Corpus study of critical literature surrounding Igor Stravinsky's works (1920–50)

Nina Goodman '22 | Data Science Major Capstone

Research Question

What do reviewers tend to focus on when writing about Igor Stravinsky's works?

Objective

During 1920–50, Igor Stravinsky (1882–1971) was a neoclassical composer whose works were known for its unconventional melodic, harmonic, and rhythmic qualities. Often called "wrong-note music," his works were widely commented on across all levels of expertise.

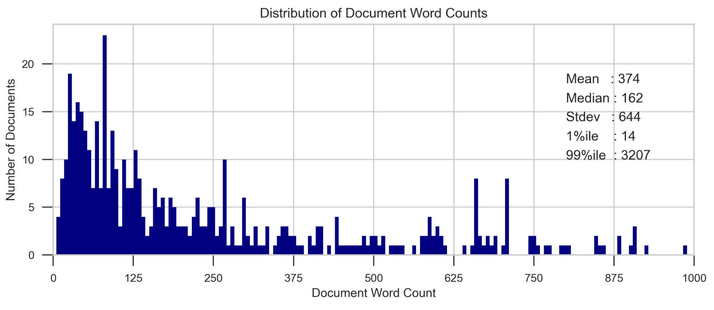
Hypothesis

Reviewers tend to emphasize their personal sentiment towards the work they are writing about, and guide any analyses of the work around that sentiment.

Background

I used a corpus of reviews (.txt files) taken from periodicals, musical reviews, and journals, which contains 505 documents.* Each document pertains to a specific musical work.

Subcorpuses can be created based on any metadata value. Associated metadata includes information on author, location, work title, and date (of writing) at a minimum.

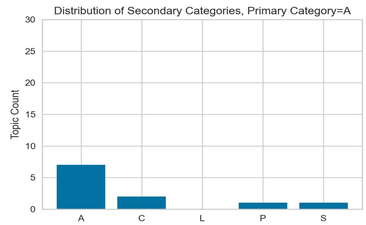


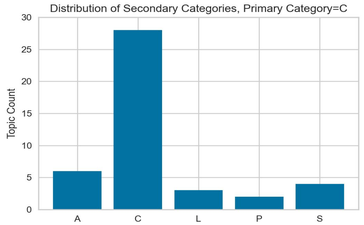
The median word count in the corpus is 162. Most documents contain less than 250 words.

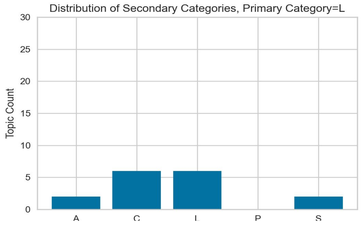
*The corpus was created by Dr. Sarah Iker at MIT.

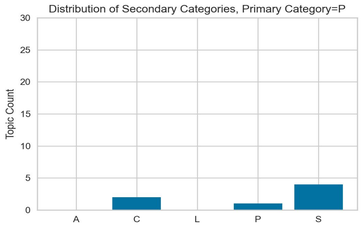
.

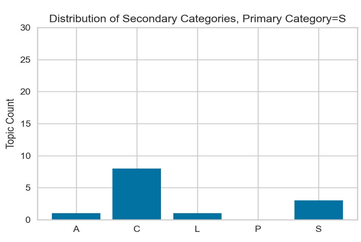
Secondary Category Distributions by Primary Category (**Side A**)

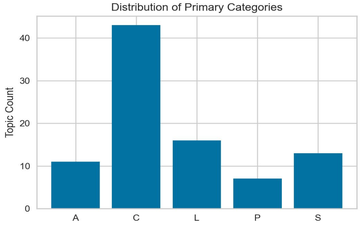




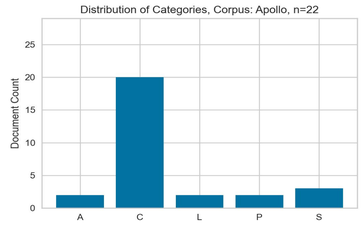


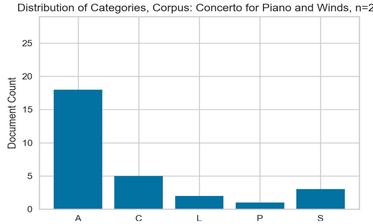


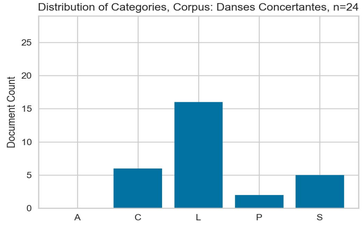


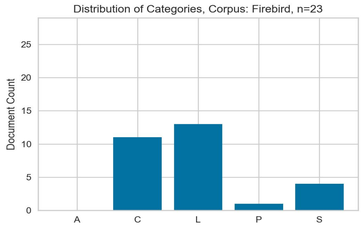


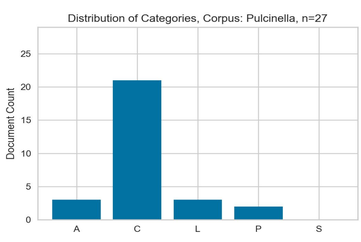
Primary Category Distributions by Corpus (**Side B**)

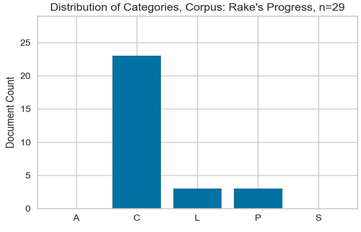












Methodology

I focused on six works whose subcorpuses contain at least twenty documents ("Apollo", "Concerto for Piano and Wind Instruments", "Danses Concertantes", "Firebird", "Pulcinella", and "Rake's Progress").

I used topic modeling with NMF (an unsupervised modelling technique) to obtain fifteen topics per corpus, consisting of ten words each, using silhouette scores to determine the appropriate number of topics.

Topics were further organized into five categories that specify type of review. The top five words in each topic were used to determine the topic's primary category, and the remaining five were used to determine the secondary category. **Side A** shows distribution of secondary categories by primary category. **Side B** shows distribution of primary categories per corpus.

The Five Categories

Analytical (A): Language pertaining to theoretical analyses of piece or performance.
Contextual (C): Observations or descriptions not necessarily analytical.
Logistical (L): References to time, place, program, performers, conductors.
Personal (P): References to Stravinsky.
Sentimental (S): Emotional language directed towards any aspect of performance.

Results

Reviewers are much less guided by sentiment than contextual information and analytical expertise (**Side B**). This goes against my initial intuition that reviewers are motivated by sentiment. Reviewers also rely heavily on context to justify analysis or emotion (**Side A**).

Discussion

The results of this study should be interpreted keeping in mind that most reviewers are white and male, and that the corpuses used in this study contain mostly non-letters. Review types (letter, program note, article, critique) in the corpus undoubtedly influence the distribution of categories. A future study on the differences in category distribution between a corpus of letters and a corpus of formal reviews and using those features to predict document type using unsupervised techniques could be fruitful.