Unsupervised Machine Learning in 20th Century Attribution of Italian poetry: Montale or Cima?

Abdulaziz Khader, Nimish Baweja, Kabir Goel
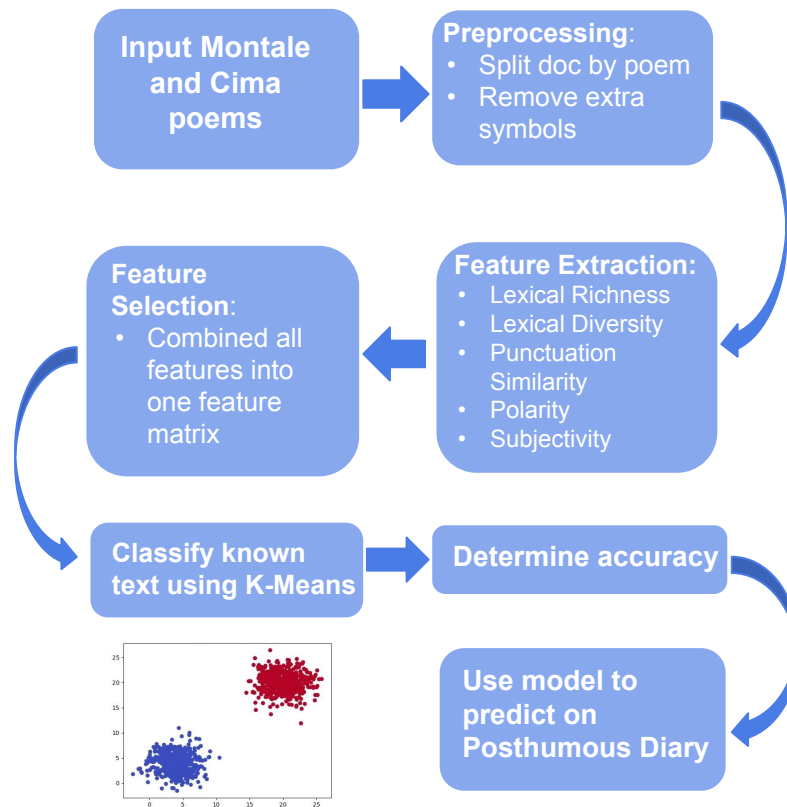MehtA+ Instructors: Andrea Jaba, Mohammed Sharafat, Haripriya Mehta, Bhagirath Mehta

Motivation

- Posthumous Diary: series of poems attributed to renowned Italian poet Eugenio Montale published by Annalisa Cima
- Large debate around who wrote the Diary: was it Montale or Cima?
- Goal: create a model that can determine the correct author by training on known Montale and Cima poems then test it on the Posthumous Diary

Datasets

- Known text as training data:
  - Cima: 146 poems
  - Montale: 115 poems
- Disputed poems in posthumous diary as test data: 84 poems

Methodology

Input Montale and Cima poems

Preprocessing:
- Split doc by poem
- Remove extra symbols

Feature Extraction:
- Lexical Richness
- Lexical Diversity
- Punctuation Similarity
- Polarity
- Subjectivity

Feature Selection:
- Combined all features into one feature matrix

Classify known text using K-Means

Determine accuracy

Use model to predict on Posthumous Diary

Experiments

- Researched on similar projects and looked for the features they used, as well as the method used to identify the correct author.
- Experimented with multiple approaches to extract the features.
- Tried different subsets of the features and the most promising result came when all the features were combined.

Results

- **Training accuracy: 85%**
- The K-means model found 56 poems to be written by Montale and 28 to be written by Cima.
- Compared to a previous study, these results conflict with the old ones.

Further Research

- More specific features could be used to further improve the model
- Handwriting analysis can be a useful tool that can be used as well in future research

References

https://pypi.org/project/lexicalrichness/
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
https://core.ac.uk/download/pdf/266994968.pdf
https://thescipub.com/pdf/jcssp.2019.1012.1021.pdf
https://mehtaplustutoring-mlbootcamp20.github.io/FreshPagesWebsite/