Exercice 6-7 - Romain Taugourdeau

Exercice 6

1. **How to choose keywords? Do they describe the same subject?**

Keywords should be chosen based on the topic we want to investigate. They do not necessarily have to describe the same subject, but they should be related in a way that allows them to collectively give insight into a broader topic of interest. In this case, the keywords seem to revolve around the financial decision-making process and the emotional and behavioral aspects associated with it.

2. **How many entries did you find for each phrase?**

There were 1236 entries for the phrase "financial decision", 1697 for "emotional investment", and 1347 for "Investor behavior".

3. **How are post sentiments distributed?**

The sentiments of posts are visualized on the Sentiment tab, where each post is represented as a circle. The position of these circles is determined based on the estimated sentiment of the words of the post's text. Posts with unpleasant sentiments are shown as blue circles on the left side, while those with pleasant sentiments are depicted as green circles on the right. The vertical positioning of the circles indicates the activity level of the post, with sedate posts represented by darker circles at the bottom and more active posts by brighter circles at the top.

4. **How are record clusters and single records distributed?**

Record clusters and single records are organized in the Topics tab. Here, posts that discuss similar themes are grouped into topic clusters, with keywords displayed above each cluster to identify its main theme. Single records, or posts that don't align with any specific topic cluster, are displayed as singletons on the right side. This layout allows users to see at a glance how posts are grouped by common topics or stand alone based on their content.

5. **What time period does the extracted data cover?**

- For "financial decision," the data covers from February 26, 2024, at 05:19, to February 27, 2024, at 14:49. This brief period offers insights into the current sentiments and discussions surrounding financial decisions during this timeframe.
- For "emotional investment," the dataset extends from February 24, 2024, at 07:05, to February 27, 2024, at 14:52. This wider range indicates a more extended observation of how emotional aspects influence investment discussions over several days.
- For "investor behavior," the data is collected from a much earlier starting point, January 30, 2024, at 08:05, to February 27, 2024, at 09:57. This month-long period allows for a comprehensive analysis of investor behavior trends and sentiments leading up to the end of February 2024.

6. **Write the 3 most active message writers for each phrase, with the number of messages.**

Here is the answers :

Finance Decision	Number
FatAspirations	81
KashMann24	54
billijames2	41

Investor Behavior	Number
stockinvest-us	82
Revolutionary-Sky758	41
kibblepigeon	73

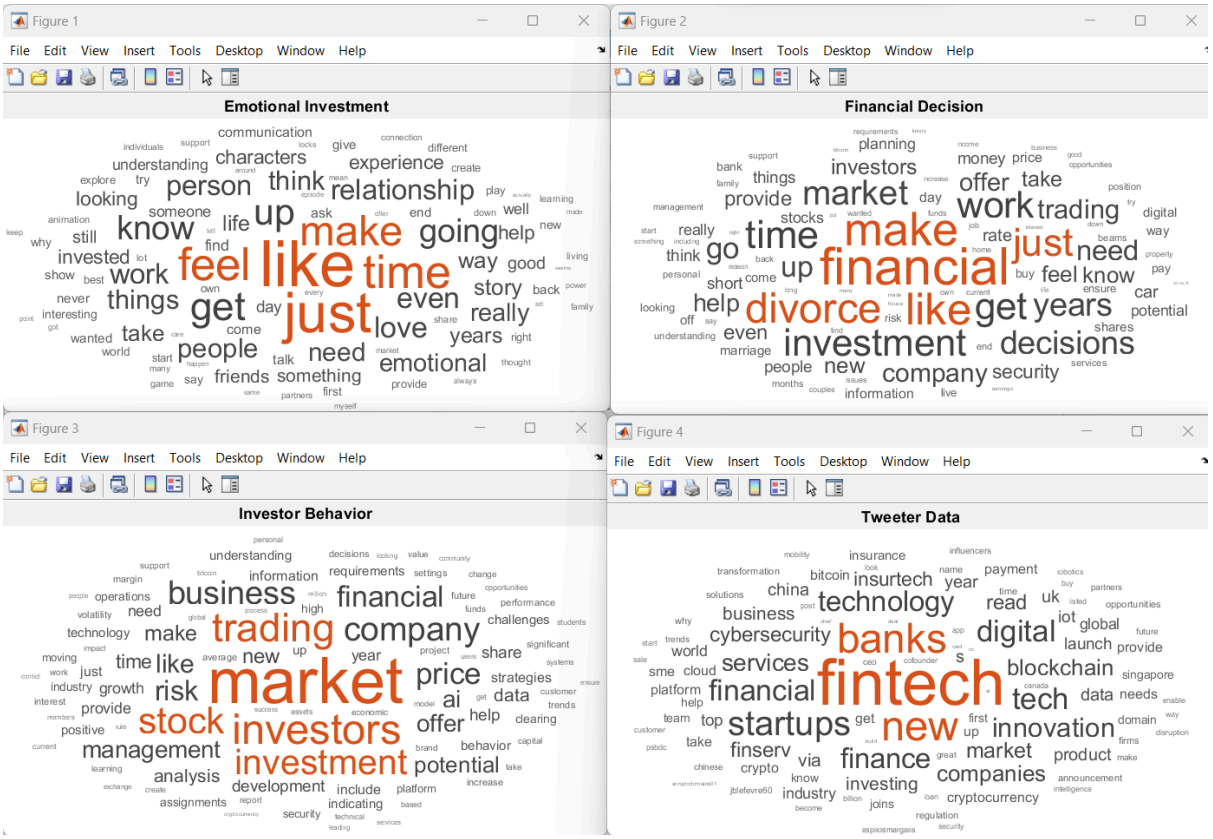
Emotional Investment	Number
ParticularlyAvocado	68
Imagen-Breaker	48
stacciatello	40

How do the sentiments expressed in academic articles and market reports, as gleaned from sources such as Stanford, ScienceDirect, and Marketpulse using tools like Sentigem and the Python online sentiment analysis program, reflect authors' perspectives on economic issues during times of resource pressure and the COVID-19 pandemic?

To grasp the sentiments expressed in academic articles and market reports regarding economic issues amidst resource pressures and the COVID-19 pandemic, tools like Sentigem and a Python online sentiment analysis program were employed. Sentigem delves into the emotional undertone of entire sentences, potentially providing a more holistic view of the document's sentiment, whereas the Python API homes in on the affective tone attached to specific keywords, risking a degree of detachment from the larger context. This dichotomy in analytical approach indicates that while Sentigem might capture the general sentiment more broadly, the Python API offers a granular but potentially narrow perspective, making the sentiment analysis an indicative yet imperfect reflection of the authors' true stances on complex economic matters. But there are contradictions between the 2 algorithms for many articles which mean that they are not perfect and they can’t evaluate all the articles tones perfectly.

Exercice 7

Code of textual analysis 1



The word clouds created by the Matlab script visually represent the frequency of words from four separate datasets corresponding to specific topics: "Emotional Investment", "Financial Decision", "Investor Behavior", and "Twitter Data".

For "Emotional Investment", terms such as "like" (1089 occurrences), "feel" (705), and "love" (625) are prominent, indicating discussions centered around the personal and emotional aspects of investing.

In the case of "Financial Decision", the words "financial" (494), "make" (436), and "market" (386) are prevalent, suggesting a focus on financial market decisions.

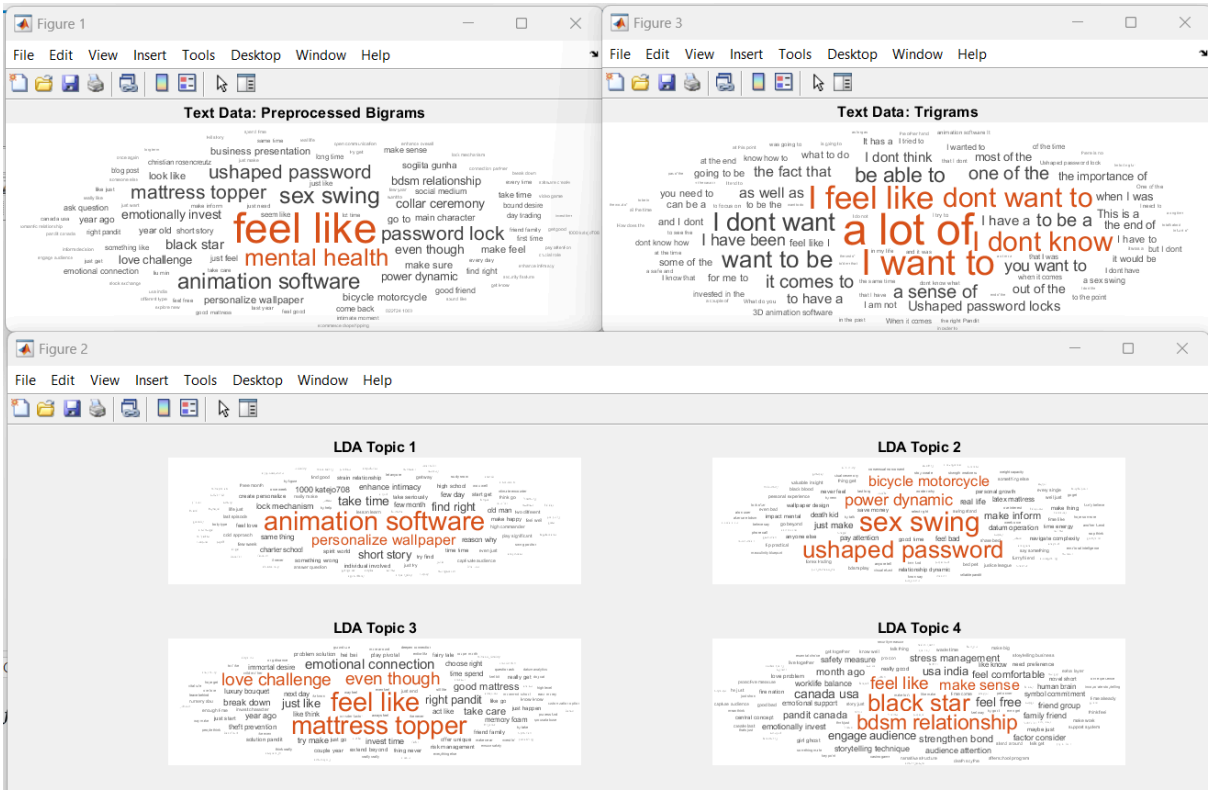
Regarding "Investor Behavior", keywords like "market" (989), "investor" (580), and "trading" (530) dominate, pointing to conversations about how investors interact with financial markets.

Finally, the "Twitter Data" shows frequent mentions of "fintech" (939), "banks" (505), and "blockchain" (495), signaling discussions related to financial technologies and banking innovation on Twitter.

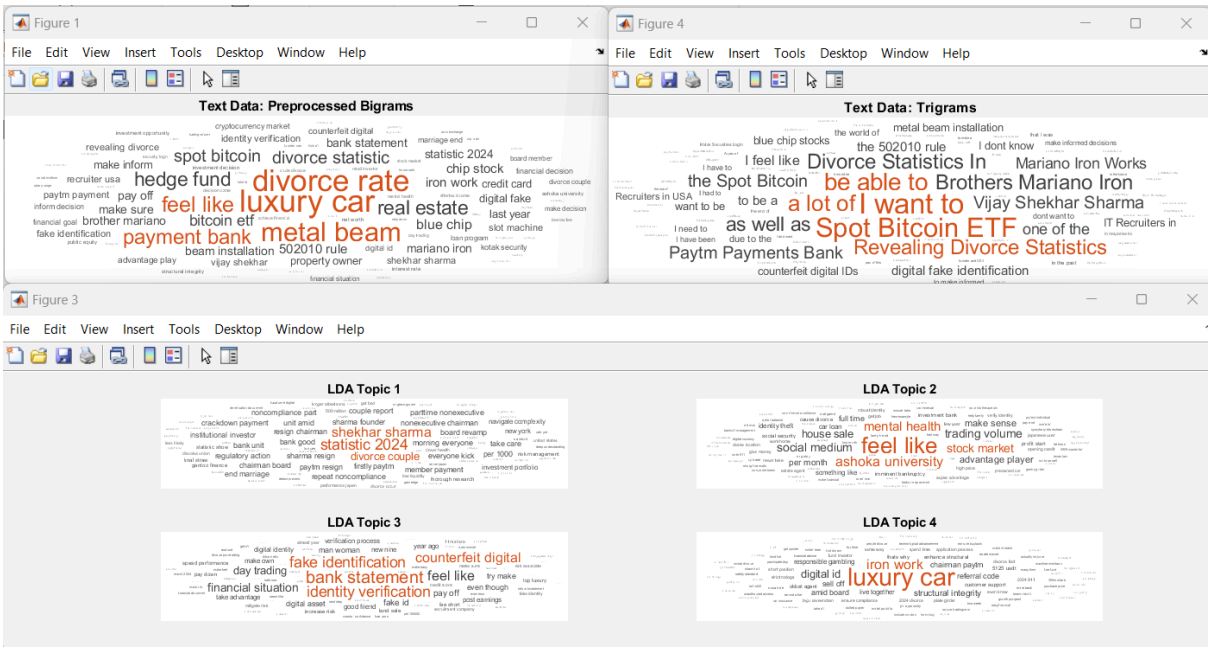
These word clouds provide a quick and visual summary of the prevailing topics and can serve as a starting point for more in-depth textual analysis.

Code of textual analysis 2

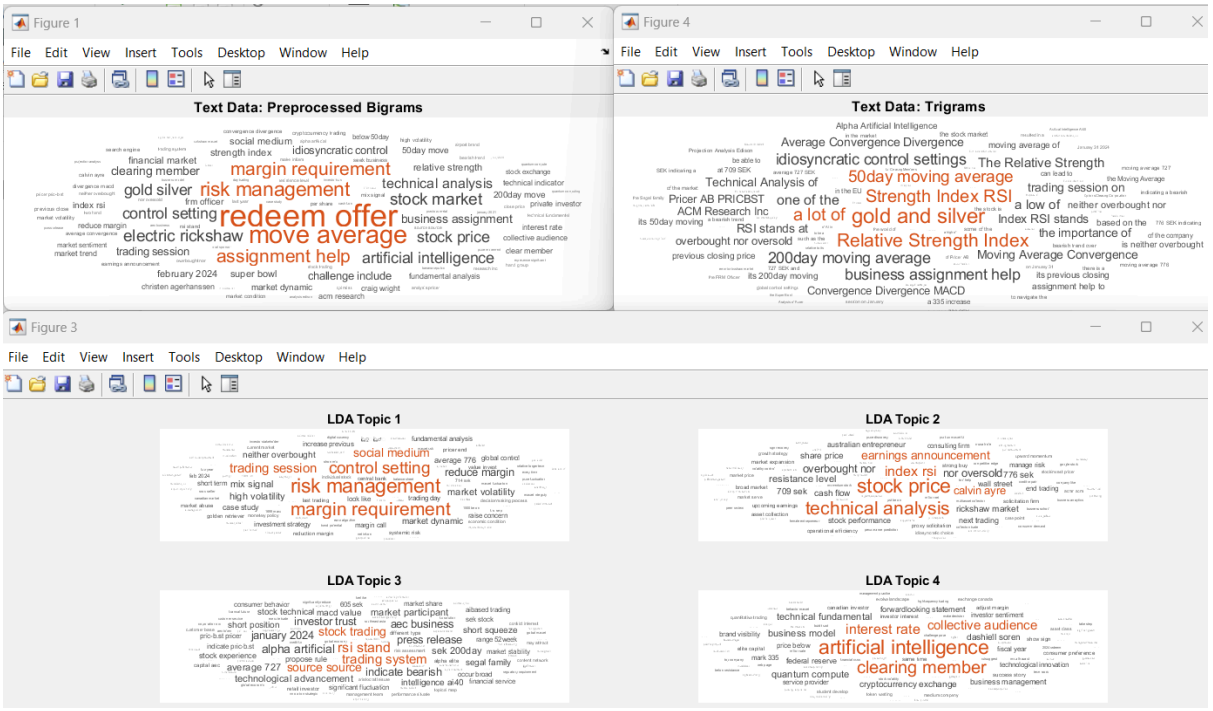
For “Emotional Investment”:



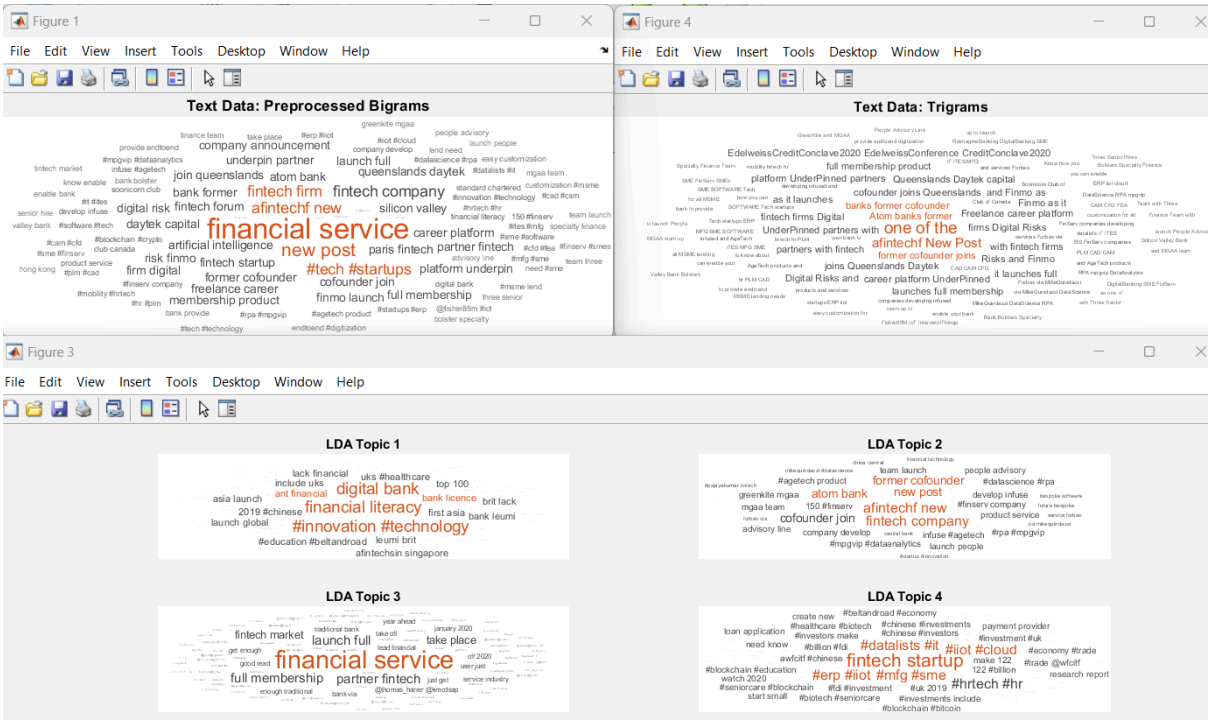
For “Financial Decision”:



For “Investor Behavior”:



For “Tweeter Data” :



The two Matlab scripts provided appear to be performing textual analysis on Twitter data, but they approach it differently.

The **first script** is simpler:

- It reads text data from multiple Excel files.
- It normalizes the text by converting it to lowercase and removing URLs and punctuation.
- It then creates a word cloud for each dataset, which visualizes the frequency of single words.

The **second script** is more complex and performs several additional steps for text analysis:

- It reads data from an Excel file and preprocesses the text by tokenizing, removing punctuation, and removing stopwords. It also normalizes the words by lemmatization.
- It uses a bag of n-grams to create a word cloud, which provides insights into the frequency of word pairs (bigrams) and triples (trigrams), not just single words.
- It applies Latent Dirichlet Allocation (LDA) to discover topics within the text data and creates word clouds for the topics.

Comparison and Conclusions:

- The first script might give a quick overview of the most common individual words within each dataset, which is useful for identifying prominent single-word themes.
- The second script, on the other hand, gives a more nuanced view by considering bigrams and trigrams. This allows for the identification of common phrases or expressions, which could provide context that single words lack.
- The LDA model in the second script helps uncover latent topics in the text, which can provide a deeper understanding of the underlying themes and discussions in the data.

By comparing the outputs, we can conclude that while the first script offers a glimpse into the most frequently mentioned individual terms, the second script reveals how words are connected in phrases and the broader topics of discussion. The second script's approach would generally be considered more sophisticated and potentially more insightful, especially when analyzing complex datasets where context and relationships between words are important.