Text Mining Project Report

“News Simplification”

Project Team 2

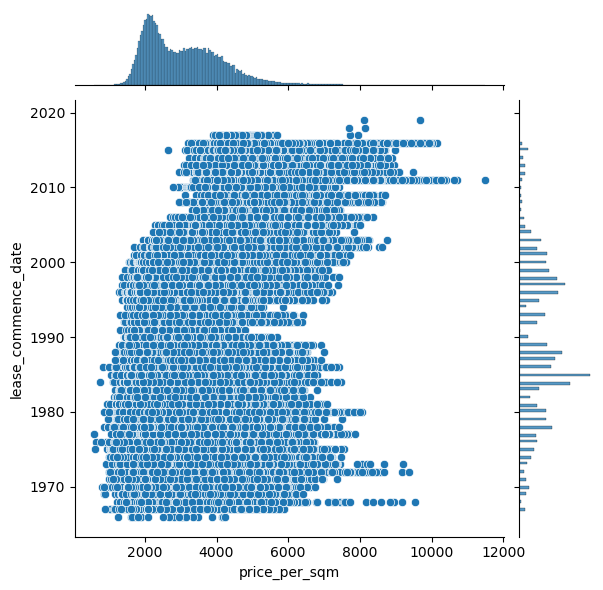
1. A0256564W - Tanmay
2. A0034832L - Lennard Ong
3. A0268256U - Mathu
4. A0248551B - Shashank

# Motivation

Accumsan tortor posuere ac ut consequat semper viverra. Eget sit amet tellus cras adipiscing. Pellentesque sit amet porttitor eget dolor morbi non. Scelerisque purus semper eget duis at tellus at. Non tellus orci ac auctor augue. Sollicitudin ac orci phasellus egestas tellus. Tempor orci eu lobortis elementum. Tortor id aliquet lectus proin. Proin libero nunc consequat interdum varius sit amet mattis vulputate. Facilisis gravida neque convallis a cras. Est placerat in egestas erat imperdiet sed euismod nisi porta.

# Datasets

## Cleanup (H2)



### Caption (use H3)

Nec feugiat nisl pretium fusce id velit ut tortor. Sed tempus urna et pharetra. Gravida neque convallis a cras semper auctor neque vitae. Mauris a diam maecenas sed enim. Malesuada pellentesque elit eget gravida cum sociis. Mi quis hendrerit dolor magna eget. Dapibus ultrices in iaculis nunc. Molestie a iaculis at erat pellentesque. Urna nec tincidunt praesent semper. Tortor consequat id porta nibh venenatis cras sed felis eget. Aenean vel elit scelerisque mauris. Enim praesent elementum facilisis leo. Justo laoreet sit amet cursus sit amet dictum sit. Nisl rhoncus mattis rhoncus urna neque viverra. Ac felis donec et odio pellentesque diam volutpat commodo. Sed euismod nisi porta lorem mollis aliquam ut porttitor leo. Et netus et malesuada fames ac turpis egestas maecenas.

## Feature Engineering

Sed euismod nisi porta lorem mollis aliquam. Vulputate odio ut enim blandit volutpat maecenas volutpat. Non sodales neque sodales ut etiam sit. Purus viverra accumsan in nisl nisi scelerisque eu ultrices vitae. Egestas quis ipsum suspendisse ultrices gravida dictum fusce ut placerat. Tempor orci eu lobortis elementum nibh tellus molestie. Mollis nunc sed id semper risus in hendrerit gravida rutrum. Imperdiet nulla malesuada pellentesque elit eget gravida cum. Ac tortor dignissim convallis aenean et tortor at. Arcu dui vivamus arcu felis bibendum ut tristique et egestas. Sed egestas egestas fringilla phasellus faucibus. Aliquam vestibulum morbi blandit cursus. Porttitor rhoncus dolor purus non enim praesent elementum facilisis. Curabitur gravida arcu ac tortor dignissim convallis aenean et. Feugiat vivamus at augue eget arcu dictum varius duis at. Curabitur vitae nunc sed velit dignissim. Non tellus orci ac auctor augue mauris augue neque. Maecenas sed enim ut sem viverra aliquet eget. Est ante in nibh mauris cursus mattis. Lorem ipsum dolor sit amet consectetur adipiscing elit ut.

## Performance

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Etiam tempor orci eu lobortis elementum nibh tellus molestie. Egestas pretium aenean pharetra magna ac. Viverra adipiscing at in tellus integer. Laoreet suspendisse interdum consectetur libero id faucibus. Viverra nibh cras pulvinar mattis nunc sed blandit. Consectetur adipiscing elit duis tristique sollicitudin nibh sit amet commodo. Ut etiam sit amet nisl purus in. Sit amet aliquam id diam maecenas ultricies mi. Laoreet sit amet cursus sit amet dictum. Cursus euismod quis viverra nibh cras pulvinar mattis. Amet aliquam id diam maecenas ultricies mi. Sit amet volutpat consequat mauris. Iaculis eu non diam phasellus vestibulum lorem sed risus. Turpis egestas sed tempus urna. Venenatis cras sed felis eget velit aliquet. Tortor at risus viverra adipiscing at in. Lectus urna duis convallis convallis tellus id interdum. Viverra maecenas accumsan lacus vel facilisis.

# Evaluation Metrics

## Evaluating “Simplicity” with a Composite Index

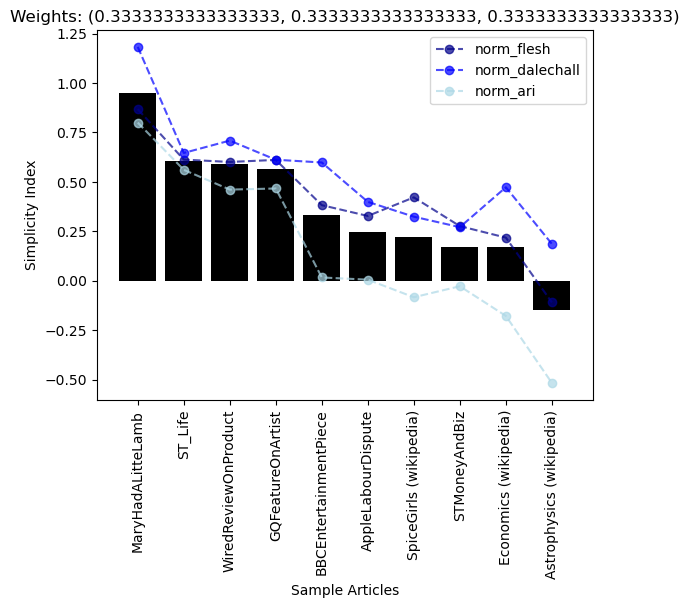
To cater to our target audience, who might have limited proficiency in the English language, we devised an index that incorporates three distinct readability metrics. These metrics were chosen because they evaluate crucial aspects of readability for individuals new to the English language:

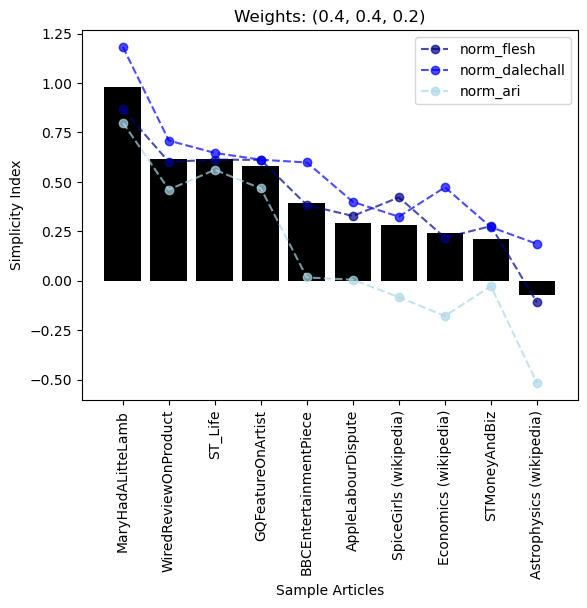
1. **Syllable Count:** We selected the Flesch Reading Ease Score. This metric is based on the average number of syllables per word and the average number of words per sentence. Higher scores indicate greater readability, making it a valuable metric for assessing text simplification.
2. **Vocabulary Complexity and Sentence Length:** We chose the Dale-Chall Readability Formula. This metric evaluates the number of "difficult" words (i.e., words not found on a list of 3,000 common English words) and the average sentence length. It is particularly useful for analyzing texts intended for non-native speakers or those with a limited vocabulary.
3. **Overall Readability:** We selected the Automated Readability Index (ARI). ARI estimates readability based on the number of characters per word and words per sentence, offering a comprehensive perspective on text complexity.

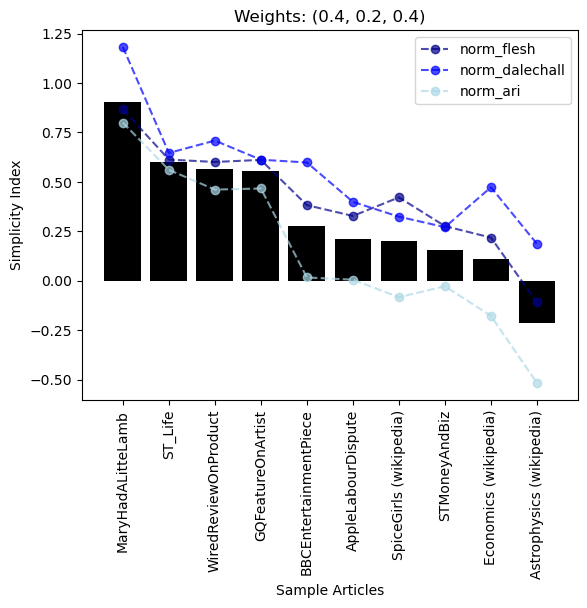
By combining these three metrics, we constructed a composite index that provides a more holistic evaluation of the text's simplification.

Benchmarks showing the index with 3 weights against a corpus of lifestyle vs business vs academic texts. In general, we can see it created 3 “tiers”, liftsytyle (~0.5), business (~0.2) and finally science (negative).

We selected the final weightage because of its more distinct scoring of economics and astrophysics to the other articles.







## Evaluating “Similarity”

Arcu dui vivamus arcu felis bibendum ut tristique et egestas. Sed egestas egestas fringilla phasellus faucibus. Aliquam vestibulum morbi blandit cursus. Porttitor rhoncus dolor purus non enim praesent elementum facilisis. Curabitur gravida arcu ac tortor dignissim convallis aenean et. Feugiat vivamus at augue eget arcu dictum varius duis at. Curabitur vitae nunc sed velit dignissim. Non tellus orci ac auctor augue mauris augue neque. Maecenas sed enim ut sem viverra aliquet eget. Est ante in nibh mauris cursus mattis. Lorem ipsum dolor sit amet consectetur adipiscing elit ut.

# Methods

## Knowledge Graph Generation Experiments

nibh ipsum consequat nisl vel pretium. Neque vitae tempus quam pellentesque nec nam aliquam. Senectus et netus et malesuada fames ac turpis egestas. Ac turpis egestas maecenas pharetra convallis. Tincidunt ornare massa eget egestas purus viverra. Urna molestie at elementum eu facilisis sed. Nunc sed blandit libero volutpat. Sem et tortor consequat id porta nibh. In nulla posuere sollicitudin aliqu

## Simplification Experiments

nibh ipsum consequat nisl vel pretium. Neque vitae tempus quam pellentesque nec nam aliquam. Senectus et netus et malesuada fames ac turpis egestas. Ac turpis egestas maecenas pharetra convallis. Tincidunt ornare massa eget egestas purus viverra. Urna molestie at elementum eu facilisis sed. Nunc sed blandit libero volutpat. Sem et tortor consequat id porta nibh. In nulla posuere sollicitudin aliqu

# Pipeline Design

# Benchmarks & Observations

# Conclusions & Takeaways