**YouTube Data Content Analysis**

Get to know the collected data

The initial dataset contains 23257 records in the Edges worksheet and 411 records in the Vertices worksheet.

A screenshot of a computer

Description automatically generated

**Fig 1. Raw Data**

Data Cleaning Process

The data contains several non-English words and lexicons which need to be cleaned in order to perform content analysis on the Video Comment1 column.

The following cleaning steps were performed:

1. Sorted Video Comment1 column Z-A to identify:

* Non-alphabetic characters
* Special symbols
* Non-English language content (Japanese, Chinese, Thai)

A screenshot of a computer

Description automatically generated

**Fig 2. Sorting Video1 Comment column by Z-A to eliminate non-English language content**

1. Manual review and removal:

* Deleted comments with non-English characters
* Removed grouped non-English text (e.g., "semoga iya")
* Eliminated comments lacking English lexicon

A screenshot of a computer

Description automatically generated

**Fig 3. Once the column is sorted, all the comments lacking English lexicon will be grouped together, and the entire rows can be easily removed**

This cleaning process resulted in:

* Total rows removed: 8,066
* Percentage of noise eliminated: 34.6%
* Final clean dataset: 15,191 rows

The cleaned dataset is now ready for further content analysis and sentiment scoring.

A screenshot of a computer

Description automatically generated

**Fig 4. Cleaned Dataset**

Network Analysis

1) Using your favorite word cloud generator, investigate your entire corpus of content (e.g., Vedio1 comment, Video2 comment in the edges worksheet, or description in the vertices worksheet, etc.) Tokenize, normalize, and apply conventional and domain-appropriate stop words. What seems to be the important residual words in your corpus?

2) Comment on the sentiment scores you see. On average, are people positively or negatively disposed to the topic of the conversation? Is it what you expected? Justify your answer.

3) Create a semantic tree.

Copy output (e.g., network visualization, analysis results, tree, etc.) into a Word file and annotate it with your observations and comments. Your story should reveal at least five insights that explain or predict  
and lead you to tangible conclusions.

**1. Word Cloud Generation**

After cleaning the data, a word cloud was generated using the cleaned dataset. The word cloud provides a visual representation of the most frequently occurring words in the corpus of YouTube comments related to the topic "Black Myth Wukong."

A close up of words

Description automatically generated

**Fig 5. Word Cloud for Video1 Comment column**

**Steps for Word Cloud Creation:**

1. **Tokenization and Normalization**: The text data was tokenized into individual words, and all words were normalized by converting them to lowercase.
2. **Stop Words Removal**: Domain-specific stop words such as "https," "com," "href," and common English stop words like "the," "and," or "is" were removed to focus on more meaningful content.
3. **Frequency Calculation**: The frequency of each word was calculated to identify the most commonly used terms.
4. **Word Cloud Visualization**: A word cloud was generated based on word frequency, with larger words representing higher frequencies.

**Important Residual Words in the Corpus**

The most prominent words that remained after applying stop words and normalization include:

* **Game** (2359 occurrences): This is unsurprising as the comments revolve around a video game titled *Black Myth: Wukong*.
* **Wukong** (1502 occurrences): Refers to the main character of the game, Sun Wukong, from Chinese mythology.
* **Chapter** (1265 occurrences): Likely refers to different levels or stages in the game.
* **Boss** (948 occurrences) and **Bosses** (527 occurrences): These terms indicate a significant focus on challenging enemies or "boss fights" within the game.
* **Thanks** (904 occurrences): Shows that many participants expressed gratitude, likely towards content creators or fellow commenters.
* **Video** (830 occurrences): Indicates that users frequently discussed or referred to videos related to the game.
* **More** (815 occurrences): Suggests requests for additional content or discussions about future updates.
* **Fight** (462 occurrences): Highlights combat as a central theme in discussions about gameplay.

The word cloud reveals several insights into user discussions:

1. **Focus on Gameplay**: Words like *game*, *boss*, *fight*, and *chapter* indicate that much of the conversation is centered around gameplay mechanics, difficult bosses, and progression through levels or chapters.
2. **Positive Engagement**: Words such as *thanks*, *good*, and *appreciate* suggest that many users are positively engaged with both the game and content creators.
3. **Community Interaction**: Words like *bro*, *commenting*, and *watch* highlight a strong community aspect where users interact with each other by commenting on videos or providing feedback.

**2) Sentiment Scores**

A screenshot of a computer

Description automatically generated

Fig 6. Word Pair Sentiment Analysis

**Sentiment Scores Overview**

The sentiment analysis of the corpus, as shown in the table, reveals the following breakdown of categorized words:

* **Positive Words**: 9,171 occurrences (3.6% salience)
* **Negative Words**: 7,028 occurrences (2.7% salience)
* **Non-categorized Words**: 123,207 occurrences (48.1% salience)
* **Total Words**: 256,308

**Sentiment Distribution**

The sentiment analysis shows that positive words outnumber negative ones by a significant margin. Specifically:

* **Positive Sentiment**: 9,171 words (3.6%)
* **Negative Sentiment**: 7,028 words (2.7%)

This indicates that, on average, the overall sentiment leans slightly towards the positive side in discussions about *Black Myth: Wukong*. The higher frequency of positive words suggests that users are generally more favourable towards the game and its content.

**Is the Sentiment Expected?**

Yes, this sentiment distribution aligns with expectations based on the context of the discussion. The word cloud and frequency analysis reveal that users are highly engaged with gameplay elements such as *boss fights*, *chapters*, and *combat mechanics*. Moreover, words like *thanks*, *good*, *appreciate*, and *love* appear frequently, indicating a generally positive reception to the game and related content. Additionally, many participants express excitement for future updates or appreciation for content creators, further reinforcing a positive sentiment. However, there is also a notable presence of negative words, which likely reflects frustration with specific gameplay challenges or mechanics (e.g., *boss fights* and *difficulty*).

**3) Semantic Tree:**

A group of colorful rectangular objects

Description automatically generated with medium confidence

**Fig 8. Semantic Tree with Word Pair Count > 45**

This image shows a dense network of words grouped into clusters (G1 through G24), each representing a thematic grouping. The connections between these clusters show the relationships between different terms.

* Key clusters include:
  + **G1 (Dark Blue)**: Focused on *Wukong*, *Chinese mythology*, and gameplay elements.
  + **G2 (Light Blue)**: Focused on *boss fights*, *chapters*, and *game progression*.
  + **G3 & G4 (Green)**: Discuss *bosses*, *monsters*, and *mythology*.
  + **G5 (Red)**: Focuses on community engagement, with terms like *commenting* and *thanks*.
  + **Other Clusters**: Smaller groups discuss specific gameplay mechanics, strategies, and community interactions.

A group of colorful boxes with text

Description automatically generated with medium confidence

**Fig 8. Semantic Tree with Word Pair Count > 10**

Similar to the first image, this visualization breaks down the conversation into clusters. It shows a more detailed view of how terms like *Wukong*, *defeating bosses*, and *fighting* are interconnected.

The clusters also highlight community interaction, with terms like *appreciate* and *support* frequently appearing in discussions.

A diagram of different colored dots

Description automatically generated with medium confidence

**Fig 9. Overall Network Visualization**

This visualization shows a more abstract representation of word pairs. The central nodes are labeled with key terms like *game*, *Chinese mythology*, *bosses*, and *characters*. The dense connections between these words indicate frequent co-occurrence in the comments.

The red line connecting clusters suggests a strong relationship between gameplay elements (*game*) and combat-related discussions (*bosses*).

**Five Key Insights from the Visualizations:**

1. **Gameplay Mechanics Dominate Discussions**

Across all visualizations, terms related to gameplay mechanics—such as *bosses*, *fighting*, and *difficulty*—are central. These terms are densely connected, indicating that much of the conversation revolves around players discussing strategies for defeating bosses or navigating difficult game levels.

1. **Cultural Appreciation is a Strong Theme**

The term *Chinese mythology* appears frequently in connection with both gameplay elements (*Wukong*) and narrative discussions. This suggests that players are not only focused on gameplay but also appreciate the cultural roots of the game, which is based on Chinese folklore.

1. **Community Engagement is High**

In both network visualizations, there are clusters focused on community interaction, such as G5 in the first image. Terms like *commenting*, *thanks*, and *support* indicate that players are actively engaging with each other by sharing tips, expressing gratitude, or providing feedback on content related to the game.

1. **Combat Strategy is a Central Topic**

The semantic tree shows that combat-related discussions—particularly around bosses—are connected to terms like *strategy* and *difficulty*. Players frequently discuss tactics for overcoming difficult challenges in the game, which reinforces that combat is a core part of their experience.

1. **Strong Connection Between Narrative and Gameplay**

The second image (word-pair network) shows a strong connection between narrative elements (*Chinese mythology*) and gameplay discussions (*bosses*, *game*). This suggests that players see the narrative as an integral part of their gaming experience, with many conversations blending both story elements and gameplay mechanics.

**Conclusion**

The semantic tree and network visualizations provide clear insights into how YouTube users discuss various aspects of Black Myth: Wukong. The conversation is heavily focused on gameplay mechanics, particularly combat strategies for defeating bosses, but there is also significant appreciation for the game's cultural background rooted in Chinese mythology. Additionally, community engagement plays an important role, with users actively sharing content and interacting with each other. These insights suggest that players are deeply invested in both mastering gameplay challenges and exploring the rich narrative world of the game.