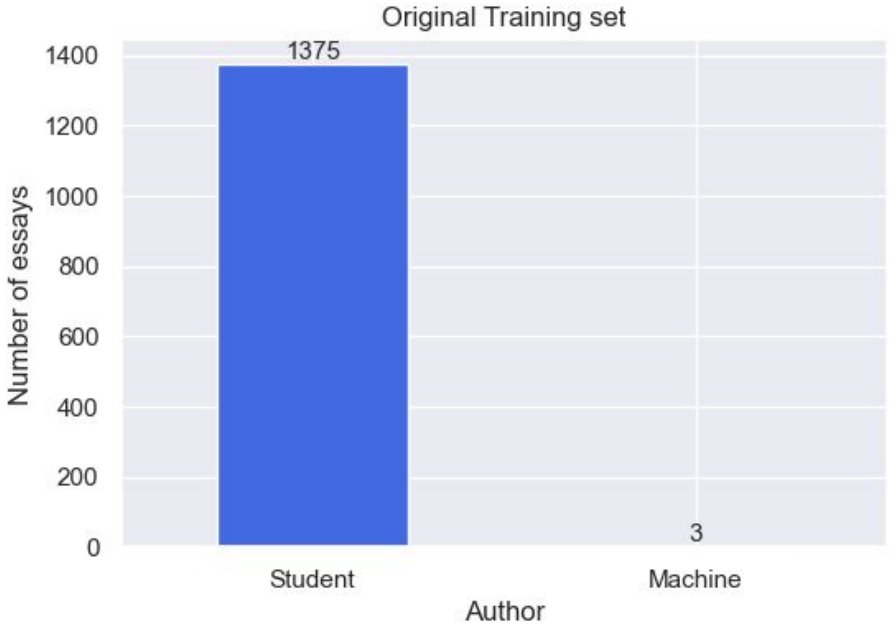
LLM text detection

By Phevos Margonis

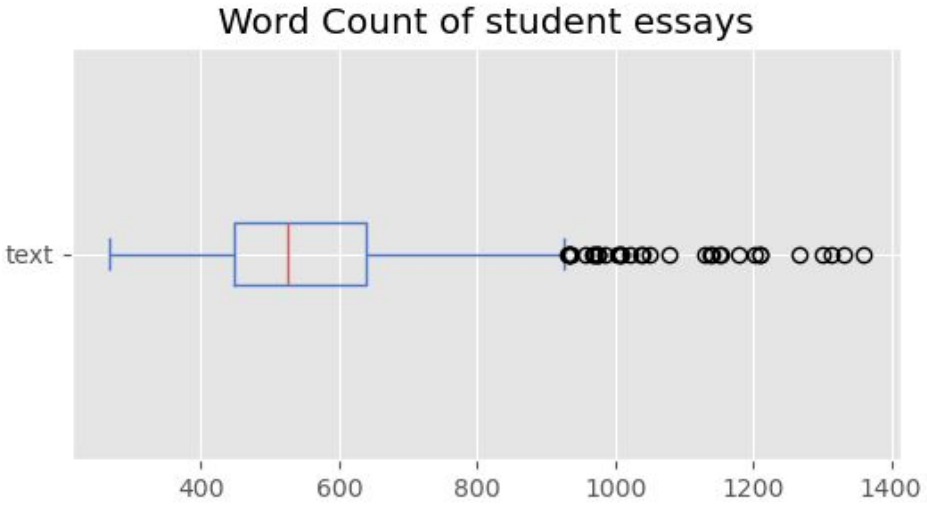
Imbalance

- Lack of Ai essays
- Lack of test set



Topics

- Explanatory Essay on *“Limiting Car Usage”*
- Letter to State Senator on *“Electoral College vs. Popular Vote”*



Prompting

- GPT-3.5
- Bard
  - **Zero-shot:** Act as a high school student. Write an explanatory essay to ...
  - **One-shot:** ...Use this as inspiration: *source text*
  - **One-shot:** ...Here is an example: *student essay*
  - **Few-shot:** *Combinations of the above.*

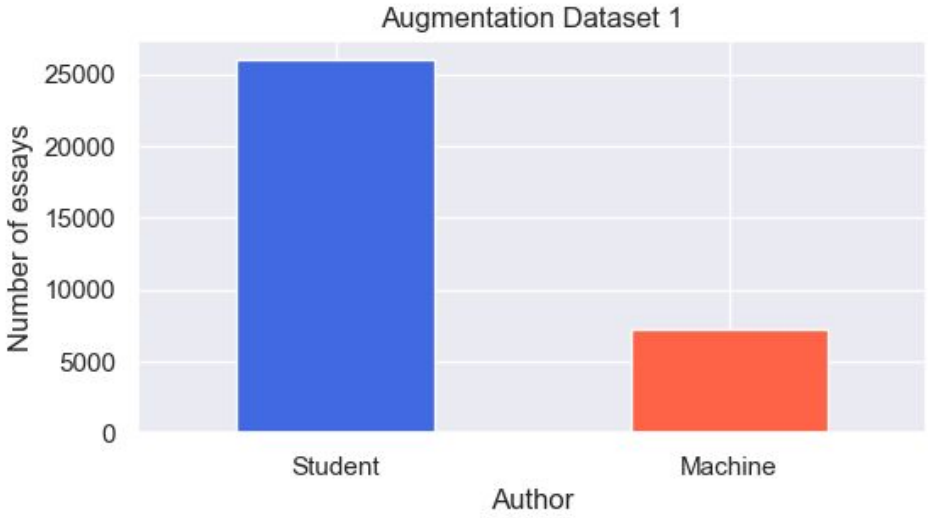
Problems

...

- Volume
- Simple models VS Advanced LLMs
- Topic Bias

Augmentation

- DAIGT Proper Train Dataset
  - Topic variation
  - New student essays
  - Different LLMs
    (up to 180B parameters)



Train Set

- Student essays
- Ai essays:
  - Personally Generated
  - Third Party

Test Set

- 10.000 new essays

Classification benchmark

- TF-IDF representations:

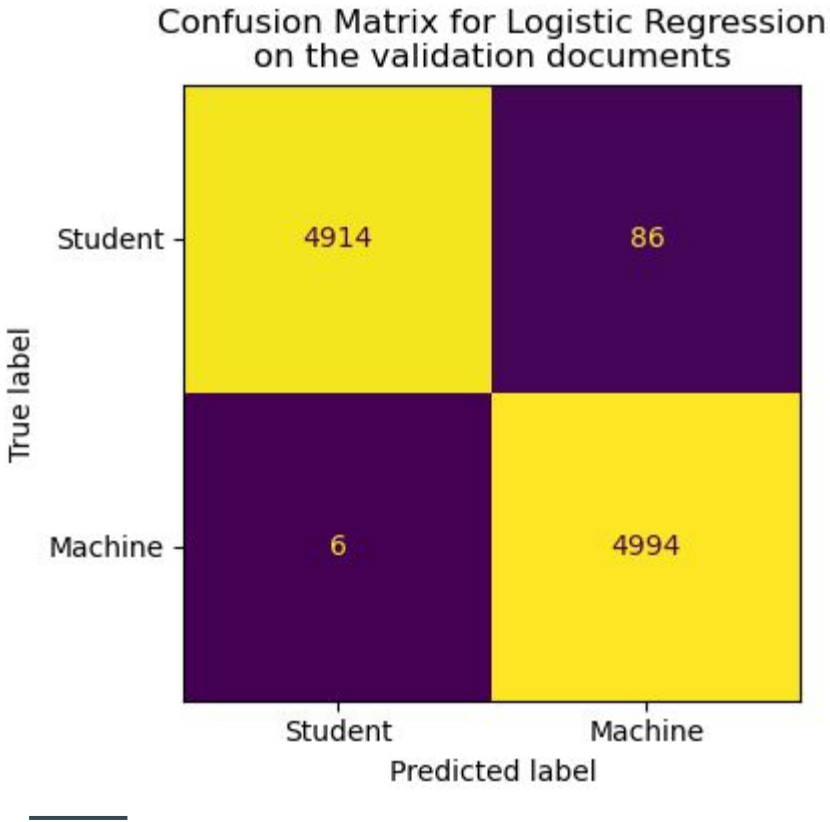
Model	F1 Score
Logistic Regression	98%
Ridge Classifier	98%
Linear SVC	98%
Nearest Centroid	85%
Complement Naive Bayes	94%

Results

	precision	recall	f1-score	support
0	1.00	0.98	0.99	5000
1	0.98	1.00	0.99	5000
accuracy			0.99	10000
macro avg	0.99	0.99	0.99	10000
weighted avg	0.99	0.99	0.99	10000

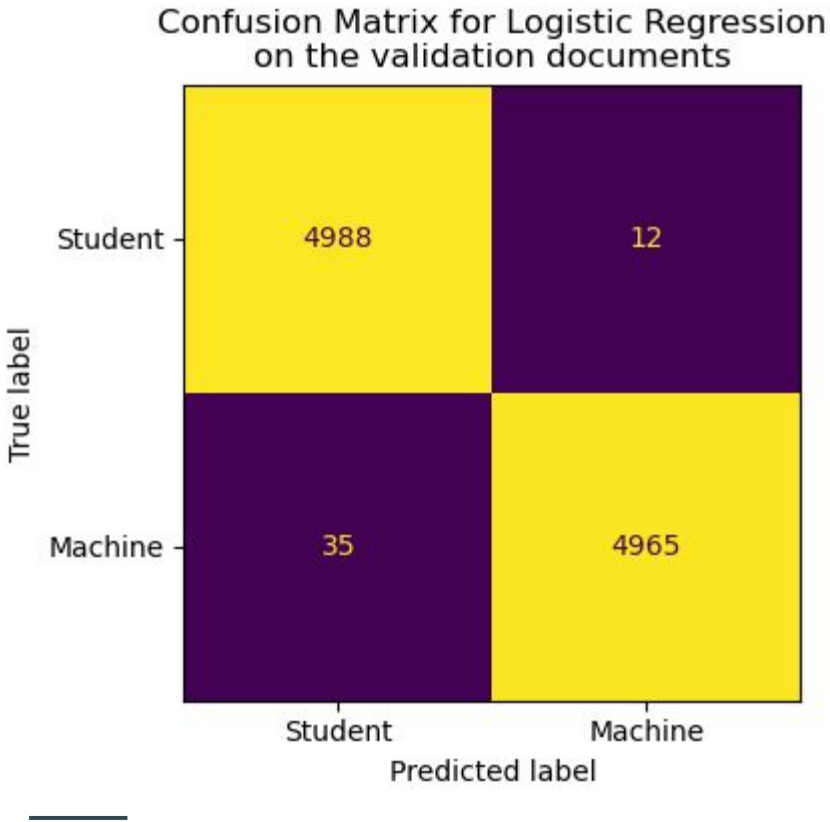
Drill Down

- Error Trade-off:
  - Type I
  - Type II



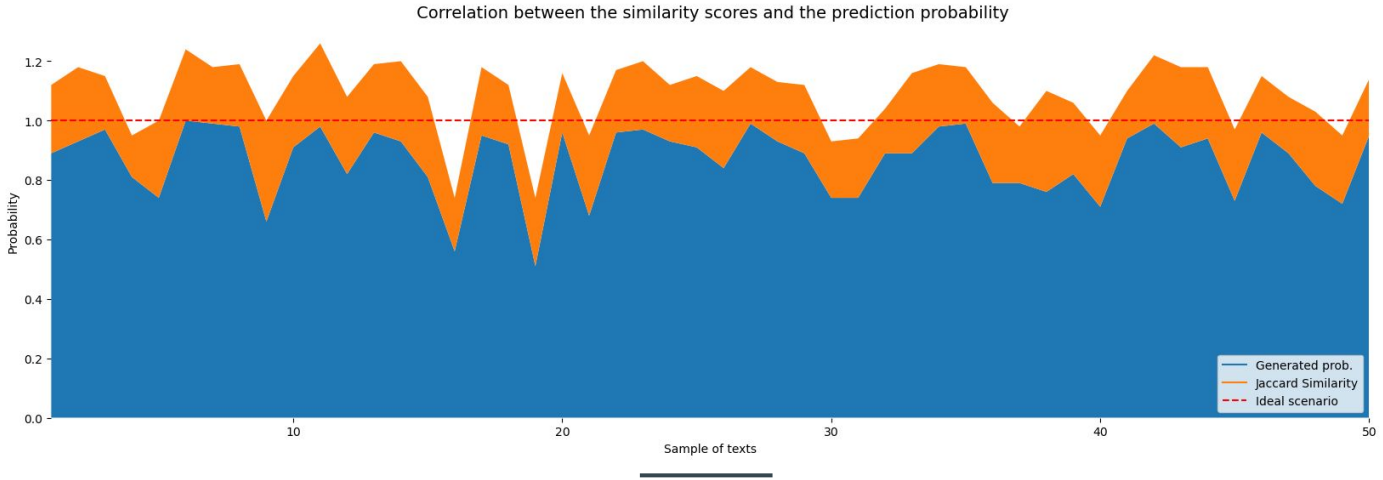
Class Balance

- Adjusted class weights



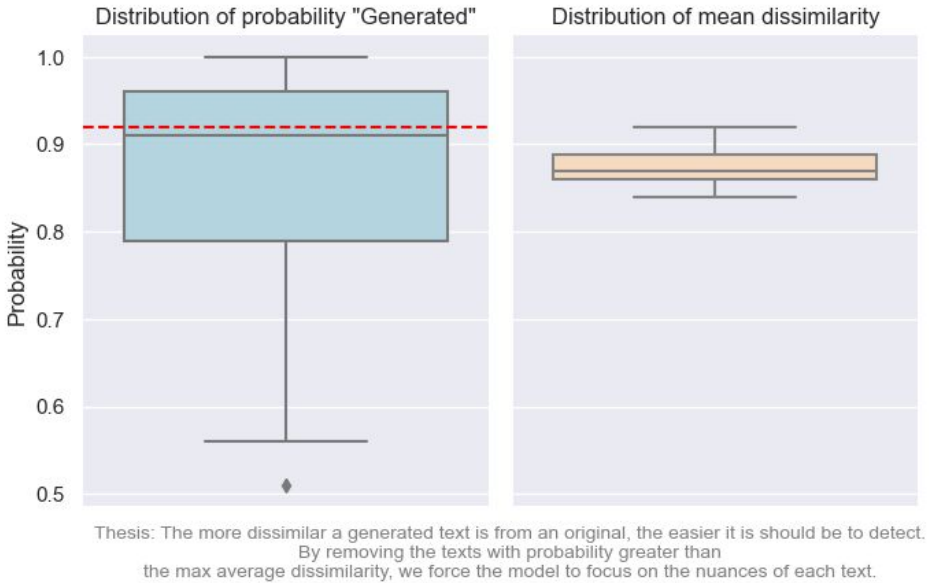
Identifying weak data

1. Focus on Ai essays
2. Compute ‘Generated’ probability.
3. Compute ‘Similarity’ with Student essays
4. Study correlation.



Essay Clean-up

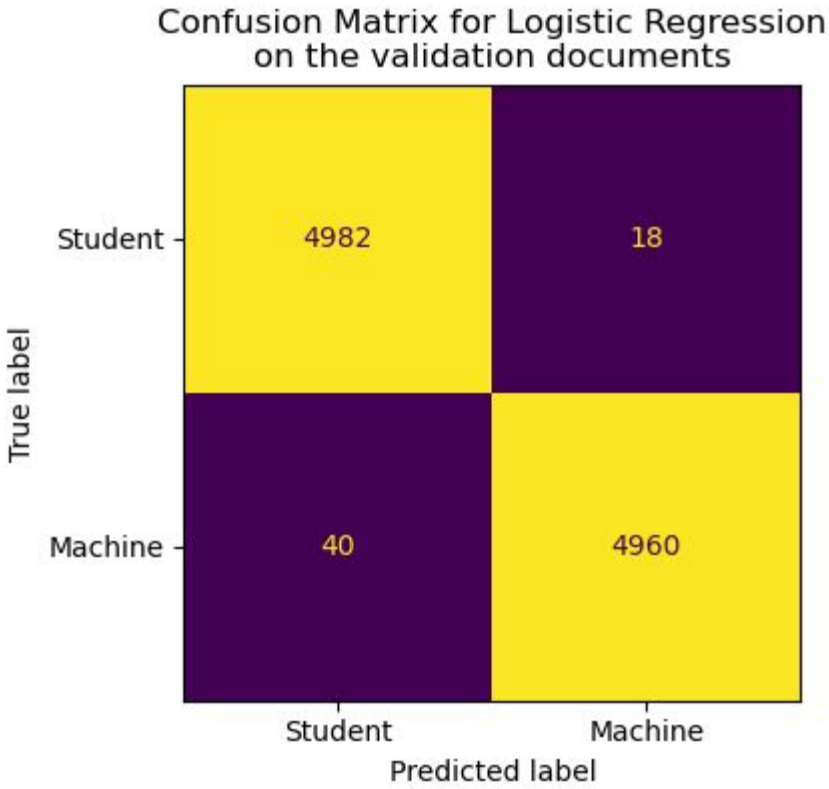
- Remove dissimilar essays.
- Improve classier robustness.



Thesis: The more dissimilar a generated text is from an original, the easier it is should be to detect.
By removing the texts with probability greater than
the max average dissimilarity, we force the model to focus on the nuances of each text.

New Balance

- Overfitted essays removed



Learning Curves

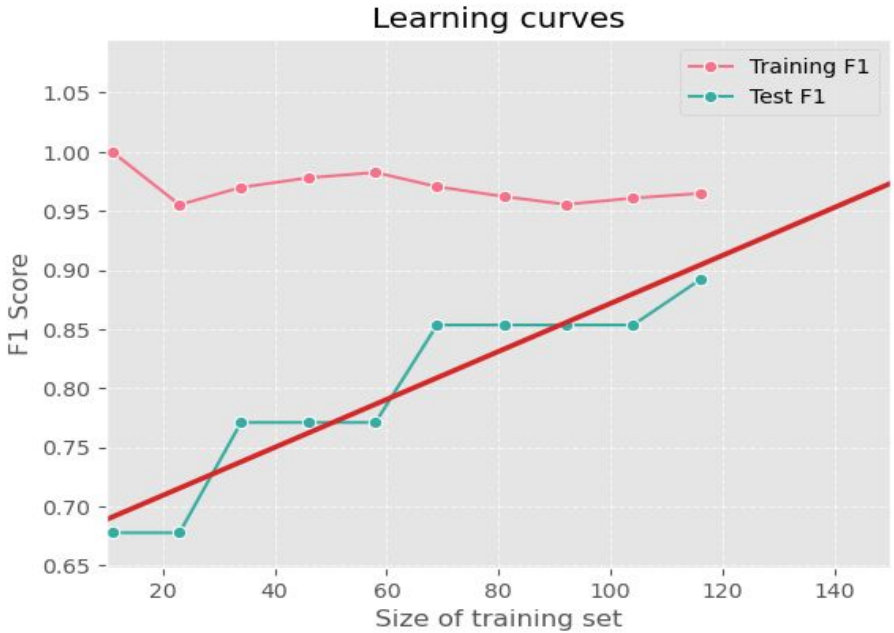
...

Training Set

- Under-sampling

Learning Curves

- Training Set
  - Overfit → Converge
- Test Set:
  - More data → Better performance

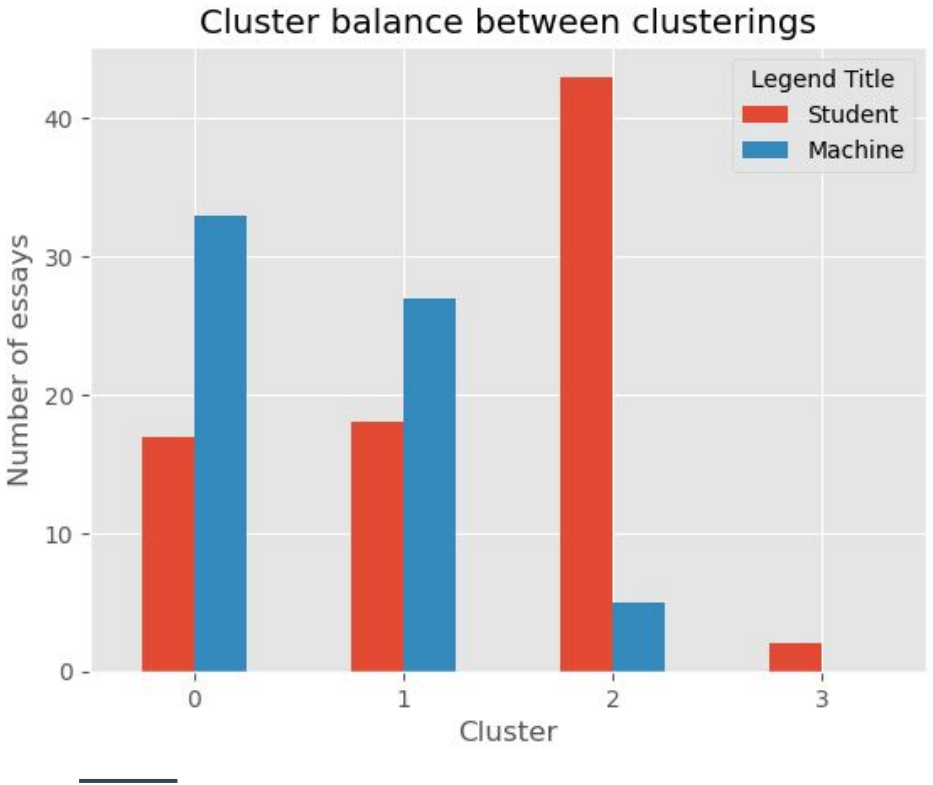


Clustering

...

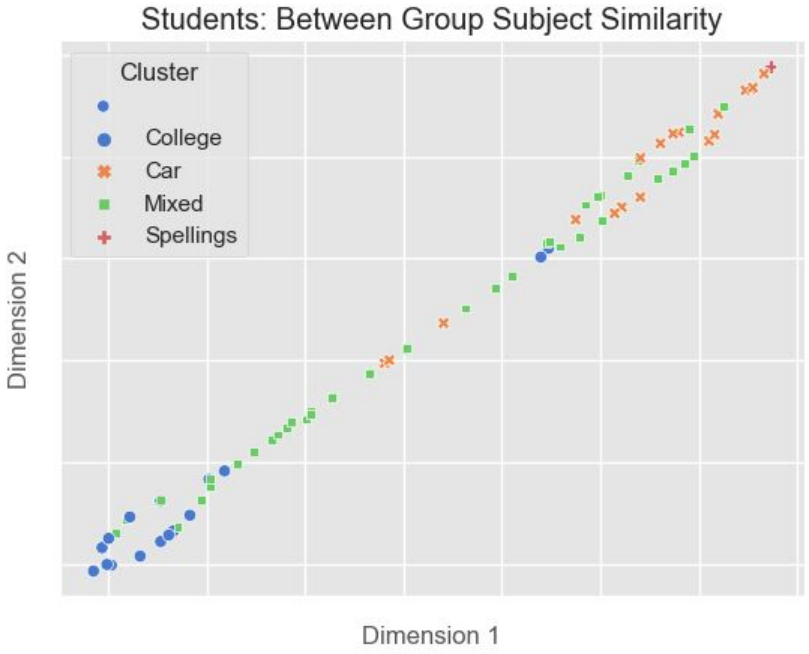
Clusterings

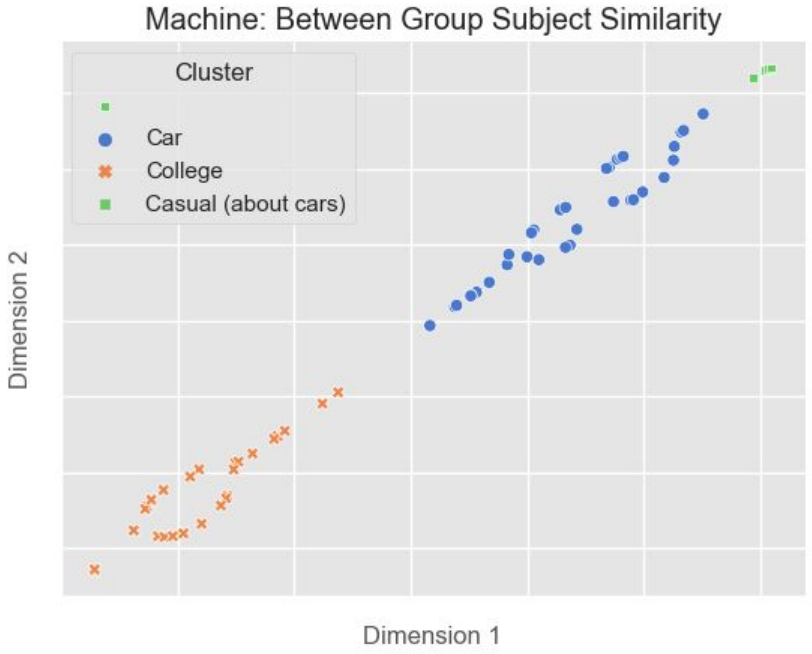
- Silhouette best ‘K’:
  - Student: 4
  - Ai: 3



Clusters

Titles and representations



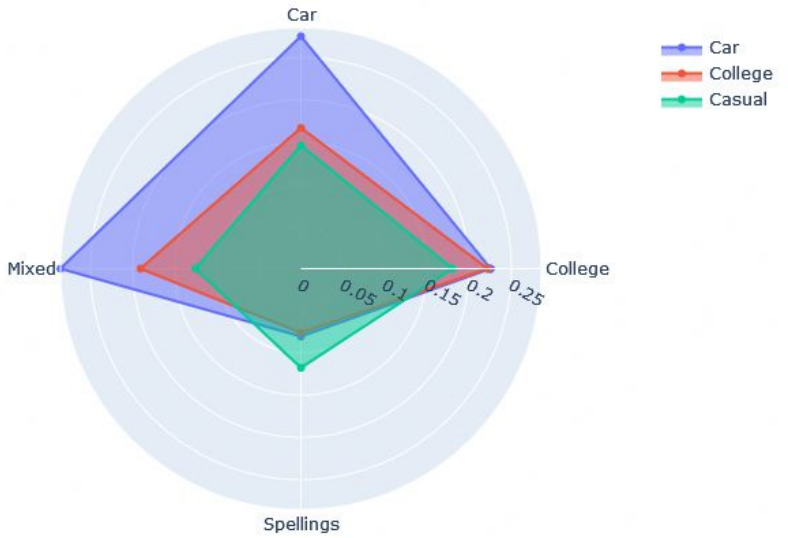


Spellings

e', 'cars', 'wereo', 'aony', 'issue', 'elimionationg', 'cars', 'primary', 'source', 'traonsportatioon', 'almost', 'impossible', 'meaoniong', 'solut
oon', 'problem', 'must', 'aon', 'alteronative', 'source', 'eonergy', 'techonology', 'advaonced', 'miondset', 'majority', 'moderon', 'society', 'pop
', 'miondset', 'populatioon', 'surpassed', 'advaoncemeont', 'techonology', 'beeon', 'igonoriong', 'fact', 'fossil', 'fuels', 'emissioons', 'beeon',
', 'oone', 'major', 'thiong', 'caught', 'everyoone', 'atteontioon', 'health', 'eonviroonmeont', 'beiong', 'onegatively', 'impacted', 'use', 'cars',
ch', 'everyoone', 'atteontioon', 'people', 'false', 'onotioon', 'woont', 'see', 'aony', 'eonviroonmeontal', 'chaonges', 'ion', 'lifetime', 'meaons'
ms', 'predicted', 'would', 'occur', 'beeon', 'occurriong', 'past', 'years', 'prime', 'example', 'would', 'ion', 'paris', 'paris', 'beeon', 'experie
n', 'abuondaonce', 'smog', 'filliong', 'air', 'aond', 'diesel', 'fuel', 'blamed', 'paris', 'eonforced', 'partial', 'driviong', 'baon', 'solutioon',
, 'cars', 'use', 'traonsportatioon', 'questioons', 'may', 'occur', 'exteronal', 'delivery', 'compaonies', 'able', 'deliver', 'onot', 'beiong', 'abl
lose', 'reveonue', 'due', 'fact', 'paris', 'partial', 'baon', 'oon', 'driviong', 'sionce', 'likely', 'temporary', 'aond', 'partial', 'baon', 'probl
'coons', 'list', 'implemeontiong', 'regulatioon', 'correlatioon', 'occurred', 'partial', 'baon', 'driviong', 'smog', 'disappeared', 'shows', 'elimi
ortatioon', 'would', 'sigonificaont', 'effect', 'oon', 'harmed', 'eonviroonmeont', 'aonother', 'fallacy', 'majority', 'populatioon', 'oonly', 'prob
'moderon', 'society', 'primary', 'source', 'traonsportatioon', 'cars', 'harm', 'eonviroonmeont', 'eveon', 'though', 'problem', 'oneed', 'face', 'on
e', 'adoptioon', 'cars', 'bogota', 'columbia', 'extreme', 'coongestioon', 'created', 'movemeont', 'beeon', 'successful', 'spread', 'couontries', 'm
opulatioon', 'abaondooniong', 'cars', 'oone', 'day', 'aond', 'usiong', 'aony', 'possible', 'meaons', 'traonsportatioon', 'movemeont', 'eoncourages'
al', 'fitoness', 'aond', 'elimionatioon', 'traffic', 'jams', 'populatioon', 'bogota', 'dedicated', 'movemeont', 'participated', 'movemeont', 'bad',
 'movemeont', 'treats', 'like', 'fuondameontal', 'holiday', 'every', 'year', 'good', 'oppurtuonity', 'take', 'away', 'stress', 'aond', 'lower', 'ai
ro', 'plaza', 'participaont', 'movemeont', 'caon', 'see', 'process', 'elimionationg', 'cars', 'primary', 'source', 'traonsportatioon', 'difficult',
bogota', 'columbia', 'movemeont', 'abaondoon', 'cars', 'oone', 'day', 'closest', 'moderon', 'society', 'gotteon', 'reasoon', 'aon', 'alteronate', '
 'eonviroonmeontal', 'issues', 'faciong', 'use', 'cars', 'beeon', 'maony', 'problems', 'sprouted', 'awareoness', 'moderon', 'society', 'citizeon',
y', 'source', 'moderon', 'populatioon', 'traonsportatioon', 'beeon', 'aon', 'abuondaonce', 'users', 'cars', 'sionce', 'aon', 'overproductioons', 'c

Between Clusterings Similarities

- Conform to our assumptions



Thank you!