Guess the Class

Richard Duong, Shreya Balaji, Benson Wan

Project Proposal

- Classify lectures to a subject using YouTube captions
- Explore relationships between words used in different classes

Data Collection

- 13 class subjects
- 150 topics
- 4,000 video captions
- Over 18 million words
- Class subjects :

subject	subject name	topic
ENGL	English	American Literature
POSC	Political Science	American Politics
CHEM	Chemistry	Analytical Chemistry
BIOL	Biology	Anatomy
HIST	History	Ancient Greek History
PHIL	Philosophy	Ancient Greek or Roman Philosophy
ENGL	English	Beginning Composition
CHEM	Chemistry	Biochemistry
CHEM	Chemistry	Bioinorganic Chemistry
PSYC	Psychology	Brain and Behavior

- BIOL
- HIST
- BUS
- MATH
- CHE
- PHIL
- CHEM
- PHYS
- CS
- POSC
- PSYC
- ECON
- ENGL

Data Cleaning

- Uncleaned Captions
  - Lecture captions from YouTube
  - Mass downloaded playlists
  - Used pytube library

Uncleaned

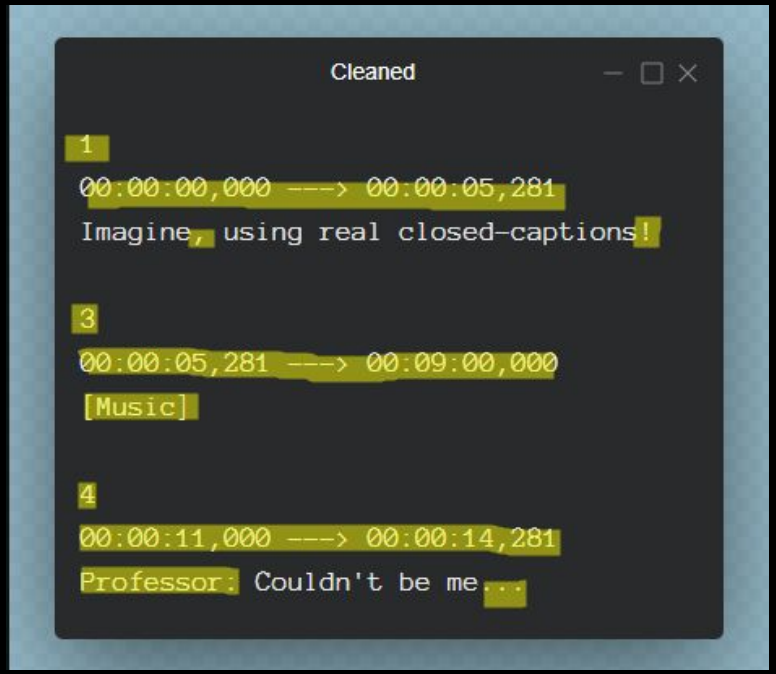
```
1
00:00:00,000 ---> 00:00:05,281
Imagine, using real closed-captions!

3
00:00:05,281 ---> 00:09:00,000
[Music]

4
00:00:11,000 ---> 00:00:14,281
Professor: Couldn't be me...
```

Data Cleaning

- Cleaned Captions
- Used regular expressions to filter
  - Caption numbers
  - Timestamps
  - Punctuation
  - Actions
  - Speakers



Data Cleaning

- Cleaned Captions
- Results:
  - Lowercase all letters
  - Preserve hyphens
  - Preserve apostrophes

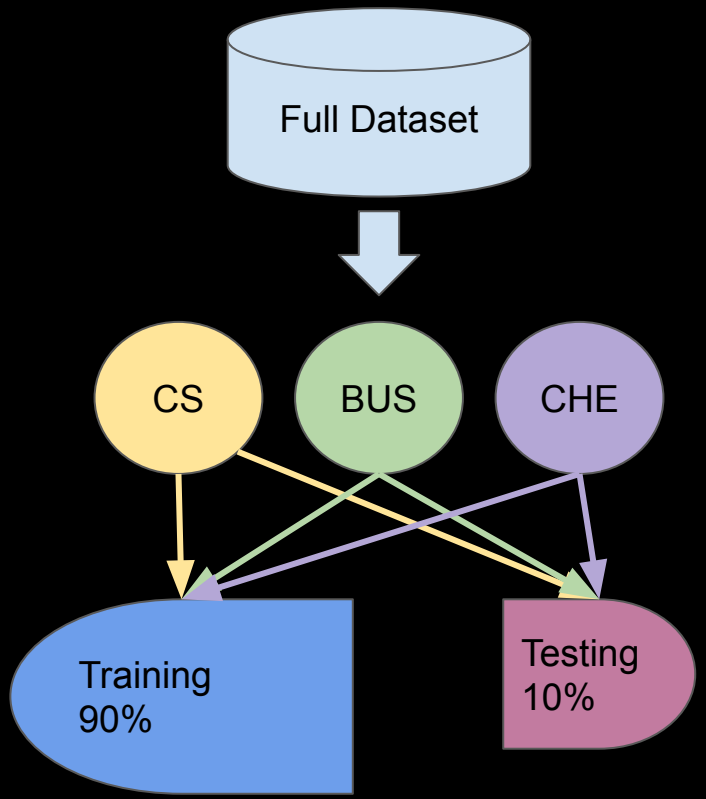
Cleaned

```
imagine using real closed-captions

couldn't be me
```

KNN Classifier

- Stratified sampling for model
  - 200 captions from each subject
    - 180 captions for training
    - 20 captions for testing
- TF-IDF Vectorization
  - Snowball Stemmer (Porter2)
  - nltk english stop words



KNN Classifier

- Developed 6 different models
  - Tuned with different parameters
- Best model can predict class subject with 90.77% accuracy!

Model Description	stop_words	ngram_range	min_df	max_df	k-value	accuracy
default (no tuning)	english	(1, 1)	1	1	100	8.40%
remove top 30% DF	english	(1, 1)	10	0.7	12	90.38%
remove top 20% DF	english	(1,1)	10	0.8	13	90.76%
unigram and bigram	english	(1, 2)	10	0.75	20	90.77%
only bigram	english	(2, 2)	5	0.75	5	89.23%

KNN Classifier

- Model #6 has accuracy of 96.11%!
  - Removed several classes
    - POSC and ENGL lacking quality captions
    - CHE and PHYS are too similar to CHEM

Model Description	stop_words	ngram_range	min_df	max_df	k-value	accuracy
illegitimate model	english	(1, 1)	10	0.75	34	96.11%

KNN Classifier

- Classifier can be used in many ways!
  - Correlate words/phrases with a subject
  - Identify linguistic choices across lecturers of different disciplines

Question:

What subject do you think the phrase "Where are your cameras" belongs to?

KNN Classifier

Answer: "Where are your cameras" is classified as BUS (Business)

- The classifier is designed to work with full transcripts
  - Failed to classify this phrase to CS

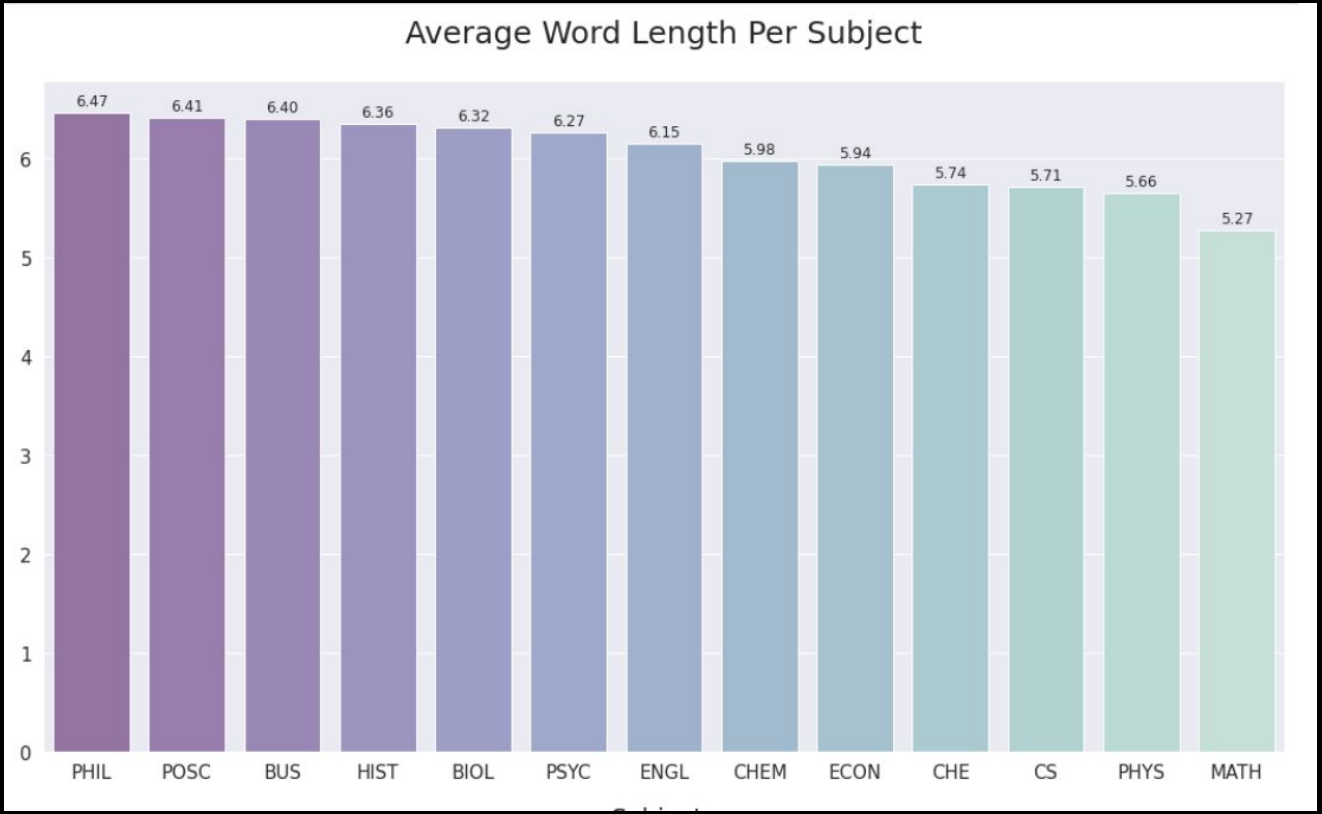
Other Interesting Predictions

phrase	linked lists	vector	nike	number	2020	matrix	celsius	uh
prediction	CS	PHYS	BUS	ENGL	POSC	MATH	CHEM	CS

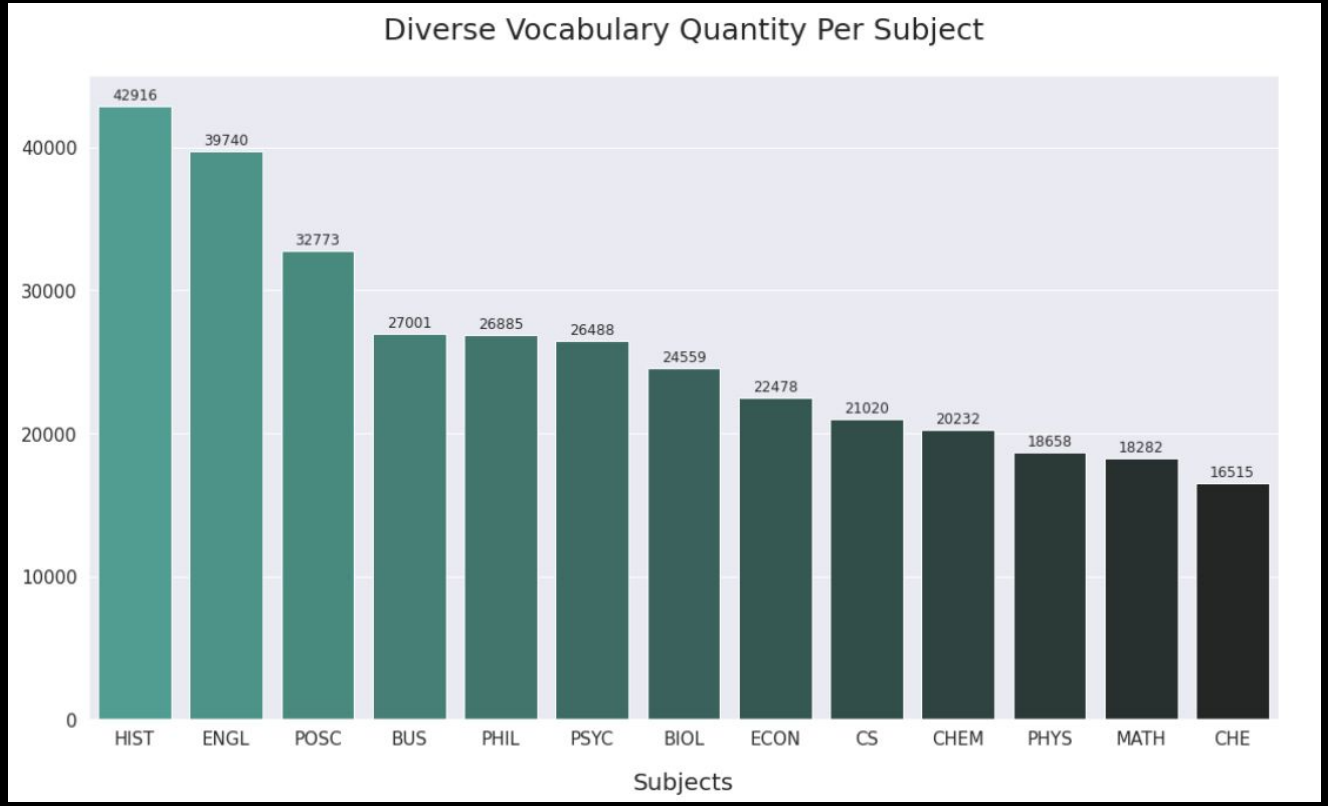
Data Analysis

- We want to know which subject uses
  - longest words?
  - most diverse vocabulary?
  - most filler words? (ah, uh, um)
  - most bad words?
  - most numbers?
  - most dates? (2020, 1950s)

Longest Words



Diverse Vocabulary

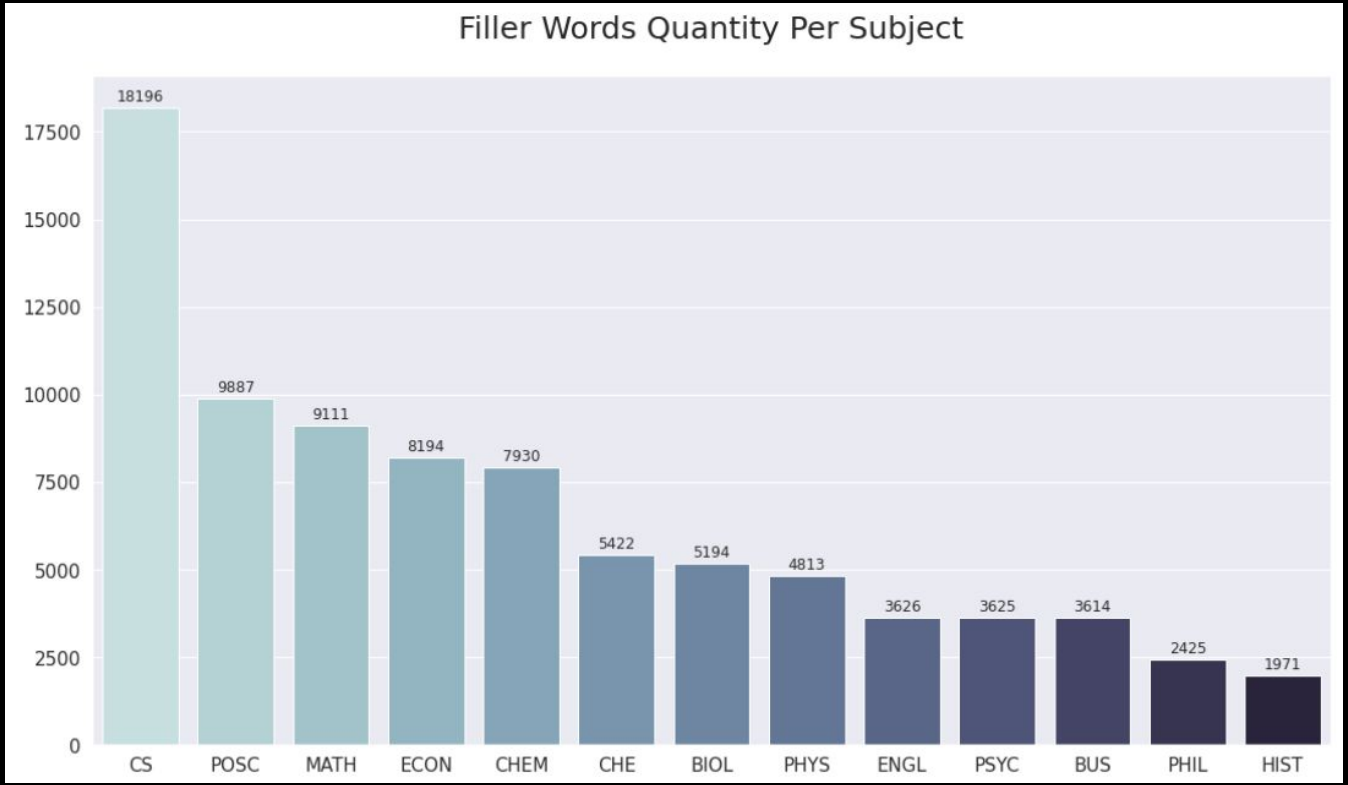


Filler Words

- Filler words can break into 2 categories:
  - “um”, “ah”, “er”
    - identifiable, no context
  - “like”, “you know”, “basically”
    - hard to identify, needs context
- We will be using no context fillers

uh	uhh
um	umm
hm	hmm
er	err
ah	ahh
eh	ehh
huh	huhh
oh	ohh

Filler Words

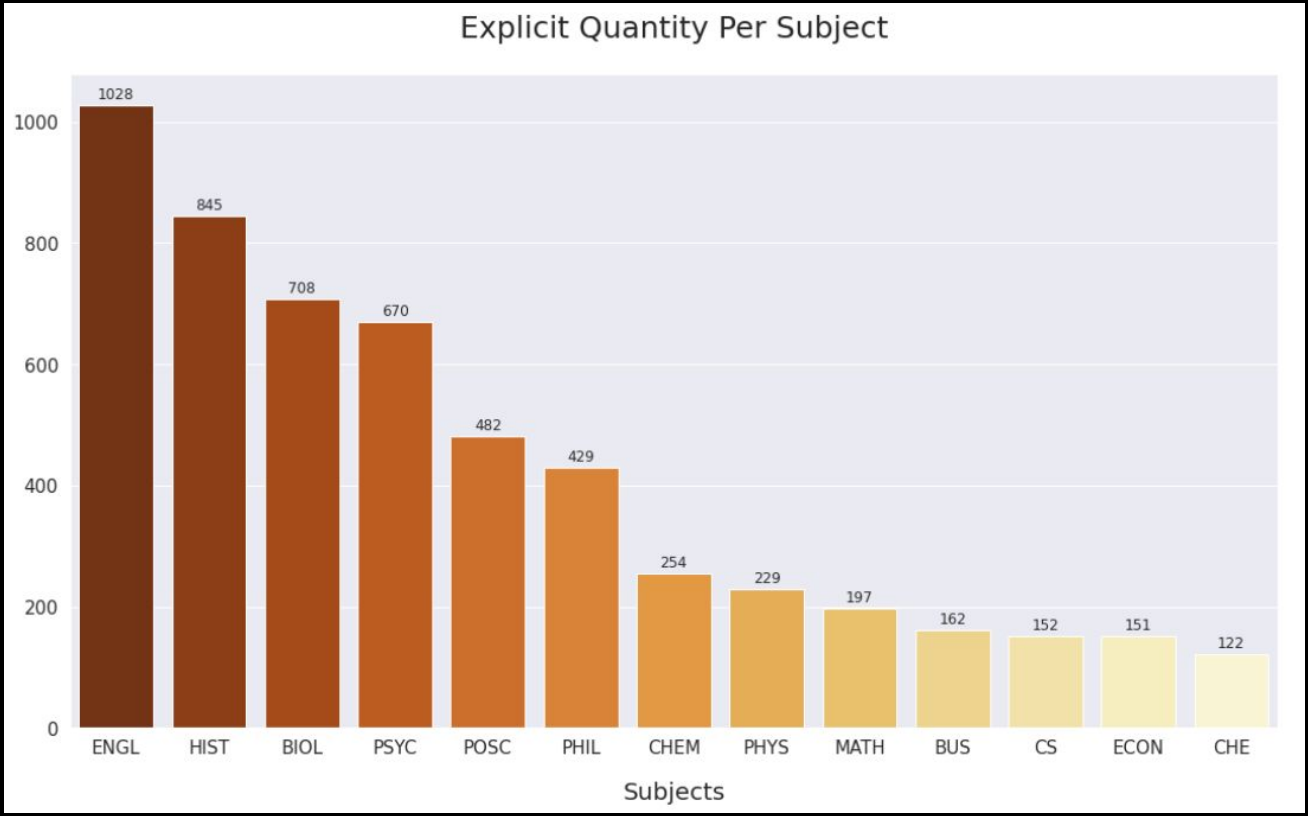


Bad Words

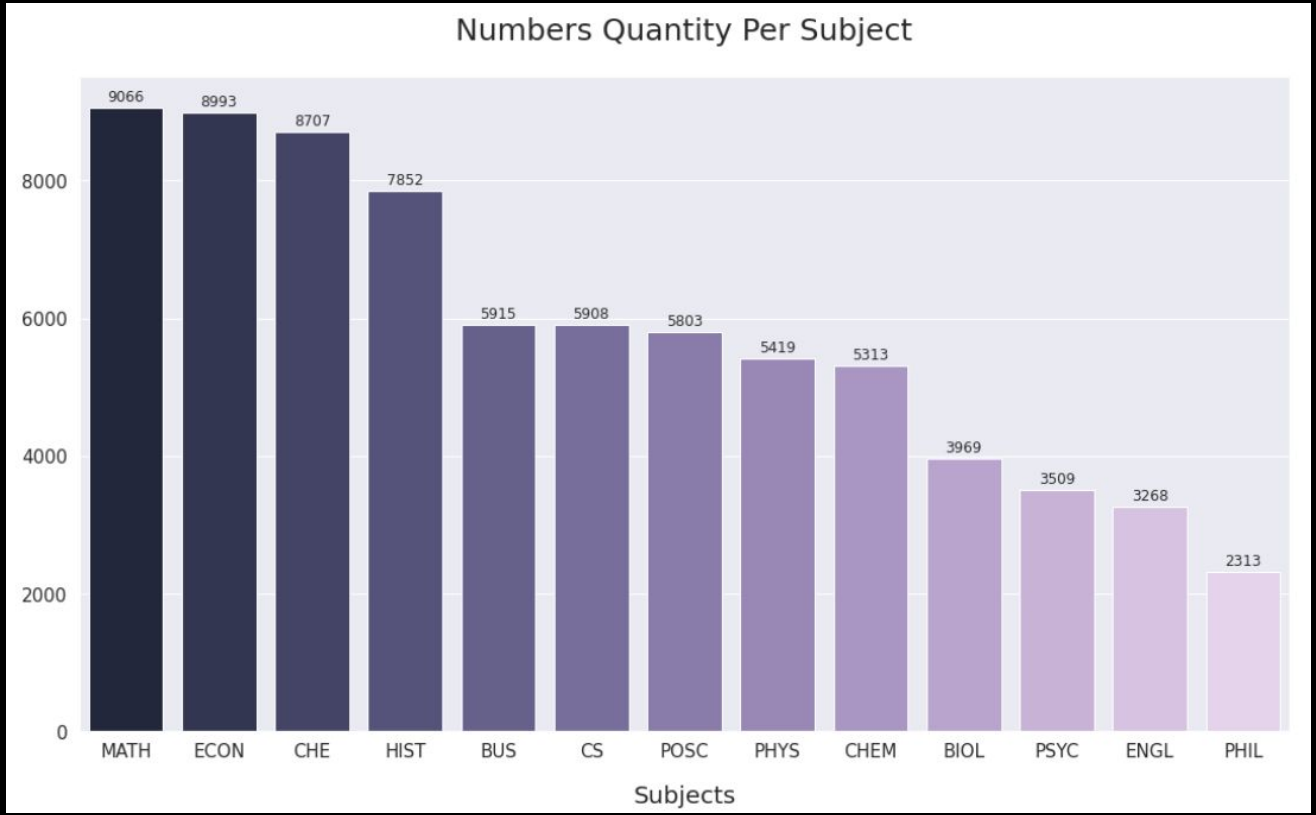
- There are an endless number of explicit words
- We used a 1200 word list compiled by bebrilliant.com
- Here's a list of some "safe" bad words
  - We apologize if this list offends anyone

sex	lmfao	flamer	damn	f4nny
crap	eunuch	xxx	erotic	ecchi

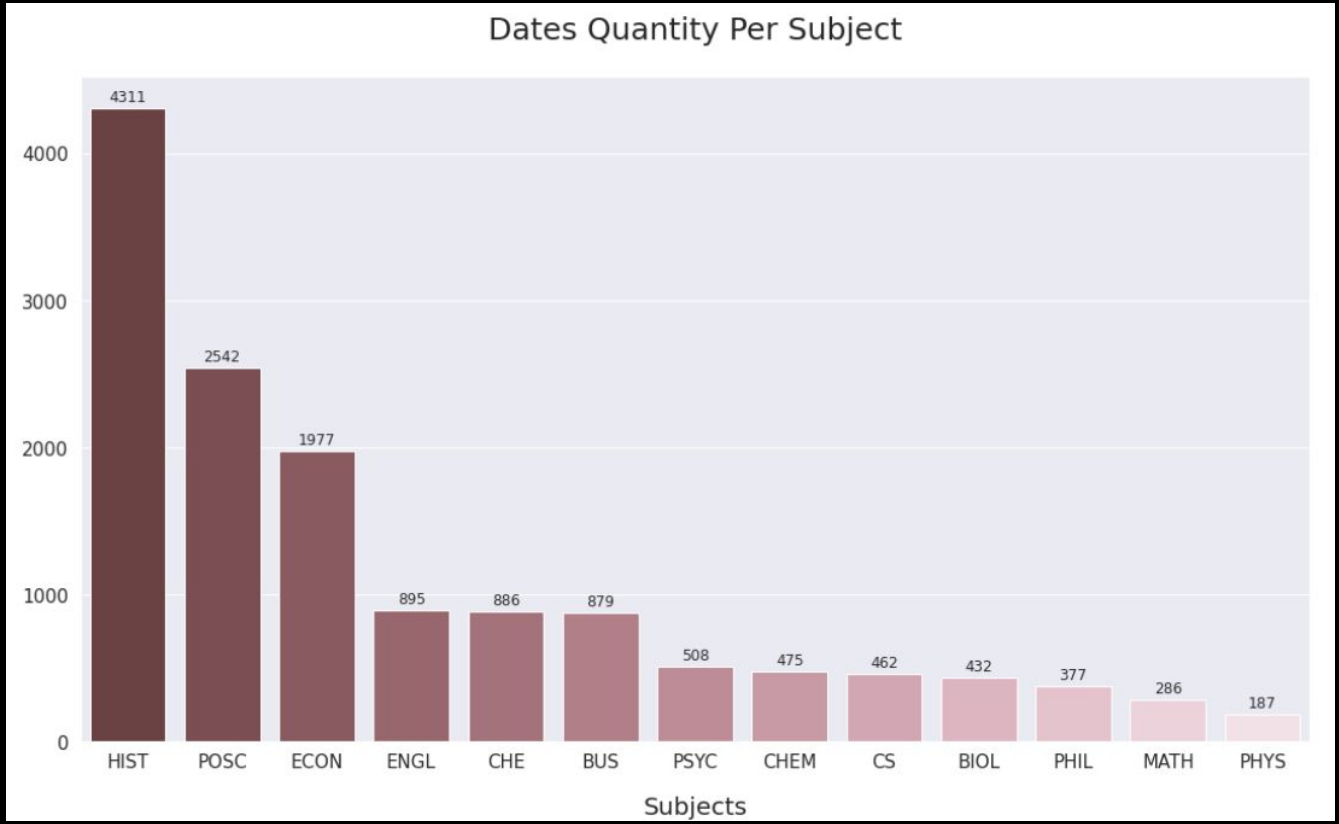
Bad Words


Explicit Quantity Per Subject
1028
845
708
670
482
429
254
229
197
162
152
151
122
1000
800
600
400
200
0
ENGL
HIST
BIOL
PSYC
POSC
PHIL
CHEM
PHYS
MATH
BUS
CS
ECON
CHE
Subjects

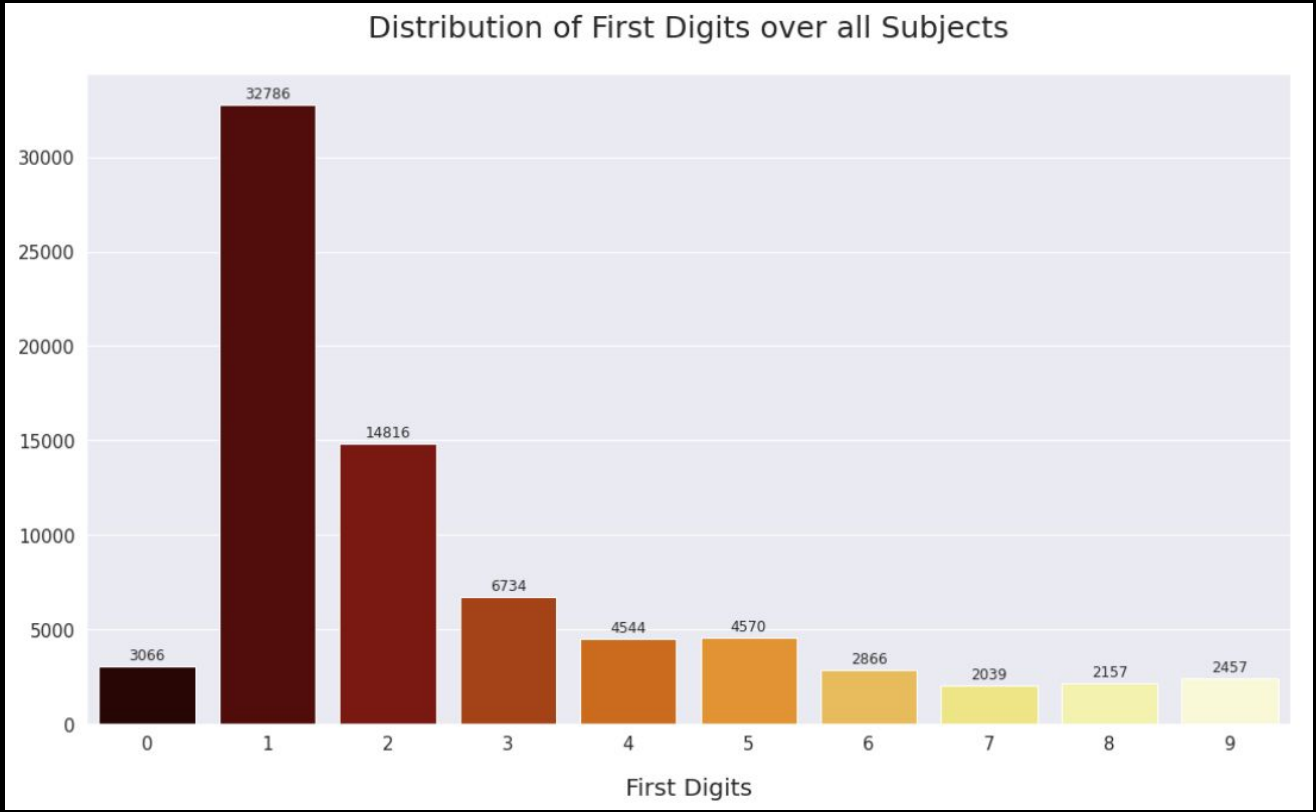
Most Numbers



Most Dates


Dates Quantity Per Subject
4311
2542
1977
895
886
879
508
475
462
432
377
286
187
4000
3000
2000
1000
0
HIST
POSC
ECON
ENGL
CHE
BUS
PSYC
CHEM
CS
BIOL
PHIL
MATH
PHYS
Subjects

Bonus! First Digits



Conclusion

- Build dataset using pytube
- Clean data with regex
- Create classifier using K-nearest neighbors
- Explore data using the bag of words model

Thank You

Richard Duong, Shreya Balaji, Benson Wan