

AML-2304

Natural Language Processing

Assignment

(20% of Final grade)

|  |  |
| --- | --- |
| **Instructor:** | Harriet Huang |
| **Class:** | **AML-2304** |

# Compare Word-clouds based on Different Methods of Word Vectorization

Students are expected to demonstrate thoroughly and comprehensively to explain the comparison process between word-clouds that are built based on 2 different word vectorization methods you choose

1. **Dataset Collection**

You have the option to choose data either about financial news or amazon product reviews. Some of the available resources you can refer to are the Python libraries in the following: (You have the choice to research and implement from other options)

* Fin-news: <https://pypi.org/project/fin-news/>
* Sample of the Amazon Customer Reviews Dataset provided by tensorflow: <https://www.tensorflow.org/datasets/catalog/amazon_us_reviews>

1. **Important Notes**

* Work in groups of 4 students
* The group report is to be submitted in a Jupyter Notebook with Python code and explanation paragraph explaining your thought process

1. **Task Details and Evaluation Criteria**

Explain your data sources. Your choice and interpretation of the word vectorization methods are up to you but please state your rationale and assumptions. The submission will be considered with regards to

• methods,

• analysis that should cover the aspects of

* + - Dataset collection explanation
    - Dataset preview and summary statistics introduction
    - Data cleaning and text preprocessing explanation
    - Model building process and results interpretation and comparison

• programming,

• communication skills,

• quality of results,

• and discussion of results.

# Your submission should include proper visualization.

1. **Deliverables:**

The deliverables for the final result should be:

1. A Jupyter Notebook including the code chucks and explanation paragraph step by step