
twoXChromosomes (7,720,661)	news (2,782,9911)	theNewRight (19,466)	conspiracy (6,767,099)
occupyWallStreet (397,538)	politics (60,354,767)	whiteRights (118,008)	911truth (79,868)
lateStageCapitalism (634,962)	energy (416,926)	Libertarian (3,886,156)	
progressive (246,435)	canada (7,225,005)	AskTrumpSupporters (1,007,590)	
socialism (1,082,305)	worldnews (38,851,904)	The_Donald (21,792,999)	
demsocialist (5269)	law (464,236)	new_right (25,166)	
Liberal (151,350)		Conservative (1,929,977)	
		tea_party (1976)	

- These data are stored on the teach.cs servers under `/u/cs401/A1/data/`. These files should only be accessed from that directory (and not copied). All data are in the JSON format.

A COMMENT, IN JSON

```
{"id":"c05os7s", "author":"[deleted]",  
"subreddit":"conspiracy", "author_flair_css_class":null,  
"ups":-1, "archived":true, "edited":true,  
"subreddit_id":"t5_2qh4r", "body":"WAIT! Are you saying  
that 9/11 was a *conspiracy*?! Like...an *inside job* or  
something?", "score_hidden":false,  
"parent_id":"t3_74xuq", "distinguished":null,  
"link_id":"t3_74xuq", "author_flair_text":null,  
"created_utc":"1223008247",  
"retrieved_on":1425887728, "gilded":0, "name":"t1_c05os7s",  
"controversiality":0, "score":-1, "downs":0},
```

- If you want to experiment a bit, there are some fields of metadata that might be interesting, but the main thing is **body**.

THREE STEPS

- In order to **infer** whether the author of a given comment leans a certain way, politically, we use three steps:
 1. **Preprocess** the data, so that we can extract meaningful information, and remove distracting 'noise'.
 2. **Extract** meaningful information.
 3. **Train** classifiers, given labeled data.



PREPROCESSING |

1. Remove all newline characters.
2. Replace HTML character codes (i.e., &...;) with their ASCII equivalent.
3. Remove all URLs (i.e., tokens beginning with *http* or *www*).
4. Split each punctuation (see `string.punctuation`) into its own token using whitespace except:
 1. Apostrophes.
 2. Periods in abbreviations (e.g., e.g.) are not split from their tokens. E.g., e.g. stays e.g.
 3. Multiple punctuation (e.g., !?!, ...) are not split internally. E.g., *Hi!!!* becomes *Hi !!!*
 4. You can handle single hyphens (-) between words as you please.
5. Split clitics using whitespace.
 1. Clitics are contracted forms of words, such as *n't*, that are concatenated with the previous word.
 2. Note: the possessive 's has its own tag and is distinct from the clitic 's, but also must be separated by a space; likewise, the possessive on plurals must be separated (e.g., dogs 's).

PREPROCESSING 2

6. Each token is tagged with its part-of-speech using spaCy (see below).
 1. A tagged token consists of a word, the '/' symbol, and the tag (e.g., *dog/NN*). See below for information on how to use the tagging module. The tagger can make mistakes.
7. Remove stopwords. See `/u/cs401/Wordlists/StopWords`.
8. Apply **lemmatization** using spaCy (see below).
9. Add a newline between each sentence.
 1. This will require detecting end-of-sentence punctuation. Some punctuation does not end a sentence; see standard abbreviations here: `/u/cs401/Wordlists/abbrev.english`.
 2. It can be difficult to detect when an abbreviation ends a sentence; e.g., in *Go to St. John's St. John is there.*, the first period is used in an abbreviation, the last period ends a sentence, and the second period is used both in an abbreviation and an end-of-sentence.
 3. You are not expected to write a 'perfect' pre-processor (none exists!), but merely to use your best judgment in writing heuristics; see section 4.2.4 of the Manning and Schütze text.
10. Convert text to lowercase.

LEMMATIZATION V STEMMING: DAWN OF SPARSENESS

- Both **lemmatization** and **stemming** are often used to transform word tokens to a more base form.
 - This helps to improve sparseness.
 - It also helps in using various resources.
 - e.g., *funkilicious* might not exist in a norm or embedding, but '*funk*' ought to).

LEMMATIZATION V STEMMING: DAWN OF SPARSENESS

- **lemma:** *n.* an abstract conceptual form of a word that has been mentally selected for utterance in the early stages of speech production.
 - E.g., $\text{lemma}(\text{best}) = \text{good}$ (degree)
 - E.g., $\text{lemma}(\text{words}) = \text{word}$ (number/amount)
- **stem:** *n.* usually, a part of a word to which affixes can be attached.
 - E.g. $\text{stem}(\text{houses}) = \text{stem}(\text{housing}) = \text{hous}$
- We use lemmatization given some of our features, but check out `nltk.stem` in the [NLTK](#) package.

PREPROCESSING: YOUR TASK

- Copy the template from `/u/cs401/A1/code/a1_preproc.py`. There are two functions you need to modify:
 - In `preproc1`, perform each preprocessing step above.
 - In `main`, replace the lines marked with **TODO** with the code they describe. Add a new **cat** field with the name of the class
- The program takes three arguments:
 1. your student ID (mandatory),
 2. the output file (mandatory), and
 3. the maximum **number of lines** to sample from each category file (optional; default=10,000).

```
python a1_preproc.py 999123456 -o preproc.json
```

SPACY.IO

NLP IN PYTHON



```
import spacy

nlp = spacy.load('en', disable=['parser', 'ner'])
utt = nlp(u"I know the best words")
for token in utt:
    ...     print(token.text, token.lemma_, token.pos_, token.tag_, token.dep_,
    ...           token.shape_, token.is_alpha, token.is_stop)
```

See next tutorial for details on how to handle tokenization in spaCy.

PREPROCESSING: SUBSAMPLING

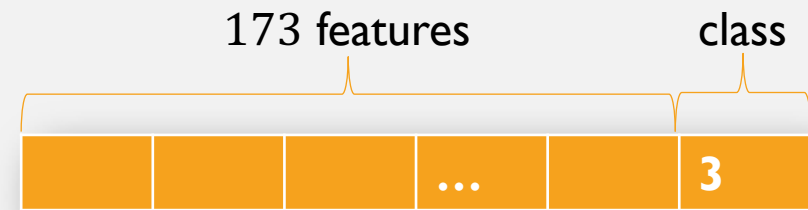
- We provide our student IDs so we each see a different part of the available data.
 - By default, you should only sample 10,000 lines from each of the `Left`, `Centre`, `Right`, and `Alt` files, for a total of 40,000 lines.
 - From each file, start sampling lines at index $[ID \% \text{len}(X)]$
- Feel free to play around with more or less data, *respectful of your peers on the servers*, but this step guarantees it's tractable (and that there's no 'desired' level of accuracy).

FEATURE EXTRACTION

- The `al_extractFeatures.py` program reads a preprocessed JSON file and extracts features for each comment therein, producing and saving a $D \times 174$ NumPy array, where the i^{th} row is the features for the i^{th} comment, followed by an integer for the class (0: Left, 1: Center, 2: Right, 3: Alt), as per the cat JSON.

```
{"id": "c05os7s",  
"body": "wait ! be you say  
that 9 / 11 be a *  
conspiracy *?! like ...  
an * inside job * or  
something ?", "cat": "Alt"}",
```

i^{th} comment in input



i^{th} row in output

1. Number of first-person pronouns
2. Number of second-person pronouns
3. Number of third-person pronouns
4. Number of coordinating conjunctions
5. Number of past-tense verbs
6. Number of future-tense verbs
7. Number of commas
8. Number of multi-character punctuation tokens
9. Number of common nouns
10. Number of proper nouns
11. Number of adverbs
12. Number of *wh*- words
13. Number of slang acronyms
14. Number of words in uppercase (≥ 3 letters long)
15. Average length of sentences, in tokens
16. Average length of tokens, excluding punctuation-only tokens, in characters
17. Number of sentences.

18. Average of AoA (100-700) from Bristol, Gilhooly, and Logie norms
19. Average of IMG from Bristol, Gilhooly, and Logie norms
20. Average of FAM from Bristol, Gilhooly, and Logie norms
21. Standard deviation of AoA (100-700) from Bristol, Gilhooly, and Logie norms
22. Standard deviation of IMG from Bristol, Gilhooly, and Logie norms
23. Standard deviation of FAM from Bristol, Gilhooly, and Logie norms

24. Average of V.Mean.Sum from Warringer norms
25. Average of A.Mean.Sum from Warringer norms
26. Average of D.Mean.Sum from Warringer norms
27. Standard deviation of V.Mean.Sum from Warringer norms
28. Standard deviation of A.Mean.Sum from Warringer norms
29. Standard deviation of D.Mean.Sum from Warringer norms

	A	B	C	D	E	F	G	H	I
1		Word	V.Mean.Sum	V.SD.Sum	V.Rat.Sum	A.Mean.Sum	A.SD.Sum	A.Rat.Sum	D.Mean.Sum
2	1	aardvark	6.26	2.21	19	2.41	1.4	22	4.27
3	2	abalone	5.3	1.59	20	2.65	1.9	20	4.95
4	3	abandon	2.84	1.54	19	3.73	2.43	22	3.32
5	4	abandonmer	2.63	1.74	19	4.95	2.64	21	2.64
6	5	abbey	5.85	1.69	20	2.2	1.7	20	5
7	6	abdomen	5.43	1.75	21	3.68	2.23	22	5.15
8	7	abdominal	4.48	1.59	23	3.5	1.82	22	5.32
9	8	abduct	2.42	1.61	19	5.9	2.57	20	2.75

Warringer: These norms measure the valence (V), arousal (A), and dominance (D) of each **lemma**, according to the VAD model of human affect and emotion.

See: Warriner, A.B., Kuperman, V., & Brysbaert, M. (2013). [Norms of valence, arousal, and dominance for 13,915 English lemmas](#). *Behavior Research Methods*, **45**:1191-1207.

	A	B	C	D	E	F	G	H	I
1	Source	WORD	AoA (Yrs)	AoA (100-70)	IMG	FAM	Length (Letters)		
2	GL	abandonmer	NA	359	348	359	11		
3	GL	abatement	NA	294	189	294	9		
4	BN	abbey	7.8	480	575	429	5		
5	GL	abdomen	NA	426	548	426	7		
6	BN	abide	8.8	533	460	387	5		

Bristol et al: measure the age-of-acquisition (AoA), imageability (IMG), and familiarity (FAM) of each word, which we can use to measure lexical complexity.

See: Gilhooly, KJ, Logie, RH (1980). [Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words](#) *Behavior Research Methods*, **12**(4):394-427.

LIWC/RECEPTIVITI |

- The **Linguistic Inquiry & Word Count (LIWC)** tool has been a standard in a variety of NLP research, especially around authorship and sentiment analysis.
 - This tool provides 85 measures mostly related to word choice.
- The company **Receptiviti** provides a superset of these features, which also includes 59 measures of personality derived from text.
- To simplify things, we have already extracted these 144 features for you. Simply copy the pre-computed features from the appropriate uncompressed `npz` files stored in `/u/cs401/A1/feats/`.

LIWC/RECEPTIVITY 2

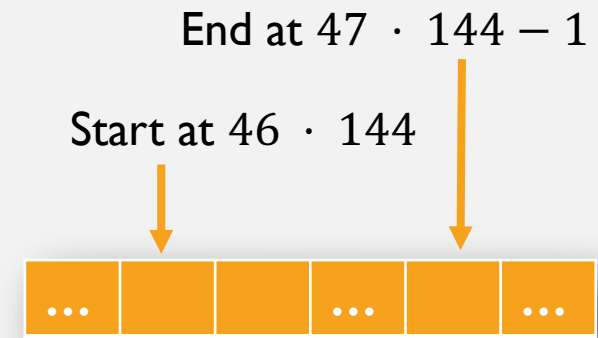
- Comment IDs are stored in `_IDs.txt` files (e.g., `Alt_Ids.txt`). When processing a comment, find the index (row) i of the ID in the appropriate ID text file, for the category, and copy the 144 elements, starting at element $i \cdot 144$, from the associated `feats.dat.npy` file.

```
"{"id": "c05os7s",  
"body": "wait ! be you  
say that 9 / 11 be a  
* conspiracy *?! like  
... an * inside job *  
or something ?",  
cat: "Alt"}",
```

comment

```
...  
c05nn92  
c05o811  
c05os7s ← 46th line  
c05p5vj  
c05pbbg  
...
```

Alt_Ids.txt



Alt_feats.dat.npy

LIWC/RECEPTIVITY 3: FEATURE NAMES

```
liwc_sexual
liwc_shehe
liwc_social
liwc_space
liwc_swear
liwc_tentat
liwc_they
liwc_time
liwc_verb
liwc_we
liwc_work
liwc_you
receptiviti_active
receptiviti_adjustment
receptiviti_adventurous
receptiviti_aggressive
receptiviti_agreeableness
receptiviti_ambitious
receptiviti_anxious
receptiviti_artistic
receptiviti_assertive
receptiviti_body_focus
```

feats.txt

CLASSIFICATION

- Four parts:
 - Compare classifiers
 - Experiment with the amount of training data used
 - Select the best features for classification
 - Do cross-fold validation

CLASSIFICATION I: COMPARE CLASSIFIERS

- *Randomly* split data into 80% training, 20% testing.



- We have 5 classification methods, which you can consider to be ‘black boxes’ (input goes in, classes come out).
 1. Support vector machine with linear kernel
 2. Support vector machine with radial basis kernel
 3. Random forest classifier
 4. Neural network
 5. Adaboost (with decision tree)

CLASSIFICATION I: COMPARE CLASSIFIERS

- **Accuracy:** the total number of correctly classified instances over all classifications: $A = \frac{\sum_i c_{i,i}}{\sum_{i,j} c_{i,j}}$.
- **Recall:** for each class κ , the fraction of cases that are truly class κ that were classified as κ : $R(\kappa) = \frac{c_{\kappa,\kappa}}{\sum_j c_{\kappa,j}}$
- **Precision:** for each class κ , the fraction of cases classified as κ that truly are κ : $P(\kappa) = \frac{c_{\kappa,\kappa}}{\sum_i c_{i,\kappa}}$

True class

$c_{i,j}$: number of times class i
was classified as class j

Predicted class				
	L	C	R	A
L				
C		8		
R				
A				

CLASSIFICATION 2: AMOUNT OF DATA

- You previously used a random $0.8 \cdot 40K = 32K$ comments to train.
- Using the classifier with the highest accuracy from Sec3.1, retrain the system using an arbitrary $1K, 5K, 10K, 15K, 20K$ samples from the original $32K$.
- One might expect something like a logarithmic growth in accuracy over training set size, but do we (see|expect) that?

CLASSIFICATION 3: FEATURE ANALYSIS

- Certain features may be more or less useful for classification, and too many can lead to various problems.
- Here, you will select the best k features for classification for $k = \{5, 10, 20, 30, 40, 50\}$.
- Train the best classifier from Sec3.1 on just $k = 5$ features on both 1K and 32K training samples.
- Are some features always useful? Are they useful to the same degree (p -value)? Why are certain features chosen and not others?

CLASSIFICATION 4: CROSS-FOLD VALIDATION

- What if the 'best' classifier from Sec3.1 only appeared to be the best because of a random accident of sampling?
- Test your claims more rigorously.

	Part 1	Part 2	Part 3	Part 4	Part 5	
Iteration 1						: Err1 %
Iteration 2						: Err2 %
Iteration 3						: Err3 %
Iteration 4						: Err4 %
Iteration 5						: Err5 %

	Testing Set
	Training Set

BONUS

- You have **complete freedom** to expand on this assignment in any way you choose.
- You should have no expectation to the *value* of such an exploration – **check with us** (privately if you want) about the appropriateness of your idea.
- Bonus marks can make up for marks lost in other sections of the assignment, but your overall mark **cannot exceed 100%**.

FESTIVAL DE MIERDA DE TORO

- If things go well, we would love to run a special ‘workshop’ where:
 1. students who did interesting **bonuses** could describe their work
 2. grad students (working around the theme) could present their **projects**
 3. we could hold a **competition** for best systems in A1, A2, A3
- Problem: the TAs are already doing a lot of work, and I’m fairly busy.
- Solution (?): If any of you are interested in spearheading such a get-together at the end of the term (and getting bonus marks), we’d be glad to support.



RULES OF LOGARITHMS

- You may need these in later assignments:
 - **Definition:** $\log_a x = N \leftrightarrow a^N = x$
 - **Product:** $\log_a(xy) = \log_a x + \log_a y$
 - **Quotient:** $\log_a\left(\frac{x}{y}\right) = \log_a x - \log_a y$
 - **Power:** $\log_a(x^p) = p \log_a x$
 - **Base change:** $\log_a x = \frac{\log_b x}{\log_b a}$
- Reminder: avoid common logarithmotechnic errors:
 - $\log_a(x + y) \neq \log_a x + \log_a y$
 - $\log_a(x - y) \neq \log_a x - \log_a y$